Published in final edited form as:

Psychol Sci. 2000 July ; 11(4): 274–279. doi:10.1111/1467-9280.00255.

What the Eyes Say about Speaking

Zenzi M. Griffin and **Kathryn Bock**

University of Illinois at Urbana-Champaign

Abstract

To study the time course of sentence formulation, we monitored the eye movements of speakers as they described simple events. The similarity between speakers' initial eye movements and those of observers performing a non-verbal event comprehension task suggested that response-relevant information was rapidly extracted from scenes, allowing speakers to select grammatical subjects based on comprehended events rather than visual salience. When speaking extemporaneously, speakers began fixating pictured elements less than a second before naming them within their descriptions, consistent with incremental lexical encoding. Eye movements anticipated the order of mention despite changes in picture orientation, in who-did-what-to-whom, and in sentence structure. The results support Wundt's theory of sentence production.

> From a psychological point of view, the sentence is both a simultaneous and a sequential structure. It is simultaneous because at each moment it is present in conscious-ness as a totality even though individual subordinate elements may occasionally disappear from it. It is sequential because the configuration changes from moment to moment in its cognitive condition as individual constituents move into the focus of attention and out again one after the other
>
> (Wundt, 1900/1970, p. 21).

Wundt's ideas about sentence production were at the center of an epic debate between a psychologist, Wundt himself, and the linguist Hermann Paul about the nature of language and its relation to thought. The claim that sentence production consists of a wholistic conceptualization followed by the sequential expression of linguistic constituents came in direct reaction to Paul's contention (1886/1970) that sentences are the sums of their parts, originating in sequential associations among individual concepts that are outwardly manifested as a series of words (see Blumenthal, 1970, for review). Wundt's arguments about sentences re-emerged at mid-century in Lashley's (1951) classic analysis of serial order in behavior, once again in reaction to associative accounts of sequenced action.

The longevity of the issues notwithstanding, we know little more than Wundt and Lashley did about the conceptual precursors of meaningful connected speech. Although considerable progress has been made in tracing the internal structure of the language production system and the cognitive and neurophysiological details of single-word production (Dell, Schwartz,

---

Correspondence should be addressed to Zenzi M. Griffin (griffin@psych.stanford.edu), who is now at the Department of Psychology, Stanford University, Stanford, CA 94305-2130.

Some of the data were presented originally at the meeting of the CUNY Sentence Processing Conference at Rutgers University in March, 1998.

Martin, Saffran, & Gagnon, 1997; Garrett, 1988; Levelt, 1989; Levelt, Roelofs, & Meyer, 1999), the transition between thinking and speaking has remained a mystery. This is for reasons that Wundt anticipated in his writings about language: Most of the processes of language production are inaccessible to the standard tactics of psychological research.

The experiment reported here was conceived as a step toward illuminating the cognitive events that give rise to simple sentences in simple situations. Speakers were asked to describe pictured events with single sentences. We used speakers' eye movements to diagnose the temporal relationships among event apprehension (extracting a coarse understanding of the event as a whole), sentence formulation (the cognitive preparation of linguistic elements, including retrieving and arranging words), and speech execution (overt production). Such processes may be reflected in eye movements because people tend to look at what they are thinking about (e.g., Rayner, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). When naming objects, speakers fixate the objects long enough to recover the sounds of the words that denote them (Meyer, Sleiderink, & Levelt, 1998). If the eye movements made while preparing a sentence reflect both the conceptual and linguistic processing of a pictured event, the trajectory and timing of eye movements relative to the content and timing of sentence constituents should strongly constrain inferences about the time course of sentence planning from apprehension to execution.

We asked four groups of participants to inspect line drawings of simple events while their eye movements were monitored. One group described the pictured events while they were in view (extemporaneous speech). Another group viewed the same events while preparing descriptions that were produced after the pictures disappeared (prepared speech). A comparison of these two groups allowed us to diagnose differences in conceptual preparation for extemporaneous and prepared speech, and in particular to infer how much and what kinds of sentence formulation preceded extemporaneous speech. To assess the amount of visual information needed for apprehension, a third group of participants viewed the pictures with the goal of finding a person or thing being acted on in each event (i.e., the patient of the action; Fillmore, 1968). Because a patient cannot be reliably identified without extracting the causal structure of an event, this silent patient detection task provided an estimate of the amount and kind of viewing needed for apprehension. The fourth group of participants viewed the pictures without any immediate or specific task requirements (inspection). Insofar as particular picture elements might strongly attract attention, the inspectors' eye movement records should reflect it. To a rough approximation, we expected apprehension to occur during viewing in the detection and inspection tasks, both apprehension and sentence formulation to occur during viewing in the prepared-speech task, and apprehension, formulation, and execution to occur during viewing in extemporaneous speech.

At issue was whether and how apprehension and formulation are interleaved with execution. On the Wundt-Lashley view, apprehension must precede formulation to provide the wholistic conception that supports the creation of a sentence. This predicts that speakers should inspect events well enough to coarsely code them prior to initiating the sentence formulation process. On Paul's view, apprehension and formulation may go hand in hand. If so, the eye movements indicating word retrieval for a grammatical subject could precede a

thorough inspection of a scene. Regarding the relationship between formulation and execution, if the part of sentence formulation that is responsible for lexical encoding is incremental (e.g., Bock, 1982; Kempen & Hoenkamp, 1987; Lindsley, 1975), eye movements should indicate that word selection and execution overlap. In contrast, if only phonological encoding is incremental (Meyer, 1996), there should be evidence that all words are selected before speech begins.

Method

Participants

Native speakers of American English aged 18 to 30 years were recruited from the University of Illinois community. All participants reported normal or corrected-to-normal vision. They received $5 or credit in introductory psychology courses. There were 20 participants in the extemporaneous speech condition, 12 in prepared speech, 8 in detection, and 8 in inspection.

Apparatus

A binocular EyeLink head-mounted eyetracker (SR Research Ltd.) controlled by a Gateway 2000 P5-120 computer recorded eye movements with a temporal resolution of 4 ms and spatial resolution of approximately $0.\bar{5}$. Another Gateway computer controlled the presentation of pictures and digital recording of speech. The digitized pictures were displayed on a 21 inch ViewSonic P815 monitor. Four reflectors at the corners of the monitor provided references for the eyetracker's head-position camera. A hand-held button box was used for manual responses, and a tie-clip microphone for voice recording.

Materials

The experimental pictures were black-and-white line drawings of simple transitive events, selected to elicit reliable descriptions in a preliminary norming study. There were four versions for each of eight events (see Figure 1). Two of the versions switched the elements that performed and underwent the actions (agents and patients, respectively), to control the perceptual, conceptual, and lexical properties of the elements. We will refer to these as the Original and Role-Traded versions. Both versions were then mirror-imaged to counterbalance left-to-right scanning preferences (Buswell, 1935).

The experimental pictures depicted two types of events. Active events elicited predominantly active sentences in the experiment, regardless of which element was the agent. Passive/Active events included a human that was consistently used as the grammatical subject, eliciting passive sentences when the human was the patient (Original version) rather than the agent (Role-traded version) but without changing the order in which event elements were mentioned. Figure 1 illustrates picture sets of the two event types. There were 5 Active and 3 Passive/Active sets. An additional 17 pictures served as example, practice, and filler items. These pictures depicted events that elicited intransitive descriptions (sentences without objects; e.g., A baseball player is bunting), although some the events included patients (e.g., the baseball).

The four versions of each experimental picture were distributed across four lists, with each list containing an equal number of mirror-imaged, original and role-traded picture versions. Each participant viewed one list. Within tasks, participants were divided equally among the four lists.

Procedure

Participants were instructed and tested on two printed example pictures before being equipped with the eye tracker. Both extemporaneous and prepared speakers were told to describe each pictured event in one sentence without pronouns, and to press a button on the button box at the end of the description. In addition, prepared speakers were instructed to press a button when they were ready to speak, which also caused the picture to disappear. To encourage normal formulation, speakers received no guidelines about the form or content of their descriptions and speed was not mentioned. Participants in the detection task were asked to locate the "victim" in each picture by fixating it and pressing a button as quickly and accurately as possible.[1] For purposes of the task, a victim was defined as something directly undergoing the action, regardless of whether it was animate or inanimate. An example illustrated the relevant notion with a picture of a man juggling, where the juggled balls constituted the victim. If no victim was present, the participants fixated a point at the top center of the computer screen and pressed the button. Inspectors were asked to examine the pictures to get a sense of their range in content and quality, ostensibly in preparation for an upcoming aesthetic judgment task. The inspectors pressed the button to move through the sequence.

After instruction, participants donned the eyetracker and went through a 9-point calibration and validation procedure. They viewed the monitor from a distance of approximately 81 cm, with the pictures subtending a $26.\bar{5}$ horizontal visual angle. After calibration, the instructions were briefly repeated and four practice trials presented, followed by the experimental trials.

To start each trial, participants fixated a point in the upper center of the computer screen and pressed a button to validate the fixation. Participants then performed their assigned tasks. Eye-movement and voice recording began at the start of each trial and continued until 500 ms after the end-of-trial button presses. Eye movements were recorded from both eyes but analyzed for the left eye only.

Speech coding

Voice recordings were transcribed and parsed by assistants blind to experimental conditions. Onsets and offsets of speech, subject nouns, and object nouns were measured for each picture description. The transcribed descriptions were categorized as modal or deviant in content and as fluent or nonfluent in execution. Deviant sentences lacked verbs or objects. Nonfluent utterances contained filled or silent pauses, stressed articles (Fox Tree & Clark, 1997), or semantically empty onsets (e.g., There is). Extemporaneous and prepared descriptions were modal and fluent on 48% and 52% of the trials, respectively. Active events

[1]Speed was emphasized in this task because, unlike simple event description, it is an unusual task from which we aimed to infer the minimum time required for event apprehension.

elicited modal actives on 84.3% of all trials. Original versions of Passive/Active events elicited modal passives or other patient-subject sentences (e.g., The mailman is running from the dog) on 83.3% of trials and their Role-traded versions elicited modal actives on 85.4% of trials.

Eye-movement analysis

The EyeLink system software identified and measured the durations of all eye fixations between saccades (defined as changes in eye position that covered more than .1$\bar{5}$ at a velocity greater than 3$\bar{0}$/sec with acceleration greater than 800$\bar{0}$/sec2). Fixation locations were categorized with respect to regions occupied by event elements, using eye-movement analysis software (Maciukenas, Althoff, Holden, Webb, & Cohen, 1997). Every experimental picture had agent and patient regions that encompassed the corresponding elements with a surrounding margin of about $\bar{2}$. Action regions included any instrument of an action or the space between an agent and patient. Within regions, individual fixations were aggregated into gazes, reflecting the cumulative fixation durations within a region before leaving it.

In the patient-detection condition, only valid trials (93.8% of the total) were analyzed. Valid trials were those in which the fixation that accompanied the end-of-trial button press was in the patient or action region, and outside the agent region.

Statistical analyses

Apart from the eye-movement measures, the dependent variables were response latencies and speech content. In the extemporaneous- and prepared-speech conditions, response latencies were speech initiation times; in patient-detection and inspection, response latencies reflected the time to press the button. Means were calculated over participants and items, pooling data from mirror-imaged pictures as single items. The stability of differences between means is indicated in terms of Tukey's HSD (95% confidence intervals for post-hoc paired comparisons; Winer, Brown, & Michels, 1991), computed using the corresponding error terms from analyses of variance for participants.

Results and Discussion

Figure 2 gives a representative example of the timing of eye fixations and accompanying speech on one trial for a single speaker in the extemporaneous condition. To infer the overall time course for extracting information about events and for using the information during sentence formulation and production, we compared the timing and trajectories of selective fixations to agents and patients across different tasks. The changes in eye fixations over time for each task can be seen in Figure 3. Our interest was in (a) whether speakers' eye-movements were guided by an overall apprehension of the event or by the salience of individual elements within it (considering the inspection task to be most likely to be sensitive to salience alone), (b) whether apprehension preceded formulation (comparing performance in the detection, extemporaneous, and prepared speech tasks), and (c) how formulation was interleaved with execution (comparing performance in the prepared and extemporaneous speaking tasks).

Inspection

The inspectors did not systematically distinguish event regions early in viewing. Only at 1300 ms after picture onset did one region attract significantly more fixation time than the other, and the difference favored patients over agents. With regions defined in terms of the elements that served as subjects and objects in modal descriptions, inspectors briefly preferred typical object over subject referents starting at 1508 ms. This implies that no event- or speech-related region systematically attracted attention early in picture viewing, and reduces the plausibility of a salience-driven account of sentence-subject selection for these events (Osgood, 1971).

Apprehension and formulation

To assess whether the extemporaneous speakers normally extracted an event's causal structure prior to initiating speech, we compared (a) the points at which fixations to the agent and patient regions began to diverge in the patient-detection task, which demanded event comprehension to (b) the corresponding divergences for the referents of the subject and object noun phrases in extemporaneous speech. During detection, fixations to the patient began to diverge from fixations to the agent at 288 ms after picture onset and reached significance (with the alpha level adjusted for multiple comparisons) at 456 ms, $z(60) > 3.10$, $p < .001$. For extemporaneous speakers, the divergence between the pictured elements began at 316 ms and reached significance at 336 ms. The overt response times for the groups were also comparable: Patient-detectors took 1690 ms to indicate that they had located the patients in the events, and speakers started to talk at 1686 ms after picture onset (HSD=370). These similarities suggest that speakers rapidly extracted the event structure of the pictures.

Following apprehension, the eye movements in extemporaneous speech appear to have been driven by a linguistic formulation process. Strong evidence for this is shown in Figure 4. The figure summarizes the eye movement patterns during fluent extemporaneous speaking relative to the onset of the subject noun. The data represent the mean proportion of time within successive 50 ms intervals spent within regions corresponding to the ensuing sentence-subjects and objects, averaged over trials.

Figure 4 suggests a very orderly linkage between successive fixations in viewing and word order in speech. The sequential dependencies were assessed by comparing the time spent fixating agents and patients prior to subject onset and during the descriptions of Original and Role-traded versions of Active events. This contrast held constant the picture elements, content words, and sentence structures, varying only the event roles played by the picture elements. Significant interactions between event roles and time period indicated that speakers spent significantly more time fixating agents before subject onset than during speech (646 and 179 ms, respectively) but more time on patients during speech than before subject onset (244 and 812 ms; HSD = 281). The absence of interactions with picture version indicated that this held for both versions ($Fs < 1$). Significant three-way interactions between event role, time period, and picture version in the analyses of Passive/Active pictures imply that speakers did not simply follow the causal structure of the events by fixating agents early and patients later. Rather when patients were encoded as sentential subjects, they were fixated longer before subject onset than after (1123 and 503 ms) whereas

agents were fixated less before subject onset than during speech (150 and 1041). Thus, the distribution of fixation times anticipated the order of mention regardless of sentence structure.

Further support for the inference that a linguistic formulation process followed the divergence in eye fixations came from comparisons between the extemporaneous and prepared speakers. Like the extemporaneous speakers, the prepared speakers' fixations to the subject and object referents diverged early (at 304 ms, becoming significant at 472 ms). For both groups, the divergence marked the onset of sustained attention to the region corresponding to the eventual sentence-subject, with corresponding inattention to the eventual object (see Figure 3).

The early onset of selective eye fixations for speakers was not attributable to making the first-fixated picture element the subject. In both extemporaneous and prepared speech, the probabilities of initially fixating regions associated with subjects, objects, or actions did not differ significantly, (2s < 1. Rather, the early divergence reflects short initial gazes to object and action regions and longer gazes to subject regions.

In summary, the overlap in onset times for selective, response-relevant eye movement activity by the speakers and patient-detectors suggests rapid apprehension of events (consistent with the results of Biederman, Mezzanotte, & Rabinowitz, 1982; Gordon, 1999; Potter, 1975). Modal subject selection for these pictures depended on knowledge of the relative humanness of the event elements and their roles, implying that speakers directed their gazes to eventual sentential subjects based on their apprehension of the events. In extemporaneous speaking, apprehension preceded a formulation process whose linguistic nature was revealed in strong dependencies between eye fixations and sentence elements. These findings support the Wundt-Lashley hypothesis about the nature of sentence production.

Formulation and execution

One measure of the temporal linkage between formulation and speech execution is the eye-voice span. This is the mean amount of time that elapsed between (a) the onset of the last gaze to an agent or patient region prior to utterance of the noun denoting the element in the region and (b) the onset of the spoken noun denoting the element. For subject and object nouns the respective spans were 902 ms and 932 ms (HSD=176). These eye-voice spans closely resemble the 910 ms mean picture-naming latencies for isolated objects (Snodgrass & Yuditsky, 1996) and suggest that extemporaneous speakers incrementally selected and phonologically encoded nouns.

Comparisons between extemporaneous and prepared speech revealed, unsurprisingly, that prepared speech tended to be more fluent. The prepared speakers had more time to formulate their utterances (with 4049 ms before speech onset, compared to 1686 ms for the extemporaneous speakers, HSD = 675) and spent as much time fixating the objects of their sentences before speaking as extemporaneous speakers spent fixating sentential subjects (890 and 824 ms, respectively, HSD = 223). Within disfluent utterances, disruptions in prepared speech were shorter than those in extemporaneous speech (279 to 554 ms, HSD =

202). This reduction in fluency suggests competition between formulation and execution in extemporaneous speech.

Conclusion

A broad view of simple, fluent sentence production is suggested by our results. The evidence that apprehension preceded formulation, seen in both in event comprehension times and the dependency of grammatical role assignments on the conceptual features of major event elements, argues that a wholistic process of conceptualization set the stage for the creation of a to-be-spoken sentence. Of course, the data reflect only the most basic kind of sentence formulation, involving perceptual input and the production of single clauses. All of the events were simple transitive actions, and there were just eight of them, selected to elicit particular utterances and sentence structures. Only one language was used, English, which has more than its fair share of idiosyncrasies. But with these restrictions, the results point to a language production process that begins with apprehension or the generation of a message and proceeds through incremental formulation of sentences. These two facets of the production process, although suspected by observers going back to Wundt, have never before been delineated as directly as in the present circumstances. The findings argue against the strongly associative accounts of language production that likewise go back to Wundt's contemporaries. The observations not only show a systematic temporal linkage between eye movements and the contents of spoken utterances, but also offer new evidence for a tight coupling between the eye and the mind, and lay the groundwork for powerful tools to explore how thought becomes speech.

Acknowledgments

This research formed part of the first author's Ph.D. dissertation at the University of Illinois at Urbana-Champaign. It was supported in part by research and training grants from the National Institutes of Health (R01 HD21011 and T32 MH18990), the National Science Foundation (SBR 94-11627, SBR 98-73450), and the Beckman Institute. We thank Neal Cohen, Gary S. Dell, Adele Goldberg, David E. Irwin, Arthur Kramer, Gordon Logan, Roger Marsh, George McConkie, and Daniel H. Spieler for their many contributions to this work, and Michael Vendel and Leslie Smith for assistance in transcribing and measuring speech.

References

Biederman I, Mezzanotte RJ, Rabinowitz JC. Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology. 1982; 14:143–177. [PubMed: 7083801]

Blumenthal, AL. Language and psychology: Historical aspects of psycholinguistics. New York: Wiley; 1970.

Bock JK. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. Psychological Review. 1982; 89:1–47.

Buswell, GT. How people look at pictures. Chicago, IL: University of Chicago Press; 1935.

Dell GS, Schwartz MF, Martin N, Saffran EM, Gagnon DA. Lexical access in aphasic and nonaphasic speakers. Psychological Review. 1997; 104:801–838. [PubMed: 9337631]

Fox Tree JE, Clark HH. Pronouncing 'the' as 'thee' to signal problems in speaking. Cognition. 1997; 62:151–167. [PubMed: 9141905]

Fillmore, CJ. The case for case. In: Bach, E., Harms, RT., editors. Universals in linguistic theory. New York: Holt, Rinehart and Winston; 1968. p. 1-88.

Gordon, R. Unpublished dissertation. University of Illinois; Urbana-Champaign: 1999. The time course of attention during scene perception.

Garrett, MF. Processes in language production. In: Newmeyer, FJ., editor. Linguistics: The Cambridge survey, III: Language: Psychological and biological aspects. Cambridge, England: Cambridge University Press; 1988. p. 69-96.

Kempen G, Hoenkamp E. An incremental procedural grammar for sentence formulation. Cognitive Science. 1987; 11:201–258.

Lashley, KS. The problem of serial order in behavior. In: Jeffress, LA., editor. Cerebral mechanisms in behavior. New York: John Wiley and Sons; 1951. p. 112-136.

Levelt, WJM. Speaking: From intention to articulation. Cambridge, MA: MIT Press; 1989.

Levelt WJM, Roelofs A, Meyer AS. A theory of lexical access in speech production. Behavioral and Brain Sciences. 1999; 22:1–75. [PubMed: 11301520]

Lindsley JR. Producing simple utterances: How far ahead do we plan? Cognitive Psychology. 1975; 7:1–19.

Maciukenas, MA., Althoff, RR., Holden, JA., Webb, AG., Cohen, NJ. EMTool (Version 2.0). Beckman Institute for Advanced Science and Technology, University of Illinois; Urbana, Illinois: 1997.

Meyer AS. Lexical access in phrase and sentence production: Results from picture-word interference experiments. Journal of Memory and Language. 1996; 35:477–496.

Meyer AS, Sleiderink A, Levelt WJM. Viewing and naming objects: Eye movements during noun phrase production. Cognition. 1998; 66:B25–B33. [PubMed: 9677766]

Osgood, CE. Where do sentences come from?. In: Steinberg, DD., Jakobovits, LA., editors. Semantics: An interdisciplinary reader in philosophy, linguistics and psychology. Cambridge, England: Cambridge University Press; 1971. p. 497-529.

Paul, H. The sentence as the expression of the combination of several ideas. In: Blumenthal, AL., translator. Language and psychology: Historical aspects of psycholinguistics. New York: Wiley; 1970. p. 34-37.(Original work published 1886)

Potter MC. Meaning in visual search. Science. 1975; 187:965–966. [PubMed: 1145183]

Rayner K. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin. 1998; 124:372–422. [PubMed: 9849112]

Snodgrass JG, Yuditsky T. Naming times for the Snodgrass and Vanderwart pictures. Behavior Research Methods, Instruments, & Computers. 1996; 28:516–536.

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. Science. 1995; 268:1632–1634. [PubMed: 7777863]

Winer, BJ., Brown, DR., Michels, KM. Statistical Principles in Experimental Design. 3. New York: McGraw-Hill; 1991.

Wundt, W. The psychology of the sentence. In: Blumenthal, AL., translator. Language and psychology: Historical aspects of psycholinguistics. New York: Wiley; 1970. p. 20-31.(Original work published 1900)

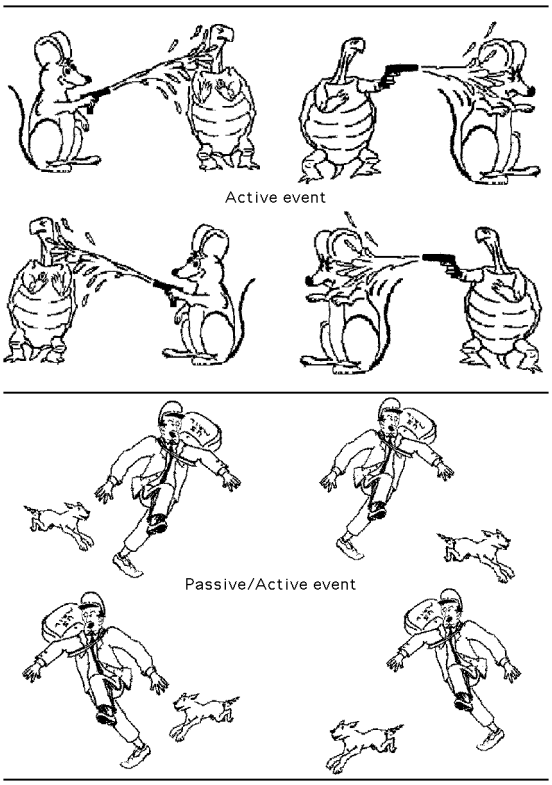

Figure 1.

Top panel shows a sample Active picture set, typically described with active sentences in all versions (*The mouse is squirting the turtle with water* and *The turtle is squirting the mouse with water*); bottom panel shows a sample Passive/Active picture set (*The mailman is being chased by the dog* and *The mailman is chasing the dog*). Within each set, the upper pictures are the Original and Role-Traded versions and the lower are their mirror-images.

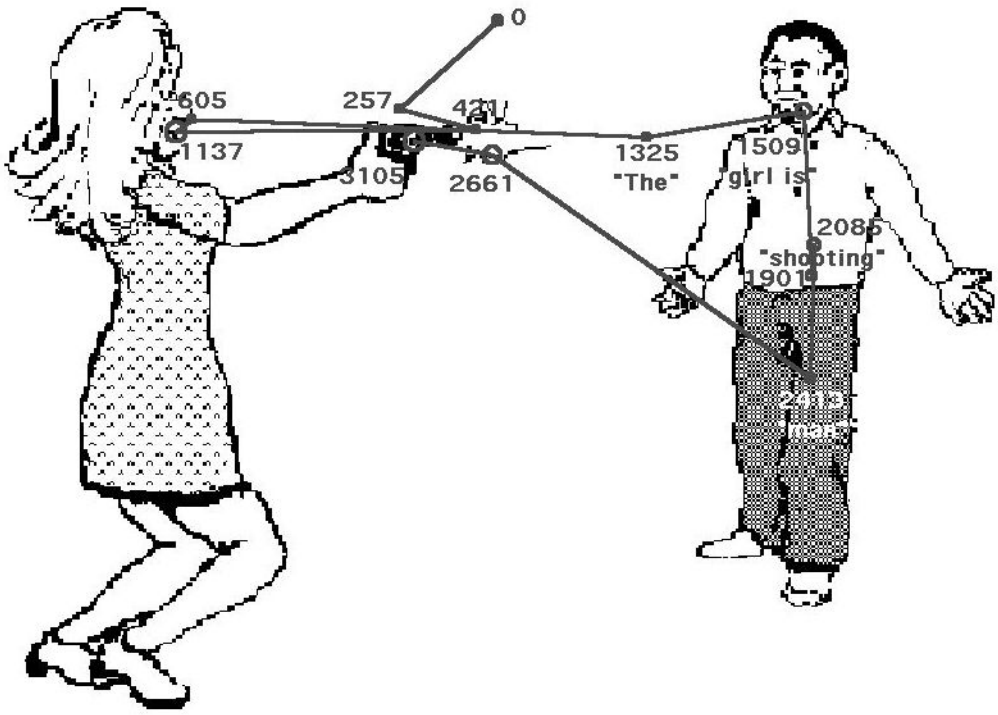

Figure 2.

One speaker's eye movements over a pictured event. Successive fixations, time-stamped in ms from picture onset, show where the eye rested with the size of the circle indicating the length of fixation (starting at the location of the fixation point that preceded picture presentation). The word "girl" began 917 ms after the first fixation on the woman and the word "man" began 843 ms after the man was first fixated.

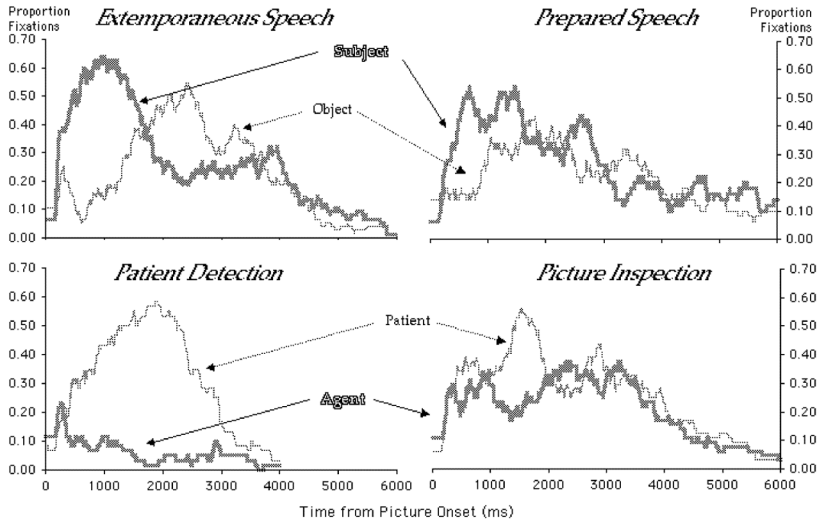

Figure 3.

Changes in viewing across successive 4 ms intervals from picture onset during (a) Extemporaneous Speech, (b) Prepared Speech, (c) Patient Detection, and (d) Picture Inspection for all picture types and versions. Shown are the proportions, for each interval, of trials in which regions were fixated.

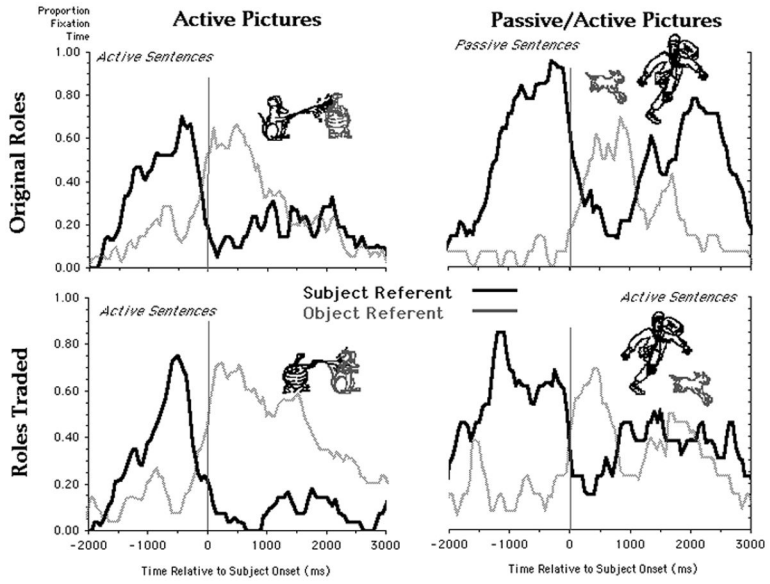

Figure 4.

Changes in viewing across successive 50 ms intervals during extemporaneous speech for the major elements of Active events (left) and Passive/Active events (right) in their Original (top) and Role-traded (bottom) forms, collapsed over mirror-imaged versions. Time is shown relative to the onset of the head of the subject noun phrase.