Combining preferences in parsing corpus items

Stephan Mehl

Univ. of Duisburg
Computational Linguistics
D-47048 Duisburg
Tel. +49-203-3792876
Fax +49-203-3792008
email: he234me@unidui.uni-duisburg.de

Abstract submitted to CPL '97

Preference: poster presentation

Combining preferences in parsing corpus items

Several recent models of psycholinguistic processing assume that a combination of syntactic and semantic factors determine the resolution of syntactic ambiguities. Such factors include

- frequency of particular syntactic rules,
- subcategorization constraints and preferences,
- lexical preferences for thematic grids,
- preferences for particular morphological, syntactic or semantic readings of a homograph, homonym, or polyseme, as well as
- semantic and referential plausibility ratings of syntactic structures

(cf. MacDonald 1994, Trueswell, Tanenhaus und Garnsey 1994, Jurafsky 1996, Trueswell 1996). However, these so-called constraint-satisfaction models are subject to contraversies; for example, syntax-first models assume that semantic analysis only comes into play after the syntactic processor has decided on one particular structural reading, while discourse-oriented models consider the overall semantic and pragmatic context as guiding syntactic analysis.

Because of the complexity of the problem, psycholinguistic studies have to confine themselves to proving the relevance of particular factors to the processing of single types of syntactic ambiguity. However, realistic sentences consist of a combination of various types of ambiguity, many of which have rarely been investigated in psycholinguistic experiments. For example, consider the ambiguities that arise during the incremental analysis of the following sentence, arbitrarily chosen from a linguistic textbook (Lyons 1977: 374):

Like ...	preposition or verb?
Like most linguists we take the view that ...	determiner, pronoun, or complementizer?
Like most linguists we take the view that at ...	preposition or part of an adverbial phrase?
Like most linguists we take the view that at least some ...	*at least* modifying verb (yet to come) or *some*?
Like most linguists we take the view that at least some part of ...	beginning of relative clause (*of which ...*) or PP?
Like most linguists we take the view that at least some part of what is ...	main verb or auxiliary?
Like most linguists we take the view that at least some part of what is covered by ...	PP attachment?
etc.	
Like most linguists we take the view that at least some part of what is covered by the term 'prosodic' should be handled in describing the structure of sentences.	

This paper investigates the assumption that all of the above-listed factors apply to the resolution of all kinds of syntactic ambiguity, and explores the applicability of these factors in parsing realistic corpus items. This is done by implementing an incremental parser that is based on these factors. The parser is intended to yield answers to the following questions that tackle important aspects of the constraint-satisfaction approach:

1. In which cases do different factors favor the same syntactic reading, in which cases do conflicts arise?

2. How can cases be handled in which information on particular factors is still missing at the point where the ambiguity occurs?

Given the multitude of factors and types of ambiguity, the first question can best be investigated by a computer simulation. This requires, however, that plausibility ratings can be defined for both large-scale dictionaries and grammars. Since such ratings are not yet available for arbitrary lexical items such as those occurring in corpus sentences, the current version of the program uses binary ratings (i.e. alternative lexical and structural readings are rated as "more/less frequent" or "equally frequent" without giving numerical weigths). As far as possible, these ratings were taken either from statistical studies in computational linguistics or from the psycholinguistic literature.

It is the second question that the current implementation focusses on. Often, crucial information is not yet available when an attachment ambiguity arises. For example, PP attachment cannot be semantically evaluated before the processing of the head noun. In other cases, it might be sensible to wait for syntactic information following close to the occurrence of an ambiguity, if the syntactic and semantic information available so far fails to be conclusive. To test the consequences of different processing assumptions, the behaviour of the parser can be directed by parameters individually for each type of ambiguity. For example, the parser can be tuned to work strictly incrementally (e.g., a PP is attached as soon as the preposition occurs), or to wait for the occurrence of the semantic head (e.g., a PP is attached as soon as the head noun occurs), or to make an attachment decision if a certain threshold of information is reached (e.g., the PP is attached as soon as the preposition occurs if and only if a PP complement is expected), or to wait for a maximum of *n* additional words. While it is desirable to choose a uniform strategy for all types of ambiguity, the model also allows for individual strategies for different types of ambiguity.

Depending on these parameters, the implementation yields different results of disambiguation. For example, an incremental attachment of PPs will only be revised if the attachment chosen is absolutely inacceptable from a semantic point of view. This may lead to different results than an explicit semantic comparison of all attachment alternatives. To take another example, attaching ambiguous relative clauses when processing the relative pronoun is possible only with respect to pragmatic criteria (i.e. if one of the prospective matrix nouns needs further specification to be referentially identifiable). Following this strategy, the system might e.g. generate the following user enquiry:

```
Sentence to be analysed:
Nor does there seem to be any other way of formulating the question that ...
Which noun needs further specification by a relative clause?
(1) way
(2) question
```

This strategy might lead to a different result than a straightforward semantic analysis at the end of the relative clause:

```
Sentence to be analysed:
Nor does there seem to be any other way of formulating the question that is not open to similar objections.

Which noun is modified by the relative clause?
(1) way
(2) question
```

The parser thus serves as a facility for testing the consequences of different processing assumptions, yielding disambiguation results that might in turn serve as hypotheses for further psycholinguistic research.

Currently, the grammar integrated into the parser covers a corpus of 30 sentences from Lyons 1977. The sentences have an average length of 25 words, and include numerous grammatical phenomena such as negation, imperatives, passive voice, gerunds, complement clauses, unbounded dependencies, and several variants of coordination. Details on the types of ambiguity handled, as well as on the preference values employed, will be given in the full paper. The parser is fully implemented in Common Lisp, and can be demonstrated at the conference.

References

Jurafsky, Daniel (1996): A Probabilistic Model of Lexical and Syntactic Access and Disambiguation. *Cognitive Science* 20, 137-194.

Lyons, John (1977): *Semantics.* 2 vols. Cambridge University Press.

MacDonald, Maryellen C. (1994): Probabilistic Constraints and Syntactic Ambiguity Resolution. *Language and Cognitive Processes* 9(2), 157-201.

Trueswell, John (1996): The Role of Lexical Frequency in Syntactic Ambiguity Resolution. *Journal of Memory and Language* 35, 566-585.

Trueswell, John; Tanenhaus, Michael; Garnsey, Susan (1994): Semantic Influences on Parsing: Use of Thematic Role Information in Syntactic Ambiguity Resolution. *Journal of Memory and Language* 33, 285-318.