Heuristisk analys med Diderichsens satsschema

Tillämpningar för svensk text

Kenneth Wilhelmsson

Institutionen för filosofi, lingvistik och vetenskapsteori, Göteborgs universitet

Akademisk avhandling för avläggande av filosofie doktorsexamen i allmän språkvetenskap vid Göteborgs universitet, som med tillstånd av humanistiska fakultetsnämnden, kommer att offentligen försvaras lördagen den 24 april, kl. 10.15 i T307, Olof Wijksgatan 6, Göteborg

Ph.D. Thesis in linguistics at University of Gothenburg 2010

Title: Heuristisk analys med Diderichsens satsschema – Tillämpningar för svensk text
English title: Heuristic Analysis with Diderichsen's Sentence Schema – Applications for Swedish Text
Author: Kenneth Wilhelmsson
Language: Swedish (including summary in English)
Department: Department of Philosophy, Linguistics and Theory of Sciences

Abstract

A heuristic method for parsing Swedish text, *heuristic schema parsing*, is described and implemented. Focusing on main clause (*primary*) analysis, a collection of licensing techniques for removing non-primary verb candidates is employed, leaving e.g. the primary verbs, particles and conjunctions (bounded key constituents) that delimit the content of the fields in Diderichsen's sentence schema. Hereby, the subsequent identification of constituents which do not have an upper bound on their length (subject, object/predicatives and adverbials) can be identified relying to a lesser on extent explicit pattern matching, and more on different heuristic rules. For phrase type identification and delimitation of these constituents, when adjacent to each other, a novel chunking technique, *rank-based chunking*, is applied. Following this, a series of further rules merge chunks into larger ones, aiming at a final number of nominal chunks compatible with the valency information of the main verb. The aim is to identify *full* nominal and adverbial constituents, including post-modifiers. The implementation uses the *Stockholm Umeå Corpus 2.0*, a corpus which is balanced for different genres in published Swedish text. SUC*'s* tagset is also used unmodified in part-of-speech tagging which enables the program to deal with input text. The functional parsing, which includes no explicit language-defining grammar component is carried out technically using an object-based representation of clause structure.

Although output formats and types of evaluations of correctness are very different in parsers for Swedish text, it is claimed that the manual approach presented can provide high accuracy, which can be improved given more time for development.

The thesis work also includes two prototype applications, both requiring high accuracy of the sort of functional syntactic analysis described here. The first one is an implementation of automatic syntactic fronting in the area of text editing for Swedish, where the user is presented with a syntactically analyzed copy of her writing, from which paraphrases easily can be generated. The second application is in the field of natural language query systems and produces questions with answers from an arbitrary declarative input text. This prototype incorporates a text database from Swedish *Wikipedia*, and investigates primarily generation of *WH*-questions formed via fronting of unbounded primary constituents. The questions are generated as a text is opened and thus permits users to only ask the available ones, thus aiming at a high precision value.

Keywords: Diderichsen's sentence schema, positional grammar, field grammar, licensing techniques, *Stockholm Umeå Corpus*, schema parsing, rank-based chunking, syntactic fronting, paraphrasing, question generation, natural language query systems, Swedish *WordNet*

See also Summary in English *at the end of this thesis*