CHAPTER 54

The mediation of eye movements by spoken language

Gerry Altmann

Abstract

Many factors can influence eye movements around a visual scene. Here, we consider the influence of spoken language. The first part of the chapter reviews a range of studies which illustrate how eye movements can illuminate aspects of spoken language understanding. The second presents a hypothesis of how spoken language mediates visual attention. The chapter then considers how this mediation might depend on the participant’s goals, including consideration of what the timing of these eye movements might, or might not, tell us about their generalizability. Methodological issues, including how to depict the dynamically changing eye movement record, and how to analyse it, are considered next. The chapter concludes with some issues that are likely to dominate the field in future years.

Introduction

L.N. Fowler is perhaps best remembered for his now-ubiquitous bust illustrating the supposed relationship between the shape and size of the human cranium and the mental abilities of the mind housed within. Fowler’s Phrenology, dating from around 1865, is of little relevance to contemporary Psychology, except for his bust’s left eye, immediately below which is inscribed a mental faculty that, until recently, was rarely associated with the eye—language. That language should have any influence on the movement of the eyes is unsurprising; just as pointing can direct attention in one direction or another, or to one location or another, so language can do the same. Of interest, however, is how language exerts this influence on eye movement control. With respect to the coordination of eye and hand movements, Neggers and Bekkering (2002) concluded that ‘ocular gaze is always forced to follow the target intended by a manual arm movement’ (p. 365), with Horstmann and Hoffman (2005) concluding further that ‘the saccadic system is no longer autonomous during a coordinated, goal-directed movement of eye and arm.’ (p. 7) A natural question, then, given this tight coordination between eye and hand, is whether there are situations where referring to an object in the immediate environment might be as constraining of eye movements as reaching towards that object would be.

The chapter proceeds in five sections: 1) an overview of findings from the spoken language processing literature on language-mediated eye movements; 2) an account developed primarily from within that literature of how, and why, spoken language mediates eye movements; 3) consideration of task-dependence and the conditions under which language-mediation of visual attention and
concomitant oculomotor control might be considered ‘automatic’; 4) a survey of methodological
issues that pervade research with language-mediated eye movements; and 5) a (brief) outline of
the future directions in which this research is likely to proceed. Throughout, the emphasis is on
spoken language; the literature on eye movements during written language processing will not be
considered here.

Psycholinguistics and the language-mediation of
eye movements
Roger Cooper first observed that as participants listen to a sentence referring to objects in a concur-
rently presented visual scene, the eyes move seemingly automatically to the objects in the scene as
expressions referring to those objects are heard (Cooper, 1974). For instance, participants were more
likely to fixate the picture of a snake when hearing ‘snake’ or part of ‘snake’ than pictures of referents
of unrelated control words. Moreover, participants were more likely to fixate pictures showing a
snake, a zebra, or a lion when hearing the semantically related word ‘Africa’ than they were to fixate
referents of semantically unrelated control words. These were the first demonstrations that the
patterns of eye movements in the context of unfolding language might reflect the online activation
of word-level representations and their semantics, and the integration of these representations with
the object-representations associated with (or activated by) a concurrent visual scene. It was not until
almost 20 years later, however, that this seminal work was followed-up, by Michael Tanenhaus and
his students (Tanenhaus et al., 1995). Whereas Cooper had demonstrated the range of issues that
might be addressed using this new methodology, Tanenhaus and his colleagues ‘validated’ the meth-
odology by showing not only how classical results found elsewhere in the psycholinguistic literature
could be ‘replicated’ within this new ‘visual world paradigm’, but how the paradigm could be applied
to the investigation of a broad range of language-related phenomena.

It should be borne in mind, throughout this chapter, that the influence of spoken language on eye
movements to visual scenes (or towards regions of space in which a scene had previously been
present) is limited; there are many factors other than the language which drive the eyes around a scene
even as the spoken language unfolds. Throughout, when language is described as mediating eye
movements, the data in fact show that it merely changes the bias to look at a part of space; on occa-
sion, these biases are large and on occasion, small (see ‘Task-sensitivity and language-mediated eye
movements’ and ‘Methodological issues’ sections). What is remarkable is when these biases change,
and what is of interest is why they change, and what cognitive representations and processes these
changes reflect.

The first study that synchronized the eye movement record with the unfolding language on a
moment-by-moment basis was reported by Allopenna et al. (Allopenna et al., 1998; neither Cooper
(1974) nor Tanenhaus et al. (1995) synchronized against the moment-by-moment unfolding of the
concurrent language). They applied the technique to what at first blush appears to be a straightfor-
ward replication of standard ‘competitor’ effects first observed in investigation of the cohort model
of spoken word recognition (Marslen-Wilson, 1987; Zwitserlood, 1989). They presented participants
with a visual display depicting a beaker (the target item), a beetle (the onset of the word ‘beetle’ over-
laps with the onset of ‘beaker’), a (loud)speaker (the word ‘speaker’ rhymes with ‘beaker’) and a baby
carriage (‘carriage’ is unrelated to any of the other words). At issue is what the eyes would do as
participants heard the instruction to ‘pick up the beaker’ (which they did by clicking with the compu-
ter mouse). For the first 400 ms after the onset of ‘beaker’, the probability of participants fixating the
beaker or the beetle rose, with no distinction between the two. After around 400 ms, the probabilities
diverged, with the probability of fixation on the beetle dropping toward zero. This point coincided,
roughly, with the end of the word ‘beaker’, and is approximately 200 ms after the point in the word
that distinguishes it from ‘beetle’. Allopenna et al. thus replicated the competitor effects predicted by
the cohort model, although unlike prior demonstrations they were able to show how these effects
unfold in real time. Interestingly, looks towards the speaker also increased in those first 400 ms
before subsiding as the speech signal became more unambiguously compatible with ‘beaker’. This was the first demonstration of rhyme competitor effects predicted by the Neighbourhood Activation Model (Luce et al., 1990; but see Marslen-Wilson, 1993 for discussion of such effects within the cohort model), which in turn demonstrated the utility of the visual world paradigm for exploring effects which hitherto had not been found using other techniques.

The ‘first wave’ of research with this paradigm focused more on language-mediated eye movements as a dependent measure used to study language processing rather than on language-mediated eye movements as an object of study in their own right (studies that are as relevant to eye movement control as they are to language processing are surveyed in subsequent sections). The survey that follows does not attempt to exhaustively summarize individual findings, nor is it an exhaustive listing—rather it is an indication of the range and breadth of psycholinguistic issues that have been elucidated with this paradigm. The vast majority of the findings support theories of language processing termed ‘constraint satisfaction’ theories (MacDonald et al., 1994; Trueswell and Tanenhaus, 1994).

These theories assume that language processing consists of the application of probabilistic constraints, in parallel, as a sentence unfolds, with no single constraint being more or less privileged than any other except in respect of its probabilistic strength. The origins of this constraints-based approach are rooted in the development of the parallel distributed processing (i.e. connectionist) models of cognitive representation and process (e.g. Rumelhart and McClelland, 1986). Thus, information that could in principle be applied to the task of interpreting a segment of the unfolding language can be applied, and at the earliest possible opportunity, subject to whatever other information might also be applied at that time, be it acoustic-phonetic, lexical, prosodic, syntactic, semantic, or pragmatic. There are some notable exceptions, and these are referred to as appropriate below.

- Sensitivity to acoustic-phonetic and prosodic variation—these studies showed that eye movements are extremely finely time-locked to the unfolding acoustic signal, reflecting graded effects on lexical access of subtle phonetic variation, including coarticulation, in the input. The essential finding is that as a word unfolds in the acoustic input, so the eyes move towards whatever in the visual scene that unfolding word could refer to, taking into account the ‘goodness of fit’ between the acoustic signal and words in the mental lexicon (Dahan et al., 2001; Magnuson et al., 2003; McMurray et al., 2002; Salverda et al., 2003, 2007).

- Lexical competition, lexical neighbourhood, and frequency effects—building on Allopenna et al. (1998), these studies used profiles of dynamically changing fixation probabilities to explore the time-course of activation of words differing in frequency (Dahan et al., 2001) or occupying ‘dense’ or ‘sparse’ neighbourhoods (Magnuson et al., 2007). They found that eye movements towards objects whose names differed in frequency, or were from differing neighbourhood types, had different dynamic fixation profiles. Thus, the lexical context (the neighbourhood) within which an unfolding word must be processed, as well as its frequency, influences the activation of the corresponding lexical representation (cf. Marslen-Wilson, 1990; Luce et al., 1990), in turn influencing the eye movements towards its referent. Some of these studies employed artificial languages so as to more precisely control word frequency and similarity to other words in the (artificial) lexicon (e.g. Magnuson et al., 2003b).

- The bilingual lexicon—only a minority of the world’s population speak a single language; the majority have to contend with word forms which, depending on the language they are hearing at the time, may mean one thing or another or nothing. Spivey and Marian (1999) used the visual world paradigm to show that hearers do not only activate lexical representations for words in the language being spoken/heard—they also activate competitor (i.e. phonologically overlapping) words from the other language

- Lexical semantics—whereas the studies described thus far have focused primarily on word form, the visual world paradigm has also been applied to individual word meaning. Yee and Sedivy (2001) showed that hearing ‘piano’ not only engenders saccades towards a piano (cf. Allopenna et al., 1998), but also towards a trumpet if present (see Yee and Sedivy, 2006, for the full report). Huettig and Altmann (2005) extended Yee and Sedivy’s design to include the case where a trumpet
was present without the piano. They also computed a measure of the 'conceptual similarity' between pianos and trumpets (and the other pairings in the study), using semantic feature norms (Cree and McRae, 2003). These similarity scores predicted the probability of launching a saccade during the target word (e.g. 'piano') towards the semantic competitor (e.g. the trumpet). They also predicted the time subsequently spent fixating the semantic competitor. We return later, in the section 'How, and why, language mediates eye movements', to the significance of these data for understanding why the eyes move in this paradigm; the Yee and Huettig data prove key to understanding the linkage between language and eye movements. More recently, the paradigm has been used also to explore differences in the lexical representation of concrete vs. abstract words (the distinction between words referring to objects that can be directly perceived vs. words that refer to concepts that cannot be perceived directly, such as 'honesty'; Duñabeitia et al., 2009).

- **Syntactic ambiguity resolution, in adults and in children**—moving 'up' into the realm of sentence processing, and building on earlier findings by Crain and Steedman (1985) and Altmann and Steedman (1988), Tanenhaus et al. (1995) and Spivey et al. (2002) reported a study showing how the real-time interpretation of syntactically ambiguous sentences depends on the context within which the sentence is interpreted. They showed participants either one apple or two (with one of the apples on a towel), and told participants to 'put the apple on the towel in the box'. Analyses of the eye and hand movements (i.e. which apple and where it was moved to) indicated that when there were two apples, 'on the towel' was interpreted as indicating which apple (i.e. it was interpreted as a modifier), but when there was just one apple, it was temporarily interpreted as where the apple should be put (i.e. it was interpreted as the goal). Trueswell et al. (1999) repeated a version of this study with children, and found that they do not use such referential context in the same way as do adults. Subsequent research with the visual world paradigm, again investigating real-time processing in adults and in children, showed that this is most likely due to children's increased sensitivity (relative to adults) to the particular syntactic contexts in which different verbs are most often found (Snedeker and Trueswell, 2004). Snedeker and Trueswell (2003) described a separate study showing also how speakers and hearers use prosody to disambiguate these kinds of ambiguity, and how this interacts with referential context. Other studies that have used this paradigm to explore effects of prosody and pitch accent, and their interactions with referential and discourse contexts, include Dahan et al. (2002) and Ito and Speer (2008).

- **Prediction in sentence comprehension, in adults and in children**—similar to Altmann (1999), who used a word-by-word reading task, Altmann and Kamide (1999) used the visual world paradigm to show how participants anticipate upcoming information in a sentence. They showed participants scenes depicting a person, a single edible object, and other distractor objects, and found that when hearing 'the boy will eat . . .' participants' eyes moved towards that edible object before the subsequent noun phrase was heard (e.g. ' . . . the cake'). Similar results have been found using other syntactic structures: Sussman and Sedivy (2003) showed that, on hearing 'what did Jodie squash the spider with?' the eyes anticipated at 'squash' the spider, but at 'the spider' they immediately anticipate a shoe (that had been used for the squashing). Arai et al. (2007) showed also that participants anticipate the ordering of referents if there will be more than one, as in 'the boy will give the girl the cake' (for related studies see Scheepers and Crocker, 2004; Thothathiri and Snedeker, 2008). Subsequently, verb-based anticipatory effects have been found in children whether skilled or less-skilled at comprehension (Nation et al., 2003), and also in infants (Anne Fernald, personal communication). Equivalent results have also been shown using other empirical methods, such as ERP (DeLong et al., 2005; van Berkum et al., 2005). Dahan and Tanenhaus (2004) used the visual world paradigm to show that verb-based constraints such as those used by Altmann and Kamide (1999) exclude from consideration semantically incompatible cohort competitors (thus, a cable would not be entertained during the earliest moments of 'cake' when preceded by 'the boy will eat'). A related result was obtained by Magnuson et al. (2008), using an artificial language; they showed that expectations with respect to the anticipated form-class
(e.g. noun versus adjective) modulated cohort competitor effects such that a competitor from the
‘wrong’ form-class was not entertained.

* Morphosyntactic variation and cross-linguistic differences in word order*—building on this last finding that participants can anticipate at the verb what may be referred to next, Kamide et al. (2003a) showed that in German, a language that permits OVS (object–verb–subject) word order rather than just SVO as in English, speakers use case-marking on the individual nouns to drive their predictions: if the first noun is marked in the nominative (indicating it refers to the agent of some event), the eyes anticipate at the verb whatever would be a suitable ‘patient’ (i.e. the object on which the agent acts); but if the first noun is marked in the accusative (indicating it refers to the patient), the eyes anticipate whatever would be a suitable agent. Related work (focusing on other aspects of grammatical marking or discourse status) has been reported in Finnish (Kaiser and Trueswell, 2004) and in Japanese (Kamide et al., 2003b), with the latter study showing that anticipation need not be verb-driven (Japanese is a verb-final language, so any predictive processing is necessarily triggered by non-verbal constituents).

* Combinatorial semantic processing*—Kamide et al. (2003) followed-up their original observation of anticipatory eye movements with a study in which they showed that it was the semantic combination of the verb with its subject that drove the anticipatory process: it was not simply whatever was edible that would be looked at after ‘the boy will eat . . .’, but rather what would most plausibly be eaten by the person doing the eating. This result was significant because it showed that it was not simply a form of semantic priming (‘eat’ causing looks toward anything edible) that drove the original effect, but was in fact the product of the syntactic and semantic processes that combine the verb with its subject, coupled with real-world knowledge of which kinds of event are more likely given the depicted participants in those events. Related data were reported by Knoeferle and colleagues (Knoeferle and Crocker, 2006; Knoeferle et al., 2005). Boland (2005) demonstrated that there are limits on anticipation during sentence processing: In the sentence ‘Chris recommended a movie to Kim in the hallway’, the movie and Kim are both arguments of the verb; the meaning of the verb entails that there is something being recommended and someone to whom the recommendation is made. But ‘in the hallway’ is an adjunct—it is not a part of the core meaning of the verb, and Boland showed that adjuncts are not anticipated to the same extent as arguments. Thus the grammatical status of the objects that might be anticipated does appear to influence the likelihood of their anticipation.

* Resolving pronominal reference during sentence processing*—most of the above studies take advantage of the fact that, as the language refers to something in a concurrent visual scene (or even beforehand), so the eyes look towards it. The paradigm is thus particularly useful for determining when, and what basis, readers or listeners interpret words such as ‘he/she/they/himself/etc.’ Reading studies, for example, can determine how such words are interpreted only indirectly (e.g. through increased reading times in certain circumstances). The visual world paradigm provides a more direct assessment of how pronouns are interpreted as they are encountered: For example, in the following sentence, ‘He’ is ambiguous: ‘Donald is bringing some mail to Mickey while a violent storm is beginning. He’s carrying an umbrella’. By observing eye movements at ‘He’ it is possible to infer which of the two possible characters the hearer attributes as the referent for the pronoun. And by changing ‘Mickey’ to ‘Minnie’ it is possible to explore the timing with which gender information (Minnie is a girl) is applied. Of course, and as mentioned at the outset of this chapter, individual participants rarely saccade to a language-relevant target 100% of the time; when they do saccade to it, it is likely that they do so because of that relevance (although there will always be some smaller likelihood that they did so for some other reason), but when they do not, aggregating across subjects and trials allows one to determine at what point relative to the language the bias to look towards the target changes. A range of studies have shown the immediacy with which information about accessibility (Donald is more accessible as he was mentioned in subject position) and gender (the Mickey/Minnie alternation) can be applied during pronoun resolution (e.g. Arnold et al., 2000). Interestingly, 3–5-year-old children show
immediate use of gender information, but appear relatively insensitive to the order of mention in
the first sentence (Arnold et al., 2007a). Other studies have focused on the processing of reflexive
pronouns (‘himself/herself’) in various contexts (e.g. Runner et al., 2003; 2006; Sekerina et al.,
2004). And most recently, Arnold and Lao (2009) have used an exogenous attentional capture
paradigm (Gleitman et al., 2007) to show how attentional mechanisms (exogenous and endog-
enous) impact on pronoun comprehension. Distinguishing between such influences will prove
crucial to a fuller understanding of how language interpretation (and production; see below)
interacts with systems that (also) serve visual processing.

The use of pronouns reflects dynamic changes in the perspectives that both hearer and speaker
take as their shared language unfolds: speakers use pronouns when they anticipate that their listeners
will be able to interpret them given the context. The studies described thus far have focused on
comprehension. Unsurprisingly, the visual world paradigm has been used also to investigate a
number of issues relating to language production (see Griffin, 2004, for review), dialogue, and
perspective-taking.

- **Name retrieval**—the first study to use eye movements to study planning processes during speech
  was reported by Antje Meyer and colleagues (Meyer et al., 1998). Participants had to
  recall the names of objects presented across a screen. This study showed that participants do not look away
  from a to-be-named object until they have recognized the object and retrieved the phonological
  form associated with its name. A number of studies, including this one, explored how the ease of
  retrieving the phonological form modulates gaze durations (see also Griffin, 2001).

- **Sentence planning**—Griffin and Bock (2000) asked participants to describe simple transitive
  events depicted by line drawings of e.g. a dog chasing a postman. They found that, irrespective of
  whether the dog or the postman is about to be named, it is fixated around 900 ms before the
  articulation of its name. But as in Meyer et al.’s (1998) study, the eyes leave the object around
  100–300 ms before articulating it, moving towards the next object to be named. Interestingly,
  participants fixate objects around 900 ms before articulating even an incorrect name for the
  object, as in ‘gira- uh zebra’ (Griffin, 2004). One important aspect of the Griffin and Bock (2000)
  study is that they managed to distinguish between event apprehension and utterance formulation
  by comparing eye movements during the event-description task with eye movements made by a
  different group of participants whose task was to discern who was acted upon in the event (the
  postman in the above example). That different patterns of eye movements can be observed as a
  function of task is something we return to in the section ‘Task-sensitivity and language-mediated
  eye movements’. Gleitman et al. (2007) reported a study which used exogenous attentional capture
to direct attention towards one protagonist or another; this influenced their order of men-
tion in a subsequent description (e.g. between ‘a dog is chasing a man’ vs. ‘a man is running from
a dog’). This study also found that early *endogenous* shifts in attention during scene apprehen-
sion also predicted order of mention, unlike in the Griffin and Bock (2000) study. This difference
between the studies may have been due to differences in both the visual and linguistic stimuli
used, and it remains unclear whether endogenous shifts in attention during early scene apprehen-
sion do indeed influence ‘accessibility’ of concepts for subsequent linguistic processing (cf. Bock,
1986, 1987). Gleitman et al. (2007) concluded that the on-line construction of descriptive utter-
ances mirrors the on-line apprehension of the events which the utterance describes.

- **Dialogue: common ground and speaker perspective**—speakers rarely engage in monologue, without
due regard for the knowledge, beliefs, and perspectives of their addressee(s). In dialogue, speaker
and hearer change turns as they converse about things and ideas that are, or become, common
knowledge to both. A lasting issue concerns how speaker and hearer keep track of the overlap
between what they know and what their interlocutor (i.e. the other) does and does not know;
i.e. what is in ‘common ground’ and what is ‘privileged’. Keysar et al. (2000) employed a com-
monly used technique in production research whereby a ‘confederate speaker’ (i.e. one of the
experimenters) instructed a hearer to manipulate objects arranged in cubbyholes—the speaker
sat on one side, and the hearer on the other, looking through the cubbyholes to one another.
Most of the objects could therefore be seen jointly by both speaker and hearer. But some of the cubbyholes were blocked off so that the hearer knew that only he/she could see them. Thus, there might be two boxes, one in a cubbyhole seen by both speaker and hearer and the other in a cubbyhole seen only by the hearer (and thus privileged). In this case, where might the hearer look on hearing ‘move the box . . .’? If the hearer tracks in real time what is and what is not in common view, eye movements during ‘the box’ should be directed only to the box in common ground (seen by both), but if the hearer does not track in real time what is common and what is privileged, the expression ‘the box’ would be ambiguous. Eye movements in fact suggested the latter, with looks toward both objects. This result has proved extremely controversial, with other studies finding that interlocutors do keep track in real time of what is privileged and what is common (e.g. in adults: Hanna et al., 2003; in children: Nadig and Sedivy, 2002. See Barr (2008a), for an example of an attempt to reconcile the different findings).

• Social referencing and common ground—Crosby et al. (2008) reported an ingenious version of the common/privileged ground work: participants viewed a screen showing four other individuals. In one condition all four were apparently able to hear what the participant could hear; in the other, only two of the four could hear, and participants knew which. Subsequently, when participants heard one of the four utter something that was potentially offensive to one of the other individuals (for example, a negative remark about affirmative action when one of the other individuals was an African American), participants looked towards that individual only when they supposed that he/she could also hear what the speaker had said. In this case, social referencing (referring to a class of individual, rather than to the specifically depicted individual) directed eye gaze, but only when that individual was in what might be called ‘communicative common ground’.

• Unscripted dialogue—a number of studies have extended the prior work into the realm of unscripted dialogue, monitoring eye movements either to just one of the interlocutors (e.g. Brown-Schmidt and Tanenhaus, 2006; Brown-Schmidt et al., 2005), or to both (Richardson and Dale, 2005; Richardson et al., 2007). The Richardson and Dale (2005) study involved one participant describing, for example, the characters from the sitcom Friends while the other listened. Both were viewing a display showing all six faces of the main characters. Listeners’ eye movements closely mirrored the eye movements of the speaker. Of course, whereas the speaker fixated the individual characters in advance of referring to them (cf. Griffin and Bock, 2000; Meyer et al., 1998), listeners’ eye movements were delayed until shortly after each was referred to (we return to timing issues later, in the section ‘Task-sensitivity and language-mediated eye movements’).

• Interpreting speakers’ disfluencies—unscripted dialogue (and even scripted dialogue) is rarely fluent. Jennifer Arnold and colleagues have used the visual world paradigm to explore how listeners interpret disfluencies such as ‘Click on thee uh red . . .’ and showed that listeners interpret such disfluencies as indicating that the speaker is having difficulty, which in turn triggers an inference about the likely cause of this difficulty (such as naming something that is new to the domain of discourse (Arnold et al., 2004); or something that is unfamiliar (Arnold et al., 2007b). This work is as important in demonstrating the role of higher-level inferences with regard to speakers’ intentions as it is in demonstrating hearer’s ability to interpret disfluencies as being informative.

This sensitivity of language-mediated eye movements to speaker intentions, and not just to the linguistic form of an utterance (i.e. sensitivity to pragmatic knowledge), renders the paradigm particularly useful with respect to determining how pragmatic information is recruited in real time as a sentence unfolds within a concurrent context. We conclude this review of psycholinguistic topics with a range of studies that have demonstrated more pragmatic influences on language-mediated eye movements.

• Gricean Pragmatics—Sedivy (2003) reviews a range of eye movement data on interactions between listeners’ expectations and the informativity of the expressions produced by the speaker (and what listeners do when the speaker is more informative than might otherwise be assumed necessary to identify a particular referent). For example, Sedivy et al. (1999) monitored eye movements as listeners heard instructions such as ‘Pick up the tall glass’. Analysis of eye movements showed
that listeners interpret scalar adjectives such as ‘tall’ by explicitly comparing against contrasting objects; a tall glass is identified through its intended contrast with a shorter glass (and in the absence of a contrasting glass would be inappropriately over-informative, just as uttering ‘the yellow banana’ would be over-informative in some contexts but appropriately informative in others (Grice, 1975)). Interestingly, listeners take account of speaker characteristics when computing such contrastive inferences (i.e. if the speaker uses the expression ‘tall glass’ it is because there must be a shorter glass available in the domain of reference); listeners did not appear to compute these inferences when they had evidence that the speaker was not adhering to the usual Gricean (cooperative) conversational maxims (Grodner and Sedivy, in press).

Visual affordances—so far we have seen that language-mediated eye movements are not simply driven by linguistic form, but also by the communicative goals of the speaker and listener (cf. Sedivy, 2003). Unsurprisingly, they are also driven by the behavioural goals of the listener (and we return to this again in the section ‘Task-sensitivity and language-mediated eye movements’). These are in turn modulated by the affordances of the objects to which that behaviour is directed. Chambers et al. (2004) showed, for example, that the instrument given to the listener with which to perform an action influenced which objects were entertained as the ‘target’ of that action; holding a hook would cause eye movements during a referring expression such as ‘the whistle’ to only those whistles to which were attached loops that could be hooked. Related findings were reported by Chambers et al. (2002). Altmann and Kamide (2007) re-interpreted their earlier findings of anticipatory eye movements as reflecting the goodness-of-fit between the interpretation of the sentence and the affordances of the objects depicted within the scene; they used a tense manipulation (‘will drink’ vs. ‘has drunk’) to show how eye movements at the verb were driven towards objects that either would afford drinking in the future (e.g. a full glass of beer) or had afforded drinking in the past (e.g. an empty glass of wine). A similar tense manipulation in the visual world is described by Knoeferle and Crocker (2007). That study was notable also because they showed how language-mediated eye movements were modulated by the interpretation of an event that unfolded across a series of scenes (cf. distinct panels of a comic strip, with earlier panels setting the context for later panels).

Linguistic relativity—different languages describe events in different ways, and since Whorf’s original observations (Whorf, 1956), researchers have questioned to what extent such differences may influence how people may conceive the world and apprehend events. Slobin (1996) proposed that speakers’ utterances as they plan to describe an event may be shaped by the language they will use for that utterance, and this in turn may influence how the event is conceived as its articulation is planned (the ‘thinking for speaking’ hypothesis). Papafragou et al. (2008) addressed this issue by comparing eye movements of Greek and English speakers as they viewed motion events and prepared to either describe them or remember them. English and Greek differ in respect of how they describe the manner and path of motion; English tends to encode manner on the verb and not path (‘slide’, ‘skip’, but less commonly ‘ascend’ favouring instead ‘go up’), while Greek tends to encode path on the verb and not manner. In support of the ‘thinking for speaking’ perspective, attention to distinct landmarks (indicating path) or to instruments indicating the kind of motion (e.g. skates) did differ as a function of native language. However, in the non-linguistic task in which participants did not have to prepare a description of the unfolding event they were viewing, there was no effect of native language on the manner in which attention was allocated as the events were apprehended. In this study at least, differences with respect to which elements of the language refer to which constituents of an event did not, contra Whorf and others, affect the way in which attention was allocated during the apprehension of the events.

Economies of space preclude a more thorough overview of what has been learned about language processing through the use of the visual world paradigm. A variant of this paradigm, the findings from which are beyond the remit of this survey, is the ‘preferential looking’ paradigm. This is in common use in developmental psychology. In this paradigm, as applied to language processes, an infant might be sat upon a carer’s lap, and different screens (e.g. one to each side of a central fixation light) may show
different objects or scenes; the infant's preference to look towards one object/scene or another, as a word or sentence is heard, is then measured. The technique has been used widely and to significant effect. Of course, here the dependent measure is not eye movements per se, as infants will most often move their heads also in this paradigm. Nonetheless, what is common to each paradigm is the language-mediated deployment of visual attention, a phenomenon that pervades the lifespan. In the following section, we consider the mechanism by which language mediates eye movements (and/or visual attention), and the data (only some of which has been previewed in this section) that reveal this mechanism.

How, and why, language mediates eye movements

A number of distinct findings are key to understanding the mechanism by which language may direct eye movements. The fact that it does so at all is not particularly surprising, as mentioned at the outset of this article. More interesting is the mechanism by which language fulfils that part of its purpose concerned with making the hearer look one way or another. And as we shall see, the actual act of looking is almost incidental—what language does is not so much direct attention towards objects in the external world, but rather towards objects as represented in an internal mental representation of that external world. This is rendered possible in the visual world paradigm because, typically, the onset of the scene and the objects it contains precedes the onset of the language, and thus the objects can be apprehended, and their locations established, before the onset of the language. As will become clear, the visual world paradigm appears to work as it does precisely because, as the language unfolds, the locations of the objects that become relevant to that language are already known (but see Moores et al., 2003, for an example of where the language preceded the scene).

The first finding that bears on the mechanism of language-mediation is simply that as the name of an object unfolds, so the likelihood of moving the eye towards that object rises (Allopenna et al., 1998). These and related data (e.g. Dahan and Tanehaus, 2004) suggest a continuous mapping between the activation of lexical representations through linguistic input and the likelihood of moving the eyes towards the object(s) in the scene that match those lexical representations. But crucially, the match need not be an exact one: Shifts in visual attention (which can be assumed to precede eye movements) are not tied only to direct reference. Thus, hearing ‘piano’ engenders looks towards a trumpet (Yee and Sedivy, 2006) and does so in proportion to the semantic relatedness of the object referred to by the word and the object depicted in the scene (Huettig and Altmann, 2005).

This second set of findings (and other related data; e.g. Myung et al., 2006) suggest that it is the conceptual overlap between the word and the object that mediates between language and eye movements. These semantic relatedness data rule out an account based solely on phonological overlap between the unfolding word and the names associated with the objects in the scene.

A third set of findings suggests that it is not even the object itself, as depicted in the concurrent scene, that is the ‘target’ to which eye movements are directed; a number of studies have demonstrated that even if the scene is removed prior to the onset of the linguistic stimulus, the eyes move back to wherever the named, or anticipated, objects had been located (Altmann, 2004; see also Hoover and Richardson, 2008; Knoeferle and Crocker, 2007. For related work see Richardson and Spivey (2000), and for eye movements during visual imagery, see Brandt and Stark, 1997; Laeng and Teodorecu, 2002). These ‘blank screen’ effects demonstrate that the eyes can be directed not towards the object per se (because in the blank screen there is no object), but rather, towards a particular location as indexed by an episodic representation of that object (cf. Richardson et al., 2009). A final set of findings demonstrates that this episodic trace is malleable. Altmann and Kamide (2009) described a study using the blank screen paradigm in which the unfolding language described a change in location for one of the objects (e.g. ‘the woman will move the glass onto the table’). Subsequent reference to that object (e.g. ‘she will pour the wine into the glass’) engendered looks during the critical referring expression (‘the glass’) towards the new location of that object, and not to the old location as had actually been seen in the previous scene (in fact, and perhaps surprisingly, the old location was looked at no more than the location of other distractor objects; this location held no residual ‘attraction’).
Unlike in earlier blank screen studies, the spatial representations that directed the eye movements were not reliant on the objects actually having occupied particular locations within the scene—in the example, the glass had never been seen on the table, and yet the eyes moved there on hearing the sentence-final ‘glass’.

These and other related data prompted Altmann and Kamide (2007) to articulate an account of the language mediation of eye movements that is based on the activation, by the scene, of a memory representation of an object—its episodic trace (see Richardson and Spivey, 2000, for a precursor of this account). This trace includes information not only about the object’s properties but its location also. When a sequence of words is subsequently heard (the assumption being that the activation of these episodic traces precedes the onset of the language), the representations engendered by the unfolding language may overlap with the pre-existing representations activated by the concurrent or prior scene (i.e. those episodic traces). This overlap causes an increase in the activation of those traces (due to the dual support they now receive), and this in turn percolates through to the spatial indices associated with those traces. Altmann and Kamide (2007) view the change in activation of an object’s representation as a change in the attentional state of the cognitive system (cf. Cohen et al., 2004), with this change either constituting, or causing, a shift in covert attention. This shift in covert attention, in turn, is accompanied by an increased likelihood of an overt eye movement (due to the increased activation to the spatial index associated with that trace). Altmann and Kamide (2007) offer one possible explanation, based on Hebbian learning, for why increased activation of the spatial index associated with the episodic trace should lead to an eye movement: Orienting towards an object increases the activation of the mental representation of that object and its encoded location. Successive pairings of this kind during development may result in the opposite pattern, with increases in the activation of a mental representation of an object and its location resulting in the increased likelihood of an orientation response towards that location. This is, of course, both simplistic and speculative. But it does account for why, irrespective of whether the scene is concurrent or absent at the time of the unfolding language, the unfolding language can cause eye movements towards objects’ current or past locations. What remains to be explained, however, are the ‘moved glass’ effects reported by Altmann and Kamide (2009). That study showed that the spatial index associated with the episodic trace of an object is not fixed, and is not determined solely by its past perceptual correlates. They proposed that the targeting of saccadic eye movements can be supported by two distinct kinds of representation; those based on perceptual properties of the configuration of objects as directly experienced in the scene, and those based on the conceptual properties of the objects and their configuration, with the latter changing dynamically as a function of the unfolding language and ensuing event representations. Both kinds of representation are required, as otherwise it would not be possible to explain how the eyes can return either to the original location of the glass (as directly perceived previously) or to the new location (as conceptually determined by the language-induced event representation).

**Task-sensitivity and language-mediated eye movements**

The previous section reviewed an account of the mechanism underlying the language-mediation of eye movements in which such mediation occurs non-consciously. But does this mean that language-mediation of eye movements will occur the same way regardless of the goals of the hearer? Certainly, language-mediation of eye movements can be non-conscious in the same way that priming can be (i.e. the recognition of a word or object being facilitated if there has been prior exposure to a related item; for review and an example of priming of picture recognition by unconsciously perceived primes, see Dell’Acqua and Grainger (1999)). Many of the findings reviewed above suggest a non-conscious quality to language-mediated eye movements: the fact that the eyes move at ‘piano’ to a trumpet, and do so in proportion to their semantic relatedness (Huettig and Altmann, 2005, Yee and Sedivy, 2006). Or the fact that language-mediated eye movements are sensitive to frequency and neighbourhood effects (Dahan et al., 2001, Magnuson et al., 2007). And so on. These effects suggest that participants are not consciously directing their eyes. But this is not the same as saying that...
presented participants with displays containing as few as just four objects (e.g. Allopenna et al., 1998; more restricted than the usual environment with which we typically interact. Many studies have the 'closed set problem' — the fact that the visual environment in typical visual world studies is very much of this processing anarchy is usually taken to be indicative of some specific, generalizable, cognitive mental psychology, not to one paradigm or another. But again, finding consistent patterns in the face many others) suffer the same problem — participants may choose to process each sentence to more patterns of eye movements can nonetheless be observed which generalize statistically across partici- the language. It is all the more remarkable, therefore, that in the face of such variance, consistent participant may maintain central fixation without attending at all to the scene or its relationship to the targets of the reaching motion or the mouse movement. One advantage is that the likelihood of fixat- the object referred to in the sentence which is acted upon by someone or something else. So if you he talk, ‘the girl rode the bicycle’, you would click on the bicycle’). Some of the earliest studies involved participants reaching for objects placed in front of them (e.g. Tanenhaus et al., 1995). Eye move- ments in such cases are in service of a behavioural task which requires the eyes to actively look at the targets of the reaching motion or the mouse movement. One advantage is that the likelihood of fixat- the object referred to in the sentence which is acted upon by someone or something else. So if you hear talk, ‘the girl rode the bicycle’, you would click on the bicycle’). Some of the earliest studies involved participants reaching for objects placed in front of them (e.g. Tanenhaus et al., 1995). Eye move- ments in such cases are in service of a behavioural task which requires the eyes to actively look at the targets of the reaching motion or the mouse movement. One advantage is that the likelihood of fixat- the object referred to in the sentence which is acted upon by someone or something else. So if you hear talk, ‘the girl rode the bicycle’, you would click on the bicycle’). Some of the earliest studies involved participants reaching for objects placed in front of them (e.g. Tanenhaus et al., 1995). Eye move- ments in such cases are in service of a behavioural task which requires the eyes to actively look at the targets of the reaching motion or the mouse movement. One advantage is that the likelihood of fixat-
Altmann, 2004; and others). It is rare that in everyday life our visual environment contains so few discernable items. The use of displays depicting only a very small set of objects could in principle lead to ‘unnatural’ processing strategies (e.g. implicit naming of the objects in the scene, although as already pointed out, such a strategy is unlikely to be widespread in view of the semantic relatedness data of Huettig and Altmann (2005), and Yee and Sedivy (2006), discussed earlier). With respect to whether this closed set problem compromises generalizability to language processing more generally, it should be noted that most often interlocutors also restrict themselves to talking about only a small number of discourse entities (most studies of reading introduce fewer entities into the domain of reference than does the typical visual world study). Similarly, observers generally attend within their visual environment to just a few objects. Thus, much (if not most) of our experience is in the context of only a limited number of objects to which we attend/refer. However, while it can be said that the modus operandi of natural language is to restrict the domain of reference to a subset of the things we could possibly know about, natural scenes are rarely restricted in the ways that are typical of visual world studies. But most often, researchers are not intending the scenes they use as surrogates of real visual environments—they are instead artificial ‘targets’ against which the unfolding language can be ‘aimed’. The paradigm is limited in some respects, but liberating in others.

Notwithstanding this last defence of the closed set problem, how we select one object from a smaller or larger set of visual alternatives may reflect quite different visual search strategies. When there are only a handful of objects in view, naming one of those objects will generally cause the eyes to move directly to that object (see the previous section for an account of how this can come about, predicated upon prior knowledge of the identities and locations of the depicted objects). But the same is unlikely upon hearing the name of one of 100 objects in view—rather, a visual search may be initiated around the area in which we conjecture that the object is located.

A further task-related issue concerns the timing of language-mediated eye movements. The timing with which language can mediate eye movements is an important issue because of the manner in which researchers typically synchronize the eye movement record with the unfolding speech (see ‘Methodological issues’ section for further detail). The standard assumption is that it takes up to 200 ms to program and launch a saccade, with many researchers assuming that any eye movement launched within 200 ms of word onset could not have been influenced by that word. The majority of studies that investigated the delay between stimulus onset and saccadic initiation used visual stimuli as both triggers (i.e. the signal to move) and targets (e.g. Rayner et al., 1983; Saslow, 1967). One oft-cited study, which attempted to measure ‘saccadic overhead’—roughly, the time to plan and launch a saccade once the target has been identified—estimated around 100 ms (Matin et al., 1993). It is perhaps curious that this study is often cited by users of the visual world paradigm in the context of the 200 ms figure, especially as this study did not even monitor eye movements! (It required participants to make judgements that either did or did not require saccadic eye movements; button press reaction times were collected and subtracted from one another across the different saccadic conditions.)

To address the lack of any data on language-mediated launch times, Altmann and Kamide (2004) reported a study in which participants were told to move their eyes to a pre-determined location as soon as they heard an auditory cue. In that case, the mean, median, and modal launch times were 254, 181, and 132 ms respectively. This highly skewed distribution (an early peak, in this case at around 130 ms, followed by a long tail) is typical of saccadic launch times (e.g. Carpenter, 1988); thus, the mean launch time in such a distribution is not an appropriate measure of the ‘central tendency’. However, in this task, participants knew in advance where they would have to move their eyes, and it is known that such prior knowledge can speed launch times (Saslow, 1967). The Altmann and Kamide (2004) study also included a condition in which participants did not have such prior knowledge; the auditory cue was the name of the target object to which they had to move their eyes. In this case, launch times were much slower, with a median response time (in error-free trials) of 484 ms. At first glance, this suggests that the 200 ms assumption is a conservative estimate. However, it needs to be borne in mind that eye movement control is, independently of linguistic mediation, sensitive to task differences: The time to launch a saccade depends on the urgency of the response...
(Reddi and Carpenter, 2000), the likelihood of having to saccade to one location or another (Carpenter and Williams, 1995), and the number and positioning of distractor objects (e.g. McSorley and Findlay, 2003; Walker et al., 1997; and references therein). Thus, in the absence of information about the distribution of launch times that would be typical for a given language-mediated task and a given set of visual configurations, it is probably safest to err on the side of caution, and to assume that some number of eye movements which were launched within 200 ms of some word’s onset were indeed influenced by that word. More recently, and in a reanalysis of the studies published by Kamide et al. (2003) and by Altmann (2004), Altmann (in press) has provided evidence that language can in fact influence the eye movement record within as little as 100 ms (notwithstanding the look-and-listen task used in these studies). Most likely, these very fast effects are due to language-mediated cancellation of already-planned saccades.

A final comment on task: A number of studies that have not utilized the visual world paradigm have also considered the influence of language on visual attention (e.g. Estes et al., 2008; Meteyard et al., 2007). Liu (2009) found that auditory words such as ‘ascend’ and ‘plummet’ either facilitated or interfered with eye movements during the smooth pursuit of a target dot moving vertically down the screen (the effects were apparent at word offset). These effects were modulated by whether the gaze position of the eye was ahead of, or behind, the dot being tracked; the directionality implied by the verb’s semantics (up vs. down) interacted with the direction in which attention had to be deployed (up vs. down) to match the velocity of the dot. This interaction with gaze position relative to the moving dot suggested that the modulatory effects of verb semantics were not under conscious control (participants are not aware that their eye is a few pixels ahead of, or behind, the target). This study is also interesting because all studies to date which have explored the influence of language on eye movements have employed situations in which eye movements are driven towards a target object on the basis of the match between the content conveyed by a word or phrase and the knowledge associated with that target object. In the smooth pursuit task, the language is incidental to the oculomotor task, and neither the semantics of the language, nor indeed the semantics of the visual object (whatever that may be in the case of a moving dot) are, a priori, relevant to the initiation and maintenance of pursuit eye movements.

Methodological issues

The visual world paradigm has engendered interest, in part, because of the very precise synchronization that is possible between the eye movement record and the linguistic input or output. It is possible, for example, to record the number of times the eyes saccade to a target location during a particular word or phrase, whose onset and offset vary on a trial-by-trial basis. Plotting such data is not straightforward however.

Plotting the data

Figs. 54.2 and 54.3 illustrate the problem. Both figures show the same data, taken from the ‘blank screen’ data reported in Altmann (2004).1 Participants heard ‘The man will eat the cake’ having previously seen an image with just four objects (one in each quadrant): a cake, a newspaper, a woman, and a man (see Fig. 54.1, which also shows the relative sparseness of the stimuli that are sometimes used in such studies—see Henderson and Ferreira, 2004, for discussion). The regions of interest were the quadrants in which each object had been located (see below for further discussion of the choice of region of interest).

1 Only the saccadic analyses, corresponding to Fig. 54.4, were reported in Altmann (2004). The same data set that generated those analyses was used to generate the fixation plots shown in Figs. 54.2, 54.3, and 54.5.
The plots in Figs. 54.2 and 54.3 both show the percentage of trials over time (quantized into successive 50 ms ‘slices’) on which one of three locations was fixated as the accompanying sentence unfolded. Figure 54.2 synchronizes time from the onset of ‘The man . . .’. The first vertical (black) line shows the average offset of ‘The man’, the second shows the average onset of ‘eat’ and the third the average onset of ‘the cake’. The shaded regions attached to each such line show the range of actual offsets/onsets, with the leftmost edge marking the earliest, and the rightmost edge marking the latest.

**Fig. 54.1** Example scene from Altmann (2004). Reprinted from *Cognition, 93*(2), Gerry T.M. Altmann, Language-mediated eye movements in the absence of a visual world: the ‘blank screen paradigm’, pp. B79–B87 © 2004, with permission from Elsevier.

**Fig. 54.2** Percentage of trials with fixations across time, synchronized from sentence onset. From Altmann (2004).
The problem that such a plot presents (assuming that the shaded regions were not present) is the indeterminacy that results from the progressive desynchronization between the eye movement record and the trial-by-trial timings of interest that occur the further into the sentence one goes. For example: the rise in looks to where the cake had been during the verb ‘eat’, which is of theoretical importance with respect to the demonstration of anticipatory eye movements even when viewing a blank screen, appears to occur during the acoustic lifetime of the verb itself. But the shaded regions show that one half of the average duration indicated for the verb includes trials in which the verb had already ended and the postverbal noun phrase had begun. Thus, what appear to be anticipatory looks towards the cake may not be anticipatory at all, but may in fact reflect those trials in which acoustic information pertaining to ‘cake’ had occurred earlier than average. Plotting the data as in Fig. 54.2 (without the shaded regions) would therefore give a misleading impression of what might actually be going on. To avoid this, the data can be plotted as in Fig. 54.3.

In Figure 54.3, the calculation of fixations has been computed separately for each interval of interest (see Altmann and Kamide, 2009, for details). For example, for the verb ‘eat’, the interval of interest begins at the onset of the verb, calculated on a trial-by-trial basis, and proceeding for 800 ms (the average duration of the verb). The vertical line at the offset of ‘eat’ in the graph represents the onset of ‘the’ in ‘the cake’ (to avoid any coarticulatory influences on the estimation of anticipatory eye movements), and the calculation of fixations during ‘the cake’ was resynchronized to the onset of this interval, again calculated on a trial-by-trial basis (in principle, there could be discontinuities between the curve just before resynchronization and just after; in practice there are few). The grey bars show the ‘uncertainty’; that is, for roughly the last 400 ms of ‘eat’ some of the rise in looks towards the cake had been may still be due to trials in which the onset of the postverbal noun phrase occurred sooner than 800 ms after the onset of the verb (the corresponding grey bar at sentence offset is not shown in the interests of clarity). However, because the vertical line at the onset of ‘the cake’ indicates resynchronization, this line (or rather, where the curves cross this line) shows accurately where the eyes were fixating at this onset and its equivalent across the different trials. The graph tells us, therefore, that the rise in looks to the cake in advance of the onset of ‘the cake’ is truly anticipatory. It is noteworthy that the two graphs are very similar; this reflects the fact that the sentences were short, had the same syntactic structure, and were designed to be similar; the greater the variability amongst the sentences, the greater the desynchronization problem.

Plotting fixations is perhaps the most common method for conveying the dynamics of eye movements in this paradigm. However, there are different ways of calculating these plots. In Figs. 54.2

---

**Fig. 54.3** Percentage of trials with fixations across time, resynchronized at each marked interval onset and at sentence offset. From Altmann (2004).
and 54.3, the plots show the percentage of trials, at each successive 50-ms time point, on which the eyes were fixating one region or the other. An alternative is to plot, as a percentage of all fixations that were counted (across all the regions of interest), the fixations within each specific region of interest. This second calculation generally yields different numbers, but it also more accurately reflects the data that is input to statistical analyses based on log likelihood or odds ratios. More recently, some researchers have taken to plotting odds ratios (or log odds) directly. In the example shown in Figs. 54.2 and 54.3, this would effectively result in a plot which, for each 50-ms time point, showed how many times more likely it was for the participant to fixate the cake region than the newspaper region (each data point is the odds ratio calculated across the target region and a control region). The advantage this presents is that the graph is then a representation of the dynamically changing effect size of the experimental manipulation.

An important issue when calculating fixational data concerns the calculation of the onset of the fixation: should it be the time when the eyes landed in the region of interest, or the time when the eyes launched towards that region (i.e. the onset of the saccade preceding the fixation)? Many researchers in fact take the onset of the preceding saccade as the onset of the fixation. As outlined in Altmann (in press) this has the advantage of eliminating the noise that is due to variable saccade duration as a function of where in the scene the eyes were before the critical fixation—fixation onset will be determined by the prior distance of the eye from the new fixation (the further away, the later the fixation onset). Saccade onset is a more accurate estimate of when covert attention switched to the new fixation location.

An alternative to plotting or analysing fixation patterns is to plot/analyse saccades (often calculated as the number of trials on which at least one saccade was launched within some interval of interest, calculated on a trial-by-trial basis, to some region of interest). Although the two dependent measures are related through reflecting different components of the oculomotor response, they can dissociate (see Altmann and Kamide, 2004, for examples). The onset of a saccade is, roughly, the earliest moment at which there is a measurable oculomotor response to some change in cognitive state, and which measure one reports depends, of course, on the specific hypothesis being tested. The saccadic equivalent to the fixation plots shown in Figs. 54.2 and 54.3 is shown in Fig. 54.4. In this case, there is no indeterminacy with respect to the relationship between the plot and the trial-by-trial onset/offset of the temporal intervals of interest. Notice the discrepancy at the interval ‘eat’ between Figs. 54.4 and 54.3: in Fig. 54.3 there are still considerable numbers of fixations on the man, whereas

![Graph](image)

**Fig. 54.4** Percentage of trials with saccades launched, during each phrase/word towards the quadrants of interest. From Altmann (2004).
Fig. 54.4 clarifies that there are no more trials on which saccades were launched to the man during ‘eat’ than there were trials on which saccades were launched towards the other unmentioned objects.

Less common dependent measures derived from the oculomotor response, and which are occasionally used in the visual world paradigm, include saccade launch times and fixation durations—either first fixation duration, gaze duration (the sum of all fixations within the region’s boundaries before the eyes first leave that region), or total fixation duration. Less is known about the factors that influence these measures, some of which are unrelated to the relationship between language per se and the eye movement response.

**Regions of interest, in time and space**

Exploring different dependent measures is just one part of analysing ‘visual world data’. Equally important is exploring different regions of interest. Unfortunately, the term ‘region’ has been applied in both the temporal and spatial domains. Henceforth, the term ‘interval’ of interest will refer to regions in time, and ‘region’ of interest to regions in space. Again, choosing the appropriate intervals of interest depends in large part on the hypotheses to be tested. Studies of word recognition (e.g. Allopenna et al., 1998) tend to focus on an interval that corresponds to the word and some (occasionally arbitrary) period beyond (but see Magnuson et al., 2007, for statistically motivated alternatives such as growth curve analysis). Here, the questions of interest concern when, relative to some part of the word (e.g. onset, uniqueness point, offset, etc.) the eyes can be seen to look more to one region than to another. The analyses of these data often assume that it takes some time from when the cognitive system ‘decides’ to move the eye to that movement becoming manifest in a saccade being launched, and some period of time is added to the onset and/or offset of the interval of interest to take this time into account (see above, and also Altmann, in press). Studies of sentence processing are often more complex in respect of which intervals to choose; in these cases, the intervals of interest will be determined in part by theoretically-driven predictions about which parts of the sentence might be relevant given the goals of the experiment.

For the most part, issues concerned with how to define the (spatial) regions of interest are the same as those encountered when studying eye movements in the absence of concurrent language. The usual caveats in such situations apply: comparisons between eye movements to one object with eye movements towards another object can be contaminated by confounds due to differences in size or the visual saliency of the different objects (e.g. colour, contrast, texture, spatial frequencies and so on). Given that most commonly in the visual world paradigm, it is the language that is manipulated rather than the objects themselves, it is possible to eliminate such confounds by ensuring that the visual scene is held constant while the language is manipulated. In such cases, each object can be its own control—in the example used for Figs. 54.1–54.4, (different) participants in fact heard both ‘the man will eat the cake’ and ‘the woman will read the newspaper’. The plotted data show the aggregated data such that looks shown in the graph towards where the cake had been are the sum of looks towards where the cake had been during ‘the man will eat the cake’ and of looks towards where newspaper had been during ‘the woman will read the newspaper’. Comparing the curves for the cake and newspaper in Figs. 54.2 and 54.3 (or the bars in Fig. 54.4) is equivalent, therefore, to comparing looks to the cake during ‘. . . eat the cake’ with looks to the cake during ‘. . . read the newspaper’ (and conversely, of course, for looks to the newspaper); any difference between the cake and the other objects within the same scene that might cause it to attract fewer, or more, looks will therefore cancel out.

Perhaps surprisingly, of less concern is how precise the definition of the (spatial) regions of interest should be. For the case where the scene is concurrent with the language, the majority of fixations on an object will indeed be on that object—that is, the majority will land within a region defined by the external boundary of the object. Defining the region of interest as extending beyond the exact boundary of the object will change the absolute number of fixations that ‘count’ as landing on the object, but the overall patterns across conditions will be virtually unchanged. Some researchers therefore
choose to report ‘pixel analyses’ (the region’s boundary is the object’s boundary) or analyses in which fixations landing some arbitrary amount beyond the object’s edge are still included (cf. a ‘blob’ or rectangle drawn around the object to define the region of interest). Neither analysis is intended to make assumptions regarding the resolution of either the human visual system (e.g. how accurately fixations can be targeted) or the eye trackers deployed in such studies (e.g. how accurate the eye tracker is in reporting the actual screen coordinates on which the fixation was located). Some studies (e.g. Allopenna et al., 1998) depicted objects on a grid, and a fixation anywhere within the corresponding cell of the grid counted as a fixation on that object. Other studies (e.g. Huettig and Altmann, 2004) also depicted objects on a virtual grid (e.g. one object in each quadrant), but still the results were based on pixel analyses. There is no reason to suppose that the data patterns would differ at all if one kind of analysis were replaced with another. However, where it might matter is in respect of the increasingly used blank screen paradigm—the data shown in Figs. 54.2–54.4 are based on quadrant analyses. But what is the result of increasing the spatial gain and computing a finer-grain analysis of these same data? Figure 54.5 shows the results of increasing the gain to capture rectangular regions of interest as well as pixel analyses. The fixation plots use the unsynchronized variant in which fixations are synchronized to sentence onset (cf. Fig. 54.2). As can be seen, it makes very little difference to the pattern of fixations, albeit not to their absolute numbers, if the regions of interest are drawn one way or another. Most striking about these data are that they are from the blank screen paradigm; the fixations plotted in Fig. 54.5 were directed towards empty space that had previously been occupied by the depicted objects.

21 Statistical issues

As with any dependent measure, how to analyse differences in that measure as a function of the experimental manipulation is an issue that warrants careful attention. A variety of methods have been used to test for differences in fixations or saccades at moments of interest (e.g. at the offset of some critical word or phrase) or during intervals of interest. These include traditional ANOVA or t-tests on untransformed proportions (cf. the percentages shown in the plots above). A problem with these tests is that proportions are often skewed, and may therefore violate the assumptions that need to be met for such tests to be appropriate; ANOVA or t-tests on arcsin transformed proportions are therefore preferred. Recently, researchers have started to use hierarchical log linear analyses on raw numbers (based on chi-square distributions, and similar in many respects to chi-square except that they allow main effects and interactions to be tested). The first studies employing hierarchical log linear analyses included Huettig and Altmann (2005) and Knoeferle et al. (2005)—see Scheepers, 2003, for discussion of hierarchical log linear analyses more generally, and Altmann and Kamide, 2007 and 2009, for further discussion of their use in visual world analyses. One important difference between the ANOVA and chi-square families of tests is that they assume a different underlying distribution of the data—they are, in effect, different classes of statistical ‘model’ (see below). Typically, ANOVA is used more with reaction times, and chi-square with frequencies or counts. Even more recently, investigators have started to use mixed effects models. These are equivalent in some respects to log linear analyses but they tend to be used when multiple predictors need to be assessed (when, for example, they cannot be included in a standard factorial design). Mixed effects models are also advantageous as they allow for simultaneous treatment of subjects and items as random effects (see Baayen et al., 2008).

An entirely different class of statistical modelling needs to be carried out for analysing time-course data. For example, and with reference to Figs. 54.2 or 54.5, how can one determine that any pair of curves are different from one another? How can one determine where the peak is located for any such curve (given that aggregating data for the purposes of such plots hides the true underlying distribution of the data across subjects and trials)? And most importantly, perhaps, how can one model the dynamic changes to fixation proportions across time when successive time points are not independent of one another? This latter question has been addressed most recently using multilevel logistic regression (Barr, 2008b), which simultaneously handles the continuous variable of time and the
The mediation of eye movements by spoken language
categorical variable of gaze location. Magnuson et al. (2007) used growth curve analysis to address
similar statistical complexities in their time-course data. This technique ‘decomposes’ each curve
into a number of parameters which can then be compared across conditions. Perhaps surprisingly,
there is no standard statistical technique for determining when, in time, the probability of fixating
one object in the scene diverges from the probability of fixating another (but see Altmann (in press),
who proposes a simple method based on a variant of cluster-based Gaussian logic as employed in
neuroimaging).

A feature of time-course plots that is sometimes overlooked is that they combine information
about when a response is made with information about the likelihood of that response being made.
Fixation plots of the kind shown in Figs. 54.2 and 54.3 are, of course, aggregated across participants
and trials. For a given region of interest on a given trial, the participant will either be fixating that
region or not. The probability plots are not depicting an underlyingly continuous measure that is
continuously changing from one moment to the next (unlike participants’ weight, for example,
which is a continuous measure that can vary from one day to the next); at a given moment in time,
the eyes either are or are not fixating the region of interest. Thus, any hypothesis that attempts to
link the patterns depicted in time-course plots to changes in cognitive state must take account of
the underlyingly discontinuous nature of the individual behaviours that are aggregated across to
generate those plots.

Which statistical test to employ with eye movement data of these kinds depends on many different
factors, including the nature of the dependent variable (and the assumptions that should be made
about its underlying distribution; see above), the experimental design, the theoretical hypotheses,
the precedents for analysing particular data in a particular way, and the limits of the individual
researchers’ (and their statistical advisors’) knowledge. It is generally agreed that there is no ‘one size
fits all’. What is most critical is that the assumptions that need to be met by the data for one or other
analysis to be appropriate are indeed met (e.g. for parametric tests such as ANOVA, that the data are
normally distributed, are continuous or interval data, and so on). On the perspective that inferential
statistics are a form of model-fit (see, e.g. Field, 2009), it really does not matter which model is fit to
the data if the aim is simply to establish that there is a model that can be fit (sometimes it might be
appropriate to fit more than one model, for example when the data are particularly sparse, and to
look for convergent patterns across the models—see Altmann (in press)). Much depends on why the
model is being fitted to the data (which translates, simply, into why the inferential statistic is being
computed).

Summary: the future for language-mediated eye movements

Whereas Roger Cooper first observed the tight temporal coupling between language and eye move-
ments (meaning that language could have a seemingly immediate influence on the bias to look
towards one part of space or another), Michael Tanenhaus and his colleagues were able to ground
that observation in developments in the vision sciences that were beginning to explore eye move-
ments as a goal-directed behaviour (e.g. Ballard et al., 1997; see also Land et al., 1999). Instead of
tracking participants’ eye movements as they manipulated objects with the goal of connecting pieces
together into some larger-scale object as determined by an instructional diagram, Tanenhaus and
colleagues tracked participants’ eye movements as they manipulated objects with the goal of moving
pieces around as determined by a verbal instruction. Tanenhaus’s insight was that language-directed
eye movements are goal-directed eye movements. The goals are determined by the task—either
explicitly given, or brought by the participant into the experiment (just as, when speaking, we might
dynamically adopt different communicative goals). What made language-mediated eye movements
so interesting for psycholinguists is that the indeterminacies of the linguistic signal manifested in
uncertainties with respect to what the immediate attentional goals should be (for example, whether
to attend to a beaker or to a beetle during the first moments of the word ‘beaker’, or whether to
attend to Mickey or to Donald on hearing ‘he’). The field blossomed with epidemic speed (but with-
out the negative connotations usually associated with epidemics), in large part because language is
indeterminate at so many different levels of analysis, with its indeterminacies changing dynamically as a word or sentence unfolds. The scope for exploring these indeterminacies, using a new paradigm that is sensitive in close to real time to psychological processes relevant to their resolution, seemed limited only by the number of eye trackers available to the field (undoubtedly, a part of the field’s blossoming was due to the availability of ever-cheaper and more user-friendly eye tracking devices, associated software, and analysis tools). As the field has matured, so researchers are beginning to reflect on why language has this mediating influence on visual attention (see the earlier section ‘How, and why, language mediates eye movements’). But the consideration of language-mediated eye movements not as a psycholinguistic tool, but as reflecting a deeper connection between language, perception, cognition, and action, is still in its infancy. Beyond its continuing development in service of psycholinguistic research, it is in this direction that the field is likely to develop in new ways—borrowing theoretical tools that already exist beyond the confines of psycholinguistics to understand better the relationship between language and cognition. Theoretical frameworks such as those developed in the context of cognitive control and attention—how task-relevant features become more activated than task-irrelevant features (e.g. Cohen et al., 2004)—are likely to provide a bridge between the many disparate features of the paradigm. Such frameworks are likely to lead also to a greater understanding not simply of the mechanisms underlying the linkage between language and eye movements (see ‘How, and why, language mediates eye movements’ section) but also of the task-specificity of this mediation (see ‘Task-sensitivity and language-mediated eye movements’ section). As suggested earlier, the fact that language mediates eye movements at all is not so surprising. What is more surprising is that ‘higher-level’ cognitive factors can penetrate so deeply into what might be presumed to be ‘lower-level’ processes controlling oculomotor behaviour (cf. Liu’s work on language-mediation of smooth-pursuit eye movements (Liu, 2009)). A natural question to consider, then, is how language interacts, in different task contexts, with other factors that constrain oculomotor control. For example, global and remote distractor effects (see Cruickshank and McSorley, 2009, and references therein), systematic biases favouring eye movements in one direction rather than another (e.g. Tatler and Vincent, 2008), or factors influencing visual salience and its impact on overt and covert attention (e.g. Itti and Koch, 2000; Parkhurst et al., 2002). One may even ask, again, the question posed at the start of this review; how, and under what conditions, do eye movements in service of language differ from those in service of reaching movements? A final question that no doubt will receive attention will concern the neural pathways implicated in the interaction between language and oculomotor control (most likely there will be more than one).

The study of language-mediated eye movements has become a fundamental tool in the study of psycholinguistic process. It is fast becoming a fundamental tool in the study of cognitive processes more generally. The past 15 years have seen extraordinary growth accompanied by real advances in our understanding of the cognitive processes with which language makes contact. The next 15 years will surely be as extraordinary.

Acknowledgements

This review would not have been possible without the original contributions of the authors whose work I refer to. Any omissions do not reflect a value judgement, but most likely reflect a lapse of memory and limited time in which to research the ever-growing field. Thanks are due to Yuki Kamide, Jelena Mirkovic, and Silvia Gennari for encouragement and suggestions, the Economic and Social Research Council (ref. RES-063-27-0138) for funding the time in which this article was written, the editors of this volume for their patience in awaiting the arrival of this manuscript, and two anonymous reviewers for their insightful comments on an earlier version of the manuscript.

References

3. Altmann, G.T.M. (in press). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. Acta Psychologica. Special Issue: Visual search and visual world: Bridging language, memory and attention.
The mediation of eye movements by spoken language


The mediation of eye movements by spoken language · 1003
