What a Rational Parser Would Do

John T. Hale

Department of Linguistics, Cornell University

Received 4 June 2009; received in revised form 30 July 2010; accepted 31 July 2010

Abstract

This article examines cognitive process models of human sentence comprehension based on the idea of informed search. These models are rational in the sense that they strive to find a good syntactic analysis quickly. Informed search derives a new account of garden pathing that handles traditional counterexamples. It supports a symbolic explanation for local coherence as well as an algorithmic account of entropy reduction. The models are expressed in a broad framework for theories of human sentence comprehension.

Keywords: Comprehension; Entropy reduction; Heuristic; Rationality; Sentence processing; Syntax

1. Introduction

As in other areas of cognitive science, the study of human sentence comprehension naturally proceeds at multiple levels. Marr (1982) draws a distinction between scientific questions posed at the computational level and questions at the algorithmic level. At his highest level, the computational level, one asks: What task does a sentence understanding carry out? What sort of mental representations are constructed, as the outputs of sentence comprehension? What features of the comprehended language are brought to bear, as inputs to the comprehension task? Is this input-output mapping “easy” or “hard” when considered as a formal problem on its own? Marr’s distinction separates these concerns from others that pertain to the algorithmic level. At this lower level, one asks: What procedure accomplishes the calculation of the function defined at the computational level? What is the time course of this calculation, and what is its resource utilization in terms of abstract resources like storage? This separation into independent levels of analysis constitutes one of the most distinctive methodological features of cognitive science.

Correspondence should be sent to John T. Hale, Department of Linguistics, Cornell University, Room 217, Morrill Hall, Ithaca, NY 14853-4701. E-mail: jthale@cornell.edu
Heuristic search in a cognitive model of human parsing

John T. Hale
Cornell University
217 Morrill Hall
Ithaca, New York 14853
jthale@cornell.edu

Abstract

We present a cognitive process model of human sentence comprehension based on generalized left-corner parsing. A search heuristic based upon previously-parsed corpora derives garden path effects, garden path paradoxes, and the local coherence effect.

1 Introduction

One of the most interesting applications of parsing technology has, for some researchers, been psycholinguistic models (Kay, 2005). Algorithmic models of language use have led in the past to a variety of cognitive insights (Kaplan, 1972; Marcus, 1980; Thibadeau et al., 1982; Pereira, 1985; Pulman, 1986; Johnson, 1989; Stabler, 1994). However they are challenged by a veritable tidal wave of new data collected during the 1990s and 2000s. Work during this later period reveals phenomena, such as the local coherence effect discussed in section 5, that have yet to be truly integrated into any particular theoretical framework.

This short paper presents a parsing system intended to serve as a model of the syntactic part of human sentence comprehension. Such a model helps make sense of sentence-difficulty data from self-paced reading, eye-tracking and other behavioral studies. It also sketches a relationship between calculations carried out in the course of automated syntactic analysis and the inferences about linguistic structure taking place in our minds during ordinary sentence-understanding.

Section 2 defines the model itself, highlighting its relationship to generalized left-corner parsing. Sections 3–5 apply this model to three controversial phenomena that are well-established in the psycholinguistics literature. Section 6 concludes.

2 Architecture of the model

2.1 Problem states and Operators

We model the human sentence comprehension mechanism as search within a problem space (Newell and Simon, 1972). We assume that all (incremental) parser states have a (partial) grammatical interpretation (Chomsky, 1965, 9). In this paper, the grammatical interpretation employs context-free grammar. An inventory of operators carries the model from one point in the problem space to another. In the interest of simplicity, we place no bound on the number of problem states the model can explore. However, we do acknowledge with Johnson-Laird (1983) and Resnik (1992) a pressure to minimize memory consumption internal to a problem state.

The model’s within-problem state memory usage should reflect human acceptability judgments with embedded sentences. These considerations motivate a generalized left-corner (GLC) parsing strategy (Demers, 1977) whose stack consumption is maximal on just the center-embedded examples that are so difficult for people to understand. To reflect the argument/adjunct distinction (Tutunjian and Boland, 2008) we adopt a mixed strategy that is bottom-up for optional postmodifiers but left-corner everywhere else. Leaving the arc-eager/arc-standard decision (Abney and Johnson, 1991) to the control policy allows four possible operators, schematized in Table 1.

2.2 Informed Search

Informed search differs from uninformed search procedures such as depth-first and breadth-first by making use of heuristic knowledge about the search domain. The strategy is to choose for expansion the node whose cost is lowest (Barr and Feigenbaum, 1981, 61). In A* search (Hart et al., 1968) this cost is divided up into a sum consisting of the known cost to reach a search node and an
shift a word $W$ | project a rule $LHS \rightarrow Trigger \downarrow Rest$
scan the sought word $W$ | project and match the sought parent $LHS$ using the rule $LHS \rightarrow Trigger \downarrow Rest$

Table 1: Four schema define the operators

<table>
<thead>
<tr>
<th>stack</th>
<th>n</th>
<th>$E[\text{steps}]$</th>
<th>standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>[VP] S [TOP]</td>
<td>55790</td>
<td>44.936</td>
<td>0.1572</td>
</tr>
<tr>
<td>S [TOP]</td>
<td>53991</td>
<td>10.542</td>
<td>0.0986</td>
</tr>
<tr>
<td>[NP] S [TOP]</td>
<td>43635</td>
<td>33.092</td>
<td>0.1633</td>
</tr>
<tr>
<td>NP [TOP]</td>
<td>38844</td>
<td>55.791</td>
<td>0.2126</td>
</tr>
<tr>
<td>NP [S] S [TOP]</td>
<td>34415</td>
<td>47.132</td>
<td>0.2122</td>
</tr>
<tr>
<td>[S] S [TOP]</td>
<td>33578</td>
<td>52.800</td>
<td>0.2195</td>
</tr>
<tr>
<td>PP S [TOP]</td>
<td>30693</td>
<td>34.454</td>
<td>0.1915</td>
</tr>
<tr>
<td>IN [PP] S [TOP]</td>
<td>27272</td>
<td>32.379</td>
<td>0.2031</td>
</tr>
<tr>
<td>DT [NP] S [TOP]</td>
<td>22375</td>
<td>34.478</td>
<td>0.2306</td>
</tr>
<tr>
<td>[AUX] [VP] S [TOP]</td>
<td>16447</td>
<td>46.536</td>
<td>0.2863</td>
</tr>
<tr>
<td>VBD [VP] S [TOP]</td>
<td>16224</td>
<td>43.057</td>
<td>0.2826</td>
</tr>
<tr>
<td>VB [VP] S [TOP]</td>
<td>13548</td>
<td>40.404</td>
<td>0.3074</td>
</tr>
<tr>
<td>the [NP] S [TOP]</td>
<td>12507</td>
<td>34.120</td>
<td>0.3046</td>
</tr>
<tr>
<td>NP [NP] S [TOP]</td>
<td>12092</td>
<td>43.821</td>
<td>0.3269</td>
</tr>
<tr>
<td>DT [TOP]</td>
<td>10440</td>
<td>66.452</td>
<td>0.3907</td>
</tr>
</tbody>
</table>

Table 2: Popular left-corner parser states. Stacks grow to the left. The categories are as described in Table 3 of Marcus et al. (1993).

The control condition 1b includes a comma which, in spoken language, would be expressed as a prosodic break (Carroll and Slowiaczek, 1991; Speer et al., 1996).

Figure 1 shows the search space explored in the experimental condition 1a. In this picture, ovals represent problem states. The number inside the oval encodes the vistation order. Arcs between ovals represent operator applications. The path (14, 22, 23, 24, 25, 29, 27) is the garden path which builds a grammatical interpretation where a sock is attached as a direct object of the verb mend. The grey line highlights the order in which $A^*$ search considers this path. At state 21 after shifting sock, experience with the Brown corpus suggests reconsidering the garden path.

Whereas the model examines 45 search nodes during the analysis of the temporally ambiguous item 1a, it dispatches the unambiguous item 1b after only 40 nodes despite that sentence having an additional token (the comma). Garden paths, on this view, are sequences of parser states explored only in a temporarily ambiguous item.

4 Garden pathing counterexamples

Purely structural attachment preferences like Right Association (Kimbail, 1973) and Minimal Attachment (Frazier and Fodor, 1978; Pereira, 1985) are threatened by paradoxical counterexamples such as 2 from Gibson (1991, 22) where no fixed principle yields correct predictions across both examples.

(2) a. I gave her earrings on her birthday.
   b. I gave her earrings to another girl.

A parser guided by Hypothesis 1 interleaves the garden path attachment and the globally-correct attachment in both cases, resulting in a search that
is strictly committed to neither analysis. In 2a, 32% of discovered states represent the globally-incorrect attachment of her. In 2b, 27% of states represent the globally-incorrect attachment of her to give as a one-word direct object. The paradox for purely structural attachment heuristics is dissolved by the observation that neither pathway fully achieves priority over the other.

5 Local Coherence

Tabor et al. (2004) discovered\(^1\) a processing difficulty phenomenon called "local coherence." Among the stimuli they considered, the locally-coherent condition is 3a where the substring the player tossed a frisbee could be analyzed as a sentence, if considered in isolation.

(3) a. The coach smiled at the player tossed a frisbee by the opposing team.
   b. The coach smiled at the player thrown a frisbee by the opposing team.
   c. The coach smiled at the player who was tossed a frisbee by the opposing team.
   d. The coach smiled at the player who was thrown a frisbee by the opposing team.

Tabor and colleagues observe an interaction between the degree of morphological ambiguity of the embedded verb (tossed or thrown) and the presence or absence of the relative-clause initial words who was. These two factors are known as ±ambiguity and ±reduction, respectively. If the human parser were making full use of the grammar, its operation would reflect the impossibility of continuing the coach smiled at... with a sentence. The ungrammaticality of a sentence in this left context would preclude any analysis of the player as a subject of active voice toss. But greater reading times observed on the ambiguous tossed as compared to the unambiguous thrown suggest contrariwise that this grammatical deduction is not made uniformly based on the left context.

Table 3 shows how an informed parser’s step counts, when guided by Hypothesis 1, derive Tabor et al.’s observed pattern. The cell predicted to be hardest is the local coherence, shaded gray. The degree of worsening due to relative clause reduction is greater in +ambiguous than in −ambiguous. This derives the observed interaction.

\(^1\)Konieczny and Müller (2006) documents a closely related form of local coherence in German.
6 Conclusion

When informed by experience with the Brown corpus, the parsing system described in this paper exhibits a pattern of "time-sharing" performance that corresponds to human behavioral difficulty in three controversial cases. The built-in elements — context-free grammar, generalized left-corner parsing and the A*-type cost function — are together adequate to address a range of comprehension difficulty phenomena without imposing an a priori memory limit. The contribution is an algorithmic-level account of the cognitive processes involved in perceiving syntactic structure.

References


Lars Kühnert and Daniel Müller. 2006. Local coherences in sentence processing. CUNY Conference on Human Sentence Processing.


