Running head: SPEECH WITH REPAIRS

Making Predictions from Speech with Repairs: Evidence from Eye Movements

Martin Corley

University of Edinburgh

Martin Corley

Psychology

School of Philosophy, Psychology, and Language Sciences

University of Edinburgh

Edinburgh EH8 9JZ, UK

(tel) +44 131 650 6682; (fax) +44 131 650 3461; Martin.Corley@ed.ac.uk

Abstract

When listeners hear a spoken utterance, they are able to predict upcoming information on the basis of what they have already heard. But what happens when the speaker changes his or her mind mid-utterance? The present paper investigates the immediate effects of repairs on listeners' linguistic predictions. Participants listened to sentences like *the boy will eat/move the cake* while viewing scenes depicting the agent, the theme, and distractor objects (which were not edible). 25% of items included conjoined verbs (*eat and move*), and 25% included repairs (*eat–uh, move*). Participants were sensitive to repairs: Where *eat* was overridden by *move*, fixations on the theme patterned with the plain *move* condition, but where there was a conjunct, fixations patterned with *eat*. However, once the theme had been heard, there were more fixations to the cake in all conditions including *eat*, showing that the first verb maintained an influence on prediction, even following a repair. The results are compatible with the view that prediction during comprehension is updated incrementally, but not completely, as the linguistic input unfolds.

Making Predictions from Speech with Repairs: Evidence from Eye Movements

It is by now uncontentious that humans engage in prediction during language comprehension. When reading, participants are quicker to identify letter-strings as words if they are predictable from the preceding context (Blank & Foss, 1978; Schwanenflugel & Shoben, 1985). When attending to spoken language, participants' eyes will fixate on depictions of likely candidate continuations to a sentence in a visual image (Altmann & Kamide, 1999). These predictions cannot be attributed to straightforward associations between words (Kamide, Altmann, & Haywood, 2003). As well as the content of what the listener hears, aspects such as the discourse context (Kaiser & Trueswell, 2004) and the prosody of the utterance (Weber, Grice, & Crocker, 2006) can drive predictive eye movements. Prediction appears to be a highly sophisticated process: Where the constraints are appropriate, people are able to predict the upcoming use of specific words, together with information such as their phonology and grammatical gender, both in written (DeLong, Urbach, & Kutas, 2005; Wicha, Moreno, & Kutas, 2004) and spoken (Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005) language. One account of prediction during comprehension is that comprehenders simulate the production processes of the speaker or writer they are attending to (Pickering & Garrod, 2007).

However, predicting what a speaker is likely to say may not be entirely straightforward. This is because, outside the laboratory, speakers are less-than-perfect communicators. Their spontaneous speech tends to be highly disfluent, with around six in 100 words affected by some kind of disfluency (Fox Tree, 1995). Despite evidence that listeners may not recognize having heard these disfluencies (Lickley, 1995; Lickley & Bard, 1996), they clearly have effects on comprehension, both immediately and in the longer term. Fillers such as *uh* improve performance in a word-identification task (Fox Tree, 2001), perhaps because attention to what follows is heightened (Collard,

Corley, MacGregor, & Donaldson, 2008; Fox Tree, 2001). They affect the ease with which words are integrated into their contexts (Corley, MacGregor, & Donaldson, 2007), and influence the parsing of garden-path sentences (Bailey & Ferreira, 2003, 2007). Listeners are more likely to remember words that occur immediately post-disfluency (Collard et al., 2008; Corley et al., 2007), and are likely to form worse impressions of speakers who use *um*s (Christenfeld, 1995). Speakers are rated as less confident of their answers to general knowledge questions if their responses are preceded by fillers (Brennan & Williams, 1995; Smith & Clark, 1993; Swerts & Krahmer, 2005). Disfluencies can also affect listeners' predictions of what a speaker is about to say. Participants following spoken instructions to manipulate objects in a visual array are more likely to fixate objects which are new to the discourse (Arnold, Tanenhaus, Altmann, & Fagnano, 2004) or difficult to describe (Arnold, Hudson Kam, & Tanenhaus, 2007) when the instructions are disfluent.

The disfluencies in these studies (typically *uh*) are presented to participants in circumstances which suggest the speakers are having difficulty retrieving the upcoming words of an utterance (cf. Clark & Fox Tree, 2002), and any changes in prediction are likely driven by the disfluent signal itself. A related but rather different problem is faced by listeners when speakers explicitly correct themselves during an utterance. Self-corrections in speech are not uncommon (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001, report 1.94 'restarts' per 100 words across a series of dialogue tasks); they require the listener to disregard part of what has been said, and any predictions made on that basis, in order to understand the speaker's potentially incompatible intended message. The focus of the present paper is on speakers' self-corrections, or *repairs*: How are listeners' predictions of what is going to be said affected when a speaker changes their mind mid-utterance?

Levelt (1983) distinguishes two types of self-correction, or *repair*, which may be made to correct errors, to change the message, or for the sake of appropriateness. Covert repairs are corrections made to inner speech, on the basis of self-monitoring.

Although they may be marked by a filler such as *uh* to introduce time for replanning (Blackmer & Mitton, 1991), the spoken message does not overtly change. Overt repairs, on the other hand, consist of a deletion of material that has been uttered and its substitution with new material. The sequence of words that comprises an overt repair can be divided into the *reparandum*, the *edit interval*, and the *repair*. The reparandum consists of the material which will be corrected, and often shows prosodic signs of the upcoming repair. The edit interval follows a suspension of speech, and may include a filler such as *uh*, typically with a long vowel duration (Shriberg, 2001). The repair comprises the information which replaces the reparandum; in more complex cases, it may be preceded by a repetition of all or part of the pre-repair utterance (see also Clark, 1994; Nakatani & Hirschberg, 1994); in some cases it is given contrastive stress, relative to the reparandum and to fluent controls (Howell & Young, 1991; Levelt & Cutler, 1983).

There is clear evidence that the comprehension of spoken language is affected by overt repairs. Lau and Ferreira (2005) asked participants to perform grammaticality ratings on spoken materials which included a main verb/reduced relative ambiguity. The experimental utterances all included an overt repair (including an *uh* edit) of the crucial verb. Relative to controls in which the reparanda were syntactically ambiguous (such as 1a), ratings for sentences with reparanda which were unambiguously incompatible with the eventual interpretation (1b) were significantly lower (Experiment 2). These findings suggest that the reparandum continues to exert an influence on interpretation (or, at least, judgements concerning the interpretive process), perhaps because the listener does not completely overwrite its representation (Ferreira, Lau, & Bailey, 2004).

(1a) the little girl picked—uh, selected the right answer...

(1b) the little girl chosen—uh, selected the right answer...

Given a suitable context, perhaps involving children answering questions in a

classroom, it would be relatively easy for a listener to predict likely continuations for the sentences in (1) at the first verb, before the repair. But what happens to predictions such as these once the repair is initiated? In the present paper, we consider three possibilities. The first is that predictions remain primarily based on the reparandum. Listeners must presumably be able to cope with mispredictions, even if there is an associated later cost. Such a cost could account for the difference reported by Lau and Ferreira (2005) between (1a), in which the repair is likely to be compatible with predictions made on the basis of the reparandum, and (1b), in which it is not. Since predictions remain unchanged following a repair, we refer to this possibility as the 'no change' hypothesis.

It may however be the case that post-process judgements (of grammaticality) do not reflect prediction during comprehension (of content). Two further hypotheses consider ways in which prediction could be affected by a repair. In a simple model, changes to predictions could be based on recency (the 'recency' hypothesis): For example, a current prediction of what is likely to be said could be based on the most recent relevant material encountered. Alternatively, prediction could be based on a model which takes the repair explicitly into account, where the reparandum is overridden by the repair in the ongoing representation of the utterance (the 'override' hypothesis). If this latter account is true, it would be compatible with a view that prediction is driven by a full consideration of current linguistic and nonlinguistic input.

In order to distinguish these possibilities, the present paper reports an experiment based on earlier work by Altmann and Kamide (1999). In Altmann and Kamide's experiments, participants viewed visual scenes while listening to spoken sentences. The verbs in each sentence were either restrictive or nonrestrictive: In the restrictive case, they selected for only one item depicted in each scene (e.g., 2a for a scene including a cake as the only edible item); in the nonrestrictive case, the verb was compatible with several depicted items (2b). After the verb but prior to the onset of the post-verbal noun, which was always compatible with either verb, participants

fixated on the appropriate object more often in the restrictive than in the nonrestrictive case. In other words, in cases where participants heard fluent sentences, they were able to make predictions about what theme was likely to be mentioned as a consequence of what they heard and what they could see.

(2a) the boy will eat the cake (*restrictive verb*)

(2b) the boy will move the cake (*nonrestrictive verb*)

(2c) the boy will eat and move the cake (*conjunct*)

(2d) the boy will eat—uh, move the cake (*repair*)

Here, we extend the design with the inclusion of two new conditions, shown (together with the original conditions) in (2). As in Altmann and Kamide (1999), we expect participants to make more fixations on the theme following the verb in sentences like (2a) than in (2b). In sentences like (2c) and (2d), the first verb should also give rise to more fixations on the theme, unless the first verb prosody in (2d) indicates an upcoming repair.

Following the conjunct or repair, the three hypotheses outlined above can be clearly distinguished. Because both the no change and the recency hypotheses represent heuristics which are not sensitive to linguistic form, (2c) and (2d) should pattern together. According to the no change hypothesis, predictions after the second verb in these cases will be driven by the first, restrictive, verb, and fixations at the theme will pattern with (2a) in both conditions. The recency hypothesis, on the other hand, suggests that predictions will be driven by the most recent relevant material encountered. In this case fixations after the second verb in (2c) and (2d) should pattern with (2b), since a nonrestrictive verb will be the most recent (or only) verb encountered in each of these examples. Finally the override hypothesis differentiates (2c) and (2d): According to this view, predictions are based on full interpretations of each utterance. The conjoined verbs in (2c) retain the selectional restrictions of (2a),

since anything that is to be eaten and moved must be edible. Thus the conjunct and restrictive conditions should pattern together. The repair in (2d) should, in contrast, be interpreted in the same way as (2b), such that the repair and nonrestrictive conditions should pattern together.

Method

Participants

Twenty-four students and other members of the University of Edinburgh community volunteered to take part in the study. All were native speakers of English and reported normal or corrected-to-normal vision.

Materials

Stimuli were freely adapted from Altmann and Kamide (1999) and Kamide et al. (2003). Twenty-four images were selected from those created by Altmann, Kamide and colleagues and were paired with sets of four sentences. Each image, which was a 640×480-pixel visual scene constructed from commercial clip-art using a 16-colour palette, depicted a number of people, animals, and inanimate objects. For each image, we selected an agent and a theme from among the items depicted. Four stimulus sentences were constructed from the agent and theme, together with a pair of verbs: A *restrictive* verb or an *nonrestrictive* verb. In the restrictive case the verb's selectional restrictions, together with the agent, dictated that only the preselected theme object was likely to serve in that role. In the nonrestrictive case, at least three of the objects in the relevant visual scene were plausible themes for the agent and verb.

The sentences were constructed as follows. First, we constructed a sentence from the agent, theme, and the restrictive verb (for example, for an image which depicted a baby, a bell, and several other objects, the sentence was *the baby will ring the bell*). In a second sentence, the nonrestrictive verb was substituted (*the baby will kick the bell*). Finally, we constructed two sentences with two verbs, as either a conjoined action (*the*

baby will ring and kick the bell) or a repair (*the baby will ring—uh, kick the bell*). In these sentences the restrictive verb always preceded the nonrestrictive verb. A full set of experimental materials, together with descriptions of the relevant visual images, can be found in the appendix.

A further set of 24 stimuli served as filler items. For these stimuli, the themes in the spoken sentences had no corresponding image in the visual scenes. Fillers corresponded to the sentence types in the experimental design: In particular, six fillers included two verbs separated by *and*, and a further six included verbs in an *uh* repair.

Spoken materials were recorded by a male native speaker of Scottish English, who spoke at a slow but plausible rate. In the case of the repair sentences, speaker was instructed to use an intonation which naturally conveyed a self-correction, and a pronunciation of *uh* appropriate to his dialect (approximately [ɜː]). Subsequent to recording, the stimuli were normalized (resulting in approximately equal subjective volumes) and resampled to create 16 kHz mono .wav files.

Four versions of the experiment were created from the 24 experimental images, such that each version contained one associated auditory stimulus for each image. Each version included equal numbers of auditory stimuli in each condition. The 24 fillers were then added to each version of the experiment.

Procedure

Before the experiment, participants were instructed to listen to each sentence and decide whether it could refer to the picture that accompanied it. They were seated approximately 70 cm from a 21" colour monitor with a 1024×768 pixel resolution. Visual stimuli were displayed centred on the screen. Audio stimuli were played via a pair of speakers situated at either side of the screen. Eye movements were monitored using a head-mounted SR Research Eyelink II eye-tracker, sampling at 500 Hz. Viewing was binocular, although only the dominant eye was tracked. Stimulus presentation, and the recording of data, were controlled by two PCs running software

developed at Saarland University.

Each participant was randomly assigned to one of the four versions of the experiment. Prior to the first item, the experimenter calibrated the eye-tracker using the Eyelink calibration routine. Participants looked at nine fixation targets presented in a random order, followed by a validation phase in which calibration accuracy was measured against the same targets. Calibration was repeated as necessary throughout the experiment. Each trial began with a central point which the participant was instructed to fixate. This allowed the eye-tracking software to perform a drift correction if necessary. Following fixation, a visual scene was presented, together with playback of the auditory stimulus which started 50 ms later. Once participants had performed the picture verification task they responded with the (left or right) arrow keys, and the next fixation point was presented. Materials were presented in a random order subject to the constraints that the first two items were fillers, and no more than two experimental items followed in sequence.

Analysis

We analyzed participants' eye movements, and the time it took them to verify each image. Eye movement data was analyzed as follows. First, we created colour-coded versions of the visual scenes, which distinguished each of the foreground elements from the background. We used different colours for the agent, theme, and distractor items. The recorded screen coordinates from the eyetracker could then be converted into distinct codes for each class of item, so that we had a record over time of which types of items were fixated in which image. Contiguous fixations of less than 80 ms were pooled and incorporated into larger fixations. The time taken by a blink or out-of-range fixation was added to the previous valid fixation.

Because the auditory stimuli were of different lengths, we could not directly compare the time-courses of fixations across conditions. Instead we analyzed fixations over a number of epochs, corresponding to the times taken to utter specific words in

the target utterances. Each epoch apart from the last was calculated using word onsets as boundaries, since these were more easily detectable in the acoustic record. Thus the first epoch, *Det1*, was measured from stimulus onset until the onset of the agent noun, and always consisted of the determiner *the*. The second epoch covered the agent noun and the auxiliary *will*, and is referred to as *N1*. The *V1* epoch consisted of the time taken by the first (or only) verb; *Det2* covered the determiner which introduced the theme; *N2* was the epoch corresponding to the final (theme) noun. For the conjunct and repair conditions, an additional two epochs intervened between V1 and Det2. *AndUh* refers to the *and* or *uh* itself, and *V2* to the second verb. The epochs and their mean durations are illustrated in Table 1. A calculation of the average rate of speech in the experiment, excluding *uh* but including all other experimental epochs, yielded a mean of 200.5 syllables per minute, well within the range of 157–357 reported for British English by Tauroza and Allison (1990).

Insert Table 1 about here

For each trial, we calculated whether fixations on the theme or other objects had been initiated during each epoch of the spoken stimulus. We did not count fixations which had been initiated in previous epochs. This prevented 'double-counting' of fixations which straddled epoch boundaries, and ensured that the antecedent conditions of relevant fixations were unambiguous. Table 2 shows the proportions of trials with fixations to each object type, in each epoch. Fixations on the agent and on distractor items are likely to have been affected by additional factors (e.g., the number of distractors in each scene) and are not considered further. Instead, we focus on fixations on the theme object as a direct index of participants' predictions as the utterances unfolded.

Insert Table 2 about here

Because our dependent variable was binomial (whether or not a fixation had been made) an ANOVA analysis would have been inappropriate (even using transformed proportions of trials as a dependent variable: cf.Jaeger, 2008). Instead we modelled fixation likelihood, using logit mixed effects models (Breslow & Clayton, 1993; DebRoy & Bates, 2004) to test competing hypotheses. Mixed models allow the simultaneous inclusion of by-participant and by-item variation, removing the need for separate F1 and F2 analyses. They can be used to analyze data with both categorical and continuous predictor variables. Importantly, this approach allowed us to directly evaluate predictions made using the three competing hypotheses outlined above.

For each epoch of interest, the analysis proceeded in two stages. First, we created a control model to account for effects that were not of experimental interest. We started with a null model including an intercept, and random by-participant and by-item variation. This model is the equivalent of claiming that there are no interesting differences to be found in the likelihood of making a fixation beyond error variance. We next tested the assumption that each epoch could be considered independently. Although our analysis rests on the claim that attention to an object will result in the initiation of fixations during the relevant epoch, it is possible that fixations which happen to have been initiated in the previous epoch are simply prolonged. In this case, there should be an interdependence between epochs such that more fixations initiated at epoch $e-1$ predict less new fixations at e. For all epochs expect the first (Det1) we accordingly determined whether the fit of the null model was significantly improved by including previous epoch fixations as a predictor of current fixations. In the one case (at N1) that this was true, we used this model rather than the null model in subsequent steps. We note however that there was no evidence for interdependence between experimentally interesting epochs.

Since we can trivially expect the probability of making a fixation at a given location to increase with time, the next step in creating a control model was to add epoch duration as a predictor. if the model fit was significantly improved, epoch duration was included in the control model against which the experimental hypotheses were evaluated.

In the second stage of analysis, our experimental hypotheses were tested separately at each epoch by comparing candidate 'full' models, including experimental variables, to the relevant control models. Each full model consisted of the control model augmented by a predictor variable which differentiated the experimental conditions in line with our hypotheses. The first predictor, derived from the no change hypothesis, differentiated the nonrestrictive verb condition from all other conditions. According to this predictor, the likelihood of making a fixation on the theme object would be affected solely by whether a nonrestrictive verb was the first verb heard. We refer to this as the *No Change* predictor. The second predictor differentiated the restrictive verb condition from all other conditions, in line with the recency hypothesis. According to this predictor, the likelihood of making a fixation on the theme would be affected by the last (or only) verb heard. We term this the *Recency* predictor. The third predictor was based on the override hypothesis that *uh* would indicate that the meaning of the first verb should be overridden in opposition to *and*. This differentiated the repair and nonrestrictive conditions from the conjunct and restrictive conditions. We refer to this as the *Override* predictor.

The selection of models was based on two criteria. First, we assessed whether successive models improved the fit to the data using log-likelihood ratio tests, calculated as $-2(l_1 - l_0)$, where l_0 and l_1 denote the maximized likelihoods of models without and with the predictors of interest respectively: For example, a null model for a given epoch and a model including epoch duration. Because this statistic has a null distribution approximating that of χ^2, with degrees of freedom representing the difference in the number of parameters, a χ^2 test can be used to assess whether a

predictor significantly improves a model. Second, in cases where two or more of the candidate full models significantly improved the control model, we selected the full model which had the smallest log-likelihood, since full models did not differ in degrees of freedom.

Each model includes an intercept and one or more slopes representing the effects of predictors in the model. Where models were selected, the Wald statistic, calculated from each estimated slope and its standard error, was used to determine whether the coefficients differed significantly from zero (see Agresti, 2002).

Results

All fixation analyses were carried out in R (R Development Core Team, 2008) using the lme4 package (Bates, Maechler, & Dai, 2008). Figure 1 is derived from the first stage of analysis, and shows the probabilities of initiating fixations in each epoch once random variation and (where significant) epoch duration are controlled for. Since these probabilities are derived from the model residuals, any remaining differences between conditions are accounted for in the second stage of analysis by adding additional (experimental) predictors to the appropriate control models.

Insert Figure 1 about here

Before V1

The content of the spoken stimuli did not differ until the first verb, V1. For Det1, the null control model was not improved by the inclusion of epoch duration (the effect of previous fixations could not be measured at this epoch). For N1, the control model was improved with the inclusion of previous fixations (at Det1: $\chi^2(1) = 4.64$, $p = .03$), but epoch duration did not improve the model further. The control model, with log-likelihood -248.27, reflected a decreased probability of making a fixation at

N1 if a fixation had been made at Det1. The coefficients of the model (and the probabilities that they differ from zero) are given in Table 3.

No experimental predictor improved the control model for either Det1 or N1, confirming that there were no differences in the likelihoods of fixating the theme ($\chi^2(1) = 1.69$ or less for each predictor; all ps $> .19$).

At V1

At the first (or only) verb, the null model (including an intercept and random variation) was not improved by the addition of previous fixations. It was, however, improved with the addition of epoch duration ($\chi^2(1) = 7.94$, $p = .005$). Taking this control model into account, fixations to the theme were most likely following restrictive verbs (*eat*: probability .51), and least likely following nonrestrictive verbs (*move*: .45). Following conjuncts the probability was similar to that of restrictive verbs (.50); following repairs, the probability was lower (.47; see Figure 1). Given this pattern of results, we tested for differences in fixation probabilities using two predictors. One predictor suggested that the residual probability of making a fixation in the repair condition was most similar to those for the restrictive and conjunct conditions; the second patterned the repair and the nonrestrictive conditions together. These predictors were equivalent to the No Change and Override predictors used in later epochs. Using the criterion of smallest log-likelihood, the best fit model included the Override-equivalent predictor ($\chi^2(1) = 9.51$, $p = .002$), with log-likelihood -280.3, suggesting that fixations to the theme were low in the repair condition. The coefficients of this model (and the probabilities that they differ from zero) are given in Table 3.

Insert Table 3 about here

At AndUh

At *uh* or *and*, the control model did not include previous fixations, but did include epoch duration ($\chi^2(1) = 25.11$, $p < .001$), resulting in a model with log-likelihood -175.3. Because the distributions of the durations for *uh* and *and* are distinct ($t(46) = 34.33$, $p < .001$), no additional predictor which distinguishes the two conditions would be expected to add explanatory power to the model. Accordingly, model fit was not improved by the addition of such a predictor ($\chi^2(1) = 0.42$, $p = .52$). In other words, there were no differences in the likelihood of fixating on the theme object during *uh* or *and* that could not be accounted for by epoch duration. Coefficients of the model are given in Table 3.

At V2

Repair and conjunct materials included a second verb. At this epoch the null model of fixation likelihood was not improved by the addition of epoch duration, previous fixations, or condition (repair vs.conjunct) as a predictor ($\chi^2(1) = 1.22$ or less, ps $\geq .27$).

At Det2

At the second determiner, the null model was not improved by the inclusion of previous fixations or epoch duration. Using the null model as the control model, clear differences emerged, such that the residual probabilities of fixating the theme in the restrictive and conjunct conditions (.51 and .50 respectively) were higher than those in the nonrestrictive and repair conditions (.44 and .45; see Figure 1). Using the criterion of smallest log-likelihood, the best fit model confirmed this pattern, including Override as a predictor ($\chi^2(1) = 16.40$, $p < .001$), with log-likelihood -225.8. Participants were less likely to fixate the theme in the nonrestrictive and repair conditions (OR $= 0.36$). Coefficients are given in Table 3.

At N2

At the final noun, the null model was not improved by previous fixations , but it was significantly improved by the inclusion of epoch duration ($\chi^2(1) = 34.8$, $p < .001$). Where the initial verb had been restrictive, the residual probability of fixating the theme was higher than when it was nonrestrictive. In line with these differences, the addition of No Change to the control model improved model fit ($\chi^2(1) = 3.91$, $p = .048$), and the effect of No Change in the full model just missed significance ($p = .0501$, see Table 3).

Potential Confounds

In order to accommodate the images used, some of the conjuncts used in the present experiment represent pragmatically odd sequences (for example, *the dog will obey and bite the boy*). To check that the results reported above were not influenced by infelicities in the materials, we ran additional analyses in which we included an additional pragmatic factor which indicated whether the conjunct sequence was odd or not. We checked for interactions with the experimental predictors which had been found to be reliable at epochs V1, Det2, and N2. There were no interactions with any of the experimental predictors ($\chi^2(1) = 2.24$ or less, ps $\geq .13$).

Although the no change and recency predictors are assumed to apply to conjuncts as well as to repairs, there is no *a priori* reason why this should be so; they could apply solely in cases where a repair has been detected.[1] For this reason, it is important to establish that the reported findings are not driven by the inclusion of the conjunct condition (which has to be included to differentiate the override from the recency predictor). A series of analyses established that there were no differences between the fixation probabilities for the restrictive and conjunct conditions across epochs (all $\chi^2(1) = 1.97$ or less, ps $\geq .16$), showing that the inclusion of the conjunct condition was highly unlikely to have affected the outcomes of the analyses.

Picture Verification

Picture-verification accuracy across the experiment (including fillers) was 89.6%, with no differences between conditions. Picture verification latencies were recorded for experimental items, relative to the onset of the last word of each utterance at N2. Of 576 responses, 66 (11.5%) were excluded because respondents hit the wrong button or failed to respond, and 13 (2.3%) because the latencies were more than 2.5 standard deviations from the condition mean. To explore the effects, we used linear mixed effects models, with Markov chain Monte Carlo sampling over 10,000 simulations to estimate coefficient probabilities (Baayen, Davidson, & Bates, 2008). As a control model, we used a null model including an intercept, and random by-participant and by-item variation. In the same way as for the analyses of fixation probability, we then compared candidate 'full' models using experimental predictor variables to establish where any difference between conditions lay.

The model which best fit the latency data differentiated the conjunct and restrictive from the repair and nonrestrictive conditions ($\chi^2(1) = 8.65$, $p = .003$), showing that participants took 151ms longer to verify pictures in the latter case. Mean response times and coefficients are given in table 4.

Insert Table 4 about here

Discussion

When participants could use information from the agent and verb in a sentence to predict a theme, fixations on the theme image were more likely than in cases where they could not. In line with previous findings (Altmann & Kamide, 1999; Kamide et al., 2003) the likelihood of making a fixation differed at the determiner before the theme noun itself had been heard, showing that the fixations were unequivocally predictive in nature. When predictions were based on two verbs, fixations following

conjoined verbs patterned with restrictive verbs, showing that participants assigned the same selectional restrictions to restrictive-nonrestrictive conjuncts as to single restrictive verbs. Importantly, fixations following repairs patterned with the nonrestrictive verbs. In other words, participants were able to override the restrictional entailments of a verb in making predictions when the speaker indicated that the verb was to be replaced with a less restrictive verb. Evidence that repairs are fully taken into account as predictions are incrementally updated lends strong support to the 'override' view, according to which the comprehension processes make use of whatever cues are available in the linguistic and nonlinguistic context to make predictions (cf. Knoeferle & Crocker, 2006) perhaps because prediction is at the heart of comprehension (Pickering & Garrod, 2007).

By the following epoch, however, the picture is a little different. As participants hear the theme noun, fixations on the relevant image are marginally more likely if the restrictive verb has been heard at any previous point, even where there has been a repair. This pattern is broadly consistent with the 'no change' hypothesis. It may be that hearing the theme noun triggers a further revision of participants' predictions, resulting in a new fixation on the relevant image if one has not already been made. However, our analyses established that there was no relationship between the likelihood of making a fixation when hearing the determiner which preceded the theme noun and that of making a fixation when hearing the theme noun itself, rendering this explanation unlikely. Instead it seems that, although listeners are clearly sensitive to the change in intended meaning, predictions, once made, are not necessarily completely abandoned.

Although they are at best an indirect measure of comprehension, the picture verification times appear to corroborate this view. Participants took longer to verify pictures following nonrestrictive verbs or repairs than they did following restrictive verbs or conjuncts. In the latter two cases only one depicted item would serve as a theme; in the case of nonrestrictive verbs, most of the objects depicted were

compatible, perhaps making verification more time-consuming. The most important finding is that repairs pattern with nonrestrictive verbs, suggesting, in line with the eye movement data, that the first verb heard retains an influence after a repair.

Taken together, the findings from the post-verb epochs appear compatible with Ferreira et al.'s (2004) suggestion that repairs do not completely replace reparanda in ongoing representations of utterances. In this case, predictions may reflect the influences of both the reparandum and the repair. Alternatively, lexical candidates may be pre-activated as a consequence of prediction (allowing specific information such as phonology to become available: cf. DeLong et al., 2005). In this case, recognition of the theme could be facilitated even if the representation of the message and the predictions made subsequently change. The present findings are compatible with either of these views. Their importance rests in the fact that they show that the reparandum in a verb-verb repair does retain some influence over predictive processes, despite the fact that the repair itself influences the predictions made.

Perhaps surprisingly, fixation likelihood at the first verb was also affected by an upcoming repair: Once epoch duration had been taken into account, the model which best fit the data was one in which the likelihood of fixating the theme object while listening to restrictive verbs such as *eat* patterned with nonrestrictive verbs such as *move*, provided *eat* was the reparandum in the repair condition. This would be worrying if overt reparanda were indistinguishable from non-repaired speech (as suggested by Ferreira et al., 2004, p. 727). In the present experiment, however, this is unlikely to have been the case: Since the utterances were rehearsed and recorded, rather than spontaneously produced, it is probable that the reparanda included prosodic elements which hinted at upcoming repairs. In fact, in spontaneous speech, (whole-word) reparanda are known to show signs of lengthening, and creaky voice (Shriberg, 2001). In the present study, the reparanda matched this pattern: verbs before repairs had reliably longer durations than those before conjuncts ($t(23) = 4.01$, $p < .001$). Research has already shown that participants are able to make predictions

on the basis of prosody (Snedeker & Trueswell, 2003; Weber et al., 2006), and it seems reasonable to suggest that listeners may have been able to use prosodic information to predict an upcoming repair, especially in the context of an experiment in which repairs were fairly frequent (4.2% of words in the present experiment were repaired).

If prosody underlies the effects at the first verb, it might also be the case that prosodic differences between materials are driving the post-repair predictions. This is entirely possible: Although there is variability in the salient prosodies of repairs (e.g., Levelt & Cutler, 1983), it is not at all clear that listeners would recognize a repair in speech which did not include appropriate prosodic changes. Prosody is crucial to the automatic identification of repairs (e.g., Nakatani & Hirschberg, 1994). A potential question for future research, particularly relevant to speech recognition systems, concerns which of the available cues (e.g., duration, pitch, manner of articulation, sequence of constituents) allow listeners to identify a repair.[2] Here the focus was on how repaired speech, once identified, affects listeners' predictions, and the materials used were accordingly 'naturalistic' in prosody.

The experiment reported here was not an orthogonal design: Instead we were able to evaluate competing hypotheses concerning which conditions would pattern together, relating the reported statistics directly to the hypotheses. Using this technique we were able to include the effects of potential confounds, such as previous fixations and epoch length, into the models (in fact several other potential confounds were considered, but no effects were found and the relevant analyses are not reported here). The difference between the proportions reported for V1 in Table 2 and the statistical analysis of this epoch point to the value of this approach. Previous studies for which information concerning epoch durations is not reported (e.g., Kamide et al., 2003), or where arbitrary adjustments have been made to epoch durations to accommodate differences in materials (e.g., Knoeferle & Crocker, 2006, Experiment 1) may have to be interpreted with a little caution.

A note of caution should also be sounded about the present study. Verb-verb

repairs of the type investigated in this paper are relatively rare (Lau & Ferreira, 2005, report that 0.7% of disfluencies in a disfluency-tagged version of the switchboard corpus fall into this category). And of course, most of our daily discourse is not conducted in the context of arrays of visual items which are relevant to what is being said (but see Dahan, Magnuson, & Tanenhaus, 2001). Thus the present experiment may have more to say about what information listeners *can* make use of than what they *do* make use of. However, any evidence that listeners can rapidly update their predictions when speakers repair what they say is an important first step in taking prediction during comprehension outside the laboratory and into the realms of everyday, imperfect, communication.

References

Agresti, A. (2002). *Categorical data analysis.* Hoboken, NJ: Wiley.

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

Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference resolution. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *33*, 914-930.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new. disfluency and reference resolution. *Psychological Science*, *15*, 578-582.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 390.

Bailey, K. G. D., & Ferreira, F. (2003). Disfluencies affect the parsing of garden-path sentences. *Journal of Memory and Language*, *49*, 183-200.

Bailey, K. G. D., & Ferreira, F. (2007). The processing of filled pause disfluencies in the visual world. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: A window on mind and brain.* Amsterdam: Elsevier.

Bates, D., Maechler, M., & Dai, B. (2008). lme4: Linear mixed-effects models using S4 classes [Computer software manual]. Available from `http://lme4.r-forge.r-project.org/` (R package version 0.999375-23)

Blackmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, *39*, 173-94.

Blank, M. A., & Foss, D. J. (1978). Semantic facilitation and lexical access during sentence processing. *Memory and Cognition*, *6*, 644-652.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and

gender. *Language and Speech*, *44*, 123-147.

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive state of speakers. *Journal of Memory and Language*, *34*, 383-398.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Society*, *88*, 9–25.

Christenfeld, N. (1995). Does it hurt to say um? *Journal of nonverbal behaviour*, *19*, 171-186.

Clark, H. H. (1994). Managing problems in speaking. *Speech communication*, *15*, 243-250.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*, 73-111.

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*, 696-702.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*, 658-668.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*, 317–367.

DebRoy, S., & Bates, D. M. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, *91*, 1–17.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabalistic word preactivation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*, 1117-1122.

Ferreira, F., Lau, E. F., & Bailey, K. G. D. (2004). Disfluencies, language comprehension, and tree adjoining grammars. *Cognitive Science*, *28*, 721-749.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*, 709-738.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory and Cognition*, *29*, 320-326.

Howell, P., & Young, K. (1991). The use of prosody in highlighting alterations in repairs from unrestricted speech. *Quarterly Journal of Experimental Psychology*, *43*, 733-758.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Kaiser, E., & Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, *94*, 113–147.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). Prediction and thematic information in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*, 133-156.

King, S., Black, A. W., Taylor, P., Caley, R., & Clark, R. (2003). Edinburgh speech tools library (1.2 ed.) [Computer software manual].

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, d world knowledge: Evidence from eye tracking. *Cognitive Science*, *30*, 481–529.

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language*, *57*, 519–543.

Lau, E. F., & Ferreira, F. (2005). Lingering effects of disfluent material on comprehension of garden path sentences. *Language and Cognitive Processes*, *20*, 633-666.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41-104.

Levelt, W. J. M., & Cutler, A. (1983). Prosodic marking in speech repair. *Journal of*

Semantics, *2*, 205-218.

Lickley, R. J. (1995). Missing disfluencies. In *Proceedings of the international congress of phonetic sciences* (Vol. 4, p. 192-195). Stockholm, Sweden.

Lickley, R. J., & Bard, E. G. (1996, October). On not recognising disfluencies in dialogue. In *Proceedings of the international conference on spoken language processing* (Vol. 3, p. 1876-1879). Philadelphia, USA.

Nakatani, C. H., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustic Society of America*, *95*, 1603-1616.

Pickering, M. J., & Garrod, M. J. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Science*, *11*, 105-110.

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org`

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of faciliation for upcoming words. *Journal of Memory and Language*, *24*, 232-252.

Shriberg, E. S. (2001). To "errrr" is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*, 153-169.

Smith, V., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*, 25-38.

Snedeker, J., & Trueswell, J. C. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*, 103–130.

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, *53*, 81-94.

Tauroza, S., & Allison, D. (1990). Speech rates in British English. *Applied Linguistics*, *11*, 90–105.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P.

(2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 443-467.

Weber, A., Grice, M., & Crocker, M. W. (2006). The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. *Cognition*, *99*, B63–B72.

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy and gender agreement in spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*, 1272-1288.

Appendix

Experimental Materials

Each material is listed with the predictive verb followed by the nonpredictive verb in square brackets. To the right we list objects other than the agent and theme which appeared in each visual image. Question marks show materials adjudged to be pragmatically odd when testing for confounds.

The baby will ring [kick] the bell	*blocks, drum, rubber duck*	
The boy will bounce [throw] the ball	*paper dart, shuttlecock, acorns, bicycle*	
The boy will eat [move] the cake	*toy car, train set, ball*	?
The dog will obey [bite] the boy	*bed, melon, dogfood*	?
The boy will walk [feed] the dog	*pick, chicken, football, bird*	
The businessman will wear [forget] the hat	*folder, wallet, chair, magnifying glass*	?
The cat will drink [knock] the water	*steps, jug and cup, builder*	
The doctor will inject [check] the child	*computer, microscope, books, soft toy*	
The dog will chase [bite] the cat	*man, apple, cup, dogfood*	
The girl will play [pick up] the tuba	*toy, rocking horse, man*	?
The hiker will climb [photograph] the mountain	*leopard, cactus, moon*	
The housewife will fry [wash] the mushrooms	*knife, jug, scales*	?
The man will sail [watch] the boat	*car, birds, sun*	?
The man will smoke [collect] the cigarettes	*glasses, briefcase, folder*	?
The man will repair [wipe] the washing machine	*machine, mirror, bin, dog*	
The monkey will eat [stand on] the bananas	*dolphin, fish, trees*	?
The policeman will arrest [search] the man	*bin, car, houses, cat*	
The tailor will roll [use] the fabric	*sewing machine, pipe, plumber, sink*	
The woman will close [watch] the door	*toy, baby, baby bottle, wine*	
The woman will eat [throw] the grapefruit	*dead bird, cat, cat flap*	?

The woman will play [dust] the piano	*table, telephone, television*	
The woman will read [throw] the book	*sweets, paperweight, coffee*	?
The woman will drink [move] the champagne	*parcel, doll, girl, baby*	?
The woman will drink [try] the wine	*lipstick, cheese, flower, chair*	?

Author Note

The author is grateful to Andrew Wilson, Gemma Devlin and Anika Fiebich for help with the creation of materials and collection of data. Yuki Kamide kindly provided the images used in these experiments, and Christoph Scheepers answered a number of queries concerning the manipulation of eyetracking data. Pia Knoeferle and two anonymous reviewers made a number of useful suggestions concerning the manuscript. Versions of this work were presented at the AMLaP 2006 and CUNY 2007 conferences.

Footnotes

[1]I am grateful to an anonymous reviewer for drawing this to my attention.

[2]In an attempt to determine whether pitch or duration played a role in repair detection, we compared durations of repair verbs (at V2) to those in conjuncts. We also automatically extracted pitch information using the pitch detection algorithm from the Edinburgh Speech Tools Library (King, Black, Taylor, Caley, & Clark, 2003). There were no differences in mean duration (see Table 1), nor in mean pitch (repairs: 139Hz; conjuncts: 156Hz).

Table 1

Mean durations, in milliseconds, of epochs in spoken materials (standard deviations in brackets).

	Det1[a]	N1	V1	AndUh	V2	Det2	N2
	the	boy will	eat/move	and/uh	move	the	cake
restrictive	400.5	760.0	491.3	—	—	180.4	397.6
	(84.5)	(128.5)	(91.7)			(72.9)	(159.3)
nonrestrictive	400.0	776.3	504.7	—	—	191.7	391.8
	(69.9)	(115.9)	(122.2)			(72.0)	(188.9)
conjunct	401.8	708.1	525.4	264.6	478.5	179.8	373.2
	(88.9)	(119.0)	(78.8)	(77.0)	(106.5)	(53.5)	(146.8)
repair	427.3	728.0	625.7	1107.7	619.5	155.6	368.2
	(109.6)	(140.3)	(105.1)	(92.4)	(137.3)	(43.9)	(137.0)

[a]Measured from onset of sound file, including any initial silence

Table 2

Probabilities of initiating fixations on the predicted theme (e.g., cake*), the agent (e.g.,* boy*), and on distractor items, by epoch and condition. Note that these probabilities may sum to more or less than one because participants may have fixated zero or more (types of) items during any one epoch.*

		Det1	N1	V1	AndUh	V2	Det2	N2
		the	boy will	eat/move	and/uh	move	the	cake
theme	restrictive	.118	.181	.278	—	—	.215	.361
	nonrestrictive	.132	.215	.160	—	—	.069	.285
	conjunct	.132	.160	.278	.222	.340	.188	.368
	repair	.132	.181	.271	.500	.333	.104	.333
agent	restrictive	.368	.681	.333	—	—	.049	.139
	nonrestrictive	.444	.674	.319	—	—	.083	.090
	conjunct	.438	.556	.306	.118	.194	.069	.160
	repair	.403	.681	.361	.368	.215	.049	.174
distractors	restrictive	.479	.528	.472	—	—	.167	.243
	nonrestrictive	.382	.556	.472	—	—	.264	.410
	conjunct	.354	.486	.507	.229	.312	.215	.278
	repair	.431	.535	.486	.688	.486	.208	.278

Table 3

Model coefficients and probabilities for each epoch

	Predictor	Coefficient Estimate	Std.Error	p(coefficient $\neq 0$)
Det1		No improvement on null model		
N1	intercept	−1.8686	0.3167	$< .0001$
	Prev. Fix	−0.7830	0.3618	.0304
V1	intercept	−3.7450	0.7380	$< .0001$
	Duration	0.0048	0.0013	.0002
	Override	−0.7516	0.2430	.0020
AndUh	intercept	−1.7028	0.2865	$< .0001$
	Duration	0.0015	0.0003	$< .0001$
V2		No improvement on null model		
Det2	intercept	−1.4966	0.2008	$< .0001$
	Override	−1.0201	0.2623	.0001
N2	intercept	−2.3796	0.3167	$< .0001$
	Duration	0.0045	0.0007	$< .0001$
	No Change	0.4454	0.2274	.0501

Table 4

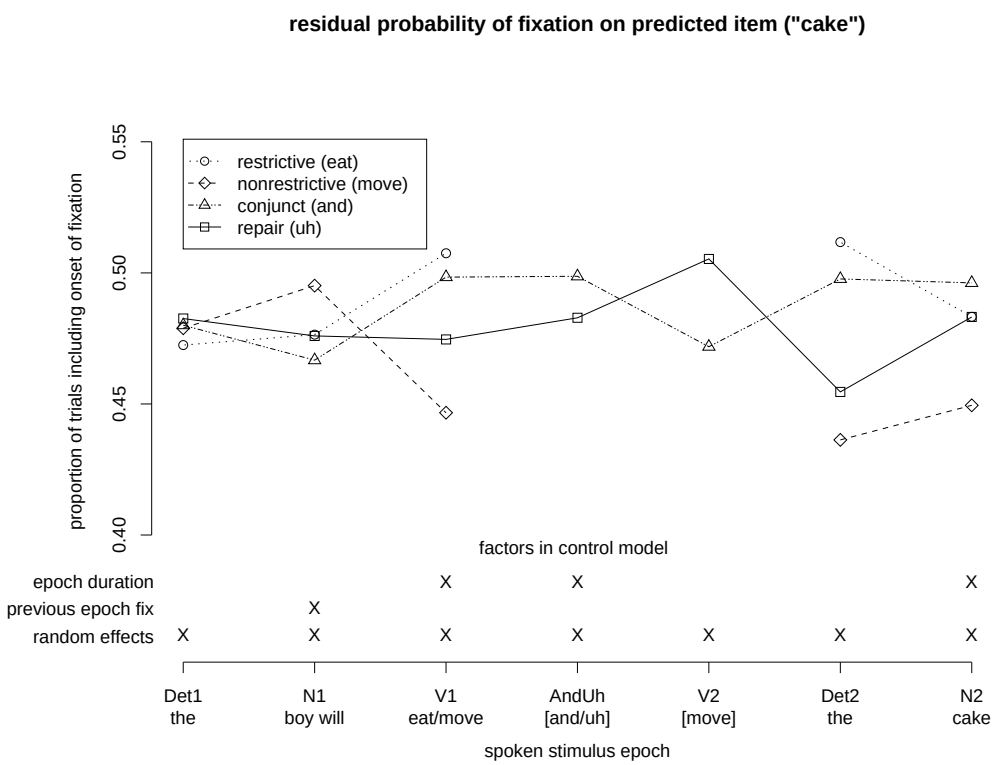
Means, standard deviations, model coefficients, and probabilities for picture verification latencies

	restrictive	nonrestrictive	conjunct	repair
mean (ms)	1318	1546	1334	1463
std.dev.(ms)	565	565	612	605
Predictor	Coefficient Estimate	95% CI (lo)[a]	95% CI (hi)[a]	p(coefficient $\neq 0$)[a]
intercept	1412.8	1255.01	1573.4	.0001
repair or nonrestrictive	151.0	49.09	248.3	.0033

[a]Estimated using 10,000 Markov chain Monte Carlo simulations

Figure Captions

Figure 1. Residual probabilities of initiating a fixation on the theme during the production of utterances, by experimental condition and epoch. Probabilities are calculated from model residuals, expressed in terms of log-likelihood R_l and converted to probabilities using the relationship $p = \frac{e^{R_l}}{1+e^{R_l}}$. Xs above the horizontal axis indicate the factors residualized against for each epoch.


residual probability of fixation on predicted item ("cake")
proportion of trials including onset of fixation
0.40
0.45
0.50
0.55
restrictive (eat)
nonrestrictive (move)
conjunct (and)
repair (uh)
factors in control model
epoch duration
previous epoch fix
random effects
Det1 the
N1 boy will
V1 eat/move
AndUh [and/uh]
V2 [move]
Det2 the
N2 cake
spoken stimulus epoch