The use of "that" in the Production and Comprehension of Object Relative Clauses

David S. Race (drace@lcnl.wisc.edu)
Department of Psychology, University of Wisconsin-Madison
1202 W. Johnson Street, WI 53706 USA

Maryellen C. MacDonald (mcmacdonald@wisc.edu)
Department of Psychology, University of Wisconsin-Madison
1202 W. Johnson Street, WI 53706 USA

Abstract

We explore the interplay between production and comprehension by investigating why producers insert or omit the function word "that" in Object Relative Clauses, and how this choice affects comprehension. We present data from three experiments which suggests that producers insert "that" to alleviate production difficulty and in doing so create a distributional pattern of "that" use. Comprehenders are shown to be sensitive to these patterns. Implications for the interaction of comprehension and production processes are discussed.

Introduction

Language inherently depends on the integration of production and comprehension, yet each of these processes is typically studied in isolation. MacDonald (1999) argued against this tendency and suggested that a consideration of production processes could shed light on ambiguity resolution during comprehension. Specifically, she argued that the incremental nature of production (in which words or phrases that are relatively easier to access and produce tend to be placed earlier in the sentence, (Bock & Levelt, 1994)) create distributional patterns of word order in the language. Sensitivity to these distributional patterns in turn create biases in ambiguity resolution during comprehension. MacDonald (1999) suggested that sensitivity to production-motivated patterns could account for patterns of ambiguity resolution preferences that were otherwise unexplained in comprehension accounts, and without recourse to specialized ambiguity resolution principles. Here, we extend this argument to a different phenomenon in both production and comprehension. Whereas MacDonald (1999) had investigated production constraints on word order and their effect on ambiguity resolution, we investigate production factors motivating the optional use of "that" in Object Relative Clauses (ORCs), and the comprehension consequences of "that" use in this syntactically unambiguous structure.

ORCs such as "The story (that) she read was long", are so called because the subject of the main clause (The story) is the object of the embedded clause. As the example illustrates, the relative pronoun "that" can be optionally inserted or omitted without changing the meaning of the sentence. Previous comprehension studies have found that inclusion of "that" reduces comprehension difficulty (Hakes, & Cairns, 1970; Hakes, & Foss, 1970; Hakes, Evans, & Brannon, 1976). However, recent evidence from production of the complementizer "that" in sentential complement constructions such as "I know (that) you missed practice" suggests the use of optional "that" in this construction is modulated by production difficulty, with increased use of "that" when upcoming material is more difficult to produce than when it is easier (Ferreira & Dell, 2000). In this view, since "that" is a highly accessible and frequent word, it can be inserted to allow extra planning time for upcoming difficult material.

We conducted two experiments to determine whether "that" use in ORCs was similarly affected by production difficulty. A corpus analysis and oral production experiment investigated "that" use in ORCs in written and oral production, respectively. Finally, a self-paced reading experiment examined whether comprehenders were sensitive to the distributional patterns of "that" use.

Experiment 1

The purpose of the corpus analysis was to extract object relative clauses and identify production factors that lead to the insertion or omission of "that". We hypothesized that "that" is inserted in response to production difficulty. Therefore, it was predicted that the use of "that" in ORCs would increase as production difficulty increased (Ferreira & Dell, 2000). Production difficulty in ORCs can stem from several different factors. We specifically looked at the factors of Embedded Subject NP type (pronoun, common noun, proper noun), the use of a determiner in the embedded subject, length in words of the main subject NP, length in words of the embedded subject NP, and length in words of the embedded clause after the embedded subject NP. A regression analysis was then carried out to observe whether they reliably accounted for any of the variance of "that" use

Method

The Penn Trebank's tgrep utility (Marcus, Marcinkiewicz, & Santorini, 1993) was used to extract all sentences from the Wall St. Journal corpus for unreduced and reduced ORCs respectively. Although the relative pronoun can be realized by a number of lexical items (*that, which, who, whom*), only ORCs containing "that" or no relative pronoun were extracted for analysis**.** This was done mainly because "that" is less subject to prescriptive claims about when it should be used relative to who vs. whom. Eventually, for reduced ORCs 1017 sentences were returned, while 323 sentences were returned for unreduced ORCs. The reduced and unreduced sentences were then coded for the following factors; (a) "that" use, (b) Main subject NP length in words, (c) Embedded subject NP length in words, (d) Rest of the relative clause length in words, (e) Embedded Subject Type (pronoun, common noun, proper noun), (f) Presence or absence of a determiner within the embedded subject.

Results and Discussion

A hierarchical regression was conducted to uncover if any production difficulty factors accounted for reliable amounts of variance in "that" use. However, since Embedded Subject Type is neither dichotomous or continuous, we first conducted a chi-square analysis to determine whether this factor could be handled within a hierarchical regression. The analysis showed that the three NP types were not distributed evenly across both levels of "that" use, $\chi^2(2) = 120.85$, $p < .05$. Pairwise comparisons, with correction for continuity, showed that pronouns appeared more frequently in reduced sentences than either common nouns $\chi^2(1) = 87.82$, $p < .05$, or proper nouns $\chi^2(1) = 85.13$, $p < .05$, but common and proper nouns did not differ in their distributions, $\chi^2(1) = 0.02$, n.s. We therefore combined common and proper nouns into a single "noun" value (to contrast with "pronoun" in NP type) for the purposes of our analysis.

Table 1 presents the five factors that entered the regression analysis, in the order they entered, and the direction of each effect. The final result after all five factors entered was $r^2 = .15$, $p < .05$. In all cases, more difficult material to produce (increased length, absence of a determiner, a full noun phrase rather than a short, high frequency pronoun) yielded increased "that" use.

It appears that at least in written production, producers can take advantage of optional "that" in ORCs to allow more time for planning the embedded clause. The relatively small amount of variance accounted for by these five factors suggests that these are not the only constraints on "that" use, but it is clear that production difficulty does play a role in use of the relative pronoun.

Table 1. Five production difficulty factors in a hierarchical regression predicting "that" use.

Entering Factors	Direction of Effect	r^2
1. Embedded Subject noun Type	"that" use associated more with full noun phrase than pronoun	.09
2. Embedded subject noun Length	"that" use increased with increasing length of the embedded subject NP.	.11
3. Determiner in embedded NP	Less "that" use if embedded subject contained determiner	.13
4. Main subject noun phrase length	Increased use of "that" with the increasing length of the main NP.	.14
5. Rest of Embedded clause Length	Increased use of "that" with increasing length of embedded clause after the embedded subject.	.15

The corpus analysis proved to be a useful tool for uncovering the distributional pattern of "that" use in written production. Factors such as prescriptive rules during the editing process of writing may influence the final product of a written corpus in ways that do not occur in oral production, however. This concern motivated the next experiment, which manipulated "that" use and Embedded Subject Type, in a read and recall oral production task.

Experiment 2

A production experiment was conducted to determine whether the results of the corpus analysis could be extended to oral production. The objective was to manipulate production difficulty in oral production and observe whether it significantly affected the degree to which speakers inserted "that" in ORCs. A sentence recall task similar to that used by Ferreira and Dell (2000) was conducted, in which participants read a number of individually presented sentences and recalled them when cued. Since noun vs. pronoun Embedded Subject Type accounted for the most variance in Experiment 1, we manipulated it here as well as the presence of "that" in the stimulus sentences, which created four conditions (+that/common noun, +that/pronoun, -that/common noun, -that/pronoun).

The recall task is useful for this study because it constrains participants to use an object relative clause, but allows the optional use of "that". Furthermore, research has indicated that the short-term recall of sentences is generated from semantic

representations and proceeds through the system in much the same manner as spontaneous production (Potter & Lombardi, 1992). These claims taken together suggest that ORCs generated in this experiment will be subject to similar constraints found outside the lab.

Method

Participants. Sixty-four students from the University of Wisconsin-Madison received extra credit for their participation in this experiment. All participants were native speakers of English, and had normal or corrected to normal vision.

Materials. The experimental sentences consisted of 28 ORCs that were manipulated for the use of "that" and type of embedded subject. Example 1 below presents an experimental item in all four conditions of "that" use and embedded subject type.

1a. +that, common noun. The pastas that chefs prepared for the reception were sticky.
b. +that, pronoun. The pastas that they prepared for the reception were sticky.
c. -that, common noun . The pastas chefs prepared for the reception were sticky.
d. -that, pronoun . The pastas they prepared for the reception were sticky.

Each sentence began with a 2-word inanimate plural noun phrase such as, "The articles" which was optionally followed by "that". Next came the embedded subject noun phrase, which consisted of a plural animate noun such as "journalists", or the plural pronoun "they". This was followed by a past tense verb (e.g. wrote), and a three-word prepositional phrase. The main clause was reintroduced with the main verb "were", after which the sentence was completed with an adjective. Seventy filler items were also constructed ranging from 5-11 words in length. Ten of these were presented as practice items at the start of the experiment. The fillers described a variety of situations and used various types of syntactic constructions, including subject relative clauses.

Design/Procedure. A trial consisted of two stages, presentation and recall. In the presentation stage, two sentences were individually presented, and were read aloud by the participant. In the recall stage, cue words from each sentence were used to prompt recall of one of the sentences. There were two types of trials, experimental and filler. Experimental trials presented one filler and one experimental sentence, while filler trials presented two filler sentences. The stimuli were presented on an iMac using Psyscope software.

For experimental trials, the critical sentence was always presented first followed by the presentation of a non-relative clause sentence. In filler trials, the first sentence never contained a relative clause of any sort, and the second sentence was always a subject relative clause. All sentences were presented in the horizontal and vertical center of the screen for 5 seconds followed by a 1-second, blank screen interval. Participants had the task of reading each sentence aloud and remembering them for later recall. After the presentation of the second sentence, there was a 1-second interval before the recall stage began.

During the recall stage, the first two words of a presented sentence in that trial were displayed to cue recall of that sentence. For both experimental and filler trials, cue words from the first presented sentence were displayed first for half of the trials, and second for the other half. The cue words remained on the screen for 800 milliseconds after which they were immediately replaced by a "timebar" made of red asterisks. The timebar began with 5 asterisks and decreased in number by one every 1-second, giving participants five seconds in total to recall the sentence. After the timebar had expired, the experimenter used a button box to initiate the next display of cue words. The recall stage of the second sentence proceeded in the exact same manner as the first sentence, except that afterward the experimenter initiated the next trial. All of the experimental and filler trials were randomized. The lists were counterbalanced such that participants experienced every condition, but saw each item in only one condition. Participants were recorded on an audio tape so that their utterances could be transcribed. The experiment lasted approximately 30 minutes.

Results

Responses to the experimental stimuli were transcribed and coded for acceptability and “that” use. Trials were deemed unacceptable for any of the following reasons: 1) During the presentation phase, participants misread the sentence by inserting extra words, omitting words, or mispronouncing words (adding or dropping inflection for example). 2) Recalling a sentence in an ungrammatical or nonsensical manner. In effect an acceptable trial had a correctly read sentence during the presentation stage, and a grammatically acceptable object relative clause during the recall stage. Overall, this method of coding yielded 62% of the trials acceptable. Occasional missing cells (because all trials for a given participant and condition yielded unacceptable sentences) were replaced by the average level of “that” use from other cells for that participant.

Participants were expected to produce ORCs with embedded subjects that corresponded to the conditions in the presentation phase (+/- that-pronoun, +/- that-common noun). However, two new

response types were also produced, in that participants sometimes inserted the definite determiner "the" in the embedded subject of the common noun conditions, both when producing and omitting "that". The embedded subject nouns in the stimulus sentences had never contained "the" and thus this insertion represents a slight error in recall. The resulting sentences are fully grammatical however, and as "the" insertion may have similar motivations from production difficulty as "that" insertion, we included all sentences with "the" for additional analysis.

An analysis of variance was conducted to determine whether there was a significant difference of "that" use based on Embedded Subject Type (noun vs. pronoun), the presence of "that" in the initial sentence presentation, and whether the critical sentence was the first or second to be recalled in a trial. This Order of Recall factor did not have a significant effect on "that" use and will not be discussed further. Since the definite determiner "the" was predicted to play a similar role as the use of "that", for this analysis those two types of responses were collapsed into the category "that". The use of "that" was significantly higher when it had been present in the stimulus sentence [$F_1(1, 63) = 196.67$, $p < .05$; $F_2(1, 27) = 10.81$, $p < .05$]. Figure 1 shows the important result in this experiment, that "that" was produced more often in the common noun condition (75% "that" use) than in the pronoun condition (41% "that" [$F_1(1, 63) = 107.194$, $p < .05$; $F_2(1, 27) = 2.67$, $p < .05$]. Finally, there was a significant interaction such that the difference in "that" use between the common noun and pronoun conditions was larger when "that" was absent from the stimulus sentence [$F_1(1, 63) = 9.22$, $p < .05$; $F_2(1, 27) = 5.02$, $p < .05$].

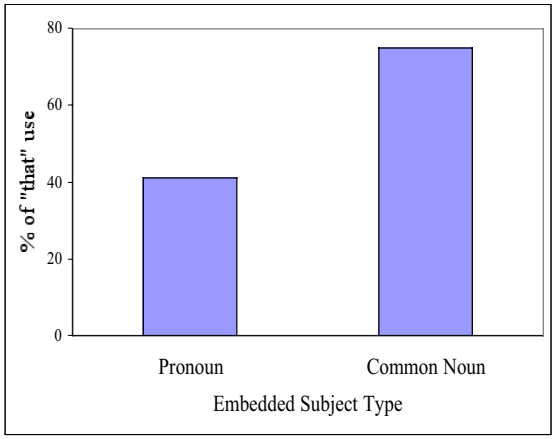

Figure 1: Percentage of "that" use for the Pronoun and Common Noun conditions.

Discussion

The results of Experiment 2 support the production difficulty hypothesis for oral production. When the embedded subject was a common noun, participants tended to produce "that" significantly more often than when the embedded subject was a highly accessible, frequent pronoun. Interestingly, participants at times inserted "the" in the embedded subject although it was not a part of any of the conditions in the presentation stage. Although the exact reason for this is not certain, as suggested earlier, this tendency may be due in part to alleviate production difficulty. Additional analyses may shed light on the exact patterns of "that" use.

The sum of the corpus and oral production studies suggests that as producers insert "that" to deal with production difficulty, a distributional pattern emerges. The use of "that" is associated with longer embedded phrases, common and proper noun embedded subjects, and lack of a determiner in the embedded subject. The next step in this study was to explore the interaction between production and comprehension by ascertaining whether comprehenders were sensitive to this distribution. The final experiment employed the self-paced reading task to observe whether participants' reading times reflected sensitivity to the probabilistic constraints imposed by "that" use and type of embedded subject.

Experiment 3

To test comprehender sensitivity to the distributional pattern of "that" in ORCs, a self-paced reading study was conducted which manipulated the use of "that" with the type of embedded subject (common noun vs. pronoun). By manipulating these variables, conditions were created that were either similar to the distributional pattern in normal production (+that with common noun, -that with pronoun) or dissimilar to that pattern (-that with common noun, +that with pronoun). If participants are sensitive to the distributional patterns of "that" use in ORCs, it is predicted that processing will be easiest when the conditions are similar to the distributional pattern, which entails that the insertion of "that" will not always be helpful. This prediction contrasts with previous claims that "that" use always helps comprehenders (Hakes, & Cairns, 1970; Hakes & Foss, 1970; Hakes, et al., 1976). These earlier studies contained a limited range of ORCs, however, and typically investigated only ones with full noun embedded subjects, that is, the subset that we predict will be helped by "that" use.

Method

Participants. Thirty-six undergraduates from the University of Southern California received extra

credit for participation in this study. Data from four participants were lost due to technical error. All participants were native speakers of English.

Materials. The items in this experiment were lengthened versions of the items used in Experiment 2. The variables of "that" (insertion vs. omission) and Embedded Subject Type (common noun vs. pronoun) were manipulated in exactly the same manner as in that study. The experimental items were lengthened by increasing the number of words after the main verb, so that factors involved with reading the end of the sentence would not influence reading times on the main verb. The experiment also contained 60 fillers, 10 of which were used for practice items. Yes/No comprehension questions were created for all items, with the correct answer being "yes" for half the items.

Procedure. The stimuli were presented on a computer screen in a moving window display. All non-whitespace characters of individually presented sentences were represented by dashes. By pressing a key, the first word of the sentence was revealed. The next keypress revealed the next word while the previous word reverted to dashes. The sentence was read in this manner until completion. After the sentence was completed, participants answered a comprehension question by pressing either a "yes" or "no" key. An error message was displayed for incorrect answers. The stimuli were presented in random order in one of four counterbalanced lists. Participants were instructed to read at a normal pace and to answer questions as quickly and accurately as possible. The experiment lasted approximately 25 minutes.

Results

The analyses of reading times included only sentences in which the participant answered the comprehension question correctly, which affected 11% of the trials. An analysis of variance (ANOVA) revealed that there was no significant difference in accuracy between conditions $F's < 1$.

Reading times were adjusted for length using a regression equation predicting reading time from word length, constructed for each subject, using both filler and experimental items (Ferreira and Clifton, 1986; Trueswell, Tanenhaus, & Garnsey, 1994). Length-adjusted reading times beyond 2 SD above the condition mean for that word were trimmed, affecting 5% of the data.

For the experimental sentences, reading times were examined from five regions (1) embedded subject, (2) embedded PP, (3) main verb, (4) embedded verb, and (5) the final phrase after the main verb. Reading times for regions 1-4 are shown in Figures 2 and 3. Our Hypothesis was that the use of "that" would be helpful in conditions similar to the distributional pattern (common noun embedded subjects) and detrimental in conditions where it was not similar (pronoun embedded subject). An ANOVA revealed an interaction of That, Embedded Subject Type, and Region Position [$F_1(9, 279) = 3.65$, $p < .01$; $F_2(9, 243) = 3.98$, $p < .01$]. For the common noun condition presented in Figure 2, unreduced sentences had faster reading times at every region. For the pronoun condition presented in Figure 3, reading times were similar at all regions except the main verb, where reduced sentences were significantly faster.

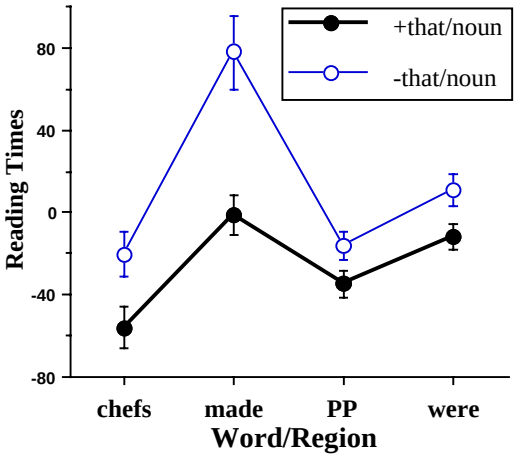

Figure 2: Length Adjusted Reading Times for the common noun conditions.

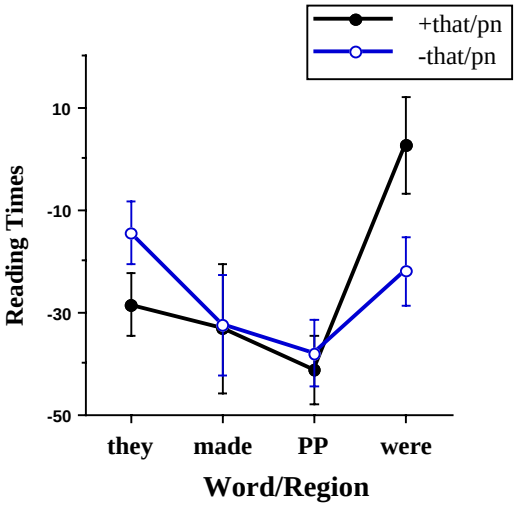

Figure 3: Length adjusted reading times for the pronoun conditions.

Discussion

The results of the comprehension experiment indicate that comprehenders are sensitive to the distributional properties of "that" in language production. The inclusion of "that" is helpful to comprehenders, but only in environments that parallel its use in language production. Experiments 1 and 2 converged to show that ORCs with a common noun embedded subject tended to be produced with "that" while those with pronoun embedded subjects tended to be produced without "that". Participants were sensitive to this pattern such that reaction times were shorter throughout unreduced ORCs with common noun embedded subjects, and at the main verb for reduced ORCs with pronoun embedded subjects. Importantly, for both types of embedded subject, "that" had an effect at the main verb, which has traditionally been a point to find experimental effects. In accordance with previous research, the comprehension of ORCs is dependant on the satisfaction of multiple probabilistic constraints.

General Discussion

The results of experiments 1-3 converge in supporting the production-constraint approach to the interaction between the production and comprehension systems. Results from Experiment 1 indicated that production difficulty factors, such as embedded subject type, and main NP length, can influence the accessibility of the embedded subject and therefore "that" use. Experiment 2 confirmed that production difficulty factors found to influence "that" use in written text, carry over to oral production. Finally, Experiment 3 determined that comprehenders are sensitive to the distributional pattern of "that" use in ORCs.

These results suggest that as with the complex syntactic ambiguities that MacDonald (1999) investigated, production difficulty in ORCs also creates clear distributional patterns that comprehenders can detect. The patterns in MacDonald's original investigation were dramatic changes in word order as a function of production difficulty, while the inclusion or omission of "that" seems much more subtle. Comprehenders nonetheless showed clear effects of "that" use in interpreting the ORCs. This suggests that the link between production constraints, distributional patterns, and comprehension processes is a potentially important area in comprehension and production work. More broadly, this work suggests that the integrated study of comprehension and production processes offers insight into the two systems that is not easily obtained when each is investigated individually.

References

Bock, J. K., & Levelt, W. (1994). Language production: Grammatical encoding. Gernsbacher, Morton Ann (Ed). (1994). Handbook of psycholinguistics. (pp. 945-984). San Diego, CA, US: Academic Press.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. Cognitive Psychology, 40, 296-340.

Ferreira, F., & Clifton, C., Jr., (1986). The independence of syntactic processing. Journal of Memory and Language, 25, 348-368.

Hakes, D. T., Evans, J. S., & Brannon, L. L (1976). Understanding sentences with relative clauses. Memory & Cognition, 4(3), 283-290.

Hakes, D. T., & Cairns H. S. (1970). Sentence comprehension and relative pronouns. Perception & Psychophysics, 8, 5-8.

Hakes, D. T., & Foss, D. J. (1970). Decision Processes during sentence comprehension: Effects of surface structure reconsidered. Perception & Psychophysics, 8(6), 413-416.

MacDonald, M. C., (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. MacWhinney, Brian (Ed). (1999). The emergence of language. (pp. 177-196). Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. Language and Cognitive Processes, 9, 157-201.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (19**93**). Building a large annotated corpus of English: The Penn Treebank.

Potter, M. C. & Lombardi, L. (1992). Theregeneration of syntax in short term memory. Journal of Memory and Language, 31, 713-733.

Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. Journal of Memory and Language, 35, 566-585.

Trueswell J. C., Tanenhaus M. K., & Garnsey, S, M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. Journal of Memory and Language, 33, 285-318.