SINTAX AS CLIPPING BLOCKS: A PARSER AS EXPERIMENTAL DEVICE

J. Vergne

Laboratoire d'Informatique, Université de Caen
F-14032 Caen cedex France

Research topic

My research topic is into **syntax of natural languages**, with, as a guideline, the very high formal redundancy of natural languages. This redundancy allows to make a morpho-syntactic parser working without an exhaustive dictionary.

Methodology

The methodology comes from the vision of linguistics as a science: corpus observation, experiments upon this corpus, using a **morpho-syntactic parser as an experimental device** to validate the theory.

Parsing strategy

A sentence is parsed progressively block by block to limit the combinatory of words possible categories in a segment as small as possible. That is why the global complexity of the parser is linear in time for 95% of sentences.

The sentence pattern validation is a sequential process **from the outside to the inside** of the sentence structure by removing the block patterns. During this validation, nominal sequences are processed as a whole.

Features of the parser

Parsing quality on a corpus of about 7 000 words

The parsing quality is automatically observed with a statistic tool which allows to collect the syntactic forms and measure the gap between the expected and the observed behavior of the model:

- gap on categories: ≈ 1.3% / words
- gap on **algorithmic** dependencies: ≈ 1.2% / algorithmic dependencies (75.5% / all dependencies)
- gap on **heuristic** dependencies: ≈ 12.4% / heuristic dependencies (24.5% / all dependencies)

Technical realization

- programming language: Think Pascal
- machine: Macintosh II
- source size: ≈ 20 000 lines, 860 Ko
- code size: ≈ 400 Ko
- research and development: ≈ 4 man-years

Input

The possible categories of words are computed **without an exhaustive dictionary**, only with a lexicon of grammatical words (400), and with ending rules (300).

Output

The parser outputs the results into following files:

- by sentence: features of each word, relations (dependencies, co-ordinations, references), and the determination tree;
- other results, grouped and classed on the whole text:
  - the lexicon of the text, lemmatized, classed by category, with statistics upon categories,
  - relations of the text, classed by syntactic type,
  - problems met during the parsing,
  - statistics about text size, processing speed, tested patterns, categories, deduction modes, relations and patterns of the text and parsing quality.