

## A Psycholinguistics-Based Parsing Prototype

Hai Doan-Nguyen - Université du Québec à Montréal

### *Abstract*

This paper presents a natural language parsing prototype which applies psycholinguistic knowledge on human sentence parsing into the computing process. Linguistically the parser is based on the Asymmetry Theory (Di Sciullo, 1999; 2000). I will concentrate here on the computational aspect, especially the parsing strategy.

Traditional parsers are generally based on context-free grammar rules. Well-known strategies such as CYK, Earley, left-corner, Tomita algorithms build all solutions whose number is exponential to the length of the input. The great number of output parses makes it impractical to use them in real applications (machine translation, natural language understanding, information extraction, etc.) It is very difficult, not to say impossible, to define the best parse and to output it as the first parse. Many approaches have been developed to improve the situation: probabilistic, connectionist, etc. One promising approach is to try to imitate human parsing process. Recent developments of psycholinguistics (see eg. Frazier, 1998, for an overview) give hopes for this. However, it seems that few psycholinguistics-based computer parsing prototypes have been implemented (Marcus, 1980, Lombardo, 1998).

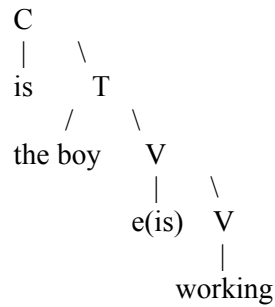
The work I develop here goes in this direction. I choose to implement a serial model, and not a (limited) parallel one. (See eg. Frazier 1998 for reasons against the later.) Parsing comprises of two main phases: first analysis and reanalysis, similar to many serial models proposed recently on human sentence processing (Fodor & Inoue, 1998; Frazier & Clifton, 1998). The best parse, ie. the most plausible parse computed using all the linguistic knowledge the parser has in the analysis process, is output. The parser goes under psycholinguistic established principles, such as Minimal Attachment, Late Closure, Minimal Revision, etc. In cases where psycholinguistic knowledge is not available, or even when it is still vague or controversial, particular computing techniques will be implemented for the benefit of the performance of the system.

The first analysis is generally data-driven (bottom-up), but also includes top-down control and development when necessary. (This agrees with Frazier, 1998). The parser tries to attach the incoming element into the current partial parse, using extensively information from the context to define the attachment position. For some ambiguous cases, it may delay the attachment to carry out searching forward, in a top-down control fashion, to get enough information for a good attachment. Reanalysis is evoked if the attachment fails, if the parse is not complete at the end of input, or if the incoming element suggests the current parse needs to be revised for a better configuration.

As a quick example, let's consider the parsing of:

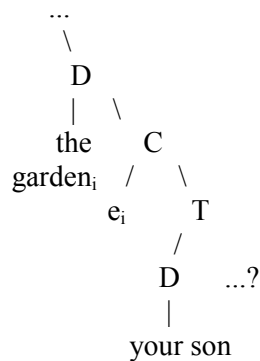
*Is the boy working in the garden your son?*

At the time '*working*' arrives, it is preferred to make it the complement of '*be*', rather than an adjunct of '*the boy*' (Minimal Attachment):

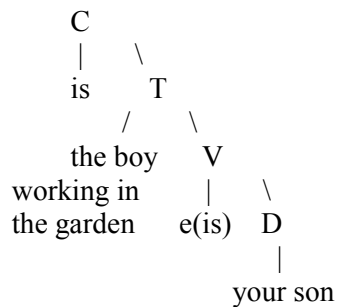


'In the garden' will then be adjoined to 'working'. Now when 'your son' comes, according to Late Closure, it will be attached to 'the garden' as the beginner of a potential relative clause, as in:

*Is the boy working in the garden your son mowed yesterday?*



However, if input stops here, the parser meets a trouble because the T is not completed. This will provoke a reanalysis which tries to eliminate the incomplete T node. The parser has to move 'your son' to some other position, the only one is the complement position of *e(is)*. This is acceptable if 'working in the garden' can be lowered to attach to 'the boy', and fortunately it can. The final structure is:



The prototype can now treat a relatively wide coverage of English grammar: noun phrases (including possessive, numeral, present/past participles as adjectives), verb phrases (subcategorizations, auxiliaries, tenses, passive, infinitive), relative clauses, questions (yes-no and wh-questions), movements and empty category elements, etc. However, many well-known difficult issues in parsing natural language are not yet treated, such as lexical ambiguities (my parser currently just treats subcategorization ambiguities), conjunctions, punctuations, etc.

The prototype shows that a psycholinguistic-based parser may be implementable, realistic, and beneficial. The approach is promising in that it tries to combine human's amazing performance with computer-specific

techniques. An interesting research direction to continue in the future is to apply psycholinguistic studies on semantic and pragmatic treatment in human sentence processing into the computer parser.

### ***Bibliography***

Di Sciullo, A.M., 1999. *An Integrated Competence-Performance Model, A Prototype for Morpho-Conceptual Parsing and Consequences for Information Processing*. In Proceedings of VEXTAL. Università Ca'Foscari Venezia.

Di Sciullo, A.M., 2000. *Parsing Asymmetries*. Second International Conference on Natural Language Processing NLP 2000: Filling the Gap between Theory and Practice. Patras, Grèce.

Fodor, J.D. and Inoue, A., 1998. *Attach Anyway*. In J.D. Fodor and F. Ferreira (eds) *Reanalysis in Sentence Processing*, 101-141. Kluwer Academic.

Frazier, L., 1998. *Getting There (Slowly)*. Journal of Psycholinguistics Research, Vol 27, No. 2, 1998.

Frazier, L. and Clifton, C., 1996. *Construal*. Cambridge, MA: MIT Press.

Frazier, L. and Clifton, C., 1998. *Sentence Reanalysis, and Visibility*. In J.D. Fodor and F. Ferreira (eds) *Reanalysis in Sentence Processing*, 143-176. Kluwer Academic.

Lombardo, V., 1998. *A Computational Model of Recovery*. In J.D. Fodor and F. Ferreira (eds) *Reanalysis in Sentence Processing*, 287-325. Kluwer Academic.

Marcus, M., 1980. *A Theory of Syntactic Recognition for Natural Language*. Cambridge, MA: MIT Press.