The Time Course of Grammaticality Judgment

Arshavir Blackwell, Elizabeth Bates, and Daniel Fisher

Center for Research in Language
U.C. San Diego
La Jolla, CA 92093-0526
e-mail: arshavir@crl.ucsd.edu

Abstract

Two experiments investigating the time course of grammaticality judgment are presented, using sentences that vary in error type (agreement, movement, omission of function words), part of speech (auxiliaries vs. determiners) and location (early vs. late sentence placement). Experiment 1 is a word-by-word "gating" experiment, similar to the word-level gating paradigm of Grosjean (1980). Results show that some error types elicit a broad and variable "decision region" instead of a "decision point," analogous to results for word-level gating. Experiment 2 looks at on-line judgments of the same stimuli in an RSVP (Rapid Serial Visual Presentation) paradigm, with reaction times measured from several different points within each sentence based on the results of Experiment 1. Qualitatively different results are obtained depending on how and where the error point is defined. Results are discussed in terms of interaction activation models (which do not assume a single resolution point) and discrete parsing models.

This research was supported by NIH/NIDCD DC00216-10.

Introduction

For close to fifty years, grammaticality judgments by trained native speakers have been the method of choice for linguists working within the generative tradition. And yet we still know very little about the cognitive processes that underlie such judgments. Two experiments on the time course of grammaticality judgment are presented below. The two methods that we have chosen for these experiments (word-by-word gating, and rapid serial visual presentation or RSVP) are motivated by recent findings in grammaticality judgment in aphasia (Linebarger, Schwartz, & Saffran, 1983; Shankweiler, Crain, Gorrell, & Tuller, 1989; Wulfeck & Bates, 1991; Wulfeck, 1987), and by a particular interactive activation model called the Competition Model (MacWhinney & Bates, 1989). Because these models lead us to expect probabilistic changes in grammaticality across the course of the sentence, we need methods that permit us to evaluate degrees of perceived grammaticality on a word-by-word basis.

Experiment 1

Subjects: Subjects were thirty-five college students (five left-handed; twenty-two female and thirteen male) who participated in the experiment for course credit, or for a payment of $7.00. All subjects stated that they were native speakers of English.

Grammaticality Judgment Stimuli: Stimuli for the grammaticality judgment task include a total of 168 sentences: 84 ungrammatical target sentences, 40 grammatical control sentences matched for length and grammatical structure, and 44 distractors (22 grammatical and 22 ungrammatical). The design of the experiment is focused on the ungrammatical targets, which vary in the part of speech involved in the error (auxiliary vs. determiner), the position of the error within the sentence (early vs. late), and the kind of violation created from a common pool of grammatical types (i.e. errors of omission, agreement and transposition). The ungrammatical target sentences fall within a 2 x 2 x 3 design (with error type, error location, and part of speech as within-subjects variables). For each of these ungrammatical sentences, subjects also see a grammatical control sentence matched for length and grammatical structure. To keep the length of the experiment within reasonable bounds, some of these grammatical sentences were used as controls for more than one particular ungrammatical sentence. Sentences were randomly pulled from lists of sentences of different structure types. Because omission, agreement and transposition errors were all created from the same basic sentence types, it can be argued that these stimuli represent a set of minimal contrasts.

Procedure: A trial began with a "READY" cue appearing near the bottom center of the screen. The subject pressed the middle button, corresponding to "not sure," to bring the first word of the sentence to the screen. The sentence was centered vertically and started at the left side of the screen. Each button press brought the next word onto the screen, until the entire sentence was visible. After the last word appeared the button press caused the next "READY" cue to appear. The experimenter instructed subjects to decide, after each word appeared upon the screen, whether the sentence up to that point was grammatical.

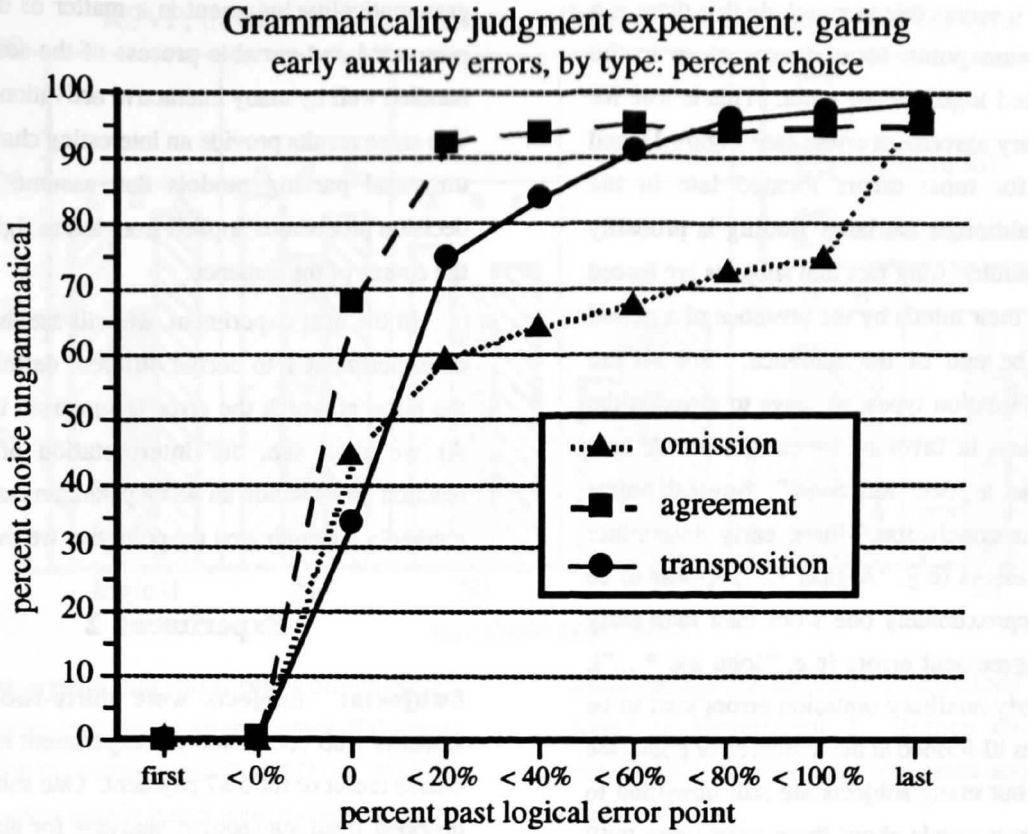
Scoring: A button press was recorded for every word of every sentence. Reaction time to each word was also recorded. The final button press was evaluated using A' to grammaticals and ungrammaticals combined. A' is a non-parametric statistic used to correct for response bias (Grier, 1971; Pollack & Norman, 1964; Grier, 1971). As such, it is similar to d'. The **logical error point** was defined as the first place in the sentence where an error could logically be detected. The **decision point** was defined as the number of words after the logical error point at which most errors were detected.

Summary of results for Experiment 1

Experiment 1 yielded a great deal of information about the time course of grammaticality judgment, summarized briefly as follows. Overall, end-of-sentence accuracy was very high in this experiment, averaging around 95% correct rejections for ungrammatical sentences and 95% correct acceptances for their grammatical controls. An analysis of variance A' scores (which corrects for response bias) yielded very few differences among the various error types, although performance was slightly worse overall for determiner omissions, especially when those omissions are located late in the sentence.

For both grammatical and ungrammatical stimuli, reaction times increase markedly on the last word of the sentence. This finding is similar

Figure 1



to the wrap-up effects reported by other investigators using word-by-word measures of reading. It may also be related to the protracted late positive voltage that is often reported for the last word in the sentence in studies using event-related brain potentials. For errors that are located late in the sentence, this means that we are faced with a confound between wrap-up effects and the increase in reaction times associated with detection/resolution of an error.

For the most part, there were striking parallels between the decision and reaction time data, suggesting that the word-by-word reaction times obtained with this gating technique can be viewed as an indirect index of the degree of confidence associated with grammaticality judgments at each point in the sentence, as well as, perhaps, a decision process in which subjects attempt to generate alternatives. However, the two data sets did diverge in some interesting ways. One example is the wrap-up effect summarized above (i.e. a rise in reaction times at the end of the sentence even though subjects do not change their minds about the grammatical status of these items). Another example comes from early determiner omissions, where subjects slow down markedly at the zero point (suggesting that they are questioning the sentence's well-formedness) even though they are still unwilling to conclude that the sentence is not well-formed. In general, we are convinced that both sources of information (word-by-word decisions and reaction times) offer useful and complementary information about the time course of grammaticality judgment.

The twelve relatively simple error types that we have manipulated here are associated with

markedly different decision functions. For some error types, it seems fair to conclude that there is a single decision point, located very close to the predetermined logical error point. This is true for early auxiliary agreement errors (see Figure 1), and it is true for most errors located late in the sentence—although the latter finding is probably due to the uninteresting fact that subjects are forced to make up their minds by the presence of a period signaling the end of the sentence. For all the remaining violation types, we have to abandon the punctate view in favor of something that is best described as a "decision zone." Several points support this conclusion. First, early determiner agreement errors (e.g. "A girls *...") appear to be resolved approximately one word later than early auxiliary agreement errors (e.g. "John are * ..."). Second, early auxiliary omission errors start to be perceived as ill-formed at the logical error point (see Figure 1), but many subjects are still unwilling to make up their minds about these error types until the very last word in the sentence. Third, early auxiliary transposition errors are resolved in at least two steps (see Figure 1): rejection rates start to go up at the logical error point (where omissions and transpositions are still equivalent), with a sharp increase at the next word (the displaced auxiliary, which serves as a second cue). Still, these errors do not reach asymptote until about 60% past the logical error point, suggesting that many subjects are unwilling to make up their minds until the end of the sentence. Finally, early determiner omissions and transpositions are still acceptable to our subjects at the logical error point. However, the reaction time data suggest that subjects are already doubtful about the grammatical status of these items.

When we put the data on individual variation together with the large "decision zones" observed for some item types (most notably early omissions and transpositions), it seems fair to conclude that grammaticality judgment is a matter of degree, a protracted and variable process of the sort that is handled well by many interactive activation models. The same results provide an interesting challenge to universal parsing models that assume discrete decision procedures applied from left to right across the course of the sentence.

In the next experiment, we will use the results of Experiment 1 to derive different definitions of the point at which the error is supposed to begin. As we shall see, our interpretation of simple reaction times within an RSVP paradigm can change markedly depending on the point that we use.

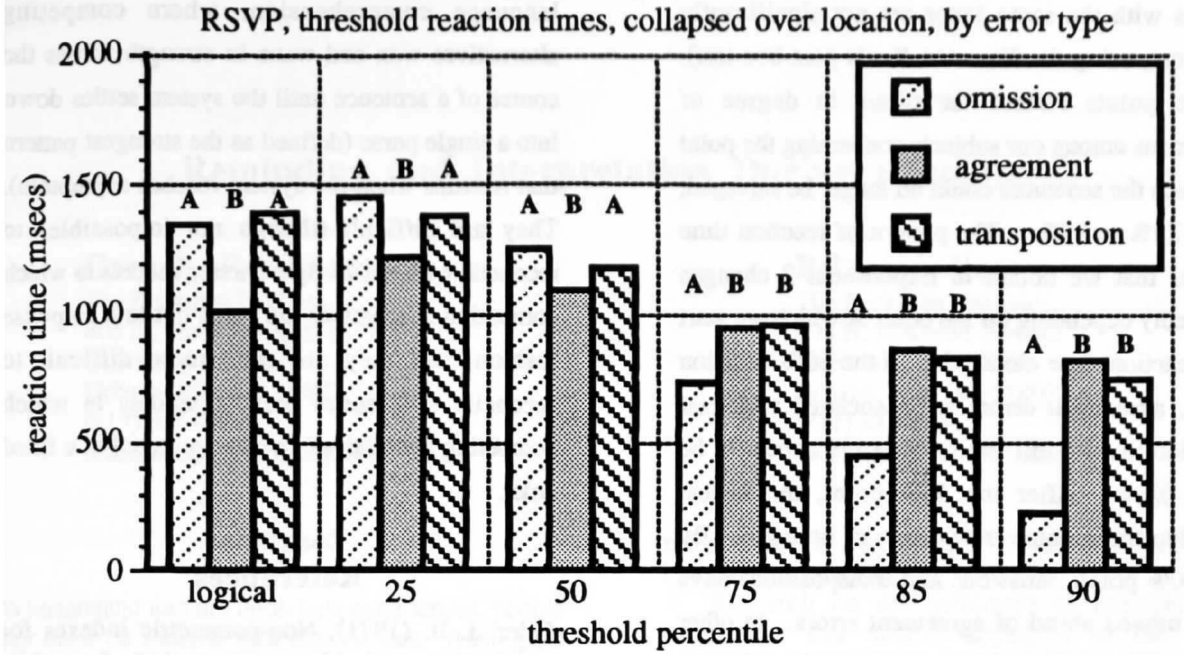
Experiment 2

Subjects: Subjects were thirty-two UCSD students who completed the experiment either for course credit or for a $7 payment. One subject was dropped from subsequent analyses for having A' scores more than 2.5 standard deviations from the mean. Of the thirty-one remaining subjects, twenty-five were male and two were left-handed. All subjects were native speakers of English.

Grammaticality Judgment Stimuli: The materials were the same as those used in Experiment 1.

Procedure: A trial consisted of the following: first, the. word "READY" appeared near the bottom center of the screen, for 1000 milliseconds (msecs). Second, the screen cleared, and a 2000-msec pause followed. Third, the sentence appeared in the middle center of the screen, one word at a time. Each word appeared for 350 msecs, without a pause between words. As soon as subjects had made the grammaticality judgment—even if the sentence was still running—they were to press the appropriate button. Finally, at the end of the sentence, the screen was blank for 3000 msecs, during which

Figure 2



time the program would still accept a button press. The following trial began after another 500-msec pause. Both the button press ("GOOD" or "BAD") and the reaction time in msecs were recorded at each trial. For ungrammatical sentences, reaction time was measured from the logical error point of Experiment 1. For grammatical sentences, reaction time was measured from sentence onset.

Summary of results for Experiment 2

The results of Experiment 2 are complementary in many respects to the results observed in Experiment 1. Overall accuracy levels were very high on Experiment 2, averaging around 93% correct rejections for ungrammatical stimuli and 91% correct acceptances for grammatical controls. An analysis of variance on A' scores (which corrects for response bias) suggests that accuracy levels are higher overall for transposition errors regardless of location or part of speech. The most vulnerable items are those that involve auxiliary agreement and determiner omission. The apparent disadvantage for determiner omissions was also found in Experiment 1. Hence the relative vulnerability of determiner omissions appears to be a robust finding.

In general, the fastest reaction times come from early violations of agreement and late violations of omission. The slowest reaction times and the largest decision zones come from early auxiliary omissions. These reaction time results are quite compatible with results from Experiment 1 on the size and shape of the decision zone for each item type. Indeed, these two indices were strongly correlated (+.91), suggesting that the reaction time results obtained in Experiment 2 are a direct reflection of the size of the decision zone for each item type.

In our view, the single most important conclusion from Experiment 2 comes from the final analysis comparing reaction time results under different definitions of the point at which an error officially begins. Results obtained with the logical error point were compared with five different decision points, all based upon the decision

functions observed in Experiment 1 (see Figure 2; means with the same letter are not significantly different, using the Newman-Keuls post-hoc test). These points reflect variations in degree of consensus among our subjects concerning the point at which the sentences could no longer be salvaged, from 25% to 90%. The pattern of reaction time results that we obtain in Experiment 2 changes markedly depending on the point at which we start the reaction time clock. Up to the 50% decision point, agreement errors are associated with fast reaction times while omission errors appear to be quite slow. After the 50% point, the fastest reaction times come from errors of omission. By the 90% point, omissions and transpositions have both moved ahead of agreement errors. In other words, the pattern of reaction time results is largely determined by the point at which we start the clock. This result is due, in turn, to there being no true decision point for most of these error types. In applying punctate reaction time techniques to a continuous and probabilistic reality, we are forced to make arbitrary decisions that can be quite misleading.

Conclusion

The purpose of this study was to investigate the time course of grammaticality judgment as a performance domain. Our sentence-level gating results were analogous in many respects to Grosjean's pioneering work (Grosjean, 1980) on lexical gating. In particular, some error types are associated with a clear-cut "decision point," while others are best described in terms of a protracted "decision region" with ample variability over items and subjects. Although these results cannot be used to rule out any particular model of sentence comprehension and parsing, the broad decision regions associated with some error types are easy to handle within interaction activation models of language comprehension, where competing alternatives wax and wane in strength across the course of a sentence until the system settles down into a single parse (defined as the strongest pattern that remains when the system reaches asymptote). They are difficult (though not impossible) to reconcile with left-to-right parsing models in which competing alternatives are ruled out in a stepwise fashion, and they are even more difficult to reconcile with serial parsing models in which competing alternatives are always tested in a fixed order.

References

Grier, J. B. (1971). Non-parametric indexes for sensitivity and bias: computing formulas. Psychological Bulletin, 75(6), 424-429.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics, 28*, 267-283.

Linebarger, M., Schwartz, M., & Saffran, E. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition, 13*, 361-392.

MacWhinney, B., & Bates, E. (Ed.). (1989). *The Crosslinguistic Study of Sentence Processing*. Cambridge: Cambridge Univ. Press.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of signal detection experiments. *Psychonomic Science, 1*, 125-126.

Shankweiler, D., Crain, S., Gorrell, P., & Tuller, B. (1989). Reception of language in Broca's aphasia. *Language and Cognitive Processes, 4*(1), 1-33.

Wulfeck, B., & Bates, E. (1991). Differential sensitivity to errors of agreement and word order in Broca's aphasia. *Journal of Cognitive Neuroscience, 3*, 258-272.

Wulfeck, B. B. (1987) *Sensitivity to grammaticality in agrammatic aphasia: processing of word order and agreement violations*. Doctoral dissertation, University of California, San Diego.