CP-UDOG
An Algorithm for the Disambiguation of Compound Participles in Danish

Jens Ahlmann Hansen and Poul Søren Kjærsgaard
Odense University

Abstract

This paper describes some aspects of the linguistic analysis which has been applied to disambiguate Danish compound participles (CPs) (sect. 1-3), followed by an introduction to the implemention of the disambiguation process (sect. 4-6).

1. Introduction

A CP consists of a present or past participle, to which another word is prefixed. This compound functions as an adjectival modifier in a noun phrase (clause level) or as a predicate to the subject/direct object (sentence level). Two examples:

1. *elproducerende vindmøller*
2. *vindmølleproduceret el*

The results of the analysis consist partly of the unfolding of a CP into a finite verb construction, either as a sentence or as a relative clause. The resulting unfolded construction allows for the identification of the syntactic and semantic functions of the elements integrated in the CP, the former by means of traditional structuralist methods, the latter by means of lexical information.

Within the framework of the UDOG-project (Research into Danish Vocabulary and Grammar), Poul Søren Kjærsgaard and Bjarne Le Fevre Jacobsen have sought to infer and to formalize the syntactic, semantic and lexical rules which underlie the formation of CPs (e.g. Jacobsen & Kjærsgaard, 1995). Specifically, they have addressed the following 3 aspects:

1. A description of CPs based on the valency and the semantic selectional restrictions of CPs.
2. A contrastive analysis of CP construction patterns and the finite construction patterns of their derivations.
3. The application of valency number and selectional restrictions in the analysis of syntactic structure and distribution of semantic roles in CP constructions.

The CP-UDOG implementation disambiguates compound-participles (CPs) in Danish. Furthermore, the application generates monolingual paraphrases in which the syntactic and semantic relationships between the participle form, its prefix and the head-noun of the NP become fully explicit.

CP-UDOG has both an actual and a potential function:

I. A tool for testing linguistic hypotheses concerning the rules governing the generation of CPs in Danish.
II. A prototype module for the automatic generation of translation equivalents in Romance languages from CPs in Germanic languages.

2. Linguistic data

The data that the investigation is based on was collected primarily from 1993 to 1996 from Danish newspapers and television teletext. The corpus amounts to more than 3000 examples of simple and compound participle constructions covering approx. 1000 different verbs.

3. Algorithm

The algorithm is data-driven and consists of two operations: analysis and generation (see appendix 1). The starting point is, first, the observation, resulting from several studies, that there exist some archetypical syntactic and semantic relations between the participle and a prefix on the one hand and between the (compound) participle and the head of a NP (clause level) or the subject/object (sentence level) on the other. These relations can be inferred from examples 1-2:

syntax:	direct object-present participle		subject
semantics:	patient		agent
1.	*el-*	*producerende*	*vindmøller*
syntax:	subject-	past participle	direct object
semantics:	agent		patient
2.	*vindmølle-*	*produceret*	*el*

The starting point is, second, the observation, resulting from the UDOG project, that there exist a number of deviating patterns which do not comply with the archetypes. The algorithm to be presented here aims at producing unfolded, hence disambiguated, paraphrases of both groups.

The analysis module consists of five parameters:

The first parameter the algorithm operates on is the type of participle: present or past. Danish present participles have an invariant ending: *-ende;* while past participles have a limited number of morphemes: *-et, -t, -te, -ede, -ne*. In either case, a second condition to be satisfied is that the root of the word under consideration can be identified as a verb. This operation is carried out by a dictionary lookup.

The second parameter is the valency of the verb that the participle is derived from. It is determined by a dictionary lookup.
Verbal valency is defined as the minimal number of dependents (including the subject) a verb presupposes in order for the verbal action to take place. Apart from avalent verbs, we distinguish rather traditionally between monovalent, bivalent and trivalent verbs. The number of dependents is crucial to the disambiguation process since it may identify for instance a nominal prefix which does not fill the function which the archetype would otherwise predict. Syntactic valency is also

important since it contributes to distinguish between direct and prepositional objects (the preposition belonging to the latter is deleted during the formation of compound participle).

The third parameter is the class the prefix belongs to. Most prefixes are lexical items in their own right, e.g. *el* and *vindmølle*. A few, however, are not, e.g. *be-*, *u-*). In either case, the class is determined by a dictionary lookup.
The archetypical patterns take into account only nominal prefixes, but the set of possible prefixes includes, among others, adjectives, adverbs, prepositions, particles and bound morphemes. If the prefix does not belong to the class of nouns, it cannot fill a nominal function such as subject or direct object. Further, the third parameter may cooperate with the second parameter to identify the function of the prefix. This happens when a nominal prefix represents an ellipsis of a PP (the preposition being elided) or vice versa when a preposition represents a PP whose head was elided.

The fourth parameter examines the compatibility between the selectional restrictions of the verb under consideration and semantic features belonging to the dependent. The metaphor that may best illustrate this phenomenon is two wheels whose gears may or may not mesh. Incompatibility between the selectional restrictions of a verb and the semantic features of a dependent overrules the archetypical relation. In such instances, the algorithm assigns syntactic and semantic functions which are different from the archetypical ones. The check is performed by using a rather primitive set of semantic features (eight).

A fifth parameter intervenes when past participles are considered. This is ergativity. An inaccusative verb inverts the normal semantic relationship between the participle and its dependents (see appendix 1).

The generation module of the algorithm assigns syntactic and semantic functions to the prefix and the head. This enables the production of an unfolded equivalent of the CP. Two aspects deserve mentioning. First, the tense of the unfolded present participle is, by default, set to the present. This is true as far as a context-free PP is concerned. But it would not hold if the CP to be unfolded was embedded in a sentence in the past tense. The tense of the unfolded past participle is set to present or present perfect depending on the verb's aktionsart (imperfective verbs deriving present tense and perfective verbs present perfect).
Second, the paraphrase is by default set to active voice. Hence, subject and object are linearized in an order that does not necessarily coincide with the order that would prevail if the CP occurred in a context-sensitive environment. This means that elided dependents are indicated formally by X1, X2,...

4. Database implementering

For the purposes of systematic analysis, categorization and hypothesis formation, a database for the collected CP examples has been constructed. The manual data-handling consists of 3 sections:

1. Actual reading
2. Lexeme
3. Paraphrase

Section 1 contains information concerning the form and function of the particular CP. Section 2 describes the valency and selectional restrictions of the related infinitive from an LFG perspective. In section 3, the CP construction is analyzed: syntactic structures and semantic relationships are formulated in terms of LFG. Finally, the CP example is transcribed into a semantically equivalent construction with a finite VP, e.g.

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

The paraphrase resòlves any potential ambiguous relationships between the participle form, its prefix and the NP head-noun. From the paraphrase, it is relatively less problematic to generate translation-equivalents in Romance languages:

II. *vindmøller , der producerer el →*
des éoliennes / qui produisent de l'électricité / productrices d'électricité/

5. Architecture of the CP-UDOG algorithm

In order to test the potential of the aforementioned rule sets for analysis of CP constructions, the CP-UDOG algorithm, which further formalizes the linguistic analysis and the associated terminology, was implemented. Due to concerns for transparency of the program code and compatibility with other language processing systems, the algorithm was implemented in the programming language Pascal.

The static structure of the program consists - in a rather simplified description - of dictionary files, a database file, an NP-parser and a module for enhanced analysis with ensuing rewrite rules.

In addition to standard morphological values, nouns are categorized within a coarse-grained semantic hierarchy. The database file - a reduced version of the database mentioned in section 4 - contains mainly information concerning the morphology, valency (selectional restrictions), aktionsart and ergativity of verbs.

The NP-parser reads the input, identifies the word units, confirms the syntax and segments the CPs. Following the identification, further semantic-syntactic information from the database file is added to the verb form. The information from the NP-parser is passed on to the rewrite module, where the data is further analyzed before the input - non-finite CPs in NPs - is transcribed into finite constructions by means of rewrite rules (see example I, section 4).
Thus, the manual treatment of data in the database and the automatic analysis of the CP-UDOG-algorithm interact on several levels:

1. The database constitutes the setting for inductive hypothesis formations, while the implementation acts as a 'top-down' hypothesis test.
2. Database information is exported directly to the database file of the implementation.
3. The manually produced paraphrases of the database are the measures by which the output of the implementation is judged.

4. The logical consistency of the implementation has contributed towards a more precise meta-terminology, in particular concerning such aspects as valency (selectional restrictions) and aktionsart.

6. Tests, maintenance and further perspectives

The implementation of the main structure of the program is concluded, though the refinement of certain rewrite rules is an ongoing process.

As mentioned in section 5, 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 categorized 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, ideally, would include a Danish parser and a full-scale dictionary in order to achieve a clearer picture of the efficiency of the algorithm.

Finally, an extension of the CP-UDOG-project will include a description and an analysis of CPs in English and German, with the primary aim of establishing analogous methods for disambiguation of CP constructions in Danish, English and German.

References

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

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.

Kjærsgaard, P.S. 1996. "Danske participialer og valens". *Adjektivernes Valens,* edited by VanDurme, 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". *Datalingvistisk Forenings 5. Årsmøde*, pp. 53-66. Odense, Institut for Sprog og Kommunikation: Odense Universitet.

Analysis					Generation				
Participle type	Valency	Prefix wordclass	Semantic Selectional Restrictions		Head	Prefix	Participle	X1	X2
			Prefix	Head					
PPI	1	+n		+S	S	ADV	V:pres		
			+S	-S	ADV	S	V:pres		
		-n			S	ADV	V:pres		
	2	+n	+O	+S	S	O	V:pres		
			-O	+S	S	ADV	V:pres	O	
		-n			S	ADV	V:pres	O	
	3	+n			S	O	V:pres	prep.O	
		-n			S	ADV	V:pres	O	prep.O

Analysis						Generation				
Participle type	Valency	Prefix wordclass	Sem.Select.Restrict.		Ergativity	Head	Prefix	Participle Aktionsart	X1	X2
			Prefix	Head						
PPII	1	+n				S	ADV	pres. / pres.perf.		
		-n				S	ADV	pres. / pres.perf.		
	2	+n	+S			O	S	pres. / pres.perf.		
			-S	+O	inergative	O	ADV	pres. / pres.perf.	S	
				+S	inaccusative	S	O	pres. / pres.perf.		
			-S, +O	-S, -O		ADV	O	pres. / pres.perf.	S	
		-n				O	ADV	pres. / pres.perf.	S	
	3	+n				O	ADV	pres. / pres.perf.	S	
		-n				O	ADV	pres. / pres.perf.	S	prep.O

Appendix 1.