Eye movement evidence that readers maintain and act on uncertainty about past linguistic input

Roger Levy,1,2 Klinton Bicknell,3 Tim Slattery,2 and Keith Rayner

Departments of 1Linguistics and 2Psychology, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093

Edited by Edward E. Smith, Columbia University, New York, NY, and approved October 20, 2009 (received for review July 14, 2009)

In prevailing approaches to human sentence comprehension, the outcome of the word recognition process is assumed to be a categorical representation with no residual uncertainty. Yet perception is inevitably uncertain, and a system making optimal use of available information might retain this uncertainty and interactively recruit grammatical analysis and subsequent perceptual input to help resolve it. To test for the possibility of such an interaction, we tracked readers’ eye movements as they read sentences constructed to vary in (i) whether an early word had near neighbors of a different grammatical category, and (ii) how strongly another word further downstream cohered grammatically with these potential near neighbors. Eye movements indicated that readers maintain uncertain beliefs about previously read word identities, revise these beliefs on the basis of relative grammatical consistency with subsequent input, and use these changing beliefs to guide saccadic behavior in ways consistent with principles of rational probabilistic inference.

A critical part of understanding a sentence is identifying the words that comprise it. It is well understood that humans achieve this goal through a combination of top-down (contextual) and bottom-up (perceptual) cues (1, 2). In spoken word recognition, comprehenders’ conclusions about a given word’s identity can be affected by phonetic and lexical cues from adjacent words (3, 4), and by subsequent semantic context when the speech stream is manipulated to make bottom-up evidence ambiguous (5, 6). It has also recently become clear that even relatively unambiguous bottom-up evidence regarding spoken word identity can be overridden by subsequent evidence within the same word (7, 8). It is unknown, however, whether word-identity uncertainty is maintained as a matter of course after perceptual cues have been encountered in full upon hearing or first reading a word in a sentence. It could be that, under normal circumstances, a categorical commitment is made to a unique candidate word, which is then integrated into an incremental sentence representation. Such a strategy could be efficient if considerable effort is required to construct and maintain syntactic/semantic relationships for multiple candidate words at a given position within the sentence (9). This possibility also coincides with the treatment of grammatical analysis in most formal models of language comprehension as a process that takes a word sequence as its input (10–12), and with the treatment in leading models of eye-movement control in reading (13, 14). On the other hand, it is also possible that the perceptual input from a given word gives rise to multiple candidate word representations entering the stream of syntactic/semantic processing. This strategy would have the advantage of greater robustness, permitting identification and correction of perceptual noise or speaker error early in a sentence on the basis of additional information obtained late in a sentence (15). In addition, maintaining and continually reassessing uncertainty about word identity could be useful in guiding behavior during real-time comprehension. In reading, eye movements could be directed to previous parts of a sentence where word-identity uncertainty is high; in spoken language comprehension, allocation of attentional resources could be updated on the basis of word uncertainty, or the listener could prompt the speaker for clarification about parts of a message that were unclear. Whether uncertainty about word identity is maintained and revised, and if so, according to what principles, is thus a question of fundamental interest in the study of language comprehension.

A Sharp Test for Word-Level Uncertainty

In the past, providing a sharp test of whether word-identity uncertainty is maintained has proven elusive because existing results could be interpreted as involving categorical word misidentification. For example, the word flour is reread more often in sentence 1a than in sentence 1b or than the word wheat in sentence 1c (16):

1a. He swept the flour that he spilled.
1b. The baker needed more flour for the special bread.
1c. He swept the wheat that he spilled.

Crucially, the word flour has an orthographically similar “near-neighbor,” floor, that is semantically incompatible with the critical word. It is possible that the distinctive rereading behavior is due to the comprehender initially maintaining some uncertainty as to whether the critical word was flour or floor, and subsequently returning to this word to confirm that it is indeed flour. This same result could, however, be obtained under a theory in which uncertainty about word identity is not maintained after initial word recognition, but there is an occasional categorical misidentification of flour as floor in sentence 1a upon first reading, and the incongruous downstream word spilled triggers a return of the eyes to the initial locus of semantic anomaly.

In the present work, we provide a sharper test of whether word-identity uncertainty is maintained by using sentences in which categorical misidentification should in principle facilitate, rather than hinder, downstream processing. We turn to sentences with a combination of grammatical constructions that has attracted considerable recent attention for the theoretically problematic comprehension behavior it elicits. Despite near meaning equivalence, sentence 2a is considerably more difficult to read and comprehend than sentences 2b and 2c (17–19):

2a. The coach smiled at the player who was tossed the frisbee.
2b. The coach smiled at the player thrown the frisbee.
2c. The coach smiled at the player who was tossed the frisbee.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence should be addressed. E-mail: rlevy@ling.ucsd.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0907664106/DCSupplemental.
Processing difficulty in sentence 2a is localized at the word tossed, whose grammatical part of speech is in principle ambiguous between finite verb and past participle (17). This difficulty pattern is similar to that obtained in classic "garden-path" sentences such as the horse raced past the barn fell (20), where the part-of-speech ambiguity of raced permits a misanalysis of raced past the barn as part of the sentence’s main clause when in fact it is a reduced relative clause that is part of the sentence’s main-clause subject. There is a critical difference in sentences like 2a, however, in that the potentially ambiguous word tossed appears at a point in the sentence where the prior context (in particular the words smiled at, which signal that the player is inside a verb-modifying prepositional phrase) should rule out this main-clause misanalysis. The comprehension difficulty observed for this and similar types of sentences (19, 21) is at odds with a long history of theoretical accounts and empirical evidence that prior context guides inferences about the grammatical analysis of new words in a sentence (10–12, 22–25), and thus poses a major challenge for most leading accounts of language comprehension.

One way in which these results could be reconciled with a strong guiding role of prior context is if the representation of prior context is less categorical than has previously been assumed. Levy (15) recently proposed a model in which these results arise from a combination of top-down grammatically driven expectations with nonveridical bottom-up input through noisy-channel Bayesian inference. This move is analogous to probabilistic and constraint-based approaches that assume probability distributions over structural analyses of a sentence (10–12, 26–29), only in this case probability distributions are extended over the content of the sentence, as well as over its analysis. In this model, the comprehender takes perceptual noise into account when making inferences about sentence form and structure, so that at all times the comprehender has a probability distribution over the sequence of words comprising the current sentence, taking into account perceptually similar and grammatically permissible variants of the sentence read thus far. In an experimental setting, researchers know the true sentence w* presented to a reader, but cannot observe the noisy perceptual input I on which the reader bases his/her inferences about sentence form and structure. However, we can marginalize over I to obtain expected inferences as follows:

\[
P(w|w^*) = P_c(w) \int P_c(I|w)P_r(I|w^*) \frac{P_c(I)}{P_r(I)} dI
\]

\[
\propto P_c(w)Q(w, w^*),
\]

where \(P_c\) is a probability distribution encoding the comprehender’s knowledge of his/her language and environment, \(P_r\) is the true distribution of perceptual noise, and \(Q(w, w^*)\) is proportional to the integral in Eq. 1 and represents the average effect of perceptual noise. To model the behavioral consequences of reading a new word \(w_i^*\) in a sentence, Levy (15) assumed that if \(w_i^*\) dramatically changes the comprehender’s beliefs about the earlier content of a sentence, then the comprehender will tend to respond behaviorally by longer fixation times and possibly making regressive saccades. If \(P(w_{i+1}|w_i)\) is defined to be the probability distribution over the sequence of words starting at the beginning of the sentence and continuing up to but not including the position occupied by \(w_i^*\), conditioning on the perceptual input obtained from words \(w_1,\ldots,w_i\), the Kullback-Leibler (K-L) divergence \(D(P_i(w_{i+1})||P(1=(w_{i+1}))\) is a natural metric quantifying the change in this probability distribution sentence induced by reading \(w_i^*\) (30).

In this model of rational probabilistic inference under uncertain input, the behavioral difficulty of sentence 2a arises because the conditional probability of tossed given the true preceding context is considerably lower than its conditional probability given visually similar but grammatically different contexts, such as those obtained by substituting for at the orthographically similar (near-neighbor) words as or and, or by inserting who after coach or player. This difference in conditional probability arises because any of these substitutions or insertions changes the sentence’s grammatical structure such that the word after player is highly likely to be a finite verb. As a result, when the model is instantiated using a probabilistic context-free grammar estimated from the parsed Brown corpus (31, 32) and a noise function \(Q\) based on the Levenshtein distance between words in \(w\) and \(w^*\), Bayesian inference upon reading tossed creates a large shift in the probability distribution over preceding-context word sequences away from the true context and toward these near-neighbor contexts; reading thrown in sentence 2b does not lead to a similar shift because it cannot be a finite verb (Fig. 1, black and green lines). If reading is influenced by principles of rational inference and decision theory, then these large changes in belief about prior context would be likely to induce comprehenders to slow their reading rate, and would be likely to trigger regressive eye movements to earlier parts of a text and the locus of uncertainty.

Crucial to this model, however, is the comprehender’s ability to maintain uncertainty about the identities of words that have already been read. Furthermore, if the prior context is changed such that there are fewer near-neighbor contexts for which the conditional probability of the critical word tossed is higher than under the true context, then in the model the ensuing shift of the probability distribution over preceding contexts at the critical word should be reduced. This observation leads to an experimental design in which comprehenders read not only sentences like 2a and 2b but also

![Fig. 1. Size of change (K-L divergence) in expected probability distribution over preceding word sequences upon encountering tossed/throw in sentences 2 and 3, under noisy-channel Bayesian inference. The parameter λ, expressing the level of perceptual noise, is free in the model; over a wide range of parameter values, the sentence types observed to be harder for humans to process are those in which the shift in probability distribution is larger. Variants of the true context have probability exponentially decreasing in λ times the Levenshtein edit distance from the true context. See SI Appendix for model details.](image)
variants of these sentences in which the word *at* is substituted by the word *toward*, as in sentences 3a and 3b:

3a. The coach smiled toward the player tossed the frisbee.
3b. The coach smiled toward the player thrown the frisbee.

Because *toward* does not have near-neighbor words under which the *player* would be a grammatical subject, the shift in belief about preceding context upon reading the verb *tossed*—but not *thrown*—is considerably reduced (Fig. 1, orange and magenta lines). In behavioral correlates of processing difficulty at this critical verb in the sentence, we should therefore see an interaction between the word used in the preceding context (at or toward) and the superficial part-of-speech ambiguity of the word in this position (tossed, ambiguous; thrown, unambiguous).

**Experiment**

Our experiment tested the prediction described above of an interaction between near-neighborhood of the preceding context and superficial ambiguity of the critical word on measures of processing difficulty in sentence reading. We constructed 24 sets of sentences on the pattern of sentences 2a and b and 3a and b; the four versions of each sentence differed only in the immediately postverbal preposition (*at* versus *toward*) and whether the critical participial verb had the same surface form as the verb’s simple-past form. We tracked the eye movements of 40 native speakers as they read these sentences on a computer screen and answered a yes/no comprehension question about each sentence.

**Results.** Following standard practice in the analysis of reading using eye tracking, we divided each sentence into seven regions as shown in example 4 and analyzed seven measures of online processing difficulty widely used in the sentence-processing literature (33).  

4. **Subj**  
   MV Prep Obj  
   /The coach/ smiled/ *at*(toward)/ the player/  
   Critical Spill Final  
   {tossed,thrown}/ a frisbee/ by the opposing team./

We report here results pertinent to our hypothesis of an interaction between preposition in preceding context (*at* versus *toward*) and superficial ambiguity of the critical verb (ambiguous versus unambiguous). (Complete results are given in SI Appendix.)

We observed precisely the predicted interaction in three measures of incremental processing difficulty: (i) the frequency with which readers’ eye movements immediately after first reading the critical verb were regressive (as opposed to progressive; significant both in by-participant and by-item ANOVAs at *P* < 0.05); (ii) the time spent reading between the first fixation on the critical verb and the first fixation after it (significant both in by-participant and by-item ANOVAs at *P* < 0.05); and (iii) the frequency with which readers fixated on the preposition (*at* or *toward*) after first fixating on the critical verb and before fixating beyond it (*P* = 0.087 in a mixed-effects logit model). By all of these measures, readers experienced the most difficulty in the condition where the critical verb is superficially part-of-speech ambiguous and near-neighbor substitutions in the preceding context would license the incorrect part of speech (Fig. 2 A–C). Additionally, readers were least likely (ANOVA interactions *P* < 0.05) to correctly answer comprehension questions about these sentences in this condition (Fig. 2D). Finally, we found that readers spent longer overall during first-pass reading of the critical word in the ambiguous conditions (Fig. 2E; ANOVA main effect *P* < 0.05 by participants, *P* < 0.1 by items).

**Discussion**

These results provide evidence that readers maintain uncertainty about the identities of previously read words and that grammatical relationships that would hold between a word at a given position and the rest of the sentence are entertained for multiple candidate words. As measured by frequency and duration of regressive eye movements, comprehension is most effortful in the *at* ambiguous condition, precisely the condition where the discrepancy between the likelihood of the critical word given the true preceding context versus near-neighbor contexts most strongly favors the near-neighbor contexts. Unlike previous
results, this result cannot be explained by occasional categorical word misidentification: misreading and or as for at should make the critical region easier when it is ambiguous (tossed) than when it is unambiguous (thrown), because the misreading would make a finite-verb analysis available.

General Discussion

Our results suggest that readers, as a matter of course, maintain uncertain beliefs about the identities of previously read words, revise these beliefs as a function of grammatical coherence with subsequent linguistic input, and act rapidly on changes in these beliefs through eye movements directed toward the loci of changes in uncertainty. Any theory of language comprehension in which representations of preceding context involve veridical word representations lacking in uncertainty will have difficulty accommodating these findings. These findings also go beyond previous results demonstrating downstream effects of context in spoken word recognition (3–6) in that our materials are visual rather than auditory, were not manipulated so as to create unusual perceptual ambiguity, and require the simultaneous pursuit of multiple grammatical analyses of the sentence to obtain the belief-revision behavior we observed.

Finally, our results are consistent with a picture in which key elements of language comprehension can be viewed as instances of rational behavior through probabilistic inference (34–36): perceptual input obtained from eye movements is recruited jointly with grammatical knowledge to produce evidential inferences about linguistic form and structure, which in turn play a role in guiding subsequent eye movements. Further research will be necessary to more fully establish the full range of uncertainty that comprehenders hold out as to word identities—whether they consider only substitutions among closed-class words (which would be consistent with the present experimental results) or also entertain substitutions among open-class words (which are necessary to more fully establish the full range of uncertainty). Any theory of language comprehension as jointly recruiting perceptual input and grammatical knowledge has potentially broad consequences for the study of language and cognition. The view simultaneously unites and underscores the differences between research in spoken language understanding and reading, raising new questions about the relative weighting of sensory input and prior knowledge, and how noise (environmental, neural, and speaker error driven) is accounted for, in the two modalities. Answers to these questions would be of both fundamental value in cognitive science and potential benefit in furthering our understanding of disorders and age-related changes in both reading and spoken language comprehension.

Methods

Materials and Procedure. We used 24 experimental items in the study and constructed four stimulus lists, rotating items among these four conditions in a Latin Square; complete experimental materials can be found in SI Appendix. These 24 experimental stimuli were interleaved with 36 fillers. Order of presentation was randomized differently for each participant, subject to the constraint that no two experimental items appeared consecutively. Forty native-English speaking undergraduate students at the University of California at San Diego participated in the experiment. All had normal vision or corrected to normal vision and were naive as to the purpose of the experiment. Participants read each sentence while their eye movements were monitored by an SR Eyelink 2000 eye tracker, obtaining one eye-position sample every 1/2 ms with a spatial resolution of 0.01° (binocular viewing, recording right eye only). Each sentence was presented on a single line in 14-point Courier New font on a 19-in. LCD monitor positioned 55 cm in front of the participants (1° visual angle ~3 characters). The eye tracker was calibrated before beginning the experiment and subsequently was recalibrated between trials as necessary.

Regions of Analysis and Data Processing. Each trial was inspected by hand using the University of Massachusetts EyeDoctor software suite (37). We discarded any trial in which there was track loss before some fixation in any region other than the final region. This process resulted in loss of 15.3% of trials. Most of these track losses were caused by the participant blinking.

We examined a number of standard eye movement measures (33) including: (i) the frequency with which a region was skipped on first reading, (ii) first fixation duration (the duration of the first fixation on a region when no material to the right of the region had yet been fixated), (iii) first pass reading time (the total fixation time on a region the first time it was entered, when no material to the right of the region had yet been fixated; also called gaze duration for regions consisting of only one word), (iv) go-pass time (the accumulated time from when a reader first fixated on a region until their first fixation to the right of the region; this measure includes any regressions the reader made before moving forward past the word), (v) total reading time (the summed time of all fixations on a region), (vi) regressions out of a region immediately after first-pass reading, and (vii) regressions into a region. These measures were computed for the regions of each sentence as delineated in example 4. In addition, we defined a go-first regression to have occurred from region Y to region X if the reader had a first-pass regression from region Y and subsequently fixated on region X before saccading past region Y. Fig. 2C shows frequencies of go-first regression from the critical region to the preposition.

Statistical Analysis Method. We report most results using traditional by-participants (F1) and by-items (F2) ANOVAs, with proportions for binary-valued outcomes arcsine transformed before statistical analysis. Consistent with standard practice, reading times > 4 SDs outside the mean for each condition in each region were discarded as outliers. In cases where the assumptions of ANOVA were badly violated (heavily imbalanced data and/or binary responses with by-subject or by-item means close to 0 or 1), we used mixed-effects models with crossed random effects of subject and item (38) using the lme4 package in R (39). Complete details can be found in SI Appendix.

Acknowledgments. We thank E. Bakovic for assistance with norming studies, L. Frazier for helpful discussion, and E. H. Levy and E. Mukamel for helpful feedback on the manuscript. This work was supported by National Institutes of Health Grant HD26765 and the University of California at San Diego Academic Senate.

37. Stracuzzi D, Kinsey J (2009) Eyetrack Software for the SR Eyelink II and Eyelink 1000, Version 0.7.10k (University of Massachusetts Amherst, Amherst, MA).