Available online at www.sciencedirect.com
SCIENCE DIRECT®
COGNITION
ELSEVIER
Cognition 98 (2006) 245–272
www.elsevier.com/locate/COGNIT

Why is *that?* Structural prediction and ambiguity resolution in a very large corpus of English sentences☆

Douglas Roland*, Jeffrey L. Elman, Victor S. Ferreira

Center for Research in Language, University of California, San Diego, 9500 Gilman Drive, La Jolla, California, CA 92093-0526, USA

Received 11 February 2004; accepted 19 November 2004

Abstract

Previous psycholinguistic research has shown that a variety of contextual factors can influence the interpretation of syntactically ambiguous structures, but psycholinguistic experimentation inherently does not allow for the investigation of the role that these factors play in natural (uncontrolled) language use. We use regression modeling in conjunction with data from the British National Corpus to measure the amount and specificity of the information available for disambiguation in natural language use. We examine the Direct Object/Sentential Complement ambiguity and the closely related issue of complementizer use in sentential complements, and find that both ambiguity resolution and complementizer use can be predicted from contextual information.

From a certain perspective, linguistic expressions include massive lexical, structural, and acoustic ambiguity. Even when the ambiguities are ultimately resolved by subsequent information (e.g. at the end of a sentence), the incremental nature of most language processing suggests that comprehenders must deal with even temporary ambiguities during the course of sentence comprehension. Yet we seem to understand linguistic expressions with comparative ease, scarcely even noticing any ambiguities at all.

☆This research was funded by NIH/NIDCD5T32DC00041, NIH/NIMHR01-MH60517, and NIH/NIMHR01-MH64733.
* Corresponding author.
E-mail address: droland@crl.ucsd.edu (D. Roland).

0022-2860/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2004.11.008

The ways in which comprehenders resolve ambiguity has been a major focus of much research in sentence processing. According to constraint-based accounts of language processing (e.g. Altmann, 1998, 1999; MacDonald, Pearlmutter, & Seidenberg, 1994; MacWhinney & Bates, 1989; Spivey & Tanenhaus, 1998), ambiguity is resolved through the interaction of multiple sources of information contained in linguistic expressions. Alternatively, serial accounts of processing (e.g. Frazier, 1978) argue that the initial interpretations of utterances are based solely on syntactic information, and that these interpretations are later revised based on subsequent consideration of other information in the input.

Much of the evidence about the availability and use of information comes from controlled psycholinguistic experiments that typically examine one or two factors at a time. The high degree of control used in experimental designs is essential for determining if and when some theoretically important potential source of information can influence ambiguity resolution.

However, this sort of methodology leaves open a number of questions: Just how ambiguous are linguistic expressions? How much information is available for resolving ambiguity in typical naturally occurring contexts? How large a role do the experimentally studied factors play in natural (uncontrolled) language use and comprehension? Are there other factors that play a significant role in processing? What sorts of interactions occur when a larger number of factors are explored?

Here, we adopt an approach that is complementary to an experimental one. Rather than addressing the question of if and when a particular source of information is used during comprehension, we use corpus data to examine the quantity and variety of information available in normal language use, and to investigate sources of information that would otherwise be missed in the carefully controlled environment of psycholinguistic experiments. Indeed, this approach has the potential to provide information that is relevant to both constraint-based models of sentence processing and syntax-first models of sentence processing. For a syntax-first approach, our results will indicate the degree to which syntax-only heuristics such as minimal attachment can correctly resolve ambiguity, and the extent to which the predictions of the first stage need to be revised by a more general second stage. Our results will also provide information about the information that is used to revise the initial syntax-based predictions. For a constraint-based approach, the analysis of the information available in naturally occurring data provides an indication of the relative importance of factors that are typically investigated in separate experiments, as well as the interaction among these factors.

Knowing how much information is available during normal language comprehension also shows the extent to which the set of factors under consideration can account for language processing. If it turns out that a limited set of factors can be used to resolve nearly all naturally occurring ambiguity, it would suggest that a complex system employing a wide variety of factors may be unnecessarily complicated. On the other hand, if the naturally occurring contexts turn out to be information-poor, it could suggest that an alternate mechanism for ambiguity resolution might be needed.

In order to investigate the amount and variety of information available during natural language comprehension, we identify a relatively large number of general factors that might affect language performance, and investigate those influences using correlational