Acta Psychologica 137 (2011) 208–216

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

I see what you're saying: The integration of complex speech and scenes during language comprehension

Richard Andersson [a], Fernanda Ferreira [b], John M. Henderson [b,*]

[a] Lund University Cognitive Science, Lund University, Sweden
[b] Department of Psychology, University of South Carolina, United States

ARTICLE INFO

Article history:
Received 29 March 2010
Received in revised form 7 January 2011
Accepted 11 January 2011
Available online 8 February 2011

PsycINFO classification:
2720 Linguistics & Language & Speech
2326 Auditory & Speech Perception
2346 Attention

Keywords:
Language comprehension
Scene perception
Eye movements
Attention
Visual world

ABSTRACT

The effect of language-driven eye movements in a visual scene with concurrent speech was examined using complex linguistic stimuli and complex scenes. The processing demands were manipulated using speech rate and the temporal distance between mentioned objects. This experiment differs from previous research by using complex photographic scenes, three-sentence utterances and mentioning four target objects. The main finding was that objects that are more slowly mentioned, more evenly placed and isolated in the speech stream are more likely to be fixated after having been mentioned and are fixated faster. Surprisingly, even objects mentioned in the most demanding conditions still show an effect of language-driven eye-movements. This supports research using concurrent speech and visual scenes, and shows that the behavior of matching visual and linguistic information is likely to generalize to language situations of high information load.

One powerful method for investigating the integration of language and vision is the practice of monitoring the eye movements people make as they listen to speech while simultaneously looking at a visual world containing relevant objects. This technique allows psycholinguists to study how information sources are integrated in real-time to allow comprehenders to form interpretations and link linguistic forms to real-world referents (see Tanenhaus & Brown-Schmidt, 2008, for a review). For example, research has shown that listeners use the visual scene context to constrain the set of possible target referents (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Knoeferle, Crocker, Scheepers, & Pickering, 2005). Altmann and Kamide (1999) showed that listeners use verb information to anticipate a postverbal object, and they later demonstrated the use of real-world information as well (Kamide, Altmann, & Haywood, 2003; see also Ferreira & Tanenhaus, 2007).

These studies demonstrate that linguistic interpretations are used to guide the eyes almost immediately to relevant objects in the visual world. Moreover, listeners are highly likely to fixate an object within about a one-second window following the onset of a word, even when nothing about the task seems to demand that the word and the object be linked. What accounts for this tendency to fixate on objects mentioned in speech? One possibility is that this link allows the comprehender to form a much richer and detailed representation than would be possible otherwise (see e.g., Altmann & Kamide, 2007; Ferreira, Apel, & Henderson, 2008; Richardson, Altmann, Spivey, & Hoover, 2009).

To understand the nature of the eye movements in the so-called Visual World Paradigm (Tanenhaus, Spivey, Eberhard, & Sedivy, 1995) and the strength of this link, it is important to conduct investigations using stimuli that are sufficiently complex to tax the language–vision interface. This is necessary in order to see whether this link weakens in demanding language situations, for example by the comprehender prioritizing processing resources elsewhere. Up to now, most experiments have involved the presentation of a single sentence per trial, and typically only one word in that sentence is identified as a potential target of eye movements. In natural speech, of course, people often hear multiple sentences containing several objects that may be of interest and may therefore become the target of an eye movement. In addition, many of the stimuli that have been presented have been simple line drawings of scenes, or scenes created from pasting clip-art images together in such a way that an event such as a wizard painting a princess is strongly implied. A simple display may allow the participant to preview all objects and possible targets, subvocalize them, and thus pre-generate the linguistic labels that may appear in the speech (for visual search, see Zelinsky & Murphy, 2000, but see also Dahan & Tanenhaus, 2005). Conscious encoding of the objects by the participants is normally disregarded (Tanenhaus, Magnuson, Dahan, & Chambers, 2000:564), but still, typical stimuli

* Corresponding author. Tel.: +1 803 777 41 37.
E-mail address: jhender@mailbox.sc.edu (J.M. Henderson).

0001-6918/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2011.01.007

displays in the visual world paradigm contain clearly identifiable objects in limited numbers, which provide every possibility to do precisely this pre-processing. As well as the flow of information can move from phonological form to visual form, it may as well move in the opposite direction (see Huettig & McQueen, 2007, for a discussion). The pre-processing may also involve memorizing the object locations or visual aspects of the objects. This would imply that simple displays have a processing advantage compared to complex scenes which do not allow this pre-processing.

However, there are studies using real-world objects as targets which have investigated the effect of somewhat complex scenes, but also with limitations to the demand on the language–vision interface. For example, a set-up by Hanna and Tanenhaus (2004) used 10 possible visual targets and referents, but allowed the participant to preview all objects and keep them highly activated. Similarly, a study by Brown-Schmidt, Campana, and Tanenhaus (2005) used a 5×56 grid of possible referential targets. However, the study used only four participant pairs (who may not be representative) and the same visual scene was used throughout the entire experiment (~2.5 h), allowing participants to become more and more familiar with the display and allowing gradually reduced complexity as portions of the display were used up. A study by Brown-Schmidt and Tanenhaus (2008) used an irregular display of 57 different objects and showed how a conversation, as opposed to merely calling out the names of the objects, helps to restrict the referential domain. The authors identify the proximity, relevance and recency of referents as helpful factors in restricting the referential domain. In this experiment, however, the display was semi-permanent in the sense that the available game board was always present and all objects to be used, except one, were also present (either as blocks or stickers). This allowed for a continuity in the visual scene and as such, the display was not as complex as an equivalent display of 57 objects where the object types are freshly generated every trial. Many real scenes are quite different (see Henderson & Ferreira, 2004, for discussion), as the reader can verify by simply looking around his or her immediate environment. Scenes may contain almost uncountable numbers of objects, some predictable, but many not, and often only temporary present never to return again. And in a situation in which objects in the scene are mentioned in speech, a very large proportion of the scene content will be irrelevant to the utterance, or at least will not be mentioned. As a result, the comprehender attempting to link words and objects in the world may have a far more demanding task than has so far been considered in visual world experiments: Utterances are multi-sentence and may contain multiple referents; and scenes are complex and may contain hundreds or thousands of objects, only a few of which are relevant at a given moment in linguistic processing. This is not to say that *all* scenes and utterances are complex, but they represent a subset of the possible scene and utterance combinations that we believe has been neglected.

Of course, it is also important to note that the properties of real-world utterances and scenes do not only make the situation for the comprehender more challenging; they may also make the task easier, because natural stimuli are constrained in ways that likely facilitate processing. For example, connected sentences tend to be coherent, and so a series of utterances may help to converge on the possibility that a particular object will soon be mentioned; and real scenes allow the rapid extraction of gist (e.g., this is a playground scene), allowing listeners to anticipate which object will be mentioned and where in the scene it is likely to be found (Castelhano & Henderson, 2007; Torralba, Oliva, Castelhano, & Henderson, 2006). Also, as shown by Brown-Schmidt and Tanenhaus (2008), a real two-way conversation may help to restrict the referential domain.

To understand to what extent people look at objects when they are mentioned in extremely complex settings, we conducted a study in which participants viewed photographs of complex real-world scenes. A representative example is shown in Fig. 1. The scenes contained a large number of objects arranged in a typically cluttered and busy manner.

The linguistic material presented to participants was also more complex than in typical studies, consisting of three sentences, the second of which was designated as the target sentence. These passages were spoken at either a slow or fast rate of speech. The purpose of this rate manipulation was as to allow the participant less or more time to navigate the scene and find the target object. This added visual search task on top of the linguistic processing task served to increase the information processing demands. Moreover, the eye movement system requires a minimum latency of about 150–170 ms to program a saccade to a fixed target (Rayner, 1998). Thus, with faster speech, the probability increases that the eye movement system will have trouble keeping up with the input because it must locate referents, program saccades to them, and fixate on them long enough for identification and integration (Gibson, Eberhard, & Bryant, 2005).

Fig. 1. A typical stimulus scene with multiple objects.

A structural manipulation was also introduced, which we will call referential density. Each target sentence contained four separate eye movement targets — noun phrases referring to an entity in the scene. These noun phrases were presented as conjoined phrases, and by varying the location of a modifier to the first and last of those phrases, we were able to manipulate whether a stretch of speech separated the first mention from the second, and the third from the last, or whether the four objects were mentioned one after the other (with only functional elements such as "the" in-between). This manipulation is a structural counterpart of the speech rate manipulation, as it also has the effect of influencing how much time pressure the oculomotor system is under to find and fixate relevant objects. The speech rate manipulation provides warning about the rate at which the objects will be mentioned already at the first (non-target) sentence, but this structural manipulation is less predictable as it becomes apparent first during the mention of the first noun phrase. It is plausible to expect that any increase in complexity may reduce language–vision integration, reflected in lower proportions of fixations for referential targets. Taken to its extreme, this means that at some point in the complexity scale, the integration may cease to occur. Our hypothesis was that the probability of fixating an object would be lower with faster speech and with greater referential density, but that even in the most difficult condition (fast speech + high referential density), we would still find evidence that the eye movement system was attempting to link linguistic expressions and depicted objects. We also expected to see longer times to locate the targets, confirming that the participants fall behind as complexity increases. These findings would demonstrate that the tight linking of linguistic expressions and eye movements to objects is not simply an outcome of any simplified linguistic and visual stimuli used in some previous studies, but rather is a fundamental property of the comprehension system.

We do not specifically predict an interaction between the speech rate and referential density manipulations; the approach we have laid out can accommodate a finding that the two sources of complexity have independent effects. On the other hand, it might be that in the most demanding condition, with fast rate and high referential density, listeners cease trying to link parts of the speech input to the scene, which would manifest itself as an interaction because doing so is too difficult. A finding of this sort might suggest that, in situations of high complexity, language–vision integration can become too challenging to be successful.

1. Methods

1.1. Participants and experimental apparatus

Thirty-two University of Edinburgh students (21 female; mean age 21.5, sd = 3.1) with normal or corrected-to-normal vision participated in exchange for £3. Post-experiment debriefing revealed that all participants were naïve to the purpose of the experiment.

Eye-movements were measured using an SR Eyelink 1000 eye-tracker, tracking at 1000 Hz. Participants were calibrated using a 9-point calibration routine, and the average calibration error was .43° (sd = .18°). Saccades were identified using a velocity threshold of 50°/s over 11 samples and an acceleration threshold of 5000°/s^2. Tracking was monocular, but participants were allowed to look with both eyes. The visual displays were presented on a 21 in. CRT-monitor 90 cm away from the participants' eyes. The auditory stimuli were presented through two speakers located near the monitor but outside the participant's field of view.

1.2. Materials and design

The visual scenes consisted of 48 full-color photographs with a resolution of 800 × 600 pixels. The photographs depicted highly cluttered scenes with many objects (Fig. 1). Similar photographs were used for 24 filler trials. Each participant saw each photograph only once.

The auditory stimuli consisted of 48 experimental utterances and 24 fillers coupled with a specific target scene. Each utterance consisted of three sentences: an introductory statement, the target sentence, and a general concluding statement. All experimental sentences referred to four objects in succession, but the fillers mentioned only a single object.

The speech rate manipulation was performed using the software Praat (Boersma & Weenink, 2008), which allowed rate of speech to be changed with no effect on pitch. The slow and fast conditions were created by decreasing and increasing the original speech rate by 20%. The density manipulation was performed by adding a non-informative pre- or post-modifier to the first and last nouns in the object sequence. In the low density condition, the first object contained a postnominal modifier (*the sailboat that is old and dust-covered*) and the last object contained a prenominal modifier (*the surprisingly mint uniform*). This placement of modifiers separated the first and last nouns from the second and third ones. In the high density condition, the positions of the modifiers were reversed so that the first object was premodified (*the old and dust-covered sailboat*) and the last object was postmodified (*the uniform that's surprisingly mint*). In the high density condition, then, the nouns appeared close together, whereas in the low density condition, they were further apart (see Table 1). The density and speech rate manipulations were complementary but not redundant because the rate of speech leading up to the critical nouns provided listeners with some warning, whereas the referential density manipulation could not be anticipated.

Speech rate and referential density were varied using a 2 × 2 mixed lists design. A typical (low density) utterance would be: "I love going to garage sales. I like the sailboat that is old and dust-covered, the plane, the sombrero, and the surprisingly mint uniform. However, I think I'll skip buying anything." (the underlined objects are the ones to be found in the corresponding scene). Utterance durations for all conditions are also shown in Table 1.

1.3. Procedure

The task was simply to look at the scene, listen to the speech, and answer a question after every trial. Every trial was preceded by a 400 ms fixation cross. The scene was then presented for a total of 22 s. First, the scene was shown for 3 s prior to the onset of speech, thus giving participants some preview. The three-sentence utterance was then presented, and lasted up to 18 s depending on the specific stimulus item. The scene remained visible throughout this time and at least an additional 1 s following the end of the last sentence, lasting until the end of the trial. The question was then presented visually.

Table 1
Examples of target sentences used in the experiment, with mean total durations and mean durations of the four-object sequence (starting from the first *the* and ending with the offset of the fourth object). Slow/fast refers to speech rate; low/high refers to referential density. The mean durations are in milliseconds, with standard deviations in parentheses.

Condition	Object sequence sentence (example)	Total utterance	Object sequence
Slow/low	I like the sailboat that's old and dust-covered, the plane, the sombrero and the surprisingly mint uniform.	14,608 (1570)	6752 (865)
Slow/high	I like the old and dust-covered sailboat, the plane, the sombrero and the uniform that's surprisingly mint.	14,509 (1496)	5266 (583)
Fast/low	I like the sailboat that's old and dust-covered, the plane, the sombrero and the surprisingly mint uniform.	9741 (1053)	4502 (576)
Fast/high	I like the old and dust-covered sailboat, the plane, the sombrero and the uniform that's surprisingly mint.	9672 (997)	3510 (389)

The question was either (70% of the time) "How was the speaker's attitude toward the scene?" or "Was the speaker talking about the scene you just saw?", and required a "positive/negative" or "yes/no" response, respectively. The questions ensured that participants paid attention to both the visual and linguistic stimuli.

The data were analyzed using a multi-level logistic regression (Barr, 2008), using R (R Development Core Team, 2008). The regression model fit the log odds of fixating a target object with the manipulations as predictors interacting with the time bin variable, plus a binary object-position factor and a main effect predictor of time. Time was included both as a linear and a quadratic term. The coding of the fixation data was a binary coding, indicating whether, for a particular point in time, a fixation was inside or outside the area of interest matched with the relevant noun. That is, if the mentioned object is "guitar", and the gaze is located within the area of interest containing the guitar, then a "1" was scored, otherwise a "0" was scored. As four objects were mentioned in each trial, four separate codings were performed for each trial. These binary values were then aggregated per time bin and computed into log odds, and this was done for every unique combination of participant, item, area of interest, and temporal bin. Our manipulations were binary/contrast coded and time had a simple eight-level (eight bins) coding. All predictors were then centered. The area of interest coding was collapsed into a binary (inner/outer) coding in order to aggregate away zero-dominated data and thus be able to compute the log odds of the fixation probabilities.

The temporal analysis window was determined by aligning all probability curves, subtracting baseline data from a temporal region before the object was mentioned, and then visually inspecting the grand average. The earliest rise and the fall of the curve determined the analysis window. The temporal analysis window was 2400 ms long, starting from 400 ms from the speech onset of the target object, lasting until 2800 ms post onset and divided up into eight 300 ms bins. This approach of determining the temporal analysis window was selected because it was unbiased towards the conditions and protected against consciously selecting an analysis window that would confirm the hypotheses. Including the fall of the grand curve as well protects us against the risk of missing any peaks for individual conditions, should they be displaced outside of a more restrictive analysis window in the grand curve. Additional analyses with other windows were explored after the primary analyses, and while this changed the absolute values, the significance of the predictors and the conclusions remained the same.

The model was fit on the log odds scale using the "lmer" function from the "lme4" package (Bates & Maechler, 2010) with separate analyses for subjects and items (to aggregate away null data), but modeling their individual intercepts in the regression model. The p-values were generated using a 10,000 sample Markov-chain Monte Carlo (MCMC) method, provided as the function "pvals.fnc" in the "languageR" package (Baayen, 2007). A significant result means that a predictor can help explain the outcome of the dependent variable and that this effect is significantly different compared to the baseline condition. The most important predictor in our analyses was the effect of time, where we try to model the change of log odds of fixations as the time after the target word onset increases. If this predictor has an estimate that significantly deviates from zero, then this means that time changes the log odds as time increases (i.e., we progress further into the trial). Other estimates may have main effects that shifts the time curve upwards or downwards (shifting the intercept), or they may modulate the slope of the time curve, which means making the growth of the time curve greater or smaller over time. Time was also modeled as a quadratic term, allowing the time curve to bend in order to model both the growth and the decline of the log odds.

The time-to-target durations were analyzed separately with a linear mixed effects model using participant and item as random factors, and speech rate, reference density and inner/outer placement as fixed effects. The durations were log-transformed and the p-values were estimated using the above-mentioned MCMC method. The reported standard errors are computed from the 95% highest posterior density (HPD) interval from the MCMC method.

All statistical models were evaluated with the "anova" function from the "stats" package and predictors that did not significantly improve both the by subject model and the by item model were excluded iteratively. The significance of the random effects was verified by a restricted likelihood ratio test, using the function "exactRLRT" from the "RLRsim" package (Scheipl, 2010).

2. Results

Fixation probabilities are shown in Figs. 2 and 3. Fig. 2 displays the probability of fixating each of the four named objects in the target sentence at the specific points in time shown on the x axis, starting from the onset of the target noun. Fig. 3 shows the cumulative probabilities for the whole trial duration. In both figures, each panel represents one of the four conditions of the experiment: Slow speech and low referential density (top left), slow speech and high referential density (top right), fast speech and low referential density (bottom left) and fast speech and high referential density (bottom right). It should be noted that the figures are down-sampled for legibility, but the analyses used data at the original resolution.

We begin by describing the patterns that can be seen in the graphs. As is clear from Figs. 2 and 3, each object was typically fixated within about 2500 ms of its verbal onset, but the tendency was greater for first and last objects, greater in the low density condition compared to the high density condition, and greater for slow speech versus fast speech. The cumulative probability graphs show that, over the whole trial, the four objects were highly likely to be fixated, reaching peak probability at about .80 for all four objects.

Some of the findings are somewhat harder to see in the graphs, but become clear from the statistical analyses summarized in Table 2. It should be noted that there is no one-to-one relationship between the terms in Table 2 and the interpretation of the effects on the log odds of fixating the targets, as is the case for a linear model. For in-depth investigations, these values should be used with a graph plotting tool, which allows better visualizations of the many possible combinations of effects.

The main analysis of the log odds of fixating the mentioned objects revealed several findings. First, after the speech onset of the target noun, the participants became more likely over time to fixate the target objects that were mentioned, which is shown by the positive estimate for time. The negative estimate of time^2 (the quadratic term of time) means that the log odds of fixating the target decreases after having reached its maximum.

Secondly, speed had a negative main effect as well as a negative interaction with time, which for this case means that the log odds of fixating the target object does not reach the same maximum in the fast condition as in the slow condition and the curve also decreases faster after the maximum in the fast condition.

Thirdly, referential density has a large negative main effect on the dependent variable, but also modulates the linear time factor negatively and the quadratic time factor positively. Taken together, this means that the high referential density condition has a significantly lower maximum for the dependent variable, but also that the growth and decline over time are less distinct, producing a somewhat flatter curve compared to the low referential density condition.

Finally, the inner objects in the sequence of mentioned objects were also negatively affected compared to the outer objects. The negative main effect resulted in a lower maximum. The positive interaction between inner placement and time, for this particular case, means that the inner objects start to attract fixations later than outer objects, but the log odds for the inner objects also decreases somewhat more slowly from their maximum than for the outer objects.

The separate analysis of the data in the most detrimental condition, with fast speech rate, high referential density, and focused only on inner

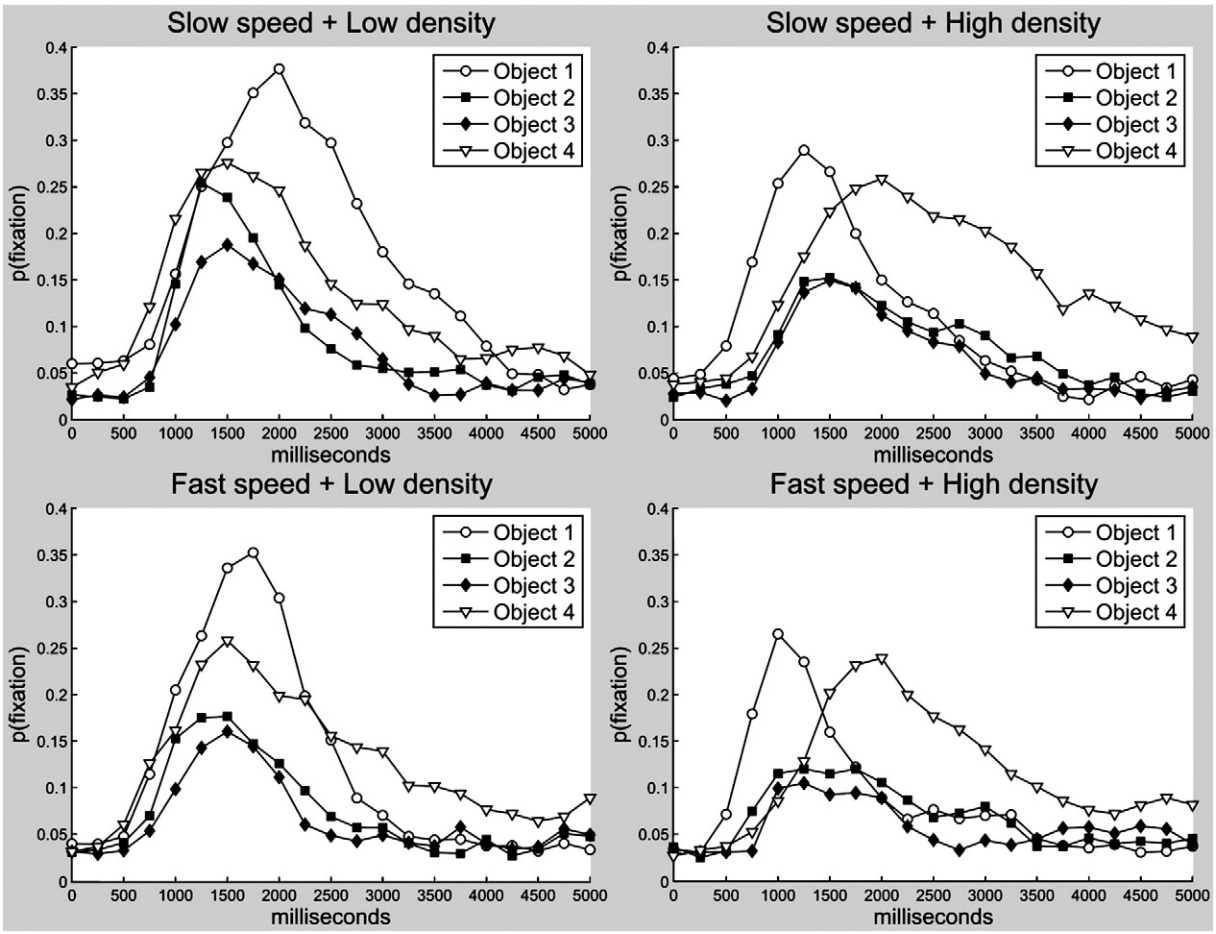

Fig. 2. Probability curves across all conditions and objects and aligned at the verbal onset of the respective objects. Data down-sampled for visibility. The x-axis represents time after onset of the referential expression and the y-axis represents the proportion of fixations to the visual counterpart of the mentioned object.

objects, still showed a significant effect of time, both as a linear (by subject: .111, se = .006; by item: .087, se = .008) and a quadratic factor (by subject: −.08, se = .003; by item: −.069, se = .003). This means that, for this particular subset of the data, the log odds of fixating the target increases over time, reaches a maximum and starts to decline again.

A separate analysis of the *time-to-target* duration from the onset of the auditory noun to the first onset of the gaze on the corresponding visual object revealed that greater complexity results in increased latency. The results are summarized in Table 3.

The time to target estimates produced by the linear mixed effects model show that for the average visual stimulus and the average participant, the time to target was 1432 ms in the condition with slow speech rate, low density and the gaze directed to one of the outer (first or last) objects. The speech rate and density manipulations as well as the object position status then provided additive effects to this default condition (the intercept). The effects were not interactive as the interaction terms were not significant nor improved the overall model significantly.

3. Discussion

In this study, participants listened to connected sentences that mentioned four objects in succession, and at the same time they viewed photographs of complex, photographic scenes. Two variables were manipulated: speech was either fast or slow, and the four objects were mentioned either in rapid succession or they were linguistically spaced.

We found that a high speech rate had a negative effect on the ability to match the referential expression with its visual referent. We observed a significantly longer time to target from the onset of the referential expression, as well as significantly lower log odds of fixating the target. Faster speech increases the rate at which new objects are presented, and the participant has to make a trade-off between trying to find and integrate the current object, or deal with the newer object — either by immediate processing or buffering it in memory for later processing. The lower maximum as well as the faster decline of the log odds of fixating the target in the fast condition suggests that for several trials, participants abandon the search for the target object, for example to focus on a newly mentioned object. The sharper decline of the log odds may be a case of participants acting in synchrony, moving away from the target at the same time, thus producing a sudden drop. However, given the slower rise of the log odds, and assuming that participants who find the targets faster need approximately the same amount of time to process the target, we would expect a rate of decline similar to the rate of growth. That is, if participants were free to process the target fully. Rather, a higher decline rate suggests that participants suddenly abandon the target, most likely to process new material. The reference density manipulation also had the purpose of increasing the processing demands of the participants, but it differed by being not quite so predictable. Whereas a fast speech rate in the first general introduction sentence provides a clue to the rate of incoming objects, the reference density becomes apparent only in the beginning of the chain of mentioned objects. This is reflected by the longer time to target durations and lower maximum log odds of fixating the target in the high density condition. This suddenly increased processing demand is also supported by the slower growth rate of the log odds over time, but it is less clear why it is also succeeded by a slower decline of the log odds. We may expect that participants have greater difficulty in finding the targets in the high density condition, and given equal processing needs

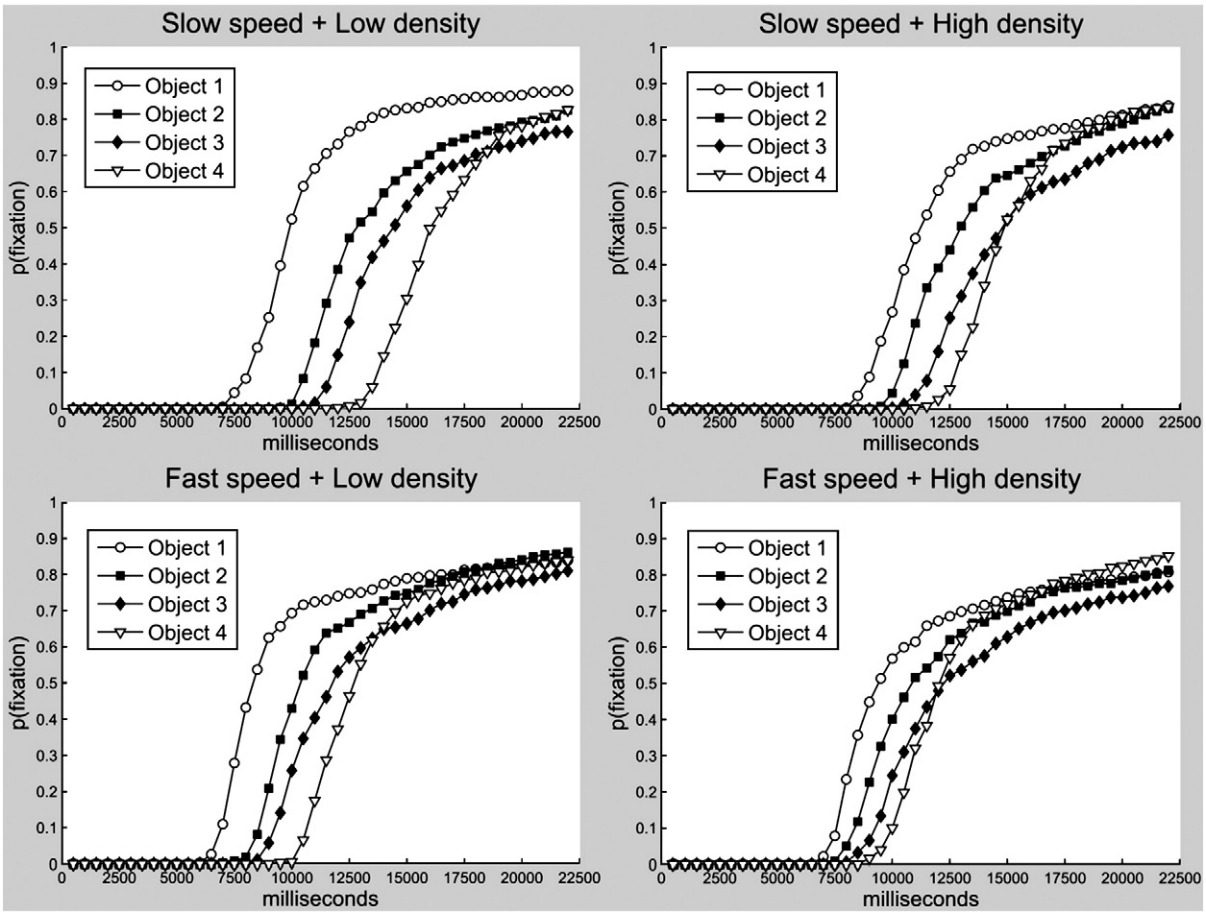

Fig. 3. Cumulative probabilities of fixating four mentioned objects across all conditions. Data down-sampled for visibility. The x-axis represents time following the start of the trial and the y-axis represents the cumulative proportion of fixations to the visual counterpart of the mentioned object.

in time across participants and items, this should be followed by a similar rate of decline. However, we found that in the fast speech rate condition, participants aborted processing to attend to upcoming material. Then, this should also be the case for the high density condition, yet to our surprise it is not. One idea which may explain this phenomenon would be if the sudden onset of several referential expressions overwhelms the comprehender to the degree that they ignore the new material. Thus, the decline rate matches the growth rate, but this is at the cost of the processing of the next object. Typically, this will be the second and third objects, i.e. the inner objects.

Indeed, we find that the inner objects are less likely to be fixated, as well as becoming fixated later compared to the outer objects in the sequence. Apart from having lower maximum log odds of fixating the targets, the decline rate is somewhat slower for the inner objects. The lower maximum is likely due to the special status of the first and the last objects. The first object is presented to a comprehension system that has not yet been challenged and as such has no unfinished processing to devote resources to. The last object is also special, because it never receives a succeeding object that forces the comprehender to prioritize between finishing the current object or attending to the new object. One explanation for the slower decline rate may be that it is largely driven by an interaction between the third object, the fourth object, and the referential manipulation, where the comprehender uses the gap between the third and the fourth object in the low density condition to fully process the third object. Unfortunately, our necessary aggregation of objects into inner and outer ones prevents a further detailed investigation of this matter. It should be noted, however, that

Table 2
Factors modulating the log odds of fixating the visual counterpart of the mentioned object. The "Intercept" is the baseline fixation probability for latency against which the individual factors are compared. "Speed" and "Density" refer to the fast speech rate and high referential density used as manipulations. Factor "Inner" refers to the status of being a middle noun (the second or third noun in a chain of four nouns). Finally, time and time^2 refer to the linear and quadratic effect of the time bins on the dependent variable.

Factor	By subjects			By items		
	Estimate (logit)	SE	p	Estimate (logit)	SE	p
Intercept	−2.082	0.196	<0.05	−2.387	0.629	<0.05
Speed	−0.143	0.021	<0.001	−0.126	0.023	<0.001
Density	−0.325	0.029	<0.001	−0.276	0.031	<0.001
Inner	−0.150	0.021	<0.001	−0.179	0.027	<0.001
Time	0.079	0.008	<0.001	0.049	0.011	<0.001
Time^2	−0.080	0.003	<0.001	−0.069	0.003	<0.001
Speed:time	−0.054	0.011	<0.001	−0.045	0.012	<0.001
Density:time	−0.040	0.011	<0.01	−0.033	0.012	<0.01
Density:time^2	0.027	0.005	<0.001	0.019	0.006	<0.001
Inner:time	0.059	0.011	<0.001	0.068	0.014	<0.001

Table 3
Factors modulating the time-to-target duration from noun onset to gaze onset on corresponding item. The "Intercept" is the baseline latency which is not explained by the factors. "Speed" and "Density" are the two manipulations, and "Inner" refers to the status of being a middle noun (the second or third noun in a chain of four nouns).

Factor	Estimate (log)	SE (log)	t-value	p-value	Estimate (linear)
Intercept	7.267	0.049	147.57	p<.001	1432 ms
Speed	0.097	0.028	3.49	p<.001	+132 ms
Density	0.099	0.028	3.55	p<.001	+132 ms
Inner	0.228	0.028	8.15	p<.001	+292 ms

the objects were not counter-balanced for position in the object sequence. This means that it is possible that, due to chance or biases, less distinct or salient objects were selected to be the inner objects in the sequence and this biased selection drives the effect.

Finally, our most important finding is that even in our highly complex settings, with fast speech rate and objects mentioned in a tight chain, we still find a signal that the participants are trying to match linguistic information with its visual counterparts. This is true also if we concern ourselves solely with the demanding inner objects.

It is tempting to assume from the development of the curves in Fig. 3 that participants recover from our manipulations and in the end attend to almost all mentioned objects anyway. The problem with this interpretation is that we can assume that participants will look around randomly in the picture after the main sentence of the utterance. By chance, their eyes will land on the defined areas of interest and register a hit for that particular area of interest. Thus, given time, every participant will attend to every object mentioned. The crux lies in that we cannot know when a participant switches from a language processing task to a general scene perception task. One solution would be to have a self-paced task, which would terminate trials as soon as the language processing is done.

The fact that our participants actively strived to visually integrate the targets of the referential expressions, even with complex scenes, fast speech rate and high referential density, is evidence that the comprehenders are highly motivated towards establishing this link. Although it is beyond the scope of this investigation, one may wonder what functional advantage in establishing these links is driving this motivation. What this motivation suggests is that the integration of language and vision is not a stimuli-specific effect, but rather a fundamental property of the language system.

An important implication of this study is that it increases the chances of previous research using simpler stimuli to generalize into more complex settings. If listeners integrate visual and linguistic information in very demanding situations, as we have seen, then effects that modulate this integration, e.g. frequency effects (Dahan & Gaskell, 2007) and real-world knowledge (Kamide et al., 2003), will have a chance to work. However, had there been no integration at all in demanding situations, then, obviously, effects that modulate the integration will have nothing to modulate. However, given our large latencies from the mentioning of the object to the fixation of the object, it seems unlikely that effects of anticipatory eye-movements will have adequate time to occur. It is possible that anticipatory language processing in some form still takes place, but that the comprehender is unable to search for and find the objects before they are mentioned because of the cluttered scenes and fast-paced utterances.

This study represents an attempt at bridging the visual search and the visual world paradigms. This bridging is necessary if we want to understand language comprehension in more complex settings. As discussed by others (e.g. Dahan & Gaskell, 2007), it is currently possible to very accurately predict the eye-movements driven by lexical activation (e.g. Allopenna, Magnuson & Tanenhaus, 1998), but this assumes an already known display of objects, and it is unclear how the lexical activation unfolds if participants are engaged in an effortful visual search. It is likely that a significant subset of language situations involves the processing of referential expressions referring to targets not immediately accessible in front of the comprehender and highly activated (c.f. Allopenna et al., 1998, where participants were asked to preview all objects and even asked to name them before the experiment). However, the experiment reported here did allow 3 s of silent preview, followed by several more seconds of a general introductory sentence that never mentioned any objects. This should be compared against a standard visual world experiment which usually has 1 s of silent preview and then around 3 s of speech before the onset of the target referential expression (e.g. Huettig & Altmann, 2005). Still, we argue that even given the long preview in this experiment, this is unlikely to pre-activate the visual referents. This is so because there are so many potential targets to activate in these scenes, that simply activating them all would produce a very large competitor set to store in memory.

However, one valid counter-argument is the fact that the target objects were not selected randomly in the display. Objects were selected primarily based on their uniqueness in the display and to some extent to provide a varied selection of objects across the different scenes. It is possible that this procedure has introduced some bias that allows the matching between referential expression and visual referent to succeed more easily than expected. However, this is unlikely, as the results show later target hits and later log odds maxima than any other visual world study, indicating that the task was indeed challenging (c.f. Yee and Sedivy (2006) and Hanna and Tanenhaus (2004) for a highly constrained and a more natural experiment, respectively). Additionally, a selection bias may not only be a problem, but also represent an ecologically valid effect, as real-world speakers may also be biased in their selections, which in turn can provide listeners with an easier integration task.

Another property of the reported study is that even though the targets are selected to be unique (to be able to determine a target area of interest), the participants do not know this. In a simple display, when the listener gazes upon a beaker after hearing the word “beaker”, the listener can be very certain that the correct target is found as it is easy to verify against the complete set of alternatives. If there ever were a visual search phase, it will now be terminated. However, for a complex scene, there may be many potential targets, but a very cluttered scene will not permit matching the concept activated by the linguistic label against every object in the scene. This is even more plausible given that the listener may match against many dimensions, e.g. phonological (Allopenna et al., 1998), visual (Dahan & Tanenhaus, 2005; Huettig & Altmann, 2007), semantic (Huettig & Altmann, 2005; Yee & Sedivy, 2006) and real-world knowledge (Kamide et al., 2003). This may trigger different behaviors, one of which would be a prolonged visual search. This is likely the case in this study, as is suggested by the relatively slow growth of the log odds curves (c.f. Allopenna et al., 1998, but see also Kronmüller & Barr, 2007). It is also possible to imagine an effect of comprehenders simply down-prioritizing the integration of linguistic and visual information, forcing an early termination of the search for a particular object, to unknown effect.

On a concluding note, in the real world it is not unusual to refer to a set of objects visually available to all interlocutors, to use or omit modifiers that have (among other things) the effect of spacing referentially linkable terms apart or compressing them together. Additionally, it is certainly not uncommon to encounter interlocutors that speak more quickly than usual. Granted, our most demanding conditions in this study are language situations we would like to avoid, perhaps by simply telling our interlocutor to slow down. Still, hectic language situations involving multiple objects are also part of the language situations that we encounter in real life. For example, a police officer receiving excited descriptions over the com radio describing the assailants that took flight through a crowd, or the computer gamer immersed in a real-time strategy game involving coordinating many units together with team members in a frantic online battle. Even as these situations make the normal integration of language and vision hard, the typical participant still struggles to keep up. Therefore, the integration of complex scenes and complex utterances, together with visual search, represents an important part of normal language processing situations.

Acknowledgements

RA gratefully acknowledges support from the Linnaeus environment Thinking in Time: Cognition, Communication and Learning, financed by the Swedish Research Council, grant no. 349-2007-8695 and grant no. 2006-24210-41867-9. We thank two anonymous reviewers for very helpful comments and Rachel Thorpe for the help during the audio recordings.

Appendix A

This table lists the low referential density sentences that were used in the experiment. The high density sentences are constructed by transforming the pre-modifier in the noun phrase into a post-modifying one, and vice versa. This was achieved by simply moving the head of the NP and inserting or removing a "that's" or "that're" were needed. Only the durations from the slow conditions are reported, but the durations for the fast condition can easily be calculated by just dividing by 1.5. The durations are measured in milliseconds, and refer to the total sentence length (low density), the object sequence in the low density condition and the object sequence in the high density condition.

Full stimuli sentences	Total	Low	High
The new restaurant isn't really posh. What gives it away is the [the lit lamps that're very retro-looking], [the bottles], [the ashtray], and [the poorly working corner fan]. I'll skip going there.	16,378	8235	6194
This is a friend's college dorm room. When he first moved in, all he had was [the coat-hangers that're really useless], [the jacket], [the desk-lamp], and [the pretty-much worthless newspapers]. He's such a pig.	16,293	8150	5671
This is a recently redecorated kitchen. The old stuff includes [the dish-washer that still works], [the mitten], [the oven] and [the really sharp and fast blender]. Too bad the kitchen is a bit messy at the moment.	15,536	6603	4827
I love going to garage sales. I like [the sailboat that's old and dust-covered], [the plane], [the sombrero] and [the surprisingly mint uniform]. However, I think I'll skip buying anything.	16,048	8261	6143
One of my flat mates is quite the do-it-yourself guy. I can see he recently used [the wall-socket that's easy to install], [the scissors], [the solvent bottle], and [the out-of-place-looking purple box]. The table's always full of his stuff.	18,000	8273	6365
This is where a friend of mine works as a graphical artist. I gave her [the radio that is old but still functional], [the plastic bottle], [the phone], and [the practically located trashbin]. She's really happy about the studio now.	16,699	7956	6404
I wish I had a country kitchen like this. I especially like [the chandelier that's really anachronistic], [the pillow], [the candelabra] and [that fairly expensive painting]. Too bad I'll never be able to afford it.	15,162	7515	6357
This is my humble flat kitchen. I recently got [the mat that's second-rate but OK], [the skull], [the clock] and [the amazingly-working coffee pot]. I know the kitchen is not much to look at.	15,258	7421	5441
This is a friend's TV room. I want [his drums that are inexpensive but fun], [the bag], [the tapestry], and [his really cool-looking cat]. Too bad he won't give them up.	13,305	7405	5645
God knows where all this stuff comes from. Take for example [the VHS tape that's completely obsolete], [the forks], [the soda] and [the folded-together flag]. I think it's time to throw it all out.	15,365	7376	6027
It's halloween night. The possible weapons were limited to [a chainsaw that's pretty much useless], [claws], [a fan] and [a more terrifying than deadly pitchfork]. I survived though.	14,244	7837	5581
This is where a friend of mine repairs his bike. [The chair that's stale from dirt], [the ladder], [the minibike] and [the stereotypical posters] make this place dirty in more ways than one. Which is kinda why I avoid it.	15,920	6215	4655
Now this is what a hotel room should look like. [The champagne that's extravagant], [the silverware], [the bowl] and [the over-the-top chandelier] is probably what jacks up the costs. If only I were rich.	14,372	6385	5168
I could gladly skip a few of these things. For starters, [the ashtray that's not necessary], [the banana], [the chewing gums] and [the nowadays obsolete internet outlet]. My desk is a lot more cleaner.	15,311	8454	6012
This is the most horrible office I know. I particularly hate [the clock in school style], [the flower], [the radio] and [the magazine that's way past old]. I would have quit my job instantly.	16,325	7420	5499
I think I know where this is. I remember [the liquor sign which flashes constantly], [the chandelier], [the tower] and [the very cool lightpoles]. Or maybe not...	12,857	7725	6183
It was some time ago I were here. What's new is [the TV that looks alright], [the fan], [the clock] and [the comfy-looking blanket]. It's a bit better now.	12,889	5962	4841
Somebody's recently been in the kitchen. They moved [the coffeemaker that's old but still working], [the lamp], [the towel] and [the newly bought chairs]. I should go see if I can find anybody.	14,597	6900	5024
This is a small school library. [The paper towels that are out of place], [the globe], [the flag] and [the old-style exit sign] strike me as not belonging here, though. Not in a library at least.	14,917	6818	5489
Looks like a child lives here. However, [the coffee mug in bright color], [the curtains], [the scarf] and [the recently placed hook] suggest otherwise. It's way too cluttered for my taste.	13,487	6055	5123
This guy really likes touring on bike. I guess it's an accident waiting to happen with [the darts that're often lying around], [the lamp], [the reflex] and [the pretty accurate velocimeter]. Or perhaps it's just my imagination running wild.	17,115	7324	5566
I guess this is an OK living room. However, I'm skeptical towards [the flowers that look plastic], [the TV], [the spray] and [the shouldn't-be-there dog]. Also, it's a bit too old-fashioned for me.	16,773	6672	4911
This small town has its own tourist office. I like [the duck that looks old], [the trumpet], [the rabbit] and [the plastic-looking eagle]. They're cool, but a bit expensive.	13,007	5150	4414
I need to fix my drain pipes. I guess I need [the duct-tape that's always useful], [the drills], [the mallet] and [the all-purpose spray]. Or I'll just do it tomorrow.	13,679	6366	5354
I'm doing exercise with the nursery. I need the jar that's almost empty, the VCR, the bag and the hard-to-remember scissors. This will be fun.	11,726	6172	4716
For my kindergarten game I need some equipment. I need the bin that's misplaced, the pencils, the truck and the heavily used blackboard. This'll be fun!	14,010	6429	4920
I think our friend was here recently. The vaccum that's still working, the clock, the suitcase and the very dusty paper roll has been moved. He'll be right back, I guess.	13,668	6421	5288
I'm going to hold a lecture here soon. Will you help me remove the stool that's very rickety, the fan, the thermos and the broken old clock? Then it'll look a bit more tidy.	13,583	6133	4994
This looks like a fun office cubicle. I like the pony that's funny, the wedding-photo, the flag and the practically located clock. A bit too much stuff overall for my taste.	14,234	6274	4544
Pretty much everything in this studio is new. Everything, but the mug that's a pr gift, the monitor, the phone, and the slightly crooked stool. I assume someone is serious about the recordings.	15,194	5920	4396
This is a cosy kitchen. Especially due to the TV that isn't working, the fruit, the toaster and the well-used baskets. I wish I had a kitchen like this.	12,868	6027	4932
This is definitely an old kitchen. Except maybe the ladle that's hardly used, the mortar, the tongs and the refurbished chest. Or have I missed something?	12,975	6301	4790
Nothing has changed in my old primary school. Even the plant that's donated, the whiteboard, the clock and the well-functioning photocopier is the same. I don't know if that's a good or a bad thing.	15,109	5955	4676
Ah, the signs of a good party. Such as the chocolate that's attractive-looking, the lights, the tv and the hilarious hat. Now all I need is a beer too.	11,822	5705	4803
What a horrible room color. Not to mention the ball that's never been used, the remote, the plant and the god-awful clock. I'm glad I don't live there.	12,452	5933	4661
I wish I had such a well-equipped workshop. I only have a screwdriver that's old, a spanner, a wastebin and a really good drill. But then again, I seldom need to repair stuff.	14,415	5865	4095
I really need to replace stuff in my dirty kitchen. For example the knife that's useless, the tap, the kettle and the really old sponge. And that is just a start.	12,559	5322	4754
This desk is chaotic! How can she possibly find her calculator that's often needed, her photos, her Garfield and her recently bought glasses? I hope she's planning on tidying it up.	14,405	6638	5565
This is the small ad-hoc office of the copy shop. The printer that'll soon break, the scissors, the speaker and the really thick books are practically never used. They should get a better office.	15,077	5707	4625

(*continued*)

Full stimuli sentences	Total	Low	High
This is a messy desk. He told me to get the hairbrush that's not too clean, the mouse, the printer and the half-empty toolbox. I can't imagine why he wants just those?	14,106	6626	5223
There's a lot of stuff in this basement. For example, a keyboard that's severely broken, a flashlight, a blowtorch and an odd-looking jar. It's fairly organized, though.	14,959	7411	5440
This is really a designer kitchen. It's emphasized by the small things, like the apples that look surprisingly real, the wine, the coffee press and the super-expensive tap. I also like the mild colors.	17,872	7974	6114
This is a really white kitchen. I love the espresso-machine that's brand new, the fan, the laptop and the well-polished tap. Perhaps a bit too minimalistic for me though.	14,308	6527	4887
I adore these old rustic kitchens. How quaint with the coat-hanger that's home-made, the cupcakes, the garlic and the nice-smelling pinecones. I'd love to move out into the country.	16,112	6820	5205
This is a rather odd office. It has a satellite dish that's budget-level, pliers, a dog and a mysterious-looking jar. I don't want to know what they are selling.	14,074	6859	5315
Not really an office I'd like to work at. The old feeling is produced by the cola of classic design, the schedule, the bag and the still working radio. Still, I've seen worse.	16,709	7074	5714
There's some pretty cool stuff here like the drum that's still working, the birdcage, the rocking-horse and the beautiful angel. I guess it's all really expensive.	11,801	5632	5158
This desk isn't too messy. Of course, the hairbrush that's ugly-looking, the water, the speaker and the misplaced paper roll should be removed. Then the desk would look OK.	13,604	5907	5042

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419—439.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247—264.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*, 502—518.

Baayen, R. H. (2007). *languageR: Data sets and functions with "analyzing linguistic data: A practical introduction to statistics"*. (R package version 1.0).

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457—474.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. http://CRAN.R-project.org/package=lme4 R package version 0.999375-37.

Boersma, P., & Weenink, D. (2008). Praat: Doing phonetics by computer (Version 4.5.17). http://www.praat.org [Computer program].

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2005). Real-time reference resolution in a referential communication task. In J. C. Trueswell, & M. K. Tanenhaus (Eds.), *Processing world-situated language: Bridging the language-as-action and language-as-product traditions*. : MIT Press.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science.*, *32*(4), 643—684.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 753—763.

Dahan, D., & Gaskell, M. G. (2007). Temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition. *Journal of Memory and Language*, *57*, 483—501.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychological Bulletin & Review*, *12*, 455—459.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409—436.

Ferreira, F., Apel, J., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in Cognitive Sciences*, *12*, 405—410.

Ferreira, F., & Tanenhaus, M. K. (2007). Introduction to the special issue on language–vision interactions. *Journal of Memory and Language*, *57*, 455—459.

Gibson, B. S., Eberhard, K. M., & Bryant, T. A. (2005). Linguistically mediated visual search: The critical role of speech rate. *Psychonomic Bulletin & Review*, *12*, 276—281.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*, 105—115.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson, & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*(1), B23—B32, doi:10.1016/j.cognition.2004.10.003.

Huettig, F., & Altmann, G. T. M. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, *15*(8), 985—1018, doi:10.1080/13506280601130875.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*(4), 460—482, doi:10.1016/j.jml.2007.02.001.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*, 133—156.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, *95*, 95—127.

Kronmüller, E., & Barr, D. (2007). Perspective-free pragmatics: Broken precedents and the recovery-from-preemption hypothesis. *The Journal of Memory and Language*, *56*, 436—455.

R Development Core Team (2008). R: A language and environment for statistical computing (Version 2.8.1). http://www.r-project.org [Computer program]. Retrieved Dec 23, 2008, from.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372—422.

Richardson, D. C., Altmann, G. T. M., Spivey, M. J., & Hoover, M. A. (2009). Much ado about eye movements to nothing: A reply to "Taking a new look at looking at nothing" by Ferreira, Apel & Henderson. *Trends in Cognitive Science*, *13*, 235—236.

Scheipl, F. (2010). RLRsim: Exact (restricted) likelihood ratio tests for mixed and additive models. http://cran.r-project.org/web/packages/RLRsim/index.html R package version 1.0.

Tanenhaus, M. K., & Brown-Schmidt, S. (2008). Language processing in the natural world. In B. C. M. Moore, L. K. Tyler, & W. D. Marslen-Wilson (Eds.), *The perception of speech: From sound to meaning. Philosophical transactions of the royal society B: Biological sciences*, *363*. (pp. 1105—1122).

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, *29*, 557—580.

Tanenhaus, M. K., Spivey, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632—1634.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. M. (2006). Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, *113*(4), 766—786.

Yee, E., & Sedivy, J. (2006). Eye movements reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning Memory & Cognition*, *32*(1), 1—14.

Zelinsky, G. J., & Murphy, G. L. (2000). Synchronizing visual and language processing: An effect of object name length on oculomotor behavior. *Psychological Science*, *11*, 125—131.