

Automatic Analysis of Compound Participle Modifiers in Germanic Languages

Part 1: Danish Compound Participles

Abstract

This paper describes:

1. The syntax and semantics of compound participle (CP) constructions in Danish.
2. The architecture of an algorithm for automatic analysis of Danish CP constructions.

These results constitute the first part of a ph.d.-project which will include a description and an analysis of CPs in English and German, with the primary aim of establishing analogous methods for automatic analysis of CP-constructions in Danish, English and German.

1. The syntax and semantics of compound participle constructions in Danish.

CPs in Danish are defined as the non-finite present participle (PP1) (-ende) or past participle (PP2) (-t, -te, -ede, -n, -ne) verbal forms which have a prefixed string, e.g.

1. *elproducerende vindmøller* (electricity-producing windmills)
2. *vindmølleproduceret el* (windmill-produced electricity)

CPs can have the following functions:

I. Premodifier in noun phrases

de elproducerende vindmøller (the electricity-producing windmills)

II. Head of a noun phrase

de mordanklagede (those who have been charged with murder)

III. Subject complement in a sentence

disse vindmøller er elproducerende (these windmills are electricity-producing)

IV. Object complement in a sentence

politikernes position gør dem mediehungrende

(the position of the politicians make them crave for media attention)

CPs are particularly frequent in journalism and therefore in HTML-documents:

3. *det lukningstruede Kellogg's* www.fyensstifttidende.dk, 5-9-1997

4. *the Hague-based International Criminal Tribunal* www.washingtonpost.com, Washington Post, 17-9-97

5. *Der angeschlagene Onlinedienst* www.spiegel.de, Der Spiegel 16-9-97

In many cases Danish CP constructions exhibit an archetypal syntactic structure and distribution of semantic roles:

A. *elproducerende vindmøller* --> vindmøller, der producerer elektricitet
O:n V:PP1 S:n (windmills that produce electricity)

B. *vindmølleproduceret el* --> elektricitet, som vindmøller producerer
S:n V:PP2 O:n (electricity that windmills produce)

C. *knivdræbt dreng* --> en dreng, som X har dræbt med en kniv
Adv:n V:PP2 O:n (a boy whom X has killed with a knife)

D. *bjørnedræbte får* --> får, som bjørne har dræbt
S:n V:PP2 O:n (sheep which bears have killed)

A: PP1 --> OVS, B and D: PP2 --> SVO. However, C is an instantiation of one class of numerous more complicated constructions, which necessitate the inclusion of semantic selectional restrictions in the analysis to arrive at a correct interpretation. In the case of C vs. D, the analysis depends on the semantic category of the prefixed string: inanimate vs. animate.

Figures 1 and 2 illustrate the difference between a traditional function-form parse tree, which generates the CP as a terminal, and our semantic-syntactic parse tree, which generates the CP as a nonterminal.

Function-form analysis:

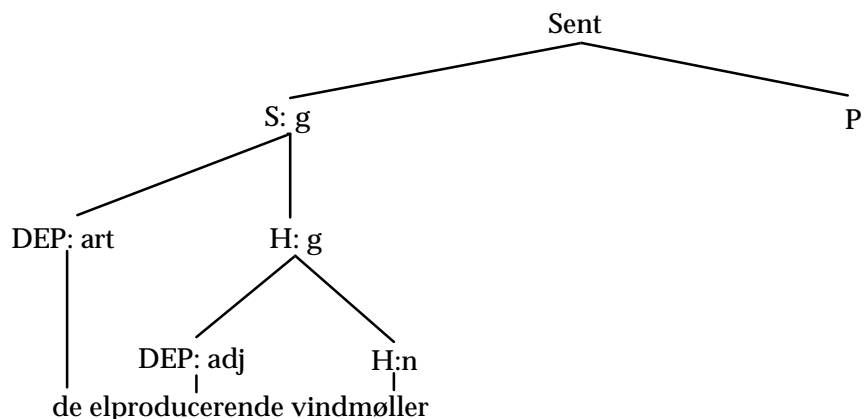
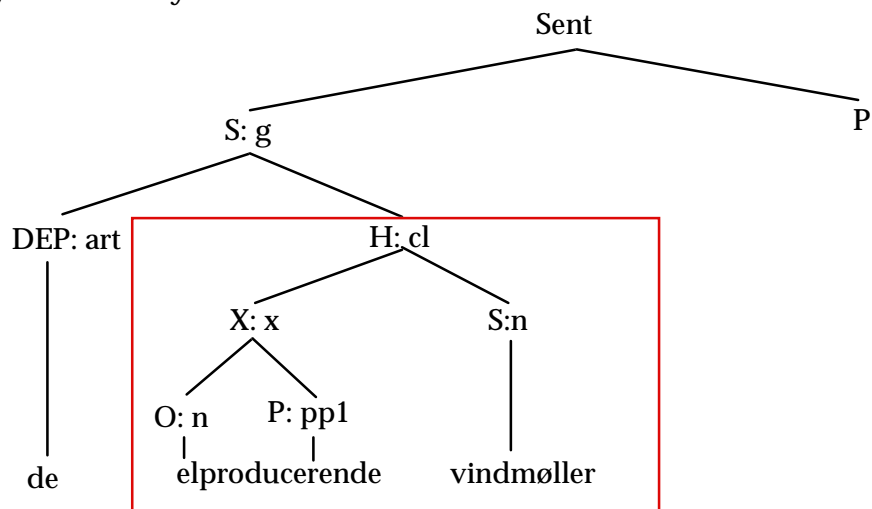


Figure 1.

Semantic-syntactic analysis:



"Semantic-syntactic structure"

Figure 2.

Since the formation of CPs (Danish: adj/adv/particle/noun + PP1/PP2) is a highly productive process in Germanic languages, CPs constitute a substantial number of the unrecognised word-forms in current NLP systems. Even if the base forms of CPs can be recognised, the semantic roles of the compound elements and the internal syntactic structure of the CP remain undefined.

As a possible solution to this problem, we have developed and implemented an algorithm that analyses CP-constructions and generates monolingual paraphrases. The paraphrase resolves any potentially ambiguous relationships between the participle form, its prefix and the NP head-noun, e.g.

6. *elproducerende vindmøller --> vindmøller, der producerer el*
 (electricity-producing windmills --> windmills, that produce electricity)

Furthermore, in a machine-translation perspective, it is relatively less problematic to generate translation-equivalents - particularly in Romance languages - from the paraphrase:

7. *vindmøller, der producerer el --> des éoliennes <qui produisent de l'électricité > | <productrices d'électricité>*

The implementation of the main structure of the program is concluded, though the refinement of certain rewrite rules is an ongoing process. Run-time tests of around 160 CP-examples have indicated a correctness rate of 97-98 % .

Detection of potentially rule based deviant CP types has been facilitated by the application's ability to process large corpora rapidly and consistently. Basically, the method consists of using files of systematically categorised CP examples as test data to achieve a better overview. Therefore, the run-time tests of the program have had

the dual function of evaluating existing linguistic hypotheses, while providing data for new analysis theories.

Future tests will include a Danish parser and a full-scale dictionary in order to achieve a clearer picture of the efficiency of the algorithm.

2. The architecture of an algorithm for automatic analysis of Danish CP constructions.

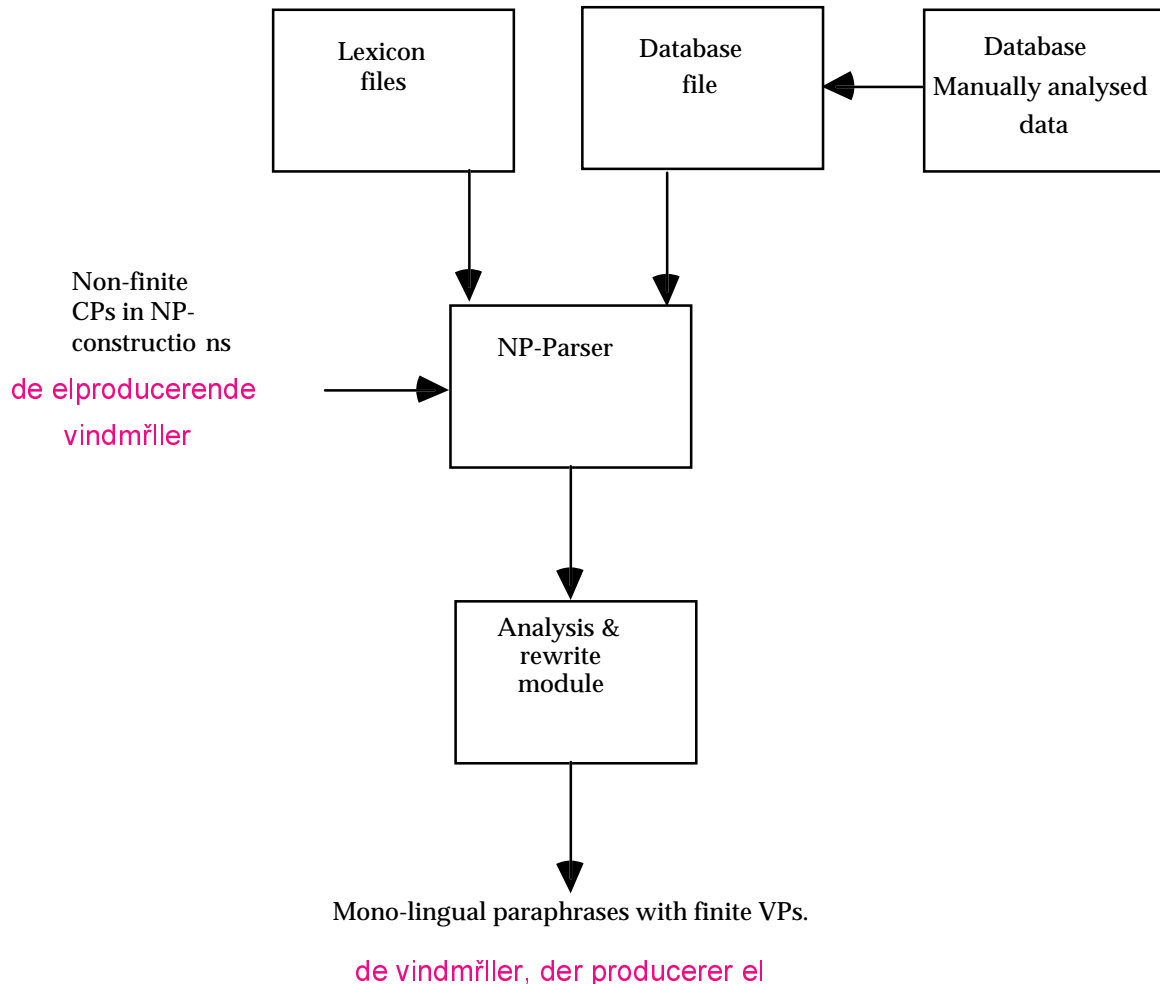


Figure 3.

The static structure of the implementation consists of lexicon files, a database file, an NP-parser and a module for semantic-syntactic analysis with ensuing rewrite rules.

Control flows from the NP-Parser - which scans the input string, segments CPs, identifies word forms and assigns a syntactic structure to the NP - to the rewrite module, where the data is further analysed before the input - non-finite CPs in NPs - is transcribed into finite constructions by means of rewrite rules, see example 6, section 1.

The relationship between the functions of the NP-Parser and the analysis part of the rewrite module is analogous to the relationship between the parse trees in figures 1 and 2.

In the database file (top middle of figure 3), verbs are tagged with information concerning valency, aktionsart and ergativity. This information is extracted from a database (top right-side corner of figure 3) in which collected examples of CPs have been manually analysed. Furthermore, verb forms are assigned semantic selectional

restrictions for arguments 1 and 2 (Subject, Direct Object), e.g. DRÆBE (KILL): arg1: animate; arg2: animate.

The lexicon files (top left-side corner of figure 3) contain standard morphological values - additionally, nouns are categorised within a coarse-grained semantic hierarchy.

REFERENCES

Bresnan, J. 1982. "The Passive in Lexical Theory". The Mental Representation of Grammatical Relations, edited by J. Bresnan. Cambridge, Mass: The MIT Press.

Copeland, C. et al (eds.). 1991. The Eurotra Linguistic Specifications. Brussels: Office for Official Publications of the European Communities.

Hansen, J. A. & Kjærsgaard, P. S. 1998 (Forthcoming). "CP-UDOG: An Algorithm for the Disambiguation of Compound Participles in Danish". Proceedings of the 11th Nordic Conference on Computational Linguistics. København: CST.

Jacobsen, B. L. F. & Kjærsgaard, P. S. 1995. "Adjektiviske deverbaler i dansk". UDOG-rapport 2, edited by Maegaard, B. and Pedersen, B. S., pp. 3-26. København: CST.

Karlsson, F. et al (eds.). 1995 Constraint grammar. A language-independent system for parsing unrestricted text. Berlin: Mouton de Gruyter.

Kjærsgaard, P. S. 1996. "Danske participialer og valens". Adjektivernes Valens, edited by Van Durme, K. , pp. 49-84. Odense, Institut for Sprog og Kommunikation: Odense Universitet.

Pümpel-Mader, M. et al. 1992. Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache , V. Adjektivkomposita und Partizipialbildungen. Innsbruck. Düsseldorf: Pädagogischer Verlag Schwann.

Stewart, P. 1995. "Brugen af en database til behandling af danske participialer". Datalogvistisk Forenings 5. Årsmøde, pp. 53-66. Odense, Institut for Sprog og Kommunikation: Odense Universitet.