A Large-Scale Parser for Unrestricted Text
by
Carl Gustave de Marcken

Submitted to the
Department of Electrical Engineering and Computer Science

May 25, 1990

In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

ABSTRACT

This thesis presents a parser that operates rapidly over large quantities of unrestricted text. The parser is currently used on the Lancaster/Oslo-Bergen corpus, a million word database of unedited English text. Rather than attempt to provide full parses of all the sentences in the corpus, the parser instead breaks the text into the largest chunks it can reliably parse. This not only increases the speed and accuracy of the parser, it also allows for selective parsing and reduced output ambiguity. The parser

- is the fastest general purpose parser known, with .04 seconds per word parsing time;
- outputs only one parse, in sharp contrast to the hundreds of parses many current parsers produce;
- can be applied to arbitrary text, giving researchers access to far greater linguistic data;
- parses substructures of complex phrases that might otherwise be difficult to analyze;
- can be configured to parse only limited substructures, for data collection purposes;
- can modify its output to conform to any number of representations;
- essentially solves the problem of accurate part-of-speech disambiguation.

Thesis Supervisor: Robert C. Berwick
Title: Professor of Computer Science

A Large-Scale Parser for Unrestricted Text

by

Carl Gustave de Marcken

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science in Computer Science and Engineering

at the Massachuesetts Institute of Technology

June 1990

Author________________________________

Department of Electrical Engineering and Computer Science
May 25, 1990

Certified by________________________________

Robert C. Berwick
Thesis Supervisor

Accepted by________________________________

Leonard A. Gould
Chairman, Department Committee on Undergraduate Theses