The Treatment of Information Structure in Machine Translation

Patrizia Paggio

University of Copenhagen and Centre for Language Technology

October 1997

Abstract

The goal of this thesis is to propose a treatment of information structure which may contribute to the generation of more varied and coherent output in an MT system than it is possible in most current approaches. The main assumption behind this work is the fact that the structure of the individual sentences in a written (technical) text is indicative of the way in which each sentence relates to the rest of the discourse. Thus, by conveniently representing the information structure of a source sentence, it is possible to choose between competing, but from a discourse perspective different syntactic structures in the target language. On the basis of a discussion of various previous approaches to the topic, a model is proposed whereby the information structure of a sentence can be represented in a unification-based formalism like HPSG by means of the three context-related features P_TOPIC, FOCUS and NEUTRAL. Specifications are then worked out for the analysis and generation of a number of linguistic constructions in Danish in terms of their syntactic, semantic and information structure properties. The issues dealt with range from the computation of the information structure of unmarked declarative sentences, which is achieved through heuristic principles based on definiteness, word order, and semantic properties of certain constituents, to the analysis and generation of clefts, existential sentences, topicalisation and extraposition. In all these cases, it is shown that adding the information structure layer of analysis to the syntactic and semantic analyses of the input make it possible to account for the linguistic facts in a more accurate and complete way. Some of the specifications proposed have been implemented in the ALEP engineering platform: the positive results of the implementation provide an additional confirmation of the soundness of the model adopted.

Contents

Acknowledgements

This research has been carried out at the Center for Sprogteknologi under the supervision of Professor Ebbe Spang-Hanssen (Institut for Almen og Anvendt Sprogvidenskab, University of Copenhagen) and Professor Bente Maegaard from the centre itself. It has partly been supported by a grant from the European Commission.

I would like to thank all those who have supported me in my work, first of all my supervisers for their willingness to read the numerous versions of this thesis, and for providing stimulating comments and advice.

I would also like to thank all my colleagues at the centre for their patience and support. In particular, I'm grateful to Bjarne Ørsnes for the many interesting discussions we have had on several of the issues dealt with in this thesis and for being a helpful informant; to Bolette Sanford Pedersen for her comments on my analysis of Danish clefts; to Nancy Underwood for the work done together in the LINDA project, which has inspired me in writing many of the syntactic analyses proposed here, as well as for her comments on this thesis; to Costanza Navarretta and to Bradley Music, without whose help I would not have managed the ALEP implementation.

Thanks also to Marius Groenendijk from Cray Systems in Luxembourg for his invaluable technical support.

Important inspiration for my work came from my stay at the Centre for Cognitive Science at the University of Edinburgh in the initial phase of my work. Thanks should go in particular to Elisabeth Engdahl and Kevin Humphreys for the interest they have shown in my work.

Finally, thanks to my family for their support throughout this work.

Chapter 1

Introduction

1.1 The aims of the research

It is a well-known fact that languages differ in the way in which the information structure of the individual sentences in a text is reflected in their syntactic form. Therefore, it seems natural to ask the question whether and how this diversity should be taken into account in Machine Translation (MT). Although the issue of information structure has been object of much study, in fact it is only quite recently that attempts have begun to be made to establish the relevance of such studies to MT. Apart from the obvious explanation that many other problems had to be dealt with first, a reason for this late start is probably the fact that it has taken MT some time to become to be acknowledged not only as a practical application domain, but also as a field of theoretical research.

The main goal of this thesis is thus to propose a treatment of information structure which may contribute to the generation of more coherent output in an MT system. This general objective subsumes a number of more specific aims, summarised below:

- To describe the detrimental effects that lack of an information structural treatment has on MT output.
- To propose a way in which the information structure of a sentence can be represented in a formalism suitable to an MT application.
- To explore the relationship between information structure and syntax, in particular for the Danish language. Translation from English into Danish is thus taken as a test-case.
- To propose strategies for the computation of information structure in analysis and for its exploitation to guide syntactic choice in generation.
- To show that the strategies proposed contribute to an enhancement of output quality with respect to the internal coherence of the translated text.

Being application-oriented, this research has to strike a sometimes difficult balance between the theoretical dimension and practical considerations of implementability. Thus, on the one hand it is an attempt at providing new insights concerning the way in which information structure is expressed through syntactic structure, with particular regard to the Danish language, and at discussing the implications this has for the treatment of a number of different syntactic phenomena. On the other hand, it makes a proposal for how these insights can be reflected in a concrete implementation. A toy implementation has been created as part of this work to support this proposal.

1.2 The data

A general remark on the Danish data quoted in the thesis is in order here. The data come from a number of different sources, namely the *DK87* corpus (Bergenholtz 1988), Danish children's books (*Peter Pedal hænger i*, *Halfdans ABC*), computer manuals (*Microsoft Windows 3.1 Brugerhåndbog*), as well as from the linguistic literature on Danish. If no indication is given as to the source of an example, this means that it has been constructed by the author of this thesis. All the Danish examples are glossed with an English translation. The literal translation, however, has only been provided where necessary to get a point across. Stress patterns are again only indicated in the translation where necessary.

1.3 Structure of this thesis

The order in which the aims of this research have been listed above corresponds rather directly to the internal organisation of this thesis.

The next chapter (Chapter 2) discusses the importance of a treatment of information structure in MT, and deals with the way in which the issue is currently approached in the field. It also provides a brief introduction to ALEP, the system used for the implementation. Chapter 3 deals with a selection of theories dealing with information structure and clarifies at least some of the different uses made of the relevant terminology within different schools of thought. Chapter 4 contains a description of the model of information structure assumed in this work, it discusses how this model relates to discourse theory on the one hand, and focus semantics on the other, and proposes a framework for representing the information structure of a sentence in a typed feature structure formalism. In Chapter 5, this framework is applied to the treatment of a number of Danish syntactic phenomena. The chapter consists of two main parts, one dealing with unmarked declarative sentences, the other with focussing and topicalising constructions. For each of the phenomena taken into consideration, a proposal is made for how it should be treated in analysis as well as in generation. A subset of the analyses proposed has been tested in a practical implementation exercise: Chapter 6 describes the principles on which the implementation builds, presents the results obtained and discusses some of the problems that have arisen. Finally, Chapter 7 contains the conclusions.

Descriptions of the implemented grammar and lexicon are enclosed as appendices, together with the test suites used, examples of representations built by the system, and the translation output.

Chapter 2

Information structure and Machine Translation

The point of departure of this project is the well-known fact that different languages may express the information structure of a sentence in different ways. By information structure we mean the organisation of the information expressed by the sentence in such a way as to fit in with what has been said in the preceding sentences and what will be said in the succeeding ones. A classic example of this difference is the contrast between Slavic languages, where the information structure is almost directly reflected in surface word order, and English, where word order is, to a large extent, grammatically determined, and information structure is rather expressed through prosody, as well as a number of focus marking syntactic constructions. However, interesting differences relative to information structure can also be observed between more closely related languages, such as English and Danish.

Looking at the PaTrans corpus of bilingual patent texts from the domain of petrochemistry[1], for example, it becomes clear that the use of passive forms, relative clause extraposition and *there* insertion is not the same in the Danish text as in the original English input. Let us discuss a few examples of such discrepancies taken from the corpus.

To take the case of the passive, both the English and the Danish texts show a high occurrence of passive forms. However, word order is often not the same in the two languages, since in Danish the passive may be combined with a pleonastic *der* (there). This is shown in our first example, where (a) is the English original, and (b) the version provided by the translator, with an English gloss in parentheses:

(1) a. In addition, alcoholamines or polyoxyalkylene polyamines can be used.

b. Der kan derudover anvendes alkoholaminer eller polyoxyalkylenpolyaminer.
(lit: There can in addition be used alcoholamines or polyoxyalkylene polyamines.)

(*PaTrans corpus*)

Although not very common, relative clause extraposition is represented in the English corpus. In the Danish translations, however, it is always avoided, and a *der* construction is again used to achieve a different word order. This is shown in the example below:

(2) a. With the increasing diversity in distillate fuels, types of fuels have emerged which cannot be treated by the existing additives or which...

b. Med destillatbrændstoffernes større forskelligartethed er der fremkommet brændstoftyper, som ikke kan behandles med de eksisterende additiver eller som...

[1]This is a corpus of English patent texts paired with their manually obtained Danish translations. The corpus has been assembled at CST in cooperation with the Danish translation company Lingtech A/S to assist in the design and implementation of the PaTrans system (cf. Maegaard & Hansen 1995).

(lit: With the increasing diversity in distillate fuels, are there emerged types of fuels which cannot be treated by the existing additives or which...)

(*PaTrans corpus*)

In both cases, the surface subject in the English sentence is a nominal phrase with no determination, constituting the new piece of information in the sentence at hand. In Danish, there is a tendency for such a constituent not to occur as the subject of the sentence. A similar strategy can be seen in the example below, where the nominal phrase which is the surface subject in English becomes the complement of an adjectival predicate, and the pleonastic *det* (it) occurs in subject position in the Danish translation:

(3) a. In copper-free oils other antioxidants in addition to the zinc dialkyldithiophosphate are sometimes required to improve the oxidative stability of the oil.

b. I kobberfrie olier er det til tider nødvendigt med andre antioxidanter udover zinkdialkyl-dithiophosphat for at forbedre oliens oxidationsstabilitet.
(lit: In copper-free oils is it sometimes necessary with other antioxidants in addition to the zinc dialkyldithiophosphate to improve the oxidative stability of the oil.)

(*PaTrans corpus*)

It is important to stress the fact that the reshuffling of constituent order which has taken place in the Danish translations is not just a question of style, but a necessity in order to yield the correct information structure in the Danish text.

Whilst all the examples discussed up to this point seem to be motivated by a general tendency in Danish not to place an indefinite constituent in subject position, differences between English and Danish can also be observed concerning thematisation, i.e. what constituent(s) may occur in the initial position of the sentence.

Following Halliday, we use the term *theme* to refer to the initial position. The constituent in theme position plays an important role as it constitutes the point of departure for the message expressed in the sentence, at the same time connecting it to the preceding discourse. Both in English and in Danish, in a declarative sentence the surface subject is the unmarked theme. However, the constraints limiting thematisation of other constituents are slightly different in the two languages.

Firstly, English allows sequences of adverbials to fill in the theme slot, whereas only one constituent is allowed in that position by the grammar of Danish. Thus, in cases where a sequence of adverbs fills the theme slot in an English sentence, in the Danish translation only one of the adverbs can fill that slot and the other(s) will have to be positioned elsewhere in the sentence. The Danish grammar must decide where. Secondly, as discussed extensively in Section 5.3.4, Scandinavian languages – with Danish among them – violate the Complex NP Constraint by allowing extraction of a constituent from a relative clause and its placement as the theme of the main clause.

Lexical differences can also be observed, as Danish seems more restrictive than English with respect to which adverbs can occur initially in the sentence. An example from our corpus is shown below: the adverb *passende* (suitably) cannot occur initially, and the translator has made the surface subject the theme in the Danish text.

(4) a. Suitably the finished lubricating oil composition may be obtained by ...

b. Den færdige smøreoliesammensætning kan passende opnås ved ...
(The finished lubricating oil composition may suitably be obtained by ...)

(*PaTrans corpus*)

A more literal translation maintaining the same theme as in the English source would have been ungrammatical:

(5) *Passende kan den færdige smøreoliesammensætning opnås ved ...
(lit: Suitably can the finished lubricating oil composition be obtained by...)

All the translation examples discussed this far have been produced manually by professional translators. The issue we want to address in this chapter, however, is whether information structure differences of the kind sketched above are taken into account in contemporary MT practice, and to what extent current MT architectures allow for the incorporation of an information structure component of some sort. Our conclusion is that the issue of information structure is largely ignored in most systems, with the exclusion of a limited number of research-oriented projects. In particular, no system aimed at translating from or into Danish is able to tackle textual coherence. This thesis is precisely an attempt at beginning to fill this gap.

In Paggio (1994), we made a preliminary attempt at treating the issue of information structure in MT. According to this proposal, the system would compute the *focus* and the *theme* of the source sentence, and try to generate the equivalent expressions in the right position in the target language. This dissertation takes up the ideas sketched out in that paper and develops them into a more comprehensive and detailed account of how information structure features can be used to enrich the interface representation of a transfer-based MT system.

Our account of current approaches to information structure issues in relation to MT starts with a discussion of simple PC-based transformer systems in Section 2.1: we show that such systems can only generate coherent text in cases where source and target sentence have roughly the same word order. As an example of a PC-based transformer system we have chosen Winger92, a system translating from and into Danish. Section 2.2 deals with linguistically more sophisticated systems, in particular transfer-based ones: here we point out that without a treatment of information structure, only the unmarked word order (subject-VP in Danish) can ever be generated, with detrimental effects on the coherence of the target text. This is exemplified by briefly discussing the generation algorithm of the English-Danish PaTrans system. In Section 2.3 we look at more research-oriented approaches proposing specific algorithms for the computation of information structure, and point out in what way this thesis contributes to the line of research embodied in the papers discussed.

2.1 PC-based translation

PC-based translation systems are based on rather simple technology. The most important component in systems of this kind is the dictionary (or dictionaries) where words and phrases of the source language are stored together with their target language equivalents, as well as limited grammatical information, such as their part-of-speech. Translation is carried out by looking up words and phrases in the dictionaries, and substituting the target language equivalents for the original strings. A comparatively simple grammar handles agreement and local word order rearrangements. The system does not parse the input sentences, let alone compute their information structure. In general, the word order of the source is maintained in the translation wherever possible. This means that the right information structure is generated if source and target language do not differ too much.

As an example we have taken Winger92, a system that has Danish as one of the languages treated. The following Danish sentences were submitted for translation into English without any previous customisation of the system's dictionaries:

(6) a. Der kan bruges en computer.
b. I dag bruges computere til mange opgaver.
c. Den opgave kan passende løses automatisk.
d. Det er sjovt at arbejde med en computer.
e. Den computer kan jeg ikke bruge.
f. Det er computeren som ikke fungerer.
g. Programmet bliver ikke brugt. Sekretæren kender det ikke.

(A computer can be used.
Nowadays computers are used for many tasks.
That task can suitably be solved automatically.
It is fun to work with a computer.

That computer I cannot use.
It is the computer that is out of order.
The program doesn't get used. The secretary doesn't know it.)

All the sentences are examples of constructions that will be discussed in detail in Chapter 5 for their special contribution to the information structure of the sentence. The first is an existential sentence, and the second an example of adjunct fronting. Sentence (c) has been included to test the system's treatment of adverbials in different sentence positions. Sentence (d) is an example of extraposition. Sentence (f) is a topicalisation example. Finally, sentence (g) displays the use of a weakened pronoun preceding the negation adverbial.

The output was as follows:

(7) a. there can be used a computer.
b. to-day computers to many opgaver. be used
c. the task can fitting automatic be solved.
d. it is droll to work with a computer.
e. the computer I can not use.
f. it is the computer which does not function.
g. the program is not brugt. the secretary know the ikke.

In (a), we can see that the system wrongly assumes that *there* constructions are used in the same way in English as in Danish. In (b), the system has tried to change the Danish word order to fit the English grammar, but has not succeeded very well. Note that the word *opgaver* (tasks) has not been translated although it was coded in the dictionary. In (c), the system has had trouble getting the right determiner in front of the subject, and translating the final adverbial, which has been rendered as an adjective and moved away from end-focus position. The extraposition example in (d), on the contrary, was translated correctly, apart from the choice of *droll* as a translation for *sjovt.* We could have changed this by editing the standard dictionary or adding a new entry in the user-defined one, but it was not very important in the light of our discussion. In (e), the object is topicalised as in the Danish original. This time (cf. b), the system has managed to generate a correct word order in English. The determiner in front of the topicalised object, however, is wrong. The translation of the cleft in (f) is acceptable. Finally, (g) shows a number of problems. The participle *brugt* (used) as not been translated: it is difficult to understand why, as the word is coded in the standard dictionary. The personal pronoun has become an article, the verb does not agree with the subject, and the negation adverbial for some reason has not been translated at all.

The conclusion of this little test is not surprising. In cases where the same grammatical construction can be used in both languages to express a certain information structure, as in (d) or (f), the system produces a correct translation (provided of course that it knows all the words and that the input is not ambiguous). Otherwise, the output can be more or less incorrect, depending on how local the necessary manipulation is. However, it can hardly be expected that a system of this kind should be equipped with a component for the treatment of information structure. The issue becomes more interesting when dealing with linguistically more sophisticated systems, as we shall see in the next section.

2.2 Linguistically sophisticated models

As mentioned above, we shall base our discussion of the way in which information structure is treated in more sophisticated MT systems on the example of PaTrans, a tranfer-based system for the translation of patent applications from English into Danish in the domain of chemistry and petrochemistry. Apart from its being tuned to specific domains especially as regards the lexical treatment of technical terms, we believe PaTrans is representative of the transfer-based model as implemented in a number of other systems[2].

[2]To discuss basic MT notions such as transfer or interlingua would lead us too far. See e.g. Hutchins & Somers (1992) or Arnold, Balkan, Humphreys, Meijer & Sadler (1994) for a general introduction to MT and a discussion of all the central

The architecture of the translation system proper[3] is based on the model implemented in the Eurotra system (Allegranza, Bennett, Durand, Eynde, Humphreys, Schmidt & Steiner 1991). PaTrans is thus a transfer-based, multistratal and, in principle, multilingual system, in which the input text is analysed sentence by sentence and assigned an Interface Structure (IS), from which transfer and generation are carried out. IS was defined in Eurotra and is still defined in PaTrans in such a way as to allow for the neutralisation of language specific variations, so that the same IS can serve as input to transfer into different languages.

More specifically, an IS representation is a dependency tree with a lowered governor, where the various constituents are labelled according to the valency relation they bear to the head of the phrase, and ordered in a canonical way, so that the head (or governor) is always the leftmost node of a phrase, followed by arguments, modifiers and sentence adjuncts in a canonical order. Consequently, surface word order is among those language specific features that are neutralised. Unlike other aspects of surface structure, such as morphosyntactic definiteness, word order is not recorded in the feature structures associated with the nodes of an IS representation. Furthermore, elements not contributing to the semantic representation, e.g. pleonastic pronouns, are not represented at IS.

Consider thus the following three Danish sentences:

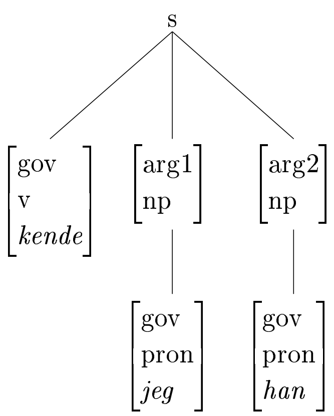
(8) a. Jeg kender ham.
(I know him.)

b. Ham kender jeg.
(lit: Him know I.
Him I know.)

c. Det er ham jeg kender.
(It's him I know.)

They would all yield the same IS representation, namely the simplified tree below (where only a few attribute values are shown):

(9)



This means that any difference there may be in the three input sentences as regards information structure is lost. In fact, as we shall argue at length in Chapter 5, sentence (b) above (an example of topicalisation) is a topic marking construction, whilst the cleft in (c) can be regarded as a focussing construction. In both cases, reducing them to the unmarked declarative sentence in (a) constitutes a loss of information. Exactly what kind of information will become clearer in the succeeding chapters.

Imagining now that the same representation has resulted from analysis and transfer of the English sentence *It's him I know*, the question is then how the Danish generation grammar would map such a structure onto a surface string. To address the issue from a more general perspective, we can distinguish two

issues.

[3] PaTrans has a number of components which can be considered peripheral to the translation system proper, such as a tagger and a pre-parser: these are not relevant to the present discussion.

different cases, depending on the way in which the coverage of the generation grammar relates to the coverage of the analysis grammar.

Case 1. In the first case, the system will have symmetric, reversible grammars, i.e. it will use the same rules to analyse and generate, and will hence be able to produce exactly the same sentences it can analyse. This approach is arguably not an optimal one in operational, large-scale systems, and is in fact not adopted in PaTrans, which is an operational system. Nevertheless, reversibility is a theoretically interesting possibility (see e.g. van Noord, Dorrepaal, van der Eijk, Florenza, Ruessink & des Tombe 1991). Clearly, in such a system, if no difference is made between the three sentences under consideration, all three of them can be generated on the basis of the IS tree. Some preference mechanism, however, may be employed to pick up one of them, perhaps on the basis of frequency of occurrence in some specific text type or domain. A consequence of such a strategy could be, as Rochemont (1989) points out, that focus marking constructions may be generated in inappropriate contexts, or never generated at all.

Case 2. In the second scenario, which is in fact the one implemented in PaTrans, generation and analysis grammars are non-symmetric, and generation has a more limited coverage than analysis. In this case generation from our IS tree would probably always result in the generation of one of the three sentences in (8), probably the unmarked version in (a). In fact this corresponds exactly to the behaviour of the PaTrans system: lacking a principled way of representing the different discourse functions of different surface versions of the same dependency structure, PaTrans generates sentences which all conform to the same default word order, and is not able to recognise contexts in which focus or topic marking constructions are appropriate or necessary.

Thus, the system is not able to generate adverbials in theme position, as would clearly be appropriate in a case such as the following one (where (a) is the input, and (b) is the output produced by PaTrans):

(10) a. In this context typical polyamines will be tetraethylene pentamine and triethylene tetramine.

b. Typiske polyaminer er tetraethylentamin og triethylentetramin i denne sammenhæng.
(Typical polyamines will be tetraethylene pentamine and triethylene tetramin in this context.)

The treatment of indefinite subjects, which in Danish tend to occur in end-focus position, also proved inadequate in both the examples we tested the system with, as in both cases the indefinite subject is not placed to the right of the main verb. The two examples (the first of which was mentioned in 1) are shown below. Again, (a) is the input, and (b) is the output generated by the system. We have also added what we consider a correct translation in (c).

(11) a. In addition alcoholamines or polyoxyalkylene polyamines can be used.

b. /- Derudover alkoholaminer eller et polyoxyalkylenpolyamin dåse anvendt. -/
(In addition alcoholamines or a polyoxyalkylene polyamine can (=tin) used.)

c. Derudover kan der anvendes alkoholaminer eller polyoxyalkylenpolyaminer.
(lit: In addition can there use(PASS) alcoholamines or polyoxyalkylene polyamines.)

(12) a. In copper free oils another antioxidant is required.

b. /- I kobber fri olier en anden antioxidant kræves. -/
(In copper free oils another antioxidant is required.)

c. I kobberfrie olier kræves der en anden antioxidant.
(lit: In copper free oils require(PASS) there another antioxidant.)

It must be noted, however, that in both cases the final result is obtained through fail-softing (as shown by the '/- -/' symbols). This means that something went wrong either during parsing or generation, and that the system has recovered the best result it could find given the error. It is not quite clear, therefore, what its intended behaviour should have been.

Finally, we show below an example where a focussed, heavy object is not recognised by the system as such, so that in the Danish translation the object is made to precede the adverbial of manner *alvorligt* (seriously). This creates an ambiguity, as the adverb in the translation seems to modify the embedded infinitival clause rather than the matrix clause.

(13) a. The sludge and varnish seriously restrict the ability of the crankcase oil to lubricate the engine.

b. Slammet og belægning begrænser krumtaphusoliens evne til at smøre maskinen alvorligt.
(The sludge and varnish restrict the ability of the crankcase oil to lubricate the engine seriously.

In a more accurate translation, the object should be in end-focus position, and the adverb (*i alvorlig grad* (in a serious manner), rather than *alvorligt* as chosen by the system) should precede it, as in:

(14) Slammet og belægningen begrænser i alvorlig grad krumtaphusoliens evne til at smøre maskinen.

2.3 Research-oriented MT

Information structure issues have not received much attention in MT research. In recent years, however, a few attempts at devising information structure components to fit the need of MT systems have been made. Hauenschild (1988), for example, claims that preserving the thematic structure of the source text is the primary goal of translation – whether human or automatic – in the case of the so-called *informative* texts. Hauenschild's main theoretical concern revolves around the possibility of using MT to simulate the way in which human translation is performed. Within such a perspective, an MT system has to be able to model the structure of a text in a way similar to the human translator's. Even though we do not share the conviction that MT can or should simulate human translation, we concur with Hauenschild that trying to formalise the notion of information structure in an MT implementation can help us shed light on an interesting aspect of linguistic theory. In fact we see MT as an important testbed for linguistic theory in general, and believe that the relation between linguistics and MT ought to be one of fruitful mutual exchange. In fact, this relationship has improved in the past ten or twenty years after decades of suspicion from both sides. To put it in Dyvik's (1991) words:

> "At least one optimistic thing can be said about the relationship between linguistics and MT: it is improving. There was a time when the two concepts were rarely conjoined, and if they were, 'but' seemed a more plausible conjunction than 'and'. Today, however, both fields have changed somewhat: linguistic theories of grammar tend to take problems of computational tractability seriously, and linguistic approaches are finding their way not only into small experimental MT projects but also into the large market-oriented ones." (p.67)

Returning to our main topic, although Hauenschild does not put forward a specific proposal for how the thematic structure of a text should be accounted for in MT, she does indicate a line of investigation which MT research should pursue. A promising strategy, she claims, is trying to predict the thematic structure of a text on the basis of the syntactic shape of the individual sentences that make up the text itself. The syntax of a sentence, in fact, in many cases can tell us quite a lot about its information structure. This insight is of course not Hauenschild's own: as we shall see in Chapter 3, the relation between information structure and syntax has been explored in great detail by the scholars of the Prague School. What is interesting about Hauenschild's observations, however, is the fact that she points out the potential relevance of the Prague School theory to MT.

In fact, a direct application of the principles of the Prague School to MT has been attempted by the very developers of those principles. Thus, Hajičová, Sgall & Skoumalová (1993) and Hajičová, Skoumalová & Sgall (1995), propose two sets of heuristics for the computation of the topic-focus articulation in English and Czech texts. This work is discussed in detail in Chapter 5, and constitutes in fact the starting point for our proposal regarding the computation of information structure in Danish sentences.

A similar attempt is described in Styś & Zemke (1995). The model developed here is applied to the automatic translation of English texts into Polish. In analysis, the different degrees of salience of the various discourse topics are calculated on the basis of centering theory (cf. Grosz, Joshi & Weinstein 1995). The syntactic and semantic contraints used for this calculation are similar to those mentioned by Hajičová et al. They include definiteness, pronominalisation, synonymity and a default ordering of grammatical functions. On the basis of the gradience of topic salience thus obtained, a complex table of preferences is then used to select the best word order(s) in generation.

Steinberger (1994) provides an algorithm for the selection of the most appropriate word order in the generation of German texts in the MT system CAT2. His account is based on a tripartite distinction between *theme*, *rheme* and (contrastive) *focus*. These notions are not defined in the paper, but they seem to correspond to the constituent in sentence-initial position, to a constituent expressing new information, and to a constituent bearing contrastive stress, respectively. Steinberger's very clear intuition is that although in a relatively free word order language like German, not all word orders are appropriate in all contexts, i.e. that the context dictates the word order of a given sentence, it is also true that given a certain word order, a given context can often be assumed. This means in short that it is not necessary to analyse the preceding context to find out what the rheme of a sentence is, for example, because the syntactic shape of the sentence itself tells us that. Given the difficulty of modelling and computing textual structure, the insight that the individual sentences in most cases provide us with the necessary information to understand their information structure, is clearly very important.

Steinberger's system is implemented in a unification-based formalism which makes it close to the model described in this thesis. From a formal point of view, his word ordering preferences are expressed as a single quite complex linear precedence statement, which orders constituents depending on their thematic functions (i.e. whether they are classified as theme, rheme or focus) as well as their syntactic functions. In the absence of a thematic assignment, constituents are assigned a default position solely on the basis of syntactic function.

The various approaches discussed in this section all share a common goal, namely that of looking at syntax as an important source of information for a computation of the information structure of the sentences in a text. The work described in this thesis clearly follows the line of research embodied in such proposals. Our goal is, in fact, to investigate to what extent the idea of computing the information structure of a sentence independent of its context can be applied to Danish data, and to implement a model of information structure that allows for the exploitation of such a computation in an MT system. As a practical testbed for our model we have chosen ALEP (Groenendijk & Simpkins 1995), a unification-based system for NLP developed for the European Commission, and presently used at the Center for Sprogteknologi for the implementation of a medium-scale grammar of Danish. The system is briefly described in the next section.

2.4 The ALEP system

ALEP (Advanced Language Engineering Platform) is a general purpose language engineering environment. One of the applications it is intended for is MT, and within the EU-funded LS-GRAM project, medium-scale grammars for a number of European languages are being implemented in the formalism supported by ALEP (cf. Schmidt, Rieder, Theofilidis & Declerc 1996).

ALEP has been chosen as the testbed for our theory for a number of reasons. First of all, ALEP supports MT: this is obviously a crucial requirement for this project.

Secondly, ALEP supports a formalism (The ET-6/1 Formalism, described in (Alshawi, Arnold, Backofen, Carter, Lindop, Netter, Pulman, Tsujii & Uszkoreit 1991)) which allows for the implementation of grammars very similar to the HPSG model. A brief overview of the formalism is found in Chapter 6. For a more detailed account, however, the reader is referred to the ALEP literature. Here it will suffice to say that linguistic descriptions – both phrasal and lexical signs – are typed feature structures. Unlike HPSG, however, ALEP makes use of a syntactic backbone represented in terms of syntactic trees. Types can be e.g. atoms, lists, tuples and functors. Structure sharing between two different features can be defined by way of variables. More or less all the principles used in HPSG can be simulated in ALEP, with the notable exception of lexical rules, which are not supported. This means that developing an information structure component in ALEP also makes it possible to test whether such a component can be made to fit in with a state-of-the-art grammar formalism. To facilitate readability, most of the linguistic representations presented in this thesis – either rules or lexical entries – follow the widely known HPSG formalism rather than the somewhat less readable ALEP notation. However, examples of ALEP rules are shown in Chapter 6. Whenever some HPSG-like strategy proposed for the treatment of a relevant linguistic phenomenon represents a problem to the ALEP formalism, this is indicated and discussed.

Furthermore, two more specific features also make ALEP a potentially interesting platform to our purposes. One is the possibility of implementing an additional *refining* module on top of the analysis grammar. A refining module is used to add features to the structures created by the parser. As will be explained in detail in Chapter 6, the refining operation provides us with an elegant means of keeping our information structure contraints separated from the rest of the grammar. This is a desirable effect because of the heuristic nature of these constraints.

The other interesting feature from the point of view of a treatment of textual structure is the availability in ALEP of a text handling system able to take the paragraph as the default linguistic unit. This means that grammar rules can be written to parse the structure of a paragraph, thus defining interactions between its various sentences. This possibility has been experimented with in the treatment of pronominal anaphora (cf. Declerck 1996). We have not implemented a grammar aimed at paragraph parsing in relation to this thesis. However, this is an interesting feature, which could provide a framework for modelling discourse-related notions relevant to our work, such as topic salience.

2.5 Summary

In this chapter we have seen that different languages express the information structure of a sentence in different ways. Even rather closely related languages like English and Danish show discrepancies e.g. as regards the position of indefinite subjects, the use of extraposition and *there* insertion, and the possibility of placing adjuncts in sentence-initial position.

These differences are not taken into account in current MT systems – either PC-based or more complex ones – with a detrimental effect on the quality of the output. In a few experimental systems, however, the information structure of the individual sentences of a text is computed on the basis of the syntactic shape of the sentence itself. This computation is the result of the application of preference rules based on a variety of different factors, e.g. word order, definiteness, pronominalisation and syntactic function.

The work described in this thesis follows up on this line of research. Its goal is to investigate to what extent the idea of computing the information structure of a sentence independent of its context can be applied to Danish data, and to implement a model of information structure that exploits this computation in an MT system.

In the final section we have provided a short description of the ALEP language engineering environment, which we have chosen for a partial implementation of the model proposed in the thesis.

Chapter 3

Some previous approaches to information structure

In the previous chapter, information structure was defined in terms of the organisation of the information expressed by the sentence in such a way as to fit in with what has been said before and what will be said afterwards. The issue of information structure has been dealt with by a variety of authors working in quite different fields, and has given rise to a number of partially overlapping terminological distinctions, such as those between *theme* and *rheme*, *topic* and *focus*, *given* and *new*, *current focus* and *potential foci*.

In this chapter, we provide a survey of some of the most influential approaches to the topic. This survey tries to meet a number of goals. Firstly, we introduce some of the linguistic problems this thesis will address; secondly, we explain some of the most used terms used in the literature and clarify our own terminology; and finally we begin to outline our standpoint with respect to the ongoing discussion of how information structure should be described.

The chapter is organised as follows. Section 3.1 deals with the theory of information structure developed by the Prague school. Section 3.2 describes Halliday's work on the same topic. Section 3.3 deals with the generative approach to information structure, in particular Chomsky's work on focus and Rochemont and Culicover's on focus marking constructions in English. In these three sections, we are particularly concerned with the way in which the various authors account for the relationship between information structure and syntax. Section 3.4 turns to the theories of focussing and centering put forth in Artificial Intelligence, and compares them to the linguistic-oriented approaches. Finally, Section 3.5 deals with Givon's and Geluyken's work, where the contribution of different syntactic constructions to the overall discourse structure is analysed on the basis of quantitative corpus analysis.

3.1 The Prague model: the topic-focus articulation

By the term Prague model we refer to the linguistic theory of functional generative description as proposed e.g. in Sgall, Hajičová & Panenova (1986). The Prague model, however, builds on Mathesius's pioneer work in the 1910's, further developed by Mathesius himself, Daneš and Firbas in the 60's[1] and is thus the first modern linguistic theory where the issue of information structure plays an important role.

In this model, a sentence consists of two parts, a *topic* and a *focus*. The topic conveys the contextually bound information expressed in the sentence, whereas the focus is the part expressing the non-bound, irrevocable information predicated of the topic. This fundamental partition is called the topic-focus articulation. Moreover, each constituent in a sentence is said to contribute to the building up of information in different degrees: this scalar order is called the *scale of Communicative Dynamism* (CD), and is reflected directly in the order of nodes in the particular type of dependency representation adopted. On the basis

[1] See e.g. Daneš (1966), or the recent editions Mathesius (1975) and Firbas (1992).

of CD, the *topic proper* is defined as the least dynamic (most given) element, and the focus as the most dynamic (newest) one.

Mathesius was the first to point out that Slavic languages such as Russian or Czech, express the topic-focus articulation in a different way from e.g. English. In fact, in Slavic languages word order reflects more or less directly the structuring of information in the sentence, with the contextually bound information expressed by the topic placed to the left of the new information conveyed by the focus. In English, where word order is constrained by rules of the grammar, topic and focus are rather indicated by intonation, or by means of particular syntactic constructions. The following Russian examples illustrate the point:

(1) Papa prines elku.
(lit: Dad brought Christmas tree.
Dad brought the CHRISTMAS tree.)

(2) Papa elku prines.
(lit: Dad Christmas tree brought.
Dad DID bring the Christmas tree.)

(3) Elku prines papa.
(lit: Christmas tree brought dad.
DAD brought the Christmas tree.)

(Gundel 1988:22)

The three Russian sentences only differ in the mutual order of the three constituents, which is determined by the fact that given information precedes new information[2]. On the basis of his comparative studies, Mathesius concluded that in English, there is a tendency to make the topic proper of the sentence coincide with the subject, since the initial position in the sentence, which is where the topic proper tends to occur, is also the default subject position according to English grammar. Naturally, it would be a simplification to claim that in English the subject always occurs in initial position or that it is always the topic. Adverbials, for example, often precede the subject, and indefinite subjects cannot belong to the topic as they refer to new information.

However, Sgall, Hajičová and Panenová assume a default ordering to be found in very simple affirmative sentences, where the subject is the topic and the rest of the sentence constitutes the focus. They call this the primary case, in which the two main structuring principles at work in the sentence, i.e. the valency patterns and the topic-focus articulation, reinforce one another. But even in simple affirmative sentences not displaying any form of syntactic topicalisation, the authors admit that the two principles may diverge. In the pair of sentences below, for example, intonation changes the default topic-focus order in (a), so that the subject expresses the new piece of information.

(4) a. TOMMY cried.
b. Tommy CRIED.

(Sgall et al. 1986:178)

According to the Prague model, the topic-focus articulation belongs to the deep-syntactic representation of the sentence, which provides the necessary basis for semantic representation. This representation, called TR (tectogrammatical representation), consists of two main types of information, the representation of the valency relations on the one hand, and the topic-focus articulation on the other. More specifically, a TR is a dependency tree where the mother node bears a label representing the governor (the main verb), while the daughter nodes represent complements and adjuncts. The linear order of the dependent nodes, as mentioned previously, is a direct reflection of the scale of communicative dynamism.

[2] Here, the terms given and new are used in the sense of *contextually* given or new, i.e. as synonymous to contextually bound and non-bound.

The notion of communicative dynamism is, in our opinion, the most controversial one of the Prague model. Clearly, it is a useful descriptive tool to account for free word order languages such as Czech, where, accepting the idea that context-boundness is a scalar dimension, the actual scale in a given sentence in most cases can simply be *read off* the relative order of the various sentence constituents[3]. But in languages with a relatively fixed word order, such as English or Danish, it is difficult to see an empirical justification for positing such a notion. The notion has been criticised as unconvincing e.g. in Chafe (1976).

The fact that the information expressed in a sentence can roughly be divided into the two categories of contextually given and new, on the other hand, seems a generally acknowledged fact, although defining exactly what it means for a piece of information to be given or new is not easy. We shall return to this issue later on. What concerns us here is the use made of the two notions to define and indentify topic and focus. Particularly, subsuming focus under the label of non-bound information does not capture the variety of pragmatic effects associated with focal prominence. This is noted by the authors themselves, who quote the following example (reported as being due to Lakoff), where the focus of the second sentence falls on a constituent referring to a contextually bound pronoun:

(5) John called Mary a REPUBLICAN. Then she insulted HIM.
(Sgall et al. 1886:221)

As we shall see in more detail, an important function of the focus of the sentence is that of signalling a contrast with another entity, either an implied one, or one explicitly mentioned in the preceding text as in the example under discussion here. This dimension is not much stressed in the Prague model. In fact, a whole variety of pragmatic functions associated with the focus of the sentence can be distinguished: we return to the issue in Chapter 4.

The distinction between topic and focus is claimed to bear a relation to pronominal anaphora and the definiteness of NPs. This particular development of the Prague theory, as embodied e.g. in Hajičová, Kuboň & Kuboň (1990) is very close to similar work in artificial intelligence (cf. Section 3.4). The idea is that the salience of a discourse referent – i.e. its degree of activation in the knowledge shared by speaker and listener at any given point in a discourse – increases or decreases differently depending on whether the referent occurs in the topic or the focus part of the sentence. A threshold of activation can then be defined to provide an operational characterisation of contextually bound items. On the basis of salience, predictions can be made regarding subsequent mentioning of a contextually bound referent – either through pronominalisation or through the use of various definite expressions. We find this approach very interesting. However, pronominal anaphora is not within the scope of this thesis. Therefore, the computation of the so-called *hierarchy of salience* will not be discussed in detail.

A strength of the Prague approach, in our opinion, is the attention given to the fact that the information structure of a sentence may contribute to its truth value. An example of this contribution is the effect that the topic-focus articulation has on quantifier scoping. Consider thus the following example:

(6) a. German is spoken by everybody in this ROOM. (or: Everybody in this ROOM speaks German.)
b. Everybody in this room speaks GERMAN.
(Sgall et al. 1986:232)

The authors explain the different quantifier scoping in the two sentences above as a function of a difference in topic-focus articulation. In the (a) versions, *German* is the topic proper (as indicated by word order in the first sentence, and by intonation in the second) and has wider scope than the quantified nominal phrase *everybody in this room*, which belongs to the focus. In (b), on the other hand, the roles are reversed, and the quantified phrase has scope over the grammatical object. In general, then, a quantifier associated with the topic of a sentence has wider scope than one associated with the focus. Note that this

[3]More precisely, the order of the constituents in the topic part of the clause is claimed to obey CD, while the order of the constituents in the focus part is in accordance to the so-called Systemic Order (SO), which is determined by the rules of grammar. SO is discussed in detail in Section 5.2.3.

is not the same as claiming that quantifier scoping directly depends on word order. The two sentences in (a), in fact, show that word order may or may not correspond to the order of quantifier scoping.

The interplay between the topic-focus articulation and negation is also relevant to semantic interpretation (cf. Vlk 1988). Thus, in the example below, the intonation shows that the main verb is contextually non-bound (part of the focus) in (a), but contextually bound (part of the topic) in (b). The two sentences have different truth values:

(7) a. John didn't go out because he was tired.
(He stayed at home)
b. John DIDN'T go out because he was tired, but because...
(He didn't stay at home)

The Prague theory is also interesting for our purposes because it has been used as a basis for operational algorithms for the computation of the information structure of a sentence. In particular, Hajičová et al. (1993) and Hajičová et al. (1995) propose heuristics for the computation of the topic-focus articulation in Czech and English technical texts. The fundamental assumption behind this work is the fact that the syntactic shape of the sentence, in English as well as in Czech, often tells us what the information structure of the sentence is without having to refer to discourse analysis. This is especially true of technical texts, where intonation tends to be unmarked, and information structure can thus be computed on the basis of word order, definiteness, lexical semantic criteria, and a language-specific default ordering of verbal complements. These heuristics are discussed in detail in Chapter 4, and used as a model for the formulation of similar criteria relevant to Danish.

3.2 The theme in Halliday's approach

The principles developed in the Prague tradition entered Western linguistics through Halliday's work. According to Halliday (1967), the English clause can be analysed along three distinct dimensions: transitivity, mood and theme. *Transitivity* is concerned with the organisation of cognitive content in the linguistic expression, i.e. with the type of process expressed in the clause, the participants in this process, and the various attributes and circumstances characterising it. *Mood* deals with the dimension of the various speech situations, i.e. with the various speaker roles. *Theme* is concerned with the information structure of the clause, i.e. the status of the elements of the clause in relation to the overall discourse. In Halliday's words, "theme is the grammar of discourse" (p.199). As will become clearer below, however, theme is also used to refer to a particular role in the information structure of the sentence.

Halliday's grammar of discourse distinguishes two different systems: one is the organisation of the discourse into units of information, the other one is the division of the clause in a theme and a rheme part.

An *information unit* is a tonic group, in other words a phonological unit containing at least one point of prominence (possibly two). In the unmarked case, a clause corresponds to one information unit. But an information unit can also be a sentence chunk, or extend across a clause boundary: all these cases are regarded as marked. Halliday gives the following example of the way in which the same (written) clause can be divided into one or more information units (the symbol '//' indicates an information unit boundary):

(8) //John saw the play yesterday//
//John// saw the play yesterday//
//John// saw the play// yesterday//
//John saw the play yesterday but said nothing about it//
//John// saw the play yesterday and is seeing it again today//
(Halliday 1967:201)

The number of information units in a clause determines the number of focus points, as each unit may contain at most two foci (an obligatory primary focus and an optional secondary one). It does not, however, specify exactly where the focus is located, as shown by the example below (where capital letters indicate focal prominence):

(9) //Mary// always goes to TOWN on Saturdays//
//Mary// always goes to town on SATURDAYS//

The pragmatic function of a *primary information focus* (or simply a focus) is characterised by Halliday as the marking of a part of the information unit as being the new, informative part of the message. *Secondary focus*, on the other hand, is associated with contingent or confirmatory information, particularly frequent on clause-initial adjuncts. As noted in the preceding section, however, it is not always the case that the information associated with a focus has not already been mentioned in the preceding text. Hallidays is aware of this:

> "What is focal is 'new' information; not in the sense that it cannot have been previously mentioned, although it is often the case that it has not been, but in the sense that the speaker presents it as not being recoverable from the preceding discourse" (p.204).

Halliday also subsumes cases of contrastive focus under the label of 'new' information, so that the 'new' category seems to be broader than the notion of contextually non-bound information used in the Prague model. It seems to us, however, that the term 'new' is a rather misleading one for information that may or may not be new in the discourse.

Generally, but not always, given precedes new in the information unit.

If focus according to Halliday is a notion linked to the information unit, the point of origin of thematisation is the clause. In other words, focality and thematisation are two independent although related systems. This constitutes a difference with respect to the Prague theory, based on one fundamental distinction between topic and focus.

Thematisation assigns the clause a structure in terms of the two functions of *theme* and *rheme*. The theme is what comes first in the clause, the rheme is the rest. Halliday claims that although given often corresponds to theme, the two categories do not necessarily coincide. The difference between given and theme is not spelled out: an obvious distinguishing feature, however, is the fact that the theme usually consists of a single constituent, whilst the given information may extend to a longer part of the clause. Halliday also claims that "thematisation takes a unit of sentence structure, the clause, and structures it in a way that is independent of what has gone before" (p. 212). This is a somewhat puzzling statement: since, as Halliday himself observes, given information tends to occur at the beginning of the sentence, and in fact to coincide with the theme, and since given information is such because of 'what has gone before', it seems to us that the theme, although defined in terms of the actual clause it occurs in, is clearly connected to the preceding text.

Thus, from a pragmatic point of view, the theme is the point of departure of the message expressed in the clause; from a syntactic point of view, it is the initial constituent of the clause. Discussing the relation between the notion of theme and the grammar of English, Halliday identifies different types of unmarked themes related to different clause types. In a non-polar interrogative, the unmarked theme is a wh-item. In a polar interrogative, it is the finite verb element carrying the realisation of polarity. In a declarative clause, the unmarked theme is claimed to correspond to the subject (and comes thus quite close to the Prague topic).

Marked themes are possible in all types of clauses. In the interrogative it is an infrequent choice but it is possible. Halliday mentions the following example:

(10) Yesterday what did you do?
(Halliday 1967:214)

In declarative clauses the thematic pressure on the subject is not as strong. Therefore, cases of marked theme in declarative clauses are quite frequent. Halliday is in fact very clear about the fact that 'unmarked' has nothing to do with frequent: an unmarked case is simply a case which is perceived as neutral, whereas a marked one conveys a kind of foregrounding of information which may be described as emphatic. Thus, he claims that the concept of theme is linked to the possibility of a choice on the part of the speaker: therefore, items that can only occur in initial position, such as coordinating conjunctions, are not considered themes.

The difference between a marked and an unmarked theme is also shown in English by the fact that a marked theme tends to constitute a separate information unit:

(11) //These HOUSES// my grandfather SOLD//
//TOMORROW// John's taking me to the THEATRE//
(Halliday 1967:214)

Halliday discusses various cases of marked themes. Here, we shall only mention the case of preposed adjuncts, an issue which will be discussed at some length in Chapter 5, in relation to the grammar of topicalisation in Danish.

The adjunct is, according to Halliday, the most frequent example of marked theme in English. An example is:

(12) Tomorrow we are going to the cinema.

It is also very frequent in Danish, where it is always associated with inversion between the subject and the finite verb:

(13) I går fik jeg besøg af en gammel ven.
(lit: Yesterday had I a visit of an old friend.
Yesterday an old friend came to see me.)

The frequency of preposed adjuncts makes it seem somewhat questionable to treat these cases as examples of marked theme: however, it has already been noted that marked does not mean infrequent. Halliday observes that the fact that a preposed adjunct is the theme of the clause is confirmed by the fact that it cannot be followed by a thematic constituent:

(14) * Yesterday the play John saw.
(Halliday 1967:219)

The same constraint holds in Danish:

(15) * I går min veninde så jeg.
(lit: Yesterday my friend saw I.)

Moreover, the phonological characterisation of a marked theme in English, namely the fact that it constitutes an independent information unit, holds in the case of a preposed adjunct. Likewise in Danish, inversion, which occurs in main clauses every time a constituent other than the subject (leaving aside coordinating conjunctions and complementisers) is placed in initial position, is also obligatory with a preposed adjunct. Therefore, Halliday's conclusion that adjunct thematisation should be considered a marked case seems a reasonable one.

To conclude, Halliday's account of information structure, in addition to its historical importance for having brought an important field of investigation to the attention of Western linguistics, in our opinion constitutes a very insightful model. Its main strength, we believe, lies in its analysis of thematisation and focality as two independent although related sentence structuring processes.

3.3 Focus in generative grammar

Syntacticians working in the paradigm of generative grammar have been particularly concerned with the interaction between focus and grammar. We shall discuss the first generative account of focus, due to Chomsky, and the work carried out by Rochemont and Culicover on English focus marking constructions.

3.3.1 Chomsky: focus and deep structure

Chomsky (1971) points out – a fact already known to the Prague scholars – that focus contributes to the semantic interpretation of the sentence, and that since focus is a feature pertaining to the surface structure of the language, certain properties of surface structure must be made available to semantic interpretation. The paper deals with a number of syntactic constructions, e.g. cleft sentences and raising adjective constructions, all bearing on the argument that focus contributes to the semantic content of the sentence.

Chomsky claims that in any isolated sentence, the focus can be any phrase that has focal stress on its rightward daughter, e.g.:

(16) He was warned to (look out for (an ex-convict (with (a red (SHIRT)))).
(Chomsky 1971:201)

By replacing the focus with a variable, the *presupposition* of the sentence is obtained. Thus, the sentence above has a number of possible presuppositions, depending on where the left-hand boundary of the focussed phrase is placed:

(17) He was warned to do something
He was warned to look out for something/somebody
etc.

As we shall see in Chapter 4, Chomsky is not alone in considering focus a presupposition-triggering factor. However, focus is not the only presupposition trigger, and it is not clear in Chomsky's approach how focus could be related to a general theory of presupposition.

Chomsky argues that both focus and presupposition must be indicated in the semantic representation of a sentence so that the different interpretations of a sentence made possible by stressing different constituents can be accommodated. His main concern is discussing whether the concept of deep structure should be changed to account for the interpretation of focus, or whether it should simply be assumed that semantic interpretation has access to surface structure. He chooses the second alternative for a number of reasons.

Firstly, in some cases the deep structure of a sentence does not even display the constituents that are necessary to account for focus and presupposition. This is for instance the case in a sentence such as the following one:

(18) John isn't (certain (to WIN)).

For a number of theory-internal reasons, the deep structure of this sentence is assumed to be something like (with, perhaps, a specification of negation):

(19) [$_S$ John win] is certain.
(Chomsky 1971:202)

Thus, the interpretation where the left-hand boundary of the focus precedes the adjective *certain*, cannot be obtained directly from the deep structure, as *certain to win* does not constitute a phrase at that level. Changing the definition of deep structure in a radical way to make it easier to represent the focus and presupposition of a sentence, on the other hand, would result in loss of generalisation and unwanted complexity.

Additional evidence of the fact that deep structure is not the most natural place for expressing the focus and presupposition of a sentence, comes from Chomsky's discussion of quantifier interaction, dative NP shift, passive voice, parallel constructions, etc. To take the case of the passive, for example, no difference is actually made in deep structure between the active and passive realisations of the same proposition. Different choices of focus, however, are associated with the active and passive forms. Therefore, also in this case the focus cannot easily be represented at the level of deep structure.

In conclusion, a weaker version of the standard theory is embraced, where deep structure defines the grammatical relations that determine semantic interpretation, and focus and presupposition are directly accessible from surface structure.

Having the semantic representation of a sentence draw directly on information provided by surface structure may be a problem in a generative framework, where surface and deep structure are separate representational levels (although of course related through transformations). Achieving the same result may be easier in a different framework, where surface information, information related to deep syntactic rules as well as the semantic representation are merged into a single representational device. Thus, Chomsky's conclusion reinforces our choice of implementing information structure in a grammar formalism reminiscent of HPSG (see Chapter 5), where the various information types are all available within the same complex feature structure.

3.3.2 Rochemont and Culicover: focus marking constructions

Rochemont and Culicover's work on English focus marking constructions has been an important source of inspiration for our work, which is particularly concerned with the relation between information structure and syntax.

According to Rochemont (1989), focus is a string containing an *accented* lexical item which may qualify as new information in the context at hand. We return to the term 'accented' below. New information is defined negatively as information which is not *c(ontext)-construable*, i.e. neither under discussion nor lexically marked as indirectly c-construable. Being under discussion seems roughly to correspond to being contextually bound in the Prague approach, whereas the notion of indirect c-construability is used to characterise classes of lexical items which tend to occur in discourse initiating contexts. Three different groups of lexically c-construable words are mentioned, namely verbs of appearance, non-deictic referential pronouns and certain temporal and locative adverbials. An example is the verb *arrive* in the following sentence:

(20) A letter has arrived.

Items belonging to these classes would be coded accordingly in the lexicon, and would never be assigned a focus analysis. Although this strategy has an immediate appeal because of its simplicity, the assumption it is based on is not uncontroversial. A verb like *arrive*, for instance, can clearly occur in focussed position, as shown below:

(21) The letter I was waiting for has finally arrived.

A similar observation can be made for time and place adverbials, which often constitute given information, describing a setting against which the focussed information should be interpreted, but which seem equally capable of constituting the focus of the sentence. However, the idea that certain lexical items tend to be associated with given information is an interesting one, and will be taken up again in Section 5.2, where we deal with heuristics for the computation of information structure in written texts.

According to Rochemont & Culicover (1990), two diagnostics can be used to determine the focus of a sentence (pp.18-19):

- in a well-formed wh-question answer sequence, all and only the information requested in the question is focussed in the response;
- in a well-formed simple question answer sequence, all and only the information provided in the response that is not contained in the question is focussed.

An example of the latter test, which will be used throughout this thesis, is shown below:

(22) Did Bill talk to his brother?
(a) No, he talked to his SISTER.
(b) ? No, he TALKED to his sister.
(c) ? No, he TALKED to his SISTER.
(Rochemont & Culicover 1990:19)

Only (a) is a well-formed response to the question, as the focus in a question answer response has to fall on the constituent(s) expressing new information. In (b) and (c), on the other hand, a given item (the verb) is focussed, in contraposition to what is expected after the question.

If, at the pragmatic level, focus is associated with information that is not c-construable, at the phonological level it is marked by sentence accent. The authors distinguish between stress, which is lexically marked and cannot be changed by focus, and accent, which indicates focality. The only context in which lexical stress can be changed is in contexts of repair. Consider thus the following three examples:

(23) a. I didn't say blueBERRYS, I said blueBIRD.
b. * We didn't find any blueBERRYS, but we did see some blueBIRDS.
c. We didn't find any BLUEberries, but we did see some BLUEbirds.
(Rochemont & Culicover 1990:23)

In (a), which is a context of repair, the lexical stress of the two focussed nouns is changed, and focus falls on a morpheme in both cases. In (b), the same change of lexical stress is not possible, even if a contrast is still involved. Finally, (c) shows that the normal lexical stress must be used instead. Thus, leaving aside contexts of repair, a focussed string is claimed always to correspond to a syntactic phrase.

We saw that Chomsky (1971) defines the focus of a sentence as any phrase that has focal stress (accent following Rochemont's terminology) on its rightward daughter. This notion is sometimes referred to as the right-branching account. Rochemont (1989) (drawing on Selkirk (1984)) criticises the right-branching account, and shows that valency-bound complements and adjuncts play different roles in the delimitation of focality. He suggests that focus spreading from a focussed lexical item to a whole phrase is subject to the following principles:

1. a head may be focussed if it has a focussed argument within its projection (thus, focus does not spread from subject to verb);
2. an adjunct may be focussed if the head it modifies is focussed;
3. the projection of a head may be focussed if its head is focussed.

The first principle says that a head may inherit focus from any of its valency-bound complements, but not from its subject. Note that the principle does not make any reference to geometrical properties. It is exemplified in the two sentences below (the second of which is quoted in Rochemont 1989:9):

(24) a. What is John doing? He's eating a SANDWICH.
b. What is John doing to the table? He is putting a new LEG on it.

In example (a) then, the whole VP is the new information in the response, and thus constitutes the focus. Focus spreads thus from the accented object – *sandwich* – to the verb, and from this to the VP (3rd principle). In example (b), the focussed phrase is still the VP. In this case, the accented complement is not the rightmost one, so that the right-branching account would fail to yield the right focus assignent[4]. This argument is only valid, however, if the VP structure in this case is assumed to be ternary rather than binary. We shall not discuss which would be the best structure, but only note that making the focus spreading principles independent of the particular geometry adopted seems a desirable choice.

The fact that focus does not spread from the subject to the verb is shown by the example below:

(25) What happened to the cat?
a. JOHN is FEEDING it.
b. ? JOHN is feeding it.

(Rochemont 1989:11)

The accent pattern in (b) makes the response odd as it forces an interpretation where the verb is not part of the focus, in contrast to what is implied in the question.

To show that complements and adjuncts behave in different ways with respect to focus spreading, Rochemont quotes the following examples:

(26) When did John disappear?
John disappeared last NIGHT.

(27) What happened to John?
? John disappeared last NIGHT.

(28) I hear there was a party at John's last weekend.
a. ? Yeah, the party was raided last NIGHT.
b. Yeah, the party lasted all NIGHT.

(Rochemont 1989:9-10)

Thus, the sequence in example (26) is correct because the verb is c-construable, and thus not in focus in the answer, whereas in (27) the verb is not c-construable, and the sequence is odd: this shows that focus does not spread from the adjunct to the VP. Finally in (28), we see a contrast between a complement and an adjunct: (a) is ill-formed because the focus does not spread from the adjunct to the VP as would be necessary since the verb is not c-construable; (b), on the other hand, is a correct response because the adverbial is subcategorised for by the verb, and focus can subsequently spread to the VP.

Finally, the following example shows that focus can spread from the verb to an adjunct:

(29) What happened to John?
He DISAPPEARED last night.

The fact that valency-bound complements and adjuncts behave differently also seems to hold for Danish, as shown by the following examples:

(30) Hvornår spiste du frokost?
Jeg spiste frokost klokken TOLV.

(When did you have lunch?
I had lunch at twelve.)

[4]Not every analyst would agree with Rochemont's analysis here. Vallduví & Engdahl (1995) claim for example that focus can only spread from the most oblique daughter.

(31) Hvad gjorde du så?
? Jeg spiste frokost klokken TOLV.

(What did you do then?
I had lunch at twelve.)

The first of the two examples is well-formed, as primary stress[5] falls on the adverbial, which is also the new information, and thus the focus, in the answer. The second example, on the other hand, is odd because the same stress pattern seems to indicate again that only the adverbial is in focus, whilst the verb is also part of the new information.

The best way to render the sequence acceptable is to move the adverbial to the beginning of the sentence, so that the object (in addition to the adverbial) bears primary stress, and the focus can spread to the whole VP:

(32) Hvad gjorde du så?
Klokken tolv spiste jeg FROKOST.

(What did you do then?
At twelve I had lunch.)

The relationship between stress and focus, however, is not a primary concern of this thesis, as our goal is to treat the information structure of written texts, where stress patterns are not available. Therefore, we shall not attempt to deal with the issue of how focus spreads from lexical items to phrases in Danish. The point we are mostly concerned with in relation to Rochemont's discussion of sentence accent and focus, is the basic assumption behind it, namely that focus is a feature relating to syntactic constituents. However, it is open to question whether the relation between syntactic constituency and information structure is one of complete isomorphy. For instance, Engdahl & Vallduví (1996) claim that the focus of a sentence does not necessarily correspond to a constituent[6]. This issue will be discussed in more detail in Chapter 5, where we propose a syntactic representation for various focus marking constructions.

Prosody is not always, however, the only way in which focus is indicated in English. Rochemont and Culicover identify a number of syntactic constructions where focus is structurally marked, and sentence accent naturally falls on the marked constituent. These are (all the examples are from Rochemont 1989:13):

- Heavy NP Shift (HNPS): *John noticed in the lobby a large group of excited reporters.*
- Presentational *There* Insertion (PTI): *There ran into the lobby a large group of excited reporters.*
- Preposing Around *Be* (PAB): *Behind him was his wife.*
- Directional/Locative Inversion (D/L): *Into the room ran his wife.*
- Extraposition from NP (EX): *A man came into the room who everybody recognised.*
- It Clefts: *It was a group of reporters that John noticed in the lobby.*
- Pseudoclefts: *What John noticed in the lobby was a group of reporters.*

[5] We are using here the terminology proposed in Hansen & Lund (1983).

[6] The example they give is:

What happened to the china set?
[The BUTLER BROKE]$_F$ the set.
(p.16)

However, it may be argued that an equally possible interpretation is:

[The BUTLER]$_{PT}$ [BROKE]$_F$ the set.

On the latter analysis, the subject is the primary topic of the sentence, and the verb alone constitutes the focus. Thus, isomorphism between syntactic and information structures is preserved.

In constructions where the focus is structurally determined, accent cannot be used to move it. To prove that the constructions under consideration really mark the focus, the authors embed relevant examples in question answer sequences, e.g.:

(33) a. Who just walked into the bathroom?
That MAN just walked into the bathroom from INDIA.
b. Which room did that man from India walk into?
?That man walked into the BATHROOM from India.
(Rochemont & Culicover 1990:26)

The first sequence shows that the focus can coincide with the extraposed constituent, the second that it must.

Rochemont claims that HNPS, PAB and even PTI are "broadly used in all sorts of texts" (p.16), although he does not claim that they are frequent. Thus, it is reasonable to expect that the analysis grammar of an MT system should be able to recognise them. But he also claims that "failure to use them [structural focus constructions] will result in texts that are stylistically inelegant and possibly even hard to read" (Rochemont 1989:16).

In fact, he mentions an example where failure to generate HPNS, in a scenario where such a construction is blocked in generation, leads to ambiguity. The ambiguous (b) version in the example below can be compared with the unambiguous one in (a):

(34) a. The committee discussed immediately a proposal to restrict media access to their deliberations.
b. The committee discussed a proposal to restrict media access to their deliberations immediately.
(Rochemont 1989:17-18)

Rochemont claims in fact that in (b), although strictly speaking the sentence is ambiguous, there is a preference for attaching the adverbial to the embedded clause, thus conveying a different meaning than in (a). Interestingly, Rochemont's example resembles one of our examples of MT output given in Chapter 2 (example 11).

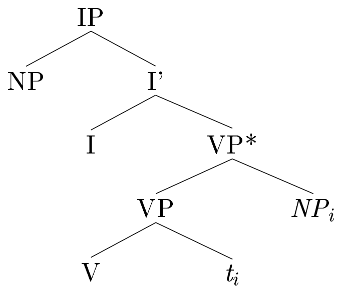
Rochemont is in fact concerned with MT, and makes a concrete implementation proposal for the treatment of focus marking constructions in an MT system. This consists of two separate sets of rules, one for the computation of focus spreading on the basis of the principles discussed above, and the other for the computation of structural focus. The first set assumes the availability of an accent feature marking the words in the input where the accent falls. It consists of the following rules (which apply in the order shown):

- Assign as a default **accent=-** to lexical nodes that are not marked by accent.
- Assign **focus=+** where **accent=+**.
- Percolate focus from head to projection.
- Percolate focus from argument to head.
- Percolate focus to specifier from its governing head.
- Assign as a default **focus=-** to nodes not marked by focus.

The second set of rules, which is more relevant to the objectives of this thesis, assigns a value to the focus feature solely on the basis of configurational properties. We shall illustrate what the basic idea is by discussing the rules proposed for HNPS and PTI, both of which are regarded as instances of *rightward move alfa*, a transformation by which an NP is moved to the right of its originating position and adjoined to another constituent.

In the case of HNPS, the NP is claimed to move from object position to a position adjoined to the VP, as shown in the tree below:

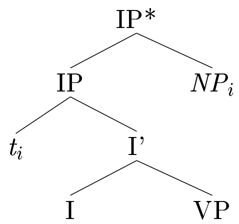
(35)



In the GB framework in which Rochemont and Culicover's theory is encouched, I (or INFL) contains tense and agreement features, and IP (which corresponds to the top node of a declarative sentence), is a projection of I (cf. Haegeman 1991). VP* is the node created by adjunction to the VP.

In the case of PTI, movement is assumed to take place from subject position to a position adjoined to IP. The fact that *there* is inserted in PTI and not in HNPS is explained as an effect of the different positions of the trace: movement from subject position, which is not a lexically governed position, triggers *there* insertion, while movement from object position, which is a lexically governed position, blocks it. The tree below shows the structure proposed for PTI, where IP* is the node created by adjunction of the moved NP:

(36)

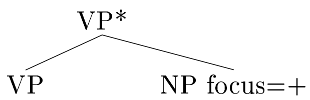


The rules for focus assignment are as follows:

- Focus assignment in HNPS:

  Assign **focus=+** to an NP in adjoined position to a VP.
  The diagram below shows the relevant configuration after the application of the rule:

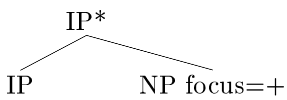
  (37)

- Focus assignment in PTI:

  Assign **focus=+** to an NP in adjoined position to an IP.
  The diagram below shows the relevant configuration after the application of the rule:

  (38)

Since our objective is to investigate to what extent the information structure of the sentences in a text can be computed on the basis of their syntactic shape, in order to generate a more coherent automatic translation of this text, Rochemont and Culicover's treatment of focus marking constructions in English is clearly relevant to our work. In Chapter 5, we deal with a number of Danish syntactic constructions in which either the focus or the primary topic[7] of the sentence can be claimed to be syntactically marked, and where the information structure of the sentence can be computed in terms of features, following a strategy similar to the one sketched out here.

[7]The term *primary topic* is explained in Section 4.1.2.

3.4 Information focus in artificial intelligence

The way in which the term 'focus' is used in Artificial Intelligence is quite different from the linguistic use we have referred to so far. In the AI sense, focus is concerned with the speaker's centre of attention at a particular point in a discourse.

Grosz (1977) distinguishes between *global focus*, which extends over a set of discourse utterances, and consists of a set of referents which are sufficiently activated during the whole discourse, and *immediate focus*, which is about the way in which the speaker's centre of attention remains constant or changes over two succeeding utterances. Grosz's notion of global focus is applied in a system for dialogue understanding. The system's knowledge base consists of a semantic network, and the matching of network structures is a crucial step in the interpretation of utterances. Global focus is represented as a partitioning of the semantic network which limits the search space of the matching algorithm. Thus, by 'focussing' on the section of the network which is relevant to a certain portion of the discourse, it is likely that the system will be quicker and more accurate in finding referents for definite NPs, or the right lexical senses of ambiguous words.

Grosz's theory is further developed in Sidner (1979) and McKeown (1985). Both are particularly concerned with immediate focus, and use a number of notions specific to that, namely:

- *current focus*: the immediate focus of a sentence;
- *potential focus list*: the items of a sentence which are potential candidates for the current focus of the succeeding sentence, ordered according to their degree of activation, or salience;
- *focus stack*: a stack of past immediate foci.

Grosz's global focus has no parallel in the linguistic theories we have discussed so far, but at the level of immediate focus a comparison between the linguistic approach to information structure and the AI concepts can be made. Thus, current focus seems to correspond to the topic proper of the Prague theory, while the most salient element on the potential focus list is equivalent to the focus in the linguistic sense. Thus, although AI workers stress the fact that focussing is about the speaker's intentions in communication rather than about functional roles played by the various sentence referents, there is at least a partial overlapping between the phenomena discussed in the AI literature on focussing and linguistic approaches to the issue of information structure (see also Hajičová 1987). A part of McKeown's work, for example, is dedicated to the syntactic realisation of different focussing strategies, in other words to the particular focus effects achieved in a discourse when using passive constructions, *there* insertion and *it* extraposition. Her most original contribution, however, regards the way in which focussing can be used to generate coherent text in response to questions about the structure of a database. Assuming that the system has to provide a certain amount of information to answer a query, its focussing mechanism provides it with a strategy for organising this information (expressed at the semantic level as a number of propositions) in a coherent way. This is done by having the generation process follow the following preference rules:

1. change focus to member of previous potential focus list if possible;
2. maintain focus if possible;
3. return to topic of previous discussion.

   (McKeown 1985:66)

McKeown's task is thus, provided a set of propositions, to organise them in a coherent text, in other words to create the most adequate information structure in the individual sentences making up this text. The problem in Machine Translation is quite different, as the input to the system is a text where the information is already organised in a presumably coherent way. The generation component, therefore, does not have to create the information structure from scratch but to *re*create the structure present in the source text following the rules of the target language.

In a more recent approach to the issue of discourse coherence (Grosz et al. 1995) and (Brennan, Friedman & Pollard 1987), the term focussing is abandoned in favour of the notion of *centering*. Centering is concerned with modelling local coherence within a discourse segment. Local discourse coherence is seen as having a relationship with the discourse participants' attentional state, i.e. it is concerned with the participants' focus of attention at any given point in the discourse.

According to this model, each utterance (U) in a discourse segment is assigned an ordered list of forward-looking centres C_f, and each utterance other than the initial one is assigned a single backward-looking centre C_b. Centres are discourse entities that link an utterance to the rest of the discourse segment. More precisely, the most salient element of $C_f(U_n)$ realised in U_{n+1} is the $C_b(U_{n+1})$. A claim of centering theory is that there is only one C_b in each utterance.

The C_f plays roughly the same role as the list of potential foci, whilst C_b does not (or not always) correspond to the current focus of the previous theory: it may in fact, but need not be, the item which the actual sentence is about. It is simply the item linking it to the previous sentence.

Thus, in the following example, the C_b of sentence (b) is expressed by the pronoun *hende* (her), referring to *naboens pige* (the neighbour's daughter), whereas the current focus would be *vores datter* (our daughter):

(39) a. Naboens pige er fire år.
b. Vores datter leger meget med hende.

(The neighbour's girl is four.
Our daughter plays a lot with her.)

Thus, a direct mapping between the linguistic notions of topic and focus and the categories used in centering theory becomes difficult[8].

Compared to McKeown's model, centering theory also makes different claims regarding the activation of the topics participating in an utterance, i.e. on their likelihood of being taken up again in the succeeding utterance. In McKeown's model, in fact, the most salient element of a sentence is its (linguistic) focus. According to centering theory, on the other hand, the order of the elements in C_f is determined i.a. by syntactic factors, and the default ranking for English depends on grammatical function, with SUBJECT > OBJECT > OTHER (where > means 'more salient than'). In other words, the most salient item on this view will often correspond to the Prague topic. Note, however, that this is in contrast to what the Prague model claims: according to Hajičová (1987), the focus and not the topic is the most salient element in a sentence.

A test is used by Grosz et al. to determine which is the most salient item in an utterance. The idea is that if a unique pronoun is present in U_2, the referent which is picked as the default antecedent in U_1 is the most salient item in U_1. The results provided by this test lead the authors to claim that the subject is in general the most salient item. One of the examples they discuss is the following one:

(40) a. Susan gave Betsy a pet hamster.
b. She reminded her that such hamsters were quite shy.
c. ? She told Susan that she really liked the gift.

(Grosz et al. 1995:212)

The authors claim that in (b) the referent of *she* could be either *Susan* or *Betsy*. In (c), however, it becomes clear that the two referents do not have the same status. The sentence is not an adequate continuation of the discourse at hand, as the pronoun does not refer to the same referent as the subject in (a). To make it acceptable, the object must be pronominalised instead:

[8] Hoffman (1996), however, presents an algorithm that tries to generate the topic in Turkish based on centering theory.

(41) a. Susan gave Betsy a pet hamster.
b. She reminded her that such hamsters were quite shy.
c. Betsy told her that she really liked the gift.
(Grosz et al. 1995:212)

This would prove that the subject is indeed the most activated item. However, in this example the focus of the first sentence is not taken up in the succeding one, and it is therefore difficult to decide on this ground whether the focus is more salient than the subject. Consider a similar example:

(42) a. Susan gave her pet hamster to her friend Betsy.
b. She told Susan that she really liked the gift.

In this case, Betsy is the focus in sentence (a), and can be referred to by the unique pronoun in (b). This would seem to imply that it is the most activated item in (a). Compare the example above with the following one:

(43) a. Susan gave her pet hamster to her friend Betsy.
b. Betsy told her that she really liked the gift.

This version of the discourse, where the object referring to the subject of the (a) sentence is pronominalised is not impossible, but is less natural. Therefore, the result of the test would seem to be that the focus is more salient than the subject of a sentence.

We shall not try to resolve the issue. To prove whether the focus or the topic of a sentence is generally more likely to become the topic of the succeeding sentence would need quite an extensive corpus analysis. However, there is a sense in which it seems to us that the position taken by Hajičová and McKeown, according to which the (linguistic) focus of a sentence is the most salient item in the discourse is right. We quote McKeown on this point:

> "The choice between current focus and items on the potential focus list corresponds to choosing between continuing to talk about the same thing or starting to talk about something introduced in the last sentence. [...] Thus, for reasons of efficiency, when one has the choice of remaining on the same topic or switching to one just introduced, I claim the preference is to switch. If the speaker has something to say about an item just introduced and he does not present it next, he will have to go to the trouble of reintroducing the topic at a later point." (p.63)

3.5 Discourse analysis

Under the heading of discourse analysis we will discuss Givon's (1983) work on topic continuity and Geluykens's (1992) on left dislocation in English. These two authors use very similar methodologies based on the study of the discourse structure of texts. Both operate with the notion of *topic*, regarding it as a discourse category rather than one related to isolated sentences. Both propose operational criteria to measure the way in which the various discourse topics of a text are related to each other, and emphasise the necessity of large-scale corpus analyses to derive such criteria.

In Givon's work, topic is practically used in the sense of discourse referent, although a clear-cut definition of the term is never provided. Thus, no difference is made between topic and focus as in the Prague school's theory. However, sometimes his topic seems to come closer to the Prague concept, meaning what the clause (but also the paragraph or the discourse as a whole) is about. For instance, it is claimed that the surface subject is often the topic of the sentence because the referent it denotes tends to have a high degree of continuity in the text.

In spite of this lack of terminological clarity, Givon's work contains some interesting suggestions regarding a number of parameters which the discourse analyst can use to measure *topic continuity* in texts, namely:

- *Referential distance*, which measures the gap between the previous occurrence of the topic (here, referent) and its current occurrence in a clause. The gap is expressed in terms of number of clauses, with a predetermined maximum distance defined.
- *Potential interference*, which expresses the disruptive effect that other referents within the immediately preceding register may have on the identification of a given topic. The immediately preceding register is set to three clauses to the left of the current clause.
- *Persistence*, which measures how long the topic is maintained in subsequent discourse. It is expressed in terms of the number of clauses to the right in which the topic in question is mentioned as a semantic argument.

Behind these measures are certain assumptions regarding the way in which discourse topics are processed. Basically, a discourse is treated as an internal filing system, possibly a double system consisting of a temporary file and a permanent one. Some topics are thus claimed to belong to the permanent file: these are typically topics referring to what is called here *generic firmament*, e.g. features of the universe, body parts, kinship terms and names. These items do not seem to obey the general topic continuity tendencies at work for other kinds of referents. Those that do obey these rules, on the other hand, reside (when active in the discourse) in the temporary file.

On the basis of the three criteria listed above, Givon sets up the following scale showing the relation between topic continuity and a number of syntactic devices which are part of the grammar of English (the scale is slightly simplified here for the sake of brevity):

most continuous/accessible topics
zero anaphora
unstressed pronouns
stressed pronouns
R-dislocated definite NPs
neutral ordered definite NPs
L-dislocated definite NPs
Y-moved NPs (topicalisation)
cleft/focus constructions
referential indefinite NPs
most discontinuous/inaccessible topics

It is interesting to note that focus marking constructions are placed towards the bottom of the scale. This confirms the widely acknowledged fact that focussed constituents often refer to items that are not recoverable from preceding context. In other words, a discontinuous topic, according to this theory, will often correspond to a focus. Hence, although no clear difference is made here between topic and focus, this distinction is somehow implicit in the scale of continuity: an item with a high degree of continuity/accessibility will coincide with topic as intended in the Prague model; an item with a low degree of continuity/accessibility, instead, will correspond to focus.

Thus, if we accept the sense in which Givon (and many others) use the term topic, namely as discourse referent, we are forced to draw the conclusion that a topic in some cases can also be the focus of the sentence. This is in fact an assumption we shall be making from here on. To refer to the topic which the sentence is about, we shall use instead the term *primary topic*. Our view of the relationship between topicality and focality will be clarified in Chapter 4.

Geluyken takes a clear position in the debate on the given/new dichotomy, by firmly criticising approaches based on the concepts of information flow, given/new information, topicality etc. because of their relying on what he calls subjective, vague and unfalsifiable assumptions. In particular, he reacts against definitions of given, recoverable and new information that refer to the speaker's intentions. The speaker's intentions are not available to analysis, and he claims therefore that they are best avoided in favour of operational criteria. The concept which is criticised most is that of shared knowledge (cf. e.g. Prince 1979), a pool of background as well as contextual information on the basis of which the listener would draw inferences and understand referents as either given or new. The main point of criticism is that this array of knowledge may be so vast as to be virtually useless for deciding which elements are given and which are not. However, although rejected in principle, the same concept re-emerges later, under the term of *contextual knowledge* to explain the inference in examples such as:

(44) I've got a job in a college; the students aren't very bright.
(Geluyken 1992:13)

It is claimed that *the students* are a recoverable referent, not directly from the context, but by virtue of a *college scenario*. Although delimiting a scenario is naturally a more manageable problem than delimiting undifferentiated background knowledge, it is still not obvious what should be included in the pool of given information and what should not.

The strength of Geluyken's approach, we believe, lies in his empirical observations about left dislocation in English, based on extensive corpus analyses. The purpose of this study is to verify whether generalisations can be made about the pragmatic function of left dislocation. The parameters he uses are *lookback* and *persistence*, which are not too different from Givon's referential distance and persistence. His conclusions show that left dislocation has a variety of communicative functions, although two of them, the introduction of new referents and the expression of a contrast, are classified as central (more frequent). In fact, it seems to be a general characteristic of the relation between syntactic constructions and pragmatic functions that in most cases no one-to-one correspondence can be postulated, although general tendencies for some correspondences to occur more frequently than others can be observed. This is a belief we share. Most of the rules for the computation of information structure we propose in Chapter 5, are based on heuristic principles rather than strict grammar rules.

3.6 Summary

In this chapter we have discussed some previous approaches to the issue of information structure.

We have seen that in the model of information structure developed by the Prague school, the sentence is segmented into a topic and a focus part, corresponding to contextually given and new information. Topic and focus are an integrated part of the deep syntactic representation of the sentence, thus contributing to semantic interpretation. An interesting development of this theory is the use of heuristic principles to compute which constituents belong to the topic and which to the focus in written technical texts. In Chapter 5, these heuristics are taken as the starting point for the definition of an algorithm for the computation of the information structure in Danish texts.

Halliday sees two different systems at work in the sentence: one is the theme-rheme articulation, the other the partition of the sentence into a given part and a new, focal one. In English, the theme is the initial constituent of a sentence, and it expresses what the sentence is about. It may or may not be given in the context at hand. The focus is what the speaker decides to present as 'new', irrespective of whether it has been mentioned previously or not. Halliday's intuition that two distinct although related articulations can be seen at work in a sentence, provides in our opinion a more accurate framework for the description of the relevant linguistic facts than the Prague model. The view of information structure we defend in the next chapter is also based on a twofold distinction between *topicality* on the one hand, and *focality* on the other.

The discussion of Chomsky's and Rochemont and Culicover's generative approaches to focus has raised a number of interesting issues. One is the relation between focus and presupposition, which will be taken up again in the next chapter. Another is the relation between information structure and syntax. This is a central issue in our work: what we are trying to do in the context of an MT application, in fact, is compute the information structure of a sentence on the basis of its syntactic shape in analysis, to generate more coherent text. In this light, we have briefly dealt with Rochemont's proposal for how focus can be represented in a GB syntactic framework. Although our work is not couched in a GB framework and the analyses of various syntactic constructions proposed in Chapter 5 are different from Rochemont's, the spirit of our work is very similar to his.

Finally, we have provided an account of two approaches which are quite different from those mentioned above. One is the theory of focussing developed in AI, recently replaced by the so-called centering theory. In this framework, described in the works of Grosz, Sidner and McKeown, the stress is not so much on the internal organisation of the individual sentence, as on the way in which different discourse strategies can be modelled on the basis of the salience of the various discourse topics. The other approach is that embraced by Givon and Geluykens, who model the relation between sentence topic and discourse topic on the basis of quantitative measures related to topic continuity and disruption. Both the AI and the discourse topic approaches use the terms focus and topic in a rather different way from the more linguistically-oriented theories mentioned above. Our approach to information structure is closer to the linguistic approaches, in the sense that it is concerned with the internal organisation of the sentence rather than with a discourse model.

Chapter 4

Information Structure Revisited

We have seen that the fact that one or more constituents in a sentence are emphasised – either by prosodic or syntactic means – has been explained in the literature by reference to the flow of information in the discourse. On this view, some mental model shared by the discourse participants is constantly updated, by adding, deleting and reorganising information as the discourse proceeds, and emphasis is used to indicate which constituents in the sentence are the repositories of new information. Thus, the flow of information in the discourse determines the information structure of the individual sentences, where given information is related to new information according to prosodic or syntactic rules that vary according to the language.

We have also seen that, although similar terms are used by scholars working within different schools of thought, there is no widely shared consensus on a general model of information structure. The very distinction between given and new information seems to elude researchers' definitional attempts. The problem is that information is not given or new in itself, not even in the context of a concrete discourse: it is rather the speaker or writer who presents a given piece of information as given or new. However, the speaker's choice is not completely free, being subject to a complex network of syntactic, semantic and pragmatic factors that regulate well-formed, cooperative discourse. Thus, in order to decide whether a certain referent or event can be classified as given or new in the context in which it occurs, the discourse analyst has to make reference to a complex system of constraints often including extratextual information.

Our aim in this work, however, is to define a model of information structure that keeps reference to world knowledge and extratextual information to a minimum, so that it can be integrated into an operational MT system without impairing the system's efficiency. The basic distinction in our model is not that between given and new information, but rather a distinction along the two pragmatic dimensions of *topicality* and *focality* as defined in Dik (1989).

Topicality characterises the entities (topics) participating in the discourse, where different topic types may be distinguished on the basis of the way in which they relate to other discourse topics. Focality, on the other hand, refers to the fact that one or more elements of an utterance are emphasised to indicate some kind of modification of the pragmatic information built by the overall discourse.

Topicality and focality are not to be seen as mutually exclusive. They overlap in the case of new topics, i.e. topics that are introduced in the discourse for the first time: these are in fact usually focussed as will be shown below.

Each of Dik's dimensions gives rise, in our model, to a primitive making up the information structure representation of the sentence. The primitive relating to topicality is the *primary topic*, which will be defined as the topic which the sentence is about. The primitive relating to focality is the *focus* of the sentence, i.e. the locus of the 'modifying' information. To these two notions, we add a third primitive, which we call the *neutral information*. Neutral information is information which can be considered peripheral to the construction of the discourse, and which serves the purpose of specifying or restricting the way in which the focal information must contribute to the discourse.

This chapter discusses each of the three primitives, and proposes a way in which they can be added to the syntactic and semantic representation of a sentence in a typed feature structure formalism such as HPSG. This model and this notation constitute the basis of the treatment of various linguistic phenomena in the subsequent chapters of this thesis.

The organisation of the chapter is as follows. Section 4.1 deals with topicality, and in particular with the notion of primary topic. Section 4.2 is about focality. First the pragmatic functions of focus are discussed, and then a unified semantic account of focus is provided following Rooth (1992). The final section contains our proposal for the representation of information structure in a typed feature structure formalism.

4.1 Topicality

Topicality is a pragmatic dimension relating to the discourse entities occurring in a discourse, which we call *discourse topics*. The notion of discourse topic should be interpreted relative to a certain discourse segment, which can be modelled in terms of an empty topic store, where topics are added as they are introduced into the discourse. According to their pragmatic functions, various topic types can be distinguished. In the following sections we deal with new topics, given topics, and resumed topics, and briefly discuss different linguistic strategies used to express them.

One of the topics of the sentence has a particular status, as it is perceived as the topic which the sentence is about. We call it the *primary topic* of a sentence. The primary topic is discussed below in relation to Halliday's theme and to the notion of contrastive topic.

4.1.1 New, given and resumed topic: three different pragmatic functions

When a discourse topic is introduced for the first time, it is called a *new topic*. After having been introduced, a topic can be considered a *given topic*. However, a discourse topic that has not been mentioned explicitly can also be presented as given if it can be inferred on the basis of another given topic. We shall call such a topic an *inferrable topic*. The issue of how to characterise inferrable topics has received a great deal of attention. Prince (1979) for example, explains inferrable items as belonging to a pool of shared information comprising general background as well as contextual information. On the basis of the pool of shared information speaker and listener draw inferences and understand referents as either given or new. Geluykens (1992) deals with the issue by way of conceptual scenarios bringing a limited set of items to the listener's attention. Dik defines an inferrable topic (which he calls a sub-topic) as a discourse topic related to another, given topic by the inference relation, which is in turn defined as an open list of more specific semantic relations, namely:

> "part of, member of, subset of, instance of, copy of, aspect of, opposite of, projection of, associated with, ..." (p.276)

No matter what precise characterisation of inference is used, it is crucial to note that the listener (or reader) does not understand a newly introduced topic as given because he is consciously aware of a certain scenario, or semantic relationship. Rather, it is the linguistic context in which the topic occurs (e.g. the fact that it is a definite nominal functioning as the grammatical subject) that leads the listener to understand the topic as given in relation to the information available at that point in the discourse.

A third pragmatic function can be distinguished to indicate a given topic which is not mentioned in the discourse for a while, so that its salience becomes very low, and which is then re-introduced into the discourse. Such a topic will be called a *resumed topic*.

Strategies for introducing a new topic

Introducing new topics is, as will become clear in our discussion of focality in Section 4.2, one of the pragmatic functions of focus. Thus, topicality and focality here overlap, and we can say that a new topic is focussed, or in focus, or a focus.

Various linguistic strategies can be used to introduce a new topic. Dik mentions the following array of syntactic and semantic means for expressing a new topic:

- meta-linguistic statements (*I'm going to tell you about a king*);
- use of the direct object to introduce a new topic (*The king suddenly saw a dragon*);
- existential or locative-existential constructions (*Once upon a time there was a king*);
- predicates expressing the notion of 'appearing' (*Suddenly a dragon appeared*);
- placing of the new topic in end-focus position (*In the middle of the forest lived a dragon*).

He correctly points out that new topics are often indefinite nominals, but they needn't be. Thus, if the identity of a new topic is known to the hearer (or is assumed to be known), a definite article or a possessive pronoun may be used also if the entity is a new topic. One of the examples he gives is:

(1) Yesterday in the pub I met *your sister Mary.*
(Dik 1989:269)

Dik's list, however, does not seem complete. It can be compared, for example, with the focus marking constructions dealt with by Rochemont and Culicover, which we list here again for clarity's sake (all examples are from Rochemont 1989:13):

- Heavy NP Shift (HNPS): *John noticed in the lobby a large group of excited reporters.*
- Presentational *There* Insertion (PTI): *There ran into the lobby a large group of excited reporters.*
- Preposing Around *Be* (PAB): *Behind him was his wife.*
- Directional/Locative Inversion (D/L): *Into the room ran his wife.*
- Extraposition from NP (EX): *A man came into the room who everybody recognised.*
- It Clefts: *It was a group of reporters that John noticed in the lobby.*
- Pseudoclefts: *What John noticed in the lobby was a group of reporters.*

In Chapter 5, we discuss what linguistic strategies are available in Danish to express new topics in the focal part of the sentence.

Strategies for maintaining a given topic

Strategies of topic maintainance have been described in various terms, e.g. centre continuation and retaining (Grosz et al. 1995), and topic continuity (Givon 1983). We shall again base our account on Dik (1989), who distinguishes a number of strategies available in different languages to maintain the same topic through a discourse segment, namely:

- Anaphoric reference.

  A given topic is maintained by subsequent anaphoric reference in a sequence of clauses.

- Syntactic parallelism.

  In certain languages a topic chain requires that the topic which is to be maintained should occur in the same syntactic positions in subsequent clauses. Also in a language like English, syntactic parallelism is at work in that the grammatical subject is also often (although not always) the primary topic of the sentence.

- Switch reference.

  Dik exemplifies switch reference with examples from the Papuan language Wambon, where there is a grammatical opposition between verbal forms signalling a change in topic and others signalling topic continuity.

- Obviation.

  This relates to a grammatical marking of third-person nominals directly mirroring the topicality status of the nominal. It is found in the Algonquian languages.

Dik's list can be compared with Givon's scale, already shown in the preceding chapter:

most continuous/accessible topics
zero anaphora
unstressed pronouns
stressed pronouns
R-dislocated definite NPs
neutral ordered definite NPs
L-dislocated definite NPs
Y-moved NPs (topicalisation)
cleft/focus constructions
referential indefinite NPs
most discontinuous/inaccessible topics

It can be seen that Givon is more explicit than Dik about what kind of anaphoric expressions are used to refer to given topics, where the heavier the pronoun used, the less given is the antecedent. In fact, it has been claimed by many (cf. e.g. Hajičová et al. 1990) that the topics of a discourse segment can be ordered according to their degree of activation, or salience, and that their gradience has an effect on the kind of expressions (either pronominal or full nominal) that can be used to refer to them in subsequent sentences. Efforts have also been made, as discussed in the preceding chapter, to define the primary topic of the sentence in terms of salience. We have seen, however, that an unequivocal characterisation is quite problematic: we will therefore abstain from it.

4.1.2 The primary topic

From an intuitive point of view, one of the topics of a sentence is often understood as the 'most important' one – the item which the sentence is about and which links it to preceding discourse. This item is often simply called the topic, or topic proper. Not to confuse it with the more general term of (discourse) topic, we shall call it the *primary topic*, and indicate it by a subscript $_{PT}$ attached to the relevant constituent.

An operational test to pick up the primary topic of a sentence is provided by the question *What about X?*, where the instantiation of X corresponds to the primary topic. Thus:

(2) *What about John?*
[John]$_{PT}$ is Mary's husband.

To give a more formal characterisation of the primary topic of the sentence does not seem possible. Dahl (1974), however, claims that there exist fairly strong intuitions amongst speakers about what a sentence is about, in other words about what its primary topic is (he calls it simply the topic). In particular, he mentions a number of psycholinguistic experiments showing that speakers of English almost uniformly identify the primary topic of a sentence with its grammatical subject when this is a definite phrase. If the subject is indefinite, on the other hand, answers show greater variation.

He then goes on to explain that the primary topic seems to be linked with the presuppositions of a sentence. Following Strawson (1964), he argues that the following two sentences carry different presuppositions:

(3) a. The king of France is bald.
b. The Exhibition was visited yesterday by the King of France.

(Dahl 1974:5)

The fact that the king of France does not exist makes (b) false, whilst (a) has no truth value. The explanation would be that the two sentences have different primary topics, and that a sentence is false only if it contains false information about its primary topic. In both examples, the grammatical subject is a definite NP, and it is perceived as the primary topic of the sentence.

If the subject is an indefinite NP, on the other hand, it does not carry an existential presupposition, and cannot easily be identified with the primary topic of the sentence. In fact, Dahl claims that in this case the sentence is topicless.

Thus, it is not obligatory for a sentence to have a primary topic: in Section 5.3.2 we discuss one such case of topicless sentences, namely existential sentences.

If we exclude well-defined cases of topicless syntactic constructions, we claim that in Danish the primary topic of the sentence often falls together with the constituent in theme position[1].

We can now discuss a few examples. Consider the following sequence:

(4) a. [Mette]$_{PT}$ er fire år.
b. [Hun]$_{PT}$ er en sød pige.
c. [Hende]$_{PT}$ leger Sofia meget med.

(Mette is four years.
She is a nice girl.
lit: Her plays Sofia a lot with.
Sofia plays a lot with her.)

In the first two sentences, the primary topic is the subject, which has the same reference in both sentences. In the third sentence, the primary topic is the object, still referring to the same discourse referent. This is a case where topicalisation is used to make the primary topic refer to the most salient discourse topic (which we take to be the focus of the preceding sentence).

The following example shows another case in which the topic is not the subject, but a topicalised constituent:

(5) a. Kan du lide dansk mad?
b. [Sild]$_{PT}$ kan jeg godt lide.
c. [Det]$_{PT}$ smager også godt.

[1]Examples where this is not so, however, can be constructed. The following is the Danish adapted version of an example quoted in Vallduví & Engdahl (1995):

Hvor kan jeg finde bestikket?
a. [Knivene]$_{PT}$ er i skuffen...
b. men jeg har glemt [gafflerne]$_{PT}$ i opvaskemaskinen.

(Where can I find the cutlery?
The knives are in the drawer...
but I've forgotten the forks in the dishwasher.)

(Lit: Do you like Danish food?
Herring I like.
It tastes good, in fact.)

In (b) the primary topic is the topicalised *sild* (herring). Note that the pragmatic function of the topicalised complement here is slightly different than in the preceding case. When the topicalised complement is a full NP, in fact, it seems to have a contrastive effect which is absent in the case of a weak pronoun[2].

Finally, in the following example of adjunct fronting, we assume again that the constituent in theme position – the fronted adjunct – is the primary topic of the sentence:

(6) I går læste jeg hele aften.
(Yesterday I studied the whole evening.)

4.1.3 Contrastive topics

We have seen that topics can be characterised as new, given or resumed on the basis of their relation to previously mentioned discourse topics, and that one of the topics of the sentence – often but not necessarily a given one – can be considered its primary topic.

In this section, we deal with an additional dimension of topicality, which cuts across the distinction between given, new and resumed. As noted i.a. by van Leusen & Kálmán (1994), some topics can be given a contrastive interpretation, a property which, as we shall see in the next section, makes them similar to the focus of the sentence. However, Leusen and Kálmán provide evidence from a number of languages showing that contrastive topic and focus are characterised by different prosodic and/or syntactic features. They mention the fact that in English, contrastive topics show a characteristic fall/rise in intonation, which they suggest can be indicated by a ˇ sign above the relevant syllable, as in:

(7) - I invited Charles for a dinner.
- I warn you: he likes few foods.
- Well, he likes HĚRRING.
(Leusen and Kálmán 1993:32)

They also point out that still in English, contrastive topics unlike foci can be topicalised:

(8) Well, HĚRRING he likes
(Leusen and Kálmán 1993:32)

Finally, they note that in Hungarian, contrastive topic and focus are subject to different positional constraints, and that Japanese marks contrastive topics with the topic morpheme *wa* and foci with the focus morpheme *ga*.

In Danish, contrastive topics are typically topicalised:

(9) - Jeg$_{PT}$ kender ikke så MANGE her.
- HAM$_{PT}$ der KENDER du da.
- Nå ja, han$_{PT}$ bor i min GADE.

(I don't know so many people here.
Him you know, surely.
Oh yes, he lives in my street.)

[2] We subsume both cases under the term topicalisation. Vallduví and Engdahl distinguish between topicalisation and topic fronting, where topic fronting refers to the fronting of weakened pronouns.

Leusen and Kálmán's characterisation of contrastive topics is not much different from the definition of focus we adopt in the next section. They claim, in fact, that a constrative topic is a prominent element providing the instantiation of a variable in the semantic interpretation of the sentence. However, it seems to us that an important additional feature is the fact that a contrastive topic always forms part of a pair together with the focus of the sentence. The contrastive topic and the focus can then be seen as a tuple, establishing a possible contrast with other, similar tuples. In our example, these would be tuples each consisting of a person and a statement of whether the subject *du* (you) knows them or not, e.g.:

(10) HAM der KENDER du.
HENDE der KENDER du (også).
DE ANDRE kender du IKKE.
etc.

(Him you know.
Her you also know.
The others you don't know.)

The fact that the presence of a contrastive topic implies a contrast between tuples explains the oddness of the following example:

(11) - Jeg tror ikke hun er gift.
? Jo, Hans er hun gift med.

(I don't believe she is married.
Yes, Hans she is married to.)

The object of the preposition, *Hans*, cannot be topicalised here, because there is no other constituent that can take over the role of focus. *Hans* is in fact the only piece of new information in the second sentence, and therefore the only candidate for focus status. In fact, the sentence is only acceptable in this context if *Hans* is placed in end-focus position:

(12) - Jeg tror ikke hun er gift.
- Jo, hun er gift med HANS.

(I don't believe she is married.
Yes, she is married to Hans.)

But in a context where another constituent, for example the verb, bears focal prominence, and where the topicalised constituent and the focus can be contrasted together with another tuple, the topicalised sentence is perfectly acceptable:

(13) Hans er hun GIFT med, men det er Peter hun bor SAMMEN med.

(Hans she is married to, but it's Peter she lives with.)

Sometimes, the contrast between tuples each containing a different contrastive topic is explicit. This is the case of the so-called parallel constructions, e.g.:

(14) [I gamle dage]$_{PT}$ FORTALTE kirken mennesker, hvilke principper de skulle leve efter. [I vore dages Danmark]$_{PT}$ betyder kirken TEMMELIG LIDT.
(Togeby, O. 1992:467, our emphasis)

(In the old days the church told people what principles they should live by. In modern Denmark the church means rather little.)

Adjuncts occurring in theme position do not necessarily indicate a contrast. In the example above, however, the topicalised adverbials in the two parallel sentences are explicitly contrasted which each other.

4.2 Focality

Focality refers to the fact that one or more elements of a sentence can be emphasised to indicate their particular pragmatic contribution. We call that part of the sentence which is singled out either through prosodic emphasis or by other grammatical means the *focus* of the sentence. The notation we use for a focussed constituent X is $[X]_F$, where $_F$ would be realised by a relevant feature in a feature structure representing the syntactic, semantic and pragmatic content of the constituent under consideration.

In what follows, we discuss the nature of the pragmatic contribution of sentence focus, and argue that the relationship between focality and new information is a problematic one. In fact, focus can be related to a number of different pragmatic functions. Finally, we try to subsume the various focus types under a uniform semantic characterisation of focality.

4.2.1 The pragmatic functions of focus

It is often mentioned that information which is new in the context at hand is always in focus, i.e. prominent either at the phonological or at some other level. As proof of this observation, the fact can be adduced that in a pragmatically well-functioning question answer sequence, only the information requested (by definition new) can be focussed in the answer. Consider:

(15) - Hvem kommer i morgen?
- HANS kommer i morgen.
? - Hans kommer i MORGEN.

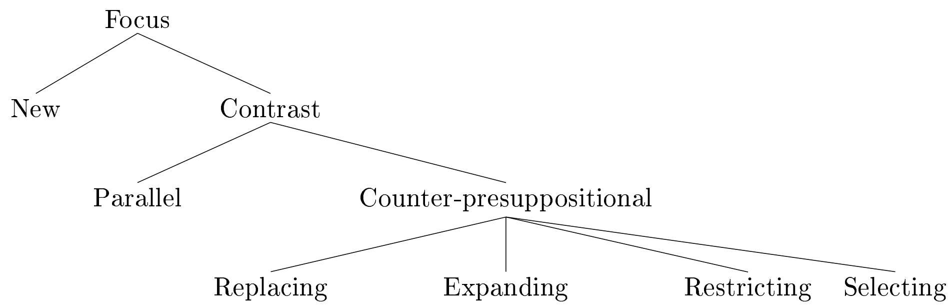
(Who is coming tomorrow?
HANS is coming tomorrow.
? Hans is coming TOMORROW.)

However, as soon as one tries to define the concepts of given and new information, clear as they seem intuitively, the simple correspondence between new information and focus becomes blurred. Thus, the following dialogue is possible, where *Hans* is given information and yet is in focus in the final line:

(16) - Har du ringet til alle?
- Jeg har ringet til HANS, ANNE og PETER.
- Og hvem kommer i morgen?
- Det gør HANS.

(Have you called everybody?
I have called Hans, Anne and Peter.
And who is coming tomorrow?
Hans is.)

In fact, introducing a new piece of information is only one of the pragmatic functions the focus of a sentence may have. On the basis of their different pragmatic functions, Dik (1989) distinguishes the following focus types[3]:



[3]Alternative typologies are of course possible. Togeby (1992) distinguishes for example *nyfokus* (new focus), *identitetsfokus* (identity focus), *modsætningsfokus* (contrastive focus), and *nægtet fokus* (negated focus).

Thus, the first basic partition is between focus introducing new information and focus of contrast. Focus of contrast is further subdivided into focus in parallel constructions and so-called counter-presuppositional focus, where the term *counter-presuppositional* points to the fact that the information which the relevant focus types modify may not be mentioned explicitly, but is somehow presupposed. Examples of the various types are given below.

- New focus

  (17) Hvad hedder du?
  Jeg hedder [SOFIA]$_F$.
  (What's your name?
  My name is Sofia.)

  The focus here introduces new information into the discourse.

- Focus in parallel constructions

  (18) Sofia og Petra er gode veninder.
  Sofia er [FEM]$_F$ og Petra er [TRE]$_F$.
  (Sofia and Petra are good friends.
  Sofia is five and Petra is three.)

  The two foci constitute a tuple establishing a contrast between the two primary topics of the coordinate sentence, namely *Sofia* and *Petra*.

- Replacing focus

  (19) Så du filmen i går aftes?
  Nej, jeg så [en politisk DEBAT]$_F$.
  (Did you watch the film last night?
  No, I watched a political debate.)

  Here the focus replaces a piece of information introduced in the preceding context.

- Expanding focus

  (20) Så du filmen i går aftes?
  Ja, og jeg så også [en politisk DEBAT]$_F$.
  (Did you watch the film last night?
  Yes, and I also watched a political debate.)

  Expanding focus presupposes that some piece of information introduced previously is correct but not complete. The missing information is thus provided.

- Restricting focus

  (21) Fik du købt det, vi havde aftalt?
  Nej, jeg købte kun [BRØD og MÆLK]$_F$.
  (Did you buy what we had agreed on?
  No, I only bought bread and milk.)

  Restricting focus modifies some previously established information by stating that only a subset of it is correct.

- Selecting focus

  (22) Drikker du kaffe eller te?
  Helst $[\text{KAFFE}]_F$, tak.
  (Would you like coffee or tea?
  Preferably coffee, please.)

  This focus type selects the correct piece of information from a list of possibly correct ones.

However, these different pragmatic functions are all characterised by the fact that the information expressed by the focus of a sentence in some sense modifies a piece of information introduced by a preceding sentence. In other words, even though different focus types may be described in terms of different pragmatic functions, they all have something in common which justifies their being understood as different aspects of the same linguistic phenomenon.

4.2.2 A semantic approach to focus

An interesting approach to the issue, is in our opinion Rooth's (1992). Rooth defines a focus as a constituent which is (prosodically) prominent, and which provides the instantiation of a variable in the semantic interpretation of the sentence. Let us consider a short Danish text, where as usual, capital letters indicate the words that are likely to be prosodically prominent:

(23) Her ser du PETER - som også hedder PETER-PEDAL, fordi han er så flink til at køre på CYKEL. Men DET må man læse om i en ANDEN bog.
(Here you see Peter - who is also called Peter-Pedal because he is so good at riding a bike. But you must read about this in another book.)
(*Peter Pedal hænger i*)

The most prominent[4] constituent in the first sentence is *Peter*, the object. Its prominence is underlined by word order in that it is placed towards the end of the sentence, before the extraposed *som også...*,

[4] Hansen & Lund (1983) distinguish three different degrees of stress in Danish, namely primary stress (*hovedtryk*), secondary stress (*bitryk*), and reduced stress (*nultryk*), indicated by ′, ˏ and ₒ respectively in the following example:

′Sagde ˏLis ₒdet?
(lit. Said Lis it?
Did Lis say it?)

Furthermore, they consider an unusually strong stress (*hovedtryk*), used to put emphasis on certain expressions, e.g.:

Det var en fan″tastisk film.
(It was a fantastic film)

A prosodically prominent constituent in our terms will at least bear primary stress in Danish. However, stress in itself does not suffice to single out the focus in a Danish sentence, as the choice of stress degree depends on a complex interaction of lexical, syntactic and pragmatic factors.

Nevertheless, focussing has an influence on the placing of primary stress in at least three ways:

- A focus bears at least primary stress.
- A sentence always contains at least one primary stress, which in this case clearly indicates the focus.
- A constituent which would only be marked by secondary or reduced stress according to lexical or syntactic rules, can be emphasised by primary stress in contrastive expressions:

  Avisen ligger ′på bordet, ikke nedenunder.
  (The newspaper is ON the table, not under it.)
  (Hansen & Lund 1983:14)

- Alternatively, a constituent can be emphasised by reducing the stress of other constituents in the same sentence:

  A: Anker ′Jørgensen rejser på ′ferie i au′gust.
  B: Nej, Kjeld ′Olesen rejser på ˌferie i auˌgust.
  (Anker Jørgensen is going on a holiday in August.
  No, Kjeld Olesen is going on a holiday in August.)
  (Hansen & Lund 1983:13)

which is separated from the rest of the sentence by a dash. For the moment, the meaning of the focussed constituent can be illustrated by way of a paraphrase:

(24) Her ser du $[\text{PETER}]_F$.
(Here you see Peter – not Anne, John, or somebody else, but Peter)

On Rooth's theory, the main function of a focus is understood as that of establishing a contrast between the focussed information and a set of alternative elements, which may be recoverable from the discourse or implied by the semantic and pragmatic content of the sentence. In other words, the contrastive nature of focus is stressed disregarding whether the information expressed is given or new in the discourse.

In fact, it can easily be seen that in all the examples of various focus types given above, a set of alternative elements to the element(s) in focus can be constructed.

Focus in Question Answer Sequences

In this section, Rooth's notion of focus as a semantic operator establishing a contrast between a semantic item and a set of alternative elements, is applied to the analysis of question answer sequences, which was mentioned above as the prototypical example of identity between focus and new information. The focussed constituent in the answer establishes a contrast between the actual answer and a set of potentially competing ones, where this set is constrained by the value of the focus itself and by semantic and pragmatic factors deriving from the context. Let us consider the following question answer sequences:

(25) - Hvem har taget bogen?
- $[\text{SOFIA}]_F$ har taget bogen.

(Who has taken the book?
Sofia has taken the book.)

(26) - Hvilken bog har Sofia taget?
- Hun har taget $[\text{PETER-PEDAL}]_F$.

(Which book has Sofia taken?
She has taken Peter-Pedal.)

The meaning of the answer in the first of the two sequences can be paraphrased as "Somebody has taken the book, and this somebody is Sofia". This can be formalised by λ-abstraction[5]:

(27) λx (take(x,book)) (Sofia)

As can be seen, a variable has been substituted for the focussed subject. The expression means that the property of having taken the book is predicated of Sofia.

Similarly, the meaning of the answer in the second sequence can be paraphrased as "Sofia has taken something, and this something is Peter-Pedal". Again, we can express this formally:

(28) λx (take(Sofia,x)) (Peter-Pedal)

The λ-expression means that the property of being taken by Sofia is predicated of Peter Pedal.

Rooth distinguishes between the ordinary semantic value of a sentence, expressed as a proposition, and its focus semantic value, where a variable is substituted for the element – whether a property, a proposition or a referent – corresponding to the focussed constituent. Thus, the focus semantic value of the answer in the first sequence will be the following set of propositions:

[5] Note that Rooth's formalisation does not make use af λ-abstraction.

(29) $\{take(x,book) \mid x \in E\}$,
where E is the domain of individuals.

Through the focus semantic value, the answer is related to a set of alternative sentences, which includes for instance:

(30) HANS har taget bogen.
Deres BROR har taget bogen.
En du KENDER har taget bogen.

(Hans has taken the book.
Their brother has taken the book.
Somebody you know has taken the book.)

Similarly, the focus semantic value of the answer in the second sequence will also be a set of propositions formalised as follows:

(31) $\{take(Sofia,x) \mid x \in E\}$

This set includes for instance:

(32) Hun har taget DIN bog.
Hun har taget den bog du havde lagt FREM til hende.
Hun har INGEN bog taget.

(She has taken your book.
She has taken the book you had put aside for her.
She has taken no book.)

Note, however, that the focus semantic values of the two answers do not describe exactly the set of propositions each of them is contrasted with in the actual question answer sequences. For example, the answer in (25) is not directly contrasted with the answer in the sequence below in spite of the fact that the question is the same in both cases:

(33) Hvem har taget bogen?
HUNDEN har taget bogen.

(Who has taken the book?
The dog has taken the book.)

The reason is of course that the question constrains the set of alternatives by indicating that the agent is a person, and that the answer in (33) overrides this presupposition. The set of potential answers contrasting with the answer in (25) can then be formalised in the following way:

(34) $\{take(x,book) \mid x \in E \,\&\, person(x)\}$

Similarly, the set of potential answers to the question in (26) is more constrained than the focus semantic value of the actual answer. It is in fact the following set:

(35) $\{take(Sofia,x) \mid x \in E \,\&\, book(x)\}$

Rooth points out that we can use the set of potential answers to identify the semantic value of the relevant question. In both the examples under discussion, the semantic value of the question is included in the set expressing the focus semantic value of the corresponding asserted answer. It appears in fact that this is a general requirement as shown by the inappropriateness of a sequence like the one below:

(36) Hvem har taget bogen?
? Sofia har taget BOGEN.

(Who has the taken the book?
? Sofia has taken the BOOK.)

The semantic value of the question, namely:

(37) $\{\text{take(x,book)} \mid \text{x} \in \text{E} \ \& \ \text{person(x)}\}$

is not included in the focus semantic value of the answer, namely:

(38) $\{\text{take(Sofia,x)} \mid \text{x} \in \text{E}\}$

It seems thus that an account of focus in terms of formal logic – i.e. without referring to pragmatic factors[6] – can lead us quite far in explaining the constraints governing focus in question answer sequences.

Question answer sequences will be used throughout this thesis as one of the means to identify the possible foci of a sentence. Thus, given a sentence, if a question answer sequence can be constructed where a constituent C is the new information instantiating the variable in the semantic value of the question, C is one of the possible foci the sentence may have. In a concrete context, of course, not all the possible foci that a sentence may have in isolation will be available.

Scalar implicature

In the preceding section, we saw that in the case of question answer sequences, the set of elements the focus in the answer should be seen as contrasting with is constrained by the semantic value of the question.

This section deals with expressions where the alternative elements in the set indicated by a focus are ordered along a scale. These are for example:

- Numbers: $1 < 2 < 3 < ... < n$
- Scalar time expressions: *second* < *minute* < *hour* < *day* < *week ...*
- Expressions of degree: *drives* < *drives quickly* < *drives very quickly*
- Hierarchically ordered lexical items: *manage* < *excel*

The elements belonging to each scale can be claimed to be ordered for informativeness, so that $A < B$ should be interpreted as *A is less informative than B*. If the focus falls on one of the items of such a scalar expression, the alternative set includes all the expressions in the same scale which are more informative than the focussed item.

Thus, *She MANAGES* contrasts with (i.e. does not entail) *She EXCELS*, whereas *She EXCELS* entails (i.e. does not contrast with) *She MANAGES*.

Certain operators can reverse the relation:

- less than a second > less than a minute > less than an hour

The interesting aspect of scalar expressions, thus, is the fact that if the focus of a sentence falls on one of those, the alternative set the element is implicitly contrasted with is lexically constrained.

[6]To define exactly what linguistic and extralinguistic factors belong to the realm of pragmatics is not an uncontroversial matter. We follow Levinson (1983) in understanding pragmatics as “the study of those relations between language and context that are grammaticalised, or encoded in the structure of a language” (p.9) . We also draw the boundary between semantics and pragmatics so that pragmatics deals with that portion of the meaning of a linguistic expression that cannot be expressed in terms of truth conditions. Thus, we understand the description of the meaning of focus as the instantiation of a variable in the semantic representation of a sentence as a semantic matter, whereas constraints related to salience and informativeness rather belong to the sphere of pragmatics.

The kinship condition

In constructions where the focus of a sentence is intended to refer to, or contrast with, another focus in the previous text – let us call it the antecedent, the set of alternative elements is obviously even more constrained than in the cases of question answer sequences or scalar expressions. It is in fact limited to the antecedent. This observation applies to all the focus types subsumed under contrast focus in Dik's focus typology.

The relationship between the actual focus and its antecedent must obey the so-called kinship condition (van Leusen & Kálmán 1994). This states that while there must hold a relation of distinction or contrast between the focus and its antecedent, the two must be akin with respect to some aspect of their meaning.

In other words, a focus and its antecedent partition a common domain in some specified way. For example, antecedent and focus may simply refer to distinct denotations, as in the example below, where the common domain is that of individuals:

(39) - Har du ringet til [PETER]$_F$?
- Nej, jeg har ringet til [PER]$_F$.

(Have you called Peter?
No, I called Per.)

The following two quotations from a book of Danish nursery rhymes (*Halfdans ABC*) are also examples where the antecedent and the focus have distinct referents although the focus type here is not contrastive, but rather an expanding focus in Dik's terminology. As for the common domain, in the first of the two examples it is semantically determined – everything that results from exposure to the sun. In the second, instead, it is rather determined by alliteration between the antecedent and the focus, with a humorous effect on the meaning of the rhyme:

(40) Sol gi'r [VARME]$_F$.
Sol gi'r [SVED]$_F$.
Sol gi'r [næsen FREGNER]$_F$...

(The sun gives [you] warmth.
The sun gives [you] sweat.
The sun gives the nose freckles...)

(41) Else elsker [PELSE]$_F$.
Else elsker [PØLSE]$_F$...

(Else loves furcoats.
Else loves sausages...)

Alternatively, the focus may refer to a subset of the denotation of the antecedent. This corresponds to Dik's restricting focus.

(42) Han elsker [ITALIENSK mad]$_F$.
Han er især vild med [FETTUCCINE]$_F$.

(He loves Italian food.
He is especially crazy about fettuccine.)

Finally, the denotation of the focus may be identical to that of the antecedent, but the linguistic expressions used must then be different, i.e. they must have different connotations. Leusen and Kálmán call these *metalinguistic arguments*.

(43) - Jeg synes, han er [lidt FRÆK]$_F$.
- Han er [en lille TERRORIST]$_F$!

(I think he is a bit naughty.
He is a little terrorist!)

A final observation to be made is that the kinship condition also applies to a contrastive topic when this is explicitly related to an antecedent, as is the case in parallel constructions which were dealt with in Section 4.1.3:

(44) *I går* skinnede SOLEN, *i dag* er vejret DÅRLIGT.

(lit: Yesterday shone the sun, today is the weather bad.)

The two topicalised adverbials here contrast with each other: the first can be considered the antecedent of the second. According to the kinship condition, they partition a common domain, which could be described here as the temporal setting of the compound sentence, into two parts, namely *i går* (yesterday) and *i dag* (today).

4.2.3 Focus and Presupposition

To conclude our discussion of focality, here we deal briefly with the relation between the notions of focus and presupposition.

We saw in Chapter 3 that in Chomsky (1971), focus is treated as a presupposition trigger. Thus, let us consider the following sentence:

(45) I går blev jeg færdig [med at revidere de første to KAPITLER]$_F$.
(Yesterday I finished editing the first two chapters.)

Replacing the focus with a variable yields the presupposition, i.e. that part of the meaning of the sentence which remains constant under negation. In this case, the presupposition would be:

(46) I går blev jeg færdig med noget.
(Yesterday I finished something.)

However, focus is not the only factor that triggers presuppositional constraints, as shown by the following simple example:

(47) Abernes skarpe skrig i den fredede skov advarer [HJORTENE]$_F$.
(The monkeys' sharp scream in the protected forest warned the deer.)

The sentence above has at least the following presuppositions:

(48) There are some monkeys.
There is a protected forest.
There are some deer.
The monkeys' sharp scream warned somebody.

The set of linguistic expressions that can be shown to trigger presupposition includes also definite articles, factive verbs, temporal conjunctions, non-restrictive relative clauses and clefts (cf. Levinson 1983).

Rooth (1992) takes a similar view on the relation between focus and presupposition. Discussing the nature of the constraints introduced by focus interpretation, he claims that:

"The important point is that they constrain the choice of antecedent for the variable introduced by focus interpretation. Based on an analogy with a similar result in a presuppositional/anaphoric theory of definite descriptions, I will tentatively identify the constraints introduced by focus interpretation as presuppositions (p.91)."

Discussing how the notion of focus as a presupposition-triggering factor can be integrated in a general theory of presupposition lies beyond the scope of this work. We shall simply subscribe to the view that focus introduces a presupposition into the semantic interpretation of the sentence where it occurs. In our discussion, this presupposition has been represented by way of λ-abstraction. According to that notation, the presupposition introduced by the focus of sentence (47) above would be:

(49) λy (warn(monkeys' sharp scream, y))

However, this is not the same as claiming that a representation in terms of λ-abstraction is the most adequate way of treating focus in MT. In fact, since focus is not the only possible presuppositional trigger, the λ operator would also be introduced in the semantic representation by other linguistic factors (e.g. definite articles). This means that a λ operator in the interface representation of a source sentence would not in itself be indicative of focus status. Therefore, if the information relative to the focality of one or more constituents is not recorded in some other way, it would be difficult to generate the correct focus in the target sentence. In other words, representing focus by way of λ-abstraction provides us with an interface representation which is both unnecessarily complex and too abstract at the same time. Thus, although we subscribe to the view of focus as a variable anaphoric to a set of alternative elements as proposed in Rooth (1992), for the purpose of an MT relevant implementation we propose to represent focus as an appropriate feature attached to the relevant constituent(s). The notation $[X]_F$ employed throughout this chapter is intended as a short-hand indicating that a focussed constituent X is conveniently marked by some focus feature.

4.2.4 Multiple and listed foci

The focus of a sentence may consist of more than a single constituent. In other words, a sentence can display multiple foci. An example is the following one:

(50) - Se, Sofia leger med $[\text{MATTHIAS}]_F$.
- Nej, det er $[\text{IDA}]_F$ der leger med $[\text{JONATHAN}]_F$.

(Look, Sofia is playing with Mathias.
No, it's Ida who is playing with Jonathan.)

Note that the two foci here can be interpreted as a tuple contrasting with the subject-complement pair of the initial sentence. Thus, it may be argued that this example comes very close to the cases of contrastive topics discussed in Section 4.1.3.

4.3 Proposal for an HPSG-like representation of information structure

In this section we discuss the way in which the information structure of a sentence can be represented in a typed feature structure formalism. As was explained in Chapter 2, our work has been implemented in ALEP. However, we shall discuss the model we propose using the more widely known HPSG notation. Typed feature structure formalisms are characterised by the interesting property of allowing a conflation of the syntactic, semantic and pragmatic analyses of a sentence into a single complex feature structure. As discussed in the previous sections, we want to represent the focus and primary topic of a sentence by way of features attached to the relevant constituents. In addition, we have chosen not to integrate a representation of focus and presupposition in the semantic representation. Therefore, here we propose to represent information structure as a feature belonging to pragmatic analysis.

We want to be able to express the fact that from the point of view of information structure, one of the constituents of the sentence can be considered the primary topic, while one or more constituents have to constitute the focus. Additionally, we want to be able to represent the fact that a part of the sentence may be neutral with respect to the building up of information in the discourse, either because it is given without constituting the primary topic, or because it is in some other sense peripheral.

Let us consider the following example:

(51) Hvad fik børnene til festen?
[De]$_{PT}$ fik [is]$_F$.
(What did the kids get at the party?
They got ice-cream.)

We assume that in the answer the subject is the primary topic, the object the focus, and the verb (which is simply repeated from the question) is neutral.

We can represent this by way of the following three features:

(52) $\begin{bmatrix}\text{INFOTYPE p\_topic}\end{bmatrix}$

(53) $\begin{bmatrix}\text{INFOTYPE focus}\end{bmatrix}$

(54) $\begin{bmatrix}\text{INFOTYPE neutral}\end{bmatrix}$

In a typed feature structure formalism such as ALEP or HPSG, the units of linguistic information are *signs*. A sign can either be phrasal or lexical. Signs are represented as complex typed feature structures expressing all the relevant phonological, syntactic, semantic and pragmatic information[7]. We propose to add the INFOTYPE feature as one of the CONTEXT features of a phrasal sign. Thus, the feature representation of the focus *is* (ice-cream) in the example above will be something like:

(55)
$$\begin{bmatrix}\text{PHON} \left\langle \text{is} \right\rangle & \\ \text{SYNSEM} & \begin{bmatrix}\text{LOCAL} & \begin{bmatrix}\text{CAT} \begin{bmatrix}\text{HEAD noun}\end{bmatrix} \\ \text{CONTENT} \begin{bmatrix}\text{RELN is}\end{bmatrix} \\ \text{CONTEXT} \begin{bmatrix}\text{INFOTYPE focus}\end{bmatrix}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

In this section, we are not concerned with the way in which the INFOTYPE feature is actually instantiated. In the next chapter, we shall discuss a number of heuristics by which the primary topic and the focus of a Danish sentence can be computed. The rest of the sentence will be assigned [INFOTYPE neutral] by default.

Thus, assuming that the INFOTYPE feature has been assigned, we need to pass the information about the roles played by the various sentence constituents in relation to topicality and focality upwards in the syntactic tree, so that the information structure of the whole sentence can be read off the sign representing the sentence itself.

To this end, we introduce another CONTEXT feature, namely:

[7] We assume here that the structure of the sign in HPSG is more or less known to the reader. For more details, cf. Pollard & Sag (1994).

(56) $$\begin{bmatrix} \text{INFOSTR} & \begin{bmatrix}\text{P\_TOPIC (singleton content)}\\ \text{FOCUS (set of contents)}\\ \text{NEUTRAL (set of contents)}\end{bmatrix}\end{bmatrix}$$

Thus, P_TOPIC takes a singleton content feature as its value, while FOCUS and NEUTRAL take sets of content features, since more than one constituent may belong to either category[8].

The actual instantiation of each of the three features will be achieved by ensuring that their values are structure-shared with the constituents bearing the relevant INFOTYPE feature. This is done by way of the following three principles:

1. In a sign marked [INFOTYPE p_topic], the value of INFOSTR|P_TOPIC is structure-shared with the content feature of the sign itself; similarly, in a sign marked [INFOTYPE focus], the value of INFOSTR|FOCUS is structure-shared with the content feature; and in a sign marked [INFOTYPE neutral], the value of INFOSTR|NEUTRAL is structure-shared with the content feature.

2. In a sign where the INFOTYPE feature is not instantiated, the value of the INFOSTR|FOCUS feature will be the union of the INFOSTR|FOCUS values of its daughters. The INFOSTR|NEUTRAL value is computed in the same way by taking the union of the corresponding values on the daugthers.

3. In a sign where one of the daughters bears the feature [INFOTYPE p_topic], the value of INFOSTR|P_TOPIC is copied from daughter to mother.

The effect of the first principle on the representation of *is* (ice-cream) shown above will be to assign a value to INFOSTR|FOCUS:

(57) $$\begin{bmatrix} \text{PHON} \left\langle \text{is} \right\rangle & \\ \text{SYNSEM} & \begin{bmatrix} \text{LOCAL} & \begin{bmatrix} \text{CAT} \begin{bmatrix}\text{HEAD noun}\end{bmatrix} & \\ \text{CONTENT } \boxed{1}\begin{bmatrix}\text{RELN is}\end{bmatrix} & \\ \text{CONTEXT} & \begin{bmatrix}\text{INFOTYPE focus}\\ \text{INFOSTR} \begin{bmatrix}\text{FOCUS} \left\{\boxed{1}\right\}\end{bmatrix}\end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

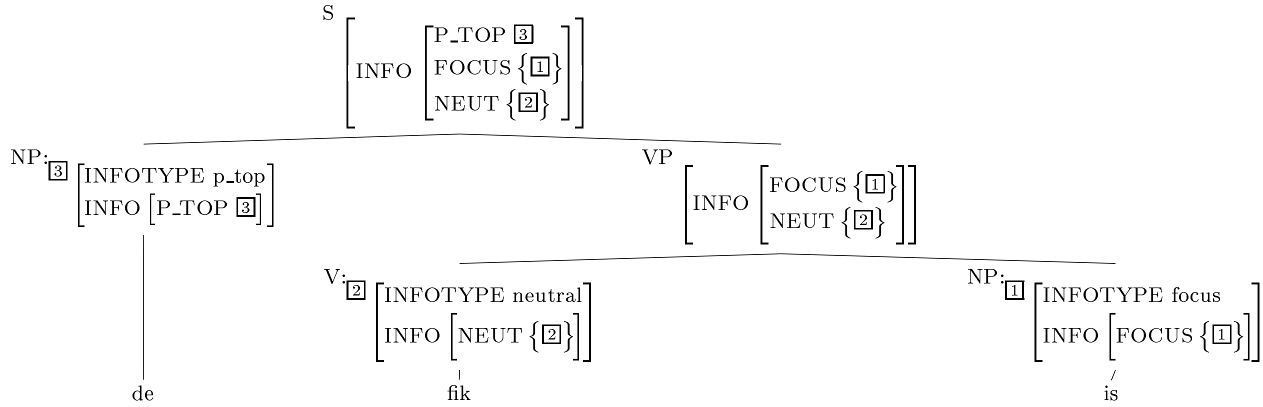
The other two principles will pass the INFOSTR features from the lower constituents up the tree to the top S node, as shown in the final representation of the information structure of our simple example[9]:

[8]In the ALEP implementation described in Chapter 6, sets are not available, and are therefore replaced by lists.

[9]Note that NP:$\boxed{i}$ is an abbreviation for:

$$\begin{bmatrix} \text{LOCAL} & \begin{bmatrix} \text{CAT} \begin{bmatrix}\text{HEAD noun}\\ \text{SUBAT} \langle\rangle\end{bmatrix} \\ \text{CONTENT } \boxed{i} \end{bmatrix}\end{bmatrix}$$

(58)



(Note that some feature names and values have been abbreviated for lack of space.)

Thus, information structure is correctly represented as a feature pertaining to the contextual characterisation of the meaning of the sentence. At the same time, it has access to the content of the entities making up the constituent structure. In other words, the constituent structure has a direct influence on the computation of the overall meaning of the sentence[10]. In the next chapters we shall discuss how this framework can be applied to the analysis of more complex examples, and deal with the implementation issues it raises.

4.4 Summary

In this chapter we have seen that the information structure of a sentence can be described along two different dimensions. The first dimension is that of *topicality*, which characterises the entities (topics) participating in the discourse, where different topic types may be distinguished on the basis of the way in which they relate to other discourse topics. The second one is that of *focality*, which deals with the fact that one or more elements of an utterance are emphasised to indicate some kind of modification of the pragmatic information built by the overall discourse.

An important notion pertaining to topicality was shown to be that of primary topic, sometimes simply called the topic, or topic proper. The primary topic was defined as 'the topic which the sentence is about'. For Danish, the primary topic was claimed often to coincide with the constituent in theme position. It was also shown that in sentences displaying topicalisation of a valency-bound constituent (with the exception of a weak pronoun), this constituent, which is considered the primary topic of the sentence, is given a contrastive interpretation. The notion of primary topic constitutes one of the two basic categories of information structure we shall make use of when discussing the relationship between information structure and syntax in the next chapter. The primary topic of a sentence is indicated in our examples by a subscript $_{PT}$ attached to the relevant constituent.

The second primitive category in our model is *focus*. We have seen that different types of focus can be described, depending on the pragmatic function they have in relation to some (implicit or explicit) discourse antecedent. However, it was pointed out that the information expressed by the focus of a sentence in some sense always modifies a piece of information introduced previously or presupposed

[10] An attempt similar to ours has been made recently by Engdahl & Vallduví (1996). Their system models the way in which the information structure of a sentence is determined by accent in English, and by word order in Catalan. It uses features similar but not entirely coincident with ours. Thus the feature they use to stack the value of the foci of a sentence, is structure-shared with the whole synsem feature of the focussed sign rather than with its content only. The kind of feature structure they propose is as follows (p.11):

$$\boxed{1}\begin{bmatrix}\text{PHON}\,|\,\text{ACCENT A}\\ \text{INFO-STRUCT}\,|\,\text{FOCUS }\boxed{1}\end{bmatrix}$$

As a result, the feature structure of the sign becomes cyclic, a property which not all computational linguists accept in a feature structure (see Pollard & Sag (1987), p. 37 for a discussion). However, the authors suggest that an alternative structure sharing between the information structure features and the content features could be used instead.

in the discourse. Thus, all the different types of focus were subsumed under the definition proposed by Rooth. On this view, a focus always provides the instantiation of a variable in the semantic interpretation of the sentence, in contraposition to a set of alternative elements. The focus of a sentence is indicated in our examples by a subscript $_F$ attached to the relevant constituent.

To the two primitives just mentioned, a third one has been added to mark the part of the sentence which is neither primary topic nor focus, namely *neutral information*. Neutral information specifies or limits the background against which the pragmatic modification expressed by the focus should be understood.

Finally, in the final section of this chapter, a proposal was made for representing the information structure of a sentence in a typed feature structure formalism like HPSG. According to this proposal, information structure is expressed as an independent feature in the phrasal sign, having values for the three primitives mentioned above.

Chapter 5

The syntactic realisation of information structure in Danish

In the preceding chapter, a framework was proposed in which the information structure of a sentence is conveniently represented. The primitives playing a role in our model – namely the primary topic, the focus, and the neutral information – have long been known and discussed by analysts working in the field, although as we have seen, a certain degree of confusion exists about the way in which these notions should be understood. We hope to have clarified what our understanding of the three terms in question is.

Our main concern, however, and the issue where we believe our work makes a contribution to the overall debate on information structure, is the way in which a model of information structure can be applied to the automatic analysis and generation of texts within an MT application. The main assumption we rely on is the fact that in written texts, the syntactic shape of the individual sentences provides us with important clues for the identification of the primary topic and the focus. Conversely, knowing what the primary topic and the focus of a sentence should be, can help a system in choosing between competing syntactic structures. Being application-oriented, our approach to the issue is a pragmatic one. This means that in many cases, our analysis will rely on the use of heuristic principles. However, we consider the theoretical dimension as important as the practical one, as a good application must build upon a sound theoretical basis. Thus, in addition to what our work may add to MT-related research results, we hope to have contributed to the more general linguistic debate with our discussion and analysis of focus marking constructions in the Danish language.

This chapter consists of two sections, the first one dealing with the notion of basic word order, and the second with focussing and thematisation processes. In Section 5.1, we claim that in Danish the primary topic and the focus have a default position: at the beginning and at the end of the sentence respectively, and we discuss how this has been dealt with in the theory of Field Grammar. The concluding part of the section consists of a proposal for the computation of the primary topic and the focus in unmarked, declarative Danish sentences. In Section 5.2, we deal with a number of focus or topic marking constructions, namely clefts, existential sentences, topicalisation and extraposition. For each construction type, we describe Danish data and contrast them with relevant English ones, we deal with previous treatments of the phenomena under consideration and propose our own implementation strategy, which takes into account syntactic, semantic and information structure issues.

5.1 Word order: theme and end-focus

One of the ways in which different communicative effects can be achieved is through word order. However, languages differ in the extent to which they allow word order to reflect the information structure of a sentence. A classic example we have already mentioned is the contrast between Slavic languages, where there is almost a direct correspondence between word order and information structure, and English, where word order is to a large extent grammatically determined, and information structure is rather expressed through sentence accent.

Also in English, however, information structure does play a role in the organisation of the constituents in the sentence. Two positions seem to constitute marked points in the English sentence (cf. e.g. Halliday 1967, Brown & Miller 1980), namely the initial position (*theme*) and the final position (*end-focus*). The theme tends in fact to be the position of the primary topic of the sentence, whilst the focus is likely to be placed in end-focus position. However, there is not, as we shall see, a complete correspondence between positions and pragmatic functions.

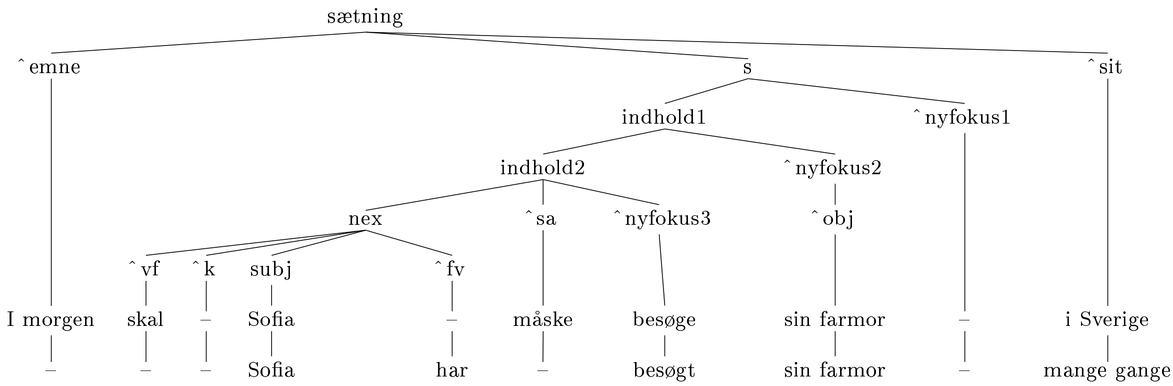
Danish resembles English in that word order is largely determined by the grammar, but also insofar as the constituents placed in the initial and final positions in the sentence seem to have special communicative functions. Thus, given information is often pronominalised and placed in initial position (thematised). The primary topic of a sentence has a tendency to coincide with the theme. In this sense, the theme constitutes the point of departure of the message expressed in the sentence, containing the constituent of which something is predicated in the rest of the sentence. However, the theme also connects the sentence to preceding text. Thus, it often contains relational adverbials. If the sentence is a question, the theme is either the wh-pronoun or a finite verb form[1]. Finally, the theme can also be filled by a contrastive topic.

Both in English and in Danish, in a declarative sentence the surface subject is the unmarked theme.

The end-focus, instead, is typically reserved either for new information, which tends to be spelt out and thus be quite long, or for contrastive information. This information also tends to be prosodically prominent and thus to constitute the focus of the sentence.

O. Togeby (cf. Togeby 1992) presents a phrase structure grammar for Danish where *emne* (topic) and *nyfokus* (new focus) are specific nodes in the phrase markers built by the grammar. *Emne* is always at the beginning of the sentence, and thus seems to correspond to the Hallidayan notion of theme. *Nyfokus* occupies one of three places at the end of the sentence. The figure below shows a slightly simplified version of Togeby's sentence schema, applied to the analysis of two Danish sentences:

(1) *Togeby's sentence schema*



(Lit: Tomorrow shall Sofia perhaps visit her grandother in Sweden.
Sofia has visited her grandmother many times.)

Togeby's model constitutes an interesting attempt at recasting Diderichsen's model of field grammar in terms of information structure. Without embarking on a discussion of the specific constituents proposed in this model, we shall mention, however, a number of problems that this strategy raises. Firstly, as noted by Ahrenberg (1990), to whom we refer the reader for a detailed discussion, merging the notion

[1] In Danish, yes/no questions are formed by inverting the finite part of the verb and the subject:

Har du været hos lægen?
(Have you been to the doctor?)

Gik det godt?
(lit: Went it well?)

.

of field and that of constituent in the way Togeby does is formally not a sound strategy. Fields and constituents have in fact different formal properties, and capture fundamentally different aspects of the language. Thus, Ahrenberg rightly notes that field structure should be seen as an independent level of linguistic description, dealing with the linearisation of independently defined constituents. Secondly, and perhaps more importantly as regards the present discussion, Togeby's modelling of the primary topic of the sentence does not seem consistent. He reserves in fact a specific node (*emne*) for the primary topic of the sentence, which he defines in the following way:

> "Den, som regel nævnte, information som frontstilles, er markeret som ytringens **emne**, dvs. det som ytringen tager udgangspunkt i og som meddelelsen handler om."
>
> (The information occurring in sentence-initial position, which is usually given, is marked as the topic of the utterance, i.e. the point of departure of the utterance and the item which the message is about.)
>
> (p.444)

However the grammatical subject, which in the default case is the primary topic of the sentence, is never dominated by the *emne* node. Thus, *emne* is reserved for relational adverbials (adverbial and prepositional phrases or subordinate clauses), or sentence arguments other than the subject, functioning as contrastive topics when placed sentence-initially. The fact that the subject occupies a different position from thematised complements and adjuncts, in our opinion undermines the transparency of the approach. In contrast to this, the Hallidayan notion of theme can be couched more easily in Diderichsen's original model. As shown by the table below, in fact, according to that theory the same field – *fundamentfeltet* (the foundation field) – is used both for the subject in declarative sentences not displaying topicalisation, and for topicalised constituents.

(2) *Diderichsen (1946)'s sentence schema*

Fundamentfelt	*Nexusfelt*			*Indholdsfelt*		
S/A	v	s	a	V	S	A
I morgen (lit: Tomorrow	skal shall	Sofia Sofia	måske perhaps	besøge visit	sin farmor her grandmother	i Sverige in Sweden)
Sofia (lit: Sofia	har has	– –	– –	besøgt visited	sin farmor her grandmother	mange gange many times)

In conclusion, we shall not adopt Togeby's approach, and rather assume that the subject of a non-topicalised declarative sentence and the topicalised constituent in a sentence where the initial constituent is not the subject, both occupy the same position in the sentence. We shall call this position the theme position, and the constituent occupying it the theme. In terms of field structure, the theme position corresponds to a field which may be termed the foundation field following Diderichsen's terminology.

As mentioned earlier, Togeby identifies three different focus positions in the Danish sentence. This proposal will not be followed here either. Instead, we shall speak of end-focus position as the position occupied by the rightmost constituent. We claim that, with the exception of certain adverbial types discussed further below, this is the default focus position in Danish.

While the definition of theme can be the same for Danish as for English, the constraints limiting thematisation of constituents other than the grammatical subject are slightly different in the two languages.

Firstly, English allows sequences of adverbials to fill in the theme slot, whereas only one constituent is allowed by the grammar of Danish. Secondly, as will be discussed in Chapter 5, Scandinavian languages violate the Complex NP Constraint by allowing extraction of a constituent out of a relative clause and its placement as the theme of the main clause. This is shown in the following sentence:

(3) Så meget spaghetti kender jeg ingen, der kan spise på en gang.
(lit: So much spaghetti know I nobody that can eat at once)

Furthermore, we saw in Chapter 2 that there is a strong tendency in Danish for indefinite NPs, typically expressing new discourse topics, not to occur in theme position. Hence, *der* and *det* insertion are often used to move such NPs away from the beginning of the sentence into end-focus position.

Some of the grammatical processes available in Danish to emphasise certain constituents either by placing them in theme or in end-focus position are discussed in more detail in the next sections.

But first, we want to conclude our discussion of the relation between word order and information structure in Danish by looking at a possible way in which the information structure of unmarked declarative sentences[2] can be computed by way of a set of heuristic principles.

5.2 Heuristics for the Computation of Information Structure

We shall take as our point of departure an algorithm presented in Hajičová et al. (1995) to compute the information structure in English technical language. The goal of the procedure, which consists of a number of heuristics, is for each sentence in a written text to determine which part constitutes the focus, and which part is not focussed. Recall that in Hajičová et al.'s framework, the non-focussed part of a sentence is actually called the *topic*, and is understood as the constituent or constituents expressing the given, or least dynamic information. Focus, in contrast, is equalled with the part of the sentence expressing the new, or most dynamic information. However, we do not use the term *topic* in this sense. Therefore, we shall simply speak of focus versus non-focus. In what follows we discuss the kind of criteria the algorithm is based on. For a detailed description of the whole procedure, the reader is referred to the original source.

A general assumption is that for languages like English or Danish, where word order is largely determined by grammar, and intonation plays an important role in determining the information structure of the sentence, the written form of a sentence is often not sufficient to ascertain in an unequivocal way which part of it is focussed and which is not. However, a number of syntactic and semantic criteria can be used to reduce the ambiguity. Furthermore, even in ambiguous cases a preferred reading can often be indicated.

The following criteria are considered by the authors (F+ means focus and F- non-focussed, F+/- means ambiguous with respect to the two categories)[3]:

- Position in the sentence and definiteness
  - If the verb is the last word of the sentence, and the subject is a definite NP, the verb is F+ and the subject F-. Conversely, if the verb is the last word of the sentence, but the subject is indefinite, the verb is F- and the subject F+.
  - Complementations[4] preceding the verb are F-, with the exception of an indefinite subject and a specific temporal or locative complementation.
  - Indefinite rightmost complementations to the right of the verb that are not temporal or locative expressions and are not assigned F- or F+ on the basis of SO (see below) are F+.
- Semantic content
  - If the verb is not final, it is F- if it has a very general meaning, F+ if it has a very specific meaning, F+/- otherwise.
  - A temporal or locative complementation to the right of the verb is F+ if it has a specific meaning, F- otherwise.

[2]By unmarked declarative sentences we mean sentences where neither the primary topic nor the focus are syntactically marked.

[3]In our formalism, F+ translates into [INFOTYPE focus], F- into [INFOTYPE $\neq$ focus], while F+/- corresponds to an undetermined INFOTYPE.

[4]The term 'Complementations' is used to refer to valency-bound complements as well as free adjuncts.

- Deviation from Systemic Ordering (SO)
  - If A is the left-hand complementation in an adjacent complementation pair which does not obey SO, then A is F-, and so are all the complementations placed between the verb and A. The righmost complementation of the whole sentence is F+, and any other complementation intervening between A and the righmost one is F+/-.
- Semantic similarity
  - The verb is F- if it is synonym to the verb in the previous sentence, or included in its meaning.
- Rhematisers
  - Complementations directly following rhematisers such as *also*, *even*, *only* are assigned F+.

In the final algorithm proposed by the authors, the last two criteria mentioned above have been left out. Note also that the verb always refers to the verb of the main clause. In fact, the algorithm does not apply to the internal information structure of subordinate clauses.

Clearly, none of the criteria – with the possible exception of the last one - should be seen as a strict constraint. They are rather preference criteria, interacting with each other and with the prosodic and grammatical component of sentence structure in a complex way. Thus, the first problem to be noted is that sentence accent can override all of these principles. However, the assumption made by the authors, which we shall subscribe to, is that in technical language this is unlikely, and prosody and word order tend to reinforce each other. Secondly, these principles may contradict one another. For example, a semantically weak verb, which should typically not be focussed according to the algorithm, can occur as the final constituent after a semantically definite NP, thus satisfying the first of the criteria above, and hence qualifying for focus status:

(4) - The bike's not here.
- Yes, it IS.

This problem is solved in the algorithm proposed by ordering the various criteria in a procedural manner. An alternative way of modelling their mutual interaction could be by assigning them weights, so that those expressing the strongest tendencies would be able to override the weaker ones.[5]

With these provisos, let us now discuss the extent to which the criteria listed above can be used in the analysis (or generation) of unmarked declarative sentences in Danish. The discussion will also point out a few general problems, and propose a few modifications to the algorithm not ony relevant to Danish.

5.2.1 Position and Definiteness

The first criterion based on position and definiteness is illustrated by the example below:

(5) Vælg "Kør" i menuen "Filer" i Programstyring eller Filhåndtering.
Dialogboksen "Kør" vises.

(Choose "Run" from the menu "Files" in Program Control or File Administration.
The dialogue box "Run" will appear – lit: is shown.)

(*Microsoft Windows 3.1 Brugerhåndbog*)

The focus of the second sentence falls on the verb, which is in end-focus position, and the definite subject is the primary topic.

Examples where the subject of a sentence-final verb is an indefinite NP, on the other hand, are not very common in Danish[6], as in this case the subject would be postponed by way of *der* insertion or by topicalising some other constituent:

[5] For a discussion of the use of preference in NLP, cf. Bennett & Paggio (1993).

[6] However, if it has generic reference, an indefinite NP can function as the subject and the primary topic of the sentence:

(6) Sådan åbnes en menu.
(lit: in this way is opened a menu)

(*Microsoft Windows 3.1 Brugerhåndbog*)

The criterion relating to complements and adjuncts placed to the right of the verb, on the other hand, is more problematic. It says that an indefinite rightmost complement is generally in focus (the two restrictions mentioned are discussed in the next section). In his account of information structure in Danish, Togeby mentions a similar criterion. He claims in fact that a definite complement in sentence-final position cannot be focussed, and that in this case the focus falls on a complement preceding it.

However, sentences where a definite NP is in final position and yet is the focus of a sentence are possible. This can be demonstrated by constructing a question answer sequence where such an NP provides the new information requested, and is thus clearly in focus, as shown below:

(7) - Hvilken kjole vil du tage på i dag?
- Jeg vil have den RØDE.

(Which dress do you want to wear today?
I want the red one.)

Using negation as a test also yields the same result, namely that the definite direct object is the focus:

(8) Jeg vil ikke have den røde.
Jeg foretrækker den med blomstrer på.

(I don't want the red one.
I prefer the one with flowers.)

Therefore, it seems difficult to maintain that complements in end-focus position can only be part of the focus when they are indefinite. Indeed, we saw already in our discussion of new topics in Section 4.1.1, that such examples are perfectly possible also in English. Thus, this is a point where the algorithm calls for a modification not only relevant to Danish.

5.2.2 Semantic Classifications

The second and third criteria in the first group mention a semantic classification of temporal and locative expressions which is claimed to constrain their possibility of functioning as the focus of the sentence. Thus, a distinction is made between adverbials establishing a general setting (e.g. *always*) or being directly determined by the sentence itself (e.g. *today, this year*), which are typically not in focus, and those expressing more specific information (e.g. *on a nice September day, on October 22nd*), which are typically part of the focus. Thus, in (9) below, the default interpretation would be that the temporal expression is included in the focus in (a), but not in (b).

(9) a. Jeg regner med at blive færdig med min PhD den 15. december 1996.
b. Jeg regner med at blive færdig med min PhD i år.

(I'm counting on finishing my PhD on 15th December 1996.
I'm counting on finishing my PhD this year.)

This is confirmed by the fact that it seems natural to topicalise the more general adverbial, thus clearly dissociating it from the focus of the sentence. Moving the specific adverbial in a similar way, on the other hand, creates an odd effect, at least out of context:

En Canon bubble-jet printer kan udskrive på almindeligt papir.
(A Canon bubble-jet printer can print on ordinary paper.)

(10) a. Den 15. december 1996 regner jeg med at blive færdig med min PhD.
b. I år regner jeg med at blive færdig med min PhD.

(On 15th December 1996 I'm counting on finishing my PhD.
This year I'm counting on finishing my PhD.)

Clearly, appropriate contexts can be constructed where the general adverbial is focussed, or where the more specific one is already presupposed in the discourse and can thus be used as a contrastive topic in theme position. However, when placed in end-focus position a specific adverbial such as *den 15. december 1996* has to bear primary stress, whereas there is a preference to read a general adverbial such as *i år* with secondary or reduced stress[7].

The distinction between general and specific temporal and locative adjuncts, therefore, also seems useful for Danish, although it may prove difficult in concrete cases to decide whether the meaning of an adverbial is general or specific.

As regards the distinction between verbs having general vs. specific meaning, this has to be studied in more detail, so it won't be considered here. Nor shall we discuss the principle of semantic similarity between the current verb and the verb of the preceding sentence.

Instead, we shall adopt Rochemont's rules for the spreading of focus from arguments to heads (see Section 3.3.2) in a slightly modified version. Thus, if an argument to the right of the verb is focussed, and the verb has not been assigned F- on the basis of Systemic Ordering, the verb will be F+. Note that an argument to the right of the verb in Danish can also be the subject in cases of subject-verb inversion. On the other hand, if the verb is followed by a focussed adjunct, it will be assigned F-[8].

The criterion regarding focussing adverbs (rhematisers in Hajičová et al.'s terminology) should be reasonably self-explanatory. The notion of Systemic Ordering, on the other hand, deserves some explanation.

5.2.3 Systemic Ordering

Systemic Ordering (SO) is according to the authors a basic ordering of complementations – valency-bound complements as well as adjuncts – corresponding to unmarked information structure. Thus, the two sentences below show that in English, Indirect Object precedes Object in SO:

(11) a. She gave several children a few apples.
b. She gave a few apples to several children.

(Hajičová et al. 1995:82)

In (11a), in fact, the information structure is unmarked, so that at least in one reading, the focus includes the verb and both its complements. Thus, the sentence could provide a well-formed answer to the question below:

(12) What did she do?

[7] It could be argued that it is more natural to include the more specific adverbial in the focus because its very specificity makes it clearer what the alternatives may be, namely in the case under consideration other dates, e.g. November 3rd, or Easter next year. In the case of the more general adverbial, on the other hand, the fact that the setting is so general makes the implicit terms of contrast more difficult to determine and the adverbial tends not to be focal. We are grateful to Bente Maegaard and Ebbe Spang-Hanssen for their comments on this point.

[8] Note that the algorithm we propose doesn't deal with a situation where some of the verbal arguments are focussed and others are not without a disruption of SO, as in the following example:

Who did John give a flower to?
He gave NANCY a flower.

However, we cannot detect such a pattern when treating written language. Therefore, unless the default order of the arguments is deviated from, we have proposed that all the arguments to the right of the verb are considered focussed.

In (11b), on the other hand, the Indirect Object occurs after the Object, and this disruption of SO indicates that the focus is restricted to the complement in end-focus position. Therefore, (11b) would not constitute an acceptable answer to the same question.

The authors suggest the following SO for English (p. 84):

(13) *SO for English, preliminary version*

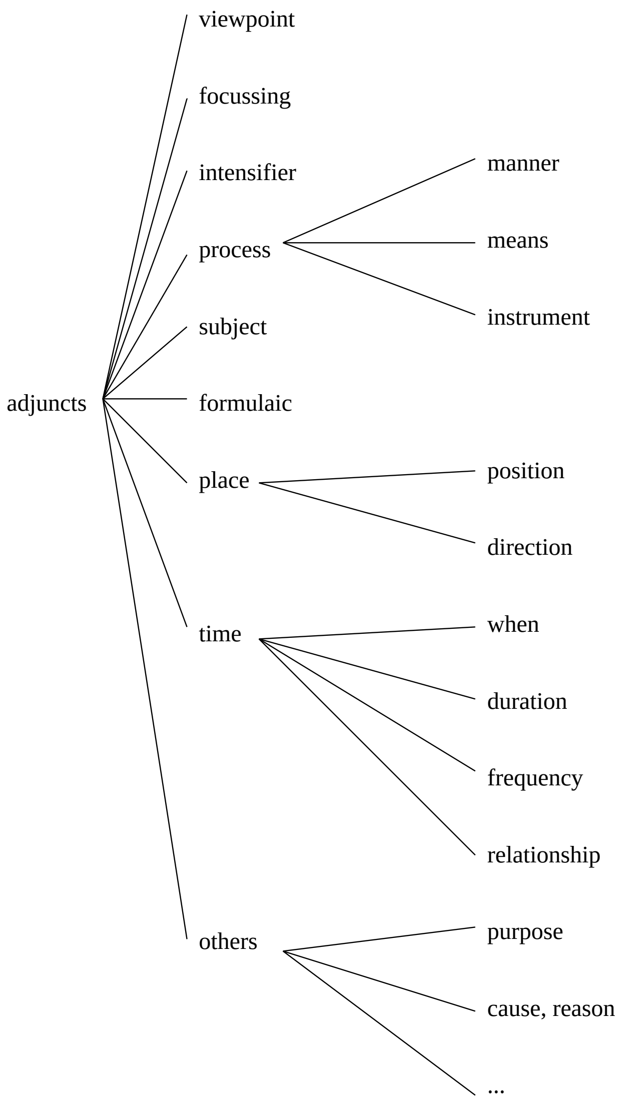
Time - Actor - Addressee - Objective – Origin – Effect – Manner – Directional *from* – Means – Directional *to* – Locative

In German and most Slavic languages, on the other hand, Objective and Effect would follow several of the adverbial adjuncts.

A few comments are necessary to clarify some of the labels used in the list above. Crucially, the roles identified by Hajičová *et al.* reflect surface grammatical functions rather closely. Thus, Actor always corresponds to the surface subject; Objective – sometimes also called Patient – to the direct object; Addressee to the indirect object; Effect to subject or object complement. We shall use the more traditional terms rather than those proposed by Hajičová et al.

What the precise inventory of complements and adjuncts in the linguistic description should be is as yet an unsettled matter. The adjunct types mentioned above, however, do not seem sufficient. A more exhaustive list is provided by Quirk et al. (1972:429), who distinguish the following types:

(14) *Adjunct types*



Quirk et al.'s view of the relative order of adjuncts in English differs slightly from Hajičová et al.'s. They claim in fact that the following principles apply to chains of adjuncts:

- An adverbial clause normally follows an adverb or prepositional phrase.
- A long adjunct normally follows a short one.
- Where adjuncts cluster in end-focus position, the default order is 'process – place – time'.
- Normally, only one adjunct occurs in theme position, but place and time adjuncts, or viewpoint and place adjuncts, may co-occur.

Adjunct length, or the mutual order between clausal and non-clausal adjuncts, can hardly be expressed in a model of SO.

Incorporating the other principles into the model proposed by Hajičová et al, however, yields the following revised SO:

(15) *SO for English, revised version*

a. Place – Time – Subject – Indirect Object – Object – Origin – Subject Complement / Object Complement – Manner / Means / Instrument – Directional *from* – Directional *to* – Locative – Time

b. Viewpoint – Place – Subject – Indirect Object – Object – Origin – Subject Complement / Object Complement – Manner / Means / Instrument – Directional *from* – Directional *to* – Locative – Time

Thus, the original SO has been split into two lists to allow for the alternation of different sequences of adjuncts before the Subject, and Means has been made to alternate with Manner and Intrument before Directional *from*. Place can be considered a superclass of Directional *from*, Directional *to* and Locative. Viewpoint adjuncts are adjuncts expressing the point of view from which the sentence is considered, e.g. *as far economics is concerned, morally speaking*.

It can be objected that this SO does not cover all the possible adjunct types identified by Quirk et al. In particular, it does not deal with adjuncts occurring in medial position, i.e. between the subject and the verb. Therefore, it does not provide a complete model of adverbial complementation in English. Furthermore, it has not undergone any testing on our part. Thus, it is only proposed as a hypothesis concerning the default mutual order of the major complementation types.

Assuming the SO above as a tentative model for English, we can now discuss whether it also holds for Danish.

For some complement pairs, a basic mutual order can be established by looking at sentence pairs, and testing their appropriateness in question answer sequences. In all the examples below, (a) displays unmarked information structure, and is thus a well-formed answer to the question, wheras in (b) the focus is limited to the complement in end-focus position and the sentence is not a natural answer to the question. The symbol > stands for "precedes in SO".

(16) Hvad gjorde manden?
a. Han gav hende æblet.
b. ? Han gav æblet til hende[9].

(What did the man do?
He gave her the apple.
He gave the apple to her.)

Indirect Object > Object

(17) Hvad leger Sofia med?
a. Hun laver spaghetti af modellervoks.
b. ? Af modellervoks laver hun spaghetti.

(What is Sofia playing with?
She is making spaghetti of play dough.
lit: Of play dough is she making spaghetti.)

Object > Origin

[9]For (b) to be an appropriate answer to the question, a context must be constructed where *til hende* receives contrastive stress, and is thus clearly the focus:

Drengen ville ikke have æblet.
Hvad gjorde manden så?
Han gav æblet til hende.

(The boy didn't want the apple.
What did the man do then?
He gave the apple to her.)

(18) Hvad skete der så?
a. Peter Larsen fra Roskilde blev valgt til formand.
b. ? Til formand blev valgt Peter Larsen fra Roskilde.

(What happened then?
Peter was elected chairman.
lit: As chairman was elected Peter.)

Subject > Subject Complement

(19) Hvor meget løber du?
a. Jeg løber hver dag fra arbejde til hvor jeg bor.
b. ? Jeg løber hver dag til hvor jeg bor fra arbejde.

(How much do you run?
I run every day from work to where I live.
I run every day to where I live from work.)

Directional-*from* > Directional-*to*

(20) Hvad skal I i jeres ferie?
a. Vi skal til Italien med toget.
b. ? Vi skal med toget til Italien.

(What are you doing for your holidays?
We are going to Italy by train.
We are going by train to Italy.)

Directional-*to* > Means

(21) Hvad bliver man syg af?
a. Man bliver syg af margarine.
b. ? Af margarine bliver man syg.

(What do you get ill from?
You get ill from margarine.
From margarine you get ill.)

Subject Complement > Origin

For other complement types, it is not clear whether SO is fixed. For instance in the example below, Time seems equally unmarked in initial as well as in final position:

(22) Hvad gjorde I i går?
a. I går gik vi i biografen.
b. Vi gik i biografen i går.

(What did you do yesterday?
lit: Yesterday went we to the cinema.
We went to the cinema yesterday.)

An interesting category is constituted by Manner adverbials, which Togeby claims are always focussed. He reserves a specific place for them in his sentence schema (cf. fig. 1), namely under the node labelled “nyfokus1”. To the right of “nyfokus1”, the schema foresees another place, reserved for attitudinal disjuncts, e.g. *måske, desværre* (perhaps, unfortunately), as well as to what the author calls “relationsadverbialer”, a class that seems to include several of the adjunct types described by Quirk et al. (e.g. time and cause). Togeby claims that such adverbials are not focussed in the normal case. Thus according to him, in the following example the accent falls on the verb, and the final adverbial is not part of the focus:

(23) Hun ARBEJDEDE i fredags.
(She worked on Friday.)

An additional indication of the fact that the adverbial is not part of the focus is that if the sentence is negated, the adverbial is left outside the scope of the negation operator. Thus, in its primary reading, the sentence below means "On Friday, she didn't work", and not "it was not on Friday that she worked".

(24) Hun arbejdede ikke i fredags.
(lit: She worked not on Friday.)

Togeby's observation confirms a claim put forward in Section 5.2.2, that certain adverbials are not focussed even if placed in end-focus position. We claimed, however, that a useful distinction can be made between temporal and locative adverbials describing the general setting for the event expressed by the sentence, and those expressing a specific temporal or locative dimension. Adverbials of the former type tend not to be in focus, while those belonging to the latter usually are.

On the basis of these observations, we suggest the following very tentative SO for Danish:

(25) *SO for Danish*

Time / Viewpoint / Attitude – Subject – Indirect Object – Object – Subject Complement / Object Complement – Origin – Directional-*from* – Directional-*to* – Locative – Manner / Means / Instrument – Time

As can be seen, Time is represented twice to account for the fact that it seems equally unmarked to have a Time expression in sentence-initial or sentence-final position (although we have noted that a specific time expression will occur in sentence-final position).

Again, it must be noted that the SO proposed here only includes a subset of the adverbials possible in Danish. However, it provides a concrete starting point to be included in our algorithm for the automatic computation of the focus of a written sentence.

5.2.4 Heuristics for Danish

In the light of the preceding discussion, we now propose a set of heuristics for the computation of information structure in Danish unmarked declarative sentences. Compared to the original heuristics proposed by Hajičová et al. for English, they have undergone a number of modifications. For the sake of shortness, we use the abbrevations F+, F-, and PT+. Note that as the final result of the application of the algorithm, some constituents will be marked as F- and PT+: these are primary topics. Others will only be marked as F-: these belong to what we have termed neutral information. Others will be marked F+: these are foci. Note also that the term 'complementation' has been replaced by 'argument' or 'adjunct' depending on the context.

1. Position in the sentence and definiteness
   (a) If the verb is the last word of the sentence, the verb is F+ and the subject F-.
   (b) Arguments and adjuncts preceding the verb are F-.
   (c) Arguments and adjuncts (excluding personal pronouns) to the right of the verb that are not general temporal or locative expressions and are not assigned F- or F+ on the basis of SO (see below) are F+.
2. Semantic content
   - A temporal or locative adjunct to the right of the verb is F+ if it has a specific meaning, F- otherwise.

3. Deviation from Systemic Ordering (SO)

    - If A is the left-hand argument in an adjacent argument pair which does not obey SO, then A is F-, and so are all the arguments placed between the verb and A. The argument following A is F+. The verb is assigned F-.

4. Spreading of focus to verbs

    - If no other criterion applies (see 1a and 3 above), and an argument to the right of the verb is focussed, the verb is assigned F+, otherwise it is assigned F-.

5. Rhematisers

    - Constituents directly following rhematisers such as *også* (also), *selv* (even), *kun* (only) are assigned F+.

6. Word order wrt. adverbials

    - Personal pronouns preceding the so-called central adverbials are always F-.

7. Theme criterion

    - The constituent in theme position is assigned PT+.

Compared to the original algorithm proposed for English by Hajičová et al. (1995), the heuristics listed here have been revised in a number of ways. All the modifications are listed below, grouped into general modifications and Danish-specific ones.

<u>General modifications:</u>

- The positional criteria have been changed so that arguments and adjuncts following the verb are considered in focus whether definite or indefinite.
- A rule for focus spreading from verbal arguments to the verb has been added. On the other hand, focus does not spread from adjuncts to the verb.
- The criterion on SO has been changed slightly. In our version, the verb is considered F-. This choice is motivated by the fact that a sentence where SO is disrupted, can only be a well-formed answer to a question where the verb is explicitly mentioned and thus not part of the focus in the answer, e.g.:

    (26) a. Hvem blev valgt i går?
    Til formand blev valgt Peter Larsen fra Roskilde.
    b. Hvad skete der i går?
    ? Til formand blev valgt Peter Larsen fra Roskilde.
    (Who was elected yesterday?
    lit: As chairman was elected Peter.
    What happened yesterday?
    lit: As chairman was elected Peter.)

- Adverbials preceding the verb are never considered focal, no matter what their semantic content.
- The primary topic is now computed. As proposed in Section 4.1.2, the primary topic corresponds to Halliday's theme.

<u>Danish-specific modifications:</u>

- A criterion has been added concerning personal pronouns preceding central adverbials (e.g. *også* (also), *ikke* (not), *aldrig* (never)), which as shown below are never focussed:

(27) a. Jeg kender ham ikke.
b. * Jeg kender HAM ikke.
c. Jeg kender ikke HAM, men HENDE.
(lit: I know him not.
I know HIM not.
I know not HIM, but HER.)

- There is no specific criterion for indefinite subjects preceding the verb, as we claim that an indefinite subject will typically not occur in theme position in Danish. In cases of generic reference, where a syntactically indefinite NP can occur in subject position, as in *En Canon bubble-jet printer kan udskrive på almindeligt papir* (A Canon bubble-jet printer can print on ordinary paper), the subject is considered the primary topic.

Let us now see what results our heuristics would produce on the analysis of the short Danish text below:

(28) Der kan køres forskellige programtyper under Windows-operativsystemet. Nogle er specielt lavet til at køre under Windows, mens andre er lavet til at køre under MS-DOS, men uden Windows. I dette afsnit beskrives følgende fire programtyper.

(lit: There can be run various program types under the Windows operating system. Some are especially made to run under Windows, while others are made to run under MS-DOS, but without Windows. In this paragraph are described the following four program types.)

(*Microsoft Windows 3.1 Brugerhåndbog*)

In the first sentence, it can be discussed whether the final adjunct should be considered general or specific. Given the fact that the text is about Microsoft Windows, we shall not consider it a specific adjunct, and claim that it establishes the setting of the sentence rather than contributing any specific new information. It is assigned F- according to (2). The postposed subject is assigned F+ (1c), the verb is F+ (4), and the initial empty pronoun is F- (1b).

Der$_{F-}$ kan køres$_{F+}$ [forskellige programtyper]$_{F+}$ [under Windows-operativsystemet]$_{F-}$.

The second sentence consists of two coordinated main clauses, each with its own focus. In both of them, the Purpose adjunct in end-focus position is assigned F+ (1c), the verb is F- (4), and the subject is F- (1b). Furthermore, in both sentences the grammatical subject in theme position is assigned PT+ (7).

Nogle$_{F-,PT+}$ er specielt lavet$_{F-}$ [til at køre under Windows]$_{F+}$, mens andre$_{F-,PT+}$ er lavet$_{F-}$ [til at køre under MS-DOS, men uden Windows]$_{F+}$.

Finally in the third sentence, the postponed subject is F+ (1c), the verb is also F+ (4), and the adverb in theme position is F- (1b) and PT+ (7).

[I dette afsnit]$_{F-,PT+}$ beskrives$_{F+}$ [følgende fire programtyper]$_{F+}$.

Disruption of SO (3) is shown in the following example, taken from a different text:

(29) Den store Flygtning 86 kampagne gav 51,4 millioner kroner til Dansk Flygtningehjælp.

(The big Refugee 86 campaign gave 51.4 million crowns to Dansk Flygtningehjælp.)

(*DK87*)

The indirect object (*til Dansk Flygtningehjælp*) follows the object (*51,4 millioner kroner*), disrupting the default order. Following the SO criterion in the algorithm (3), the object and the verb are both assigned F-, while the Indirect Object is assigned F+. Furthermore, the subject is assigned F- (1b) and PT+ (7). Thus, the analysis of the clause will be as follows:

[Den store Flygtning 86 kampagne]$_{F-,PT+}$ [gav]$_{F-}$ [51,4 millioner kroner]$_{F-}$ [til Dansk Flygtningehjælp]$_{F+}$.

To see how the same heuristics might be used to generate the correct information structure from an abstract representation of the content of the same text, let us assume that such a representation of the sentences in (28) and (29) would contain at least the following information[10]:

can$_{F+}$ (run$_{F+}$ (–, (several program types)$_{F+}$, (under Windows OS))$_{F-}$))

while(make$_{F-}$ (–, (some program types)$_{F-,PT+}$, (for running...)$_{F+}$)), (make$_{F-}$ (–, (other program types)$_{F-,PT+}$, (for running...)$_{F+}$))

describe$_{F+}$ (–, (following four program types)$_{F+}$, (in this paragraph)$_{F-,PT+}$)

give$_{F+}$ ((big refugee campaign)$_{F-,PT+}$, (51.4 million crowns)$_{F-}$, (to Dansk Flygtningehjælp)$_{F+}$)

To generate a sentence from the first representation, the grammar must take into account the fact that the deep subject is not expressed (indicated by '–' in the representation), and thus generate a passive form. Since the deep object must be focussed and placed to the right of the verb (1c), two options are available: either to generate a *der* construction, or an adverbial in theme position. Since the adverbial is marked F-, it can be thematised (1b). However, assuming that it is also marked as being a general locative, it can also be generated in end-focus position (2). Thus, an automatic generation that produced both the following sentences would do a good job (the generation of *der* constructions is treated in more detail in Section 5.3.2):

(30) a. Der kan køres forskellige programtyper under Windows-operativsystemet.
b. Under Windows-operativsystemet kan køres forskellige programtyper.

(lit: There can be run various program types under Windows OS.
Under Windows OS can be run various program types.)

In the second conjoined proposition, in both sentence representations the deep subject is again not mentioned, but the deep object is F- and PT+ and can therefore function as the grammatical subject, placed sentence-initially in its default position (1b). The adverbial, on the other hand, is F+ and must therefore be placed in end-focus position (1c). Thus, the original sentence will be generated.

The third proposition shows the same pattern as the first one. In fact, the first argument is empty, the second is F+, and the adverbial is F-. In this case, however, the adverbial does not express a general temporal or locative setting, and cannot therefore occur in end-focus position without being focussed (1c). Thus, the heuristics rightly give preference to (31a) rather (31b):

(31) a. I dette afsnit beskrives følgende fire programtyper.
b. ? Der beskrives følgende fire programtyper i dette afsnit.

(In this paragraph are described the following four program types.
lit: There are described the following four program types in this paragraph.)

In the last proposition, the default grammatical order of complements would be 'Subject, Indirect Object, Object'. However, this would generate the sequence F-, F+, F-. A good generation algorithm should reverse the places of the indirect and direct object, so that the focussed complement is shifted to the right of the complement that should normally follow it, according to SO (3). This would yield the appropriate sentence, namely:

(32) Den store Flygtning 86 kampagne gav 51,4 millioner kroner til Dansk Flygtningehjælp.

[10] The representation given here is of course greatly simplified. It ignores a number of important features such as tense or determination, and shows no labelling of complements and adjuncts.

Thus, the heuristics proposed seem to provide a useful strategy for the analysis and generation of information structure in Danish. In the next section, we discuss how these principles can be complemented by syntactic constraints applying to constructions where the focus or a contrastive topic are marked in a structural way.

5.3 Focussing and topicalising constructions

The previous discussion of the interaction of information structure and word order does not exhaust the issue of how the information structure of a sentence is reflected in its syntactic form. In this section, we look at a number of syntactic constructions that have been dealt with in the literature as expressing particular patterns of information structure. The central issue we shall be concerned with is determining to what extent the syntactic shape of the sentence alone allows us to predict where the primary topic or the focus are or are not to be found. The following syntactic constructions are discussed:

Clefts:	*Det var hende, der slog ham.*
	(It was her who hit him.)
Existential sentences:	*Der løb en kat igennem døren.*
	(lit: There ran a cat through the door.)
Topicalisation:	*Ida har vi glemt at invitere.*
	(Ida we have forgotten to invite.)
Extraposition:	*Det er rart at se dig.*
	(It's nice to see you.)

For each of the constructions treated, we discuss syntactic aspects pertaining to the Danish data, we propose a representation which tries to capture the relevant syntactic and semantic properties, and discuss how such a representation can accommodate the correct information structure. Finally, we discuss the extent to which the proposed representation provides an adequate starting point for the generation of coherent text.

These four construction types provide quite a broad spectrum for our discussion. They allow us in fact to deal with such different issues as the relationship between focus and presupposition in clefts; the quantifier effect as embodied in existential sentences; the heaviness effect and its interaction with focality in sentences exhibiting extraposition of clausal constituents; and finally the contrastive nature of topicalisation.

However, they may not exhaust the list of syntactic phenomena that are relevant to an account of the relation between syntax and information structure: pronominal anaphora for example is another pertinent phenomenon. It is not treated in this thesis, however, because our focus here is on intrasentential phenomena. Conversely, it may be objected that the constructions taken into consideration have rather different distributions in real texts. Clefts, for example, as well as some topicalisation types, belong to the realm of spoken language, and are thus perhaps not very interesting from the perspective of machine translation. Existential sentences and extraposition, and other topicalisation types, however, are also very frequent in written texts. Although we recognise this disparity, we feel that it is important to show that the framework proposed can be applied to a range of different constructions irrespective of their frequency of distribution.

A final general observation relates to our use of the terms *focussing constructions* and *topicalising constructions*, by which we refer to the fact that in certain constructions, either the focus or the primary topic **tend** to be placed on a specific constituent predictable from the syntactic structure. In certain cases (e.g. wh-clefts) the assignment of focal stress to a specific constituent will in fact be totally fixed, in others (e.g. subject extraposition), we can talk about a specific preference rather than complete predictability. In the case of extraposition in particular, the fact that the focus is not totally predictable is in fact rather surprising, as the syntactic structure at first blush would seem to mark it in a clear way. Thus, our approach is different e.g. from Rochemont's (1989), where focus marking is strictly defined in terms of phrase structure.

5.3.1 Clefts

While it is generally acknowledged that clefts syntactically mark a certain constituent in a special way, it is not clear that this marking and focus as described in the preceding chapter, are one and the same thing. In this section, an approach to Danish clefts as focus marking constructions is presented, and the relationship between the presuppositional nature of clefts and the information structure that characterises them is discussed.

Danish clefts fall into three different types. Following the terminology used by Delin (1989) for English, we shall call them it-clefts, wh-clefts and reverse wh-clefts, respectively. These are exemplified below:

(33) Det var hende, der slog ham.
(It was her who hit him.)

(34) Hvad jeg ikke forstår er, at du allerede går.
(What I don't understand is that you are already leaving.)
(Nølke 1984:85)

(35) Ja, det er lige netop hvad jeg er.
(Yes, that is precisely what I am.)
(Hansen, A. 1967: III 459)

For all three types we shall call the complement of the copula the *head* of the cleft. The embedded clause will be termed the *c-clause* (*cleft clause*).

In Danish, the c-clause in a wh-cleft or a reverse wh-cleft need not be introduced by a wh-pronoun:

(36) Sofia er den, der snakker mest.
(Sofia is the one that talks most.)

(37) Det regner altid når man skal på ferie! Ja det er det det gør.
(lit: It always rains when one must [go] on a holiday. Yes, it is that it does.)
(Hansen, A. 1967: III 460)

Thus, the term wh-cleft may seem ill-suited, at least for Danish. However, the more traditional term *pseudo-cleft* seems even more misleading to us, suggesting that such clefts have a somewhat derived status compared to it-clefts. Therefore, we have adopted Delin's terminology in the absence of a valid alternative.

The main characteristic of clefts is that, although syntactically they consist of two separate clauses, a copular clause plus an embedded one, they can be described as being somehow "equivalent" to a single clause. Thus, example (33) above is in some sense equivalent to the non-cleft:

(38) Hun slog ham.
(She hit him.)

More precisely, the two sentences are truth-conditionally equivalent.

The same relation of truth-conditional equivalence holds between different cleft types. Thus, the non-cleft sentence in (38), the it-cleft in (33), and the wh-clefts below are all clearly related:

(39) a. Den der slog ham, var hende.
(The one that hit him was her.)

b. Hun var den, der slog ham.
(She was the one that hit him.)

Therefore, in transformational approaches, various theories have been put forward to generate cleft sentences from a non-cleft base structure, or from each other (cf. Nølke (1984) for a discussion).

The fact that the meaning of a cleft can be directly related to the meaning of a non-cleft sentence, distinguishes it clearly from similar syntactic constructions where this relation does not hold. For example in the sequence below, the response is syntactically very similar to the cleft in (33). However, there are two important differences: the pronoun *det* is not a pleonastic form, as it refers back to *Jonathan* in the question, and the embedded clause is an ordinary restrictive relative clause modifying *dreng* (boy):

(40) - Hvem er Jonathan?
- Det er den dreng der går i Sofias børnehave.
(Who is Jonathan?
lit: It's that boy that goes to Sofia's kindergarten.)

Clearly, the meaning of the response in (40) is in no way equivalent to the meaning of the sentence below:

(41) Den dreng går i Sofias børnehave.
(That boy goes to Sofia's kindergaarden.)

There are other interesting differences between clefts and sentences containing a modifying relative clause. Firstly, the choice of tense is more constrained in clefts than in non-clefts:

(42) a. Det er Peter, der har gjort det.
(It's Peter that has done it.)

b. * Det er Peter, der gjorde det.
(It's Peter that did it.)

K. Togeby (cf. Togeby 1982) notes that in French, the same mood must also be used in the head and c-clause of a cleft. Thus in the following example, the subjunctive is used in both:

(43) Je préfère que ce soit son chef de cabinet qui le lui apprenne.
(I prefer that it's his head of department that tells him that.)

(Togeby, K. 1982 II:252)

In Danish, the choice of the pronoun in the c-clause is more limited than in relative clauses. In particular, if the head of the cleft does not correspond to the subject of the c-clause, there is a tendency not to use any pronoun at all in the c-clause. Therefore, in the following pair of sentences, there is a strong tendency not to interpret the second one as a cleft, since a relative pronoun with object function is used in the relative clause:

(44) a. Det er blommer, jeg har spist.
(It's plums I've eaten.)

b. Det er blommer, som man kan spise.
(It's plums that one can eat.)

Another specific syntactic feature of clefts is that the c-clause of a cleft can be preceded by a relative clause modifying the same head noun, whereas a sequence of two relative clauses modifying the same head noun is impossible:

(45) a. Det er de drenge, der har hugget æblerne, som jeg er gal på.
(It's the boys that have stolen the apples that I'm angry with.)

b. * Jeg kender ikke de drenge, der har hugget æblerne, som jeg er gal på.
I don't know the boys that have stolen the apples that I'm angry with.)

In relation to the present discussion, the tense and mood equality constraints are particularly interesting, as they seem to confirm the assumption that a cleft is equivalent to a non-cleft in the sense just made clear. Thus, only one realisation of tense and mood is possible in the cleft, corresponding to the tense and mood of its non-cleft counterpart[11].

Now, having established that the meaning of clefts is truth-conditionally equivalent to the meaning of their non-cleft counterparts, let us focus on what differentiates the two construction types. The distinguishing feature is that clefts are presupposition triggering constructions. More precisely, the c-clause introduces a variable-containing presupposition, which is instantiated by the head of the cleft. Thus, example (33) above can be paraphrased as:

Somebody hit him, and this somebody was her.

The fact that the proposition expressed by the c-clause is a presupposition and not just an entailment can be shown by embedding the cleft in a number of presupposition preserving contexts, where ordinary entailments do not survive:

(46) a. Det var ikke hende, der slog ham.
(It was not her that hit him.)

b. Det var muligvis hende, der slog ham.
(It was possibly her that hit him.)

c. Hvis det var hende, der slog ham, kan jeg forstå, at han blev vred.
(If it was her that hit him, I understand that he got angry.)

In all the three examples, it is still true that "somebody hit him". If we try the same test with the non-cleft counterpart, on the other hand, we see that although the non-cleft entails that "somebody hit him", the entailment does not survive in context:

(47) a. Hun slog ham ikke.
(She didn't hit him.)

b. Muligvis slog hun ham.
(Possibly, she hit him.)

c. Hvis hun slog ham, kan jeg forstå, at han blev vred.
(If she hit him, I understand that he got angry.)

An additional constraint on the content of clefts seems to be that the instantiation of the variable in the head of the cleft must be maximal, i.e. there must not be other referents in the discourse which can instantiate the same variable. Thus, the following sequence is infelicitous:

(48) Det er min mand, der henter vores datter i børnehaven.
? Jeg henter hende også.

(It's my husband that picks up our daughter from the kindergarten.
? I pick her up, too.)

The example below seems to constitute a counterexample. However, the presence of the conditional *hvis* (if) provides a particular context in which the constraint can be overridden:

(49) Hvis det er Peter der kommer, kommer Ulla også.
(If it's Peter that comes, Ulla is coming, too.)

(Togeby, O. 1992:465)

In the next section, we shall deal with the information structure of clefts, and discuss its relation with cleft presupposition.

[11]We are grateful to Ebbe Spang-Hanssen for many observations and examples on the difference between clefts and relative clauses.

Clefts, focus and presupposition

Chomsky (1971) points out that clefts contain an intonation centre which corresponds to the focus of the cleft sentence. However, he also points out that although the head of a cleft bears main stress under normal intonation, the intonation centre may well be shifted to a constituent within the c-clause to achieve specific expressive or contrastive effects. Thus, he claims for example that all the following intonation patterns are possible:

(50) a. Is it JOHN who writes poetry?
b. Is it John who writes POETRY?
c. Is it John who WRITES poetry?
(Chomsky 1971:199)

More recently, Quirk, Greenbaum, Leech & Svartvik (1995) have analysed English clefts as bearing two foci, one on the head and one at the end of the c-clause. Only one of them is claimed to express information which is new in context, and is therefore termed the dominant focus. In the case of the it-cleft, it is assumed that the dominant focus will fall either on the head or on the c-clause depending on the context, while in wh-clefts it is claimed always to occur within the head.

This approach seems to provide a better account for real data, which, as demonstrated by Delin, show a variety of intonation patterns. On the basis of an extensive corpus of naturally occurring examples, she claims that the following intonation patterns are typical of the three types of cleft in English:

Cleft type	*Intonational pattern*
it-clefts	double focus, or focus on the c-clause
wh-clefts	double focus
reverse wh-clefts	focus on the c-clause

Looking at the information structure of clefts, Delin observes that a variety of patterns are represented. She distinguishes among three categories of information, namely new information, old information and information which is inferrable from previous context in a number of specified ways (cf. Prince 1981). According to this three-fold distinction, she claims that the following patterns occur in her corpus:

Cleft type	*Head*	*Presuppositional Relation*
it-clefts	a. old	new/inferrable
	b. new (contrastive)	old/inferrable
wh-clefts	new	old/inferrable
reverse wh-clefts	old	new/inferrable (preferably new)

The patterns exhibited by wh-clefts and reverse wh-clefts seem to confirm the claim often made that clefts should mark a specific information structure. In wh-clefts in particular, the head always marks new information – in Quirk et al.'s terms it carries the dominant focus of the sentence – and the c-clause expresses old or inferrable information. In fact, Delin claims that it is impossible for the c-clause in this type of cleft to express new information, as indicated by the following example:

(51) - Sue, this is Marta.
- Hi! It was your flat I rented when you were in the States.
? - Hi! What I rented when you were in the States was your flat.
(Delin 1989:197)

Delin explains this constraint as the result of a number of circumstances:

> “When it comes to why such a contraint should exist, we can only propose that the combination of leftmost position in the cleft construction, traditionally the site of information already available to the hearer or reader (cf. Halliday 1967), and the fact that this position in the *wh*-cleft is marked as the site of presupposed information, together represent too strong a signal that the information is in some sense contextually available for New information to be situated there.”
> (p.198)

Also in Danish, the head of a wh-cleft expresses new information and is prosodically dominant:

(52) Hvem der kom i går var PETER.
(The one that came yesterday was Peter.)
(Togeby, O. 1992:465)

The c-clause, however, may also contain a focus, as in:

(53) Hvad jeg IKKE forstår er, at han kan være så LIGEGLAD.
(What I don’t understand is that he can care so little.)

Both in English and Danish, then, wh-clefts can be considered focus marking constructions, in as much as the head can always be considered the dominant focus. Compared to the corresponding non-cleft sentence, it appears that a wh-cleft construction provides an effective means of moving a focussed subject to end-focus position:

(54) a. PETER ryger mest.
b. Den, der ryger mest, er PETER.
(a. PETER smokes most.
b. The one that smokes most is PETER.)

In reverse wh-clefts, on the other hand, Delin claims that the head, which is placed in the sentence leftmost position, always expresses old information. The c-clause usually carries the only focus and contains new or inferrable information. This characterisation also seems to hold for Danish reverse wh-clefts. Therefore, we shall consider them focus marking constructions. However, it is difficult to predict which constituent within the c-clause is stressed. Consider thus the following examples:

(55) a. Ja, det er LIGE netop, hvad jeg mener.
(Yes, it’s JUST exactly what I mean.)
b. Ja, det er det, jeg HELE TIDEN har sagt.
(Yes, it’s what I’ve said ALL ALONG.)

A closer look at the pragmatic function of reverse wh-clefts shows that they provide an effective topicalising device. In other words, they allow for constituents referring to referents already present in the discourse to be moved towards the beginning of the sentence. The following sequences exemplify this:

(56) - Jeg har ikke inviteret Peter med.
? - Hvorfor ikke? Jeg kan bedst lide PETER.
(I haven’t invited Peter.
lit: Why not? I like best Peter.)

(57) - Jeg har ikke inviteret Peter med.
- Hvorfor ikke? Peter er den, jeg BEDST kan lide.
(I haven’t invited Peter.
Why not? Peter is the one I like best.)

The oddness of the non-cleft sentence as a response is due to a clash between the high salience of *Peter* in the discourse and the default information structure assigned by word order and accent, according to which *Peter* should be new information[12]. In the cleft, on the other hand, *Peter* occurs in sentence-initial position, which is the privileged position for salient information, and it is functioning as the grammatical subject, which is considered the primary topic of the sentence by default.

For it-clefts, the issue is more complex. In Delin's material, it-clefts either show focus on the c-clause or double focus. As far as Danish it-clefts are concerned, both Nølke (1984) and Togeby (1992) seem to assume that the head is always in focus. In fact, Nølke uses the term *focus* rather than *head*. The examples we have collected (a dozen sentences from *DK87*) all confirm that the it-cleft head is focal.

We have seen that Delin identifies two different patterns of information structure as possible in it-clefts. In the first, which is also the most frequent in her material, the head is given information, and the c-clause is either new or inferrable. This pattern is also possible and actually very common in Danish it-clefts. An example is the following text, where *den lille* (the little one) in the head of the cleft, refers back to *barn* (the child) in the second clause of the passage, while *græd* (was crying) in the c-clause is inferrable from the previously mentioned *snøften* (sniffling):

(58) Fødslen gik fint, men da mor og barn havde været hjemme et par dage hørte hendes mand lydelig snøften inde fra børneværelset, og det var ikke den lille, der græd.
(The delivery went well, but after mother and child had been at home for a couple of days, her husband heard an audible sniffling from the children's room, and it was not the little one who was crying.)
(*DK87*)

In the second pattern, the head is new information, and often establishes a contrast with some other discourse referent, while the c-clause conveys information which is either known or inferrable, as in:

(59) - Min mand går rundt og får psykiske problemer på grund af det bildæk, han har rundt om maven.
- Det går min mand også, men det er nu ventilen, han har problemer med.
(My husband's getting all worked up about that inner tube he's got around his waist.
So is mine, but it's the valve he's got problems with.)
(*DK87*)

However, it-clefts displaying this pattern of information structure seem rare in Danish[13]. This tendency seems partly to confirm Togeby's claim that the head of a cleft bears a special type of focus, characterised by the fact that it often refers to given information. He calls this focus type *identitetsfokus* (identity focus). Togeby, however, seems to assume that this is true of any kind of cleft, and hence he cannot account for the fact that the head of wh-clefts, as we have seen, typically expresses new information. Therefore, we shall not adopt Togeby's term, but simply agree that for the majority of Danish it-clefts, the head corresponds to a given referent.

An additional pragmatic characteristic of it-clefts shown by most of our text examples, is that they provide a natural way to emphasise the subject of the sentence by moving it away from sentence-initial position, i.e. by detopicalising it, as in the examples below:

(60) Det var tidligere premierminister Jacques Chirac, der dannede RPR... Det er også ham, der som topkandidat har lidt nederlag til Mitterrand ved de to sidste præsidentvalg. Det er også ham, der som leder af RPR er hovedansvarlig for partiets tilbagegang.

[12]The sentence can be made acceptable by stressing the verb rather than the object, thus avoiding focus on a highly salient referent: "Jeg kan bedst LIDE Peter".

[13]This tentative conclusion, however, is based on rather sparse material as clefts are not very frequent in written texts. Analysis of examples from an oral corpus may provide different results.

(It was former Prime Minister Jacques Chirac that founded RPR... It is also him that, as top candidate, has been defeated by Mitterand in the last two presidential elections. It is also him that as leader of RPR is mainly responsible for the party's decline.)

(*DK87*)

However, we agree with Delin that pragmatic properties are not in themselves sufficient to motivate the use of it-clefts, or of clefts in general. First of all, although the head of an it-cleft often seems to correspond to a given referent, this need not always be the case, so that the focussing of a given referent cannot be considered a defining characteristic of it-clefts in general. Secondly, the other property we have identified, namely the fact that it-clefts provide a natural way of emphasising the subject of a sentence, is also not necessarily the case. In the example below, for instance, the focus is not the subject of the sentence, and can easily bear stress in a unmarked declarative sentence. However, the example also shows that a cleft is still possible:

(61) a. Jeg talte med PETER i går.
(I talked to Peter yesterday.)

b. Det var PETER, jeg talte med i går.
(It was Peter I talked to yesterday.)

An account of it-clefts, and indeed of clefts in general, cannot ignore their presuppositional nature, which is an integral part of their meaning, and the characteristic that clearly distiguishes them from their non-cleft counterparts.

In conclusion, in Danish clefts, focus is structurally determined as it falls on the head in it-clefts and wh-clefts, and on the c-clause in reverse wh-clefts. As in non-cleft sentences, the focus of a cleft need not correspond to new information. On the contrary, an important pragmatic feature of it-clefts is that they allow and indeed most often display focussing of given information.

However, the use of clefts in a discourse cannot be accounted for solely in terms of focus, as the same focussing pattern can also be achieved in non-clefts by sentence accent. Rather, it is their presuppositional nature that clearly distinguishes clefts from non-clefts.

The grammar of Danish clefts

Danish clefts are like English clefts unbounded dependency constructions (UDCs). Pollard & Sag (1994) distinguish two classes of UDCs, which they call filler-gap constructions (or strong UDCs) and weak UDCs. The first class is characterised by the fact that a wh-pronoun 'fills' an empty gap (or trace) in the rest of the sentence[14]. This is illustrated by the examples below:

(62) a. Det var Peter, som$_i$ hun sendte bogen til $\__i$.
(It was Peter who she sent the book to.)

(Nølke 1984:81)

b. Den, som$_i$ hun sendte bogen til $\__i$, var Peter.
(lit: That who she sent the book to was Peter.)

c. Peter er den, som$_i$ hun sendte bogen til $\__i$.
(lit: Peter is that who she sent the book to.)

In constructions belonging to the second class, on the contrary, there is no explicit pronominal filler for the gap, but rather an argument in the matrix sentence can be interpreted as being coreferential with the trace in the embedded clause, as in the example below:

[14] More generally, in strong UDCs the antecendent can be either a topicalised constituent or a wh-pronoun – in Pollard & Sag's terminology an *overt nonargumental constituent*. In the case of clefts, however, only a wh-pronoun can act as the filler.

(63) Vi må håbe, at det er en$_i$, vi kan enes om $\_{i}$.
(We must hope that it is someone, we can agree on.)

(*DK87*)

In both cases, the fact that the coreference between the antecedent and the gap can extend over a theoretically unlimited number of sentence boundaries can be shown by embedding the cleft into one or more matrix sentences:

(64) a. Det var Peter, som$_i$ jeg tror, hun sagde, hun sendte bogen til $\_{i}$.
(It was Peter who I think she said she sent the book to.)

b. Vi må håbe, at det er en$_i$, vi begge synes, det er muligt, vi kan enes om $\_{i}$.
(We must hope that it is someone, we both think it is possible we can agree on.)

In practice, there is a limit to how many levels of embedding the human mind seems capable of processing. However, this limitation is usually considered external to the grammar.

Before dealing with the syntactic structure of clefts, the nature of the filler in clefts of the filler-gap dependency type deserves a more detailed discussion. It was mentioned earlier that the choice of pronoun is more limited in clefts than in relative clauses. *Som* can be used in all syntactic functions (except determiner), but there is a strong tendency to omit it when the extracted constituent is not the subject. Both sentences below are thus possible, although (b) is the preferred version:

(65) a. Det er Peter, som hun elsker.
(It is Peter who she loves.)

b. Det er Peter, hun elsker.
(lit: It is Peter she loves.)

An additional possibility is to omit *som* and use the complementiser *at*:

(66) Det var ikke ham, at jeg mente.
(lit: It was not him that I meant.
It wasn't him I meant.)

(Hansen, A. 1967 III:457)

At can also be used in colloquial speech when the head of the cleft is an adverbial phrase or clause:

(67) Det er i morgen, (at) jeg skal undervise.
(lit: It is tomorrow (that) I have to teach.
It's tomorrow I have to teach.)

According to Nølke (1984), wh-pronouns (e.g. *hvem, hvilken, hvor, hvorfra, hvortil* – who, which, where, from where, to where) are not allowed in clefts. A possible exception is *hvis* (whose), which functions as a determiner:

(68) Det er Peter, hvis sommerhus vi lånte i weekenden.
(It's Peter whose summer house we borrowed at the weekend.)

When the extracted constituent is the subject, *der* can be used instead of *som*. In fact, *der* can only function as a subject:

(69) Det er Peter, der ringede i går.
(lit: It is Peter, there rang yesterday.
It's Peter who rang yesterday.)

(70) * Det er Peter, der jeg har ringet til.
(lit: It is Peter there I have rung to.
It's Peter I've rung.)

In colloquial speech, the following combinations of *som, der* and *at* (that) can occur (although not all Danish speakers would accept them):

som der
som at der
som at

Of these, *som der* and *som at der* can only have subject function, whereas *som at* cannot be a subject:

(71) a. Det er Peter, som (at) der kommer i aften.
(lit: It is Peter who (that) there comes tonight.
It's Peter who's coming tonight.)

b. * Det er Peter, som at kommer i aften.
(lit: It is Peter who that comes tonight.
It's Peter who's coming tonight.)

The order in which they can occur is fixed, so that the following sequences are incorrect:

**som der at*
**at som*
**der som*

Nølke deals with these patterns within Government and Binding theory by proposing that *som*, as generally assumed for *wh*-phrases, is adjoined to COMP, and that COMP actually consists of two adjacent nodes. In the marked case where *som* is present, COMP will have the following shape:

$({}_{COMP}\ ({}_{COMP}\ \textit{som})\ ({}_{COMP}\ \textit{-wh}))$

where *-wh* in the second node may surface as *at* or null. Since *der* is always to the right of *at* when this is present, it cannot be placed in COMP, and is thus not a relative pronoun. Hence, Nølke suggests that *der* appears in surface structure as the result of a *der*-insertion transformation. On this approach, *der*-insertion is considered obligatory if the trace in the relative clause is not properly governed, as is the case when the trace is in subject position and COMP is empty.

An additional indication that *der* does not behave like a relative pronoun, is the fact that it cannot be the filler of a gap in an unbounded dependency:

(72) * Det var Peter, der hun sagde, hun sendte bogen til.
(lit: It was Peter there she said she sent the book to.
It was Peter that she said she sent the book to.)

The correct alternative to the sentence above would be:

(73) Det var Peter, (som) hun sagde, hun sendte bogen til[15].
(It was Peter (who) she said she sent the book to.)

[15]The different behaviour of *som* and *der* with respect to unbounded dependency constructions is in fact our main reason for treating the former but not the latter as a relative pronoun. This view, however, is not shared by all workers in the field. E. Hansen (cf. Hansen 1980) claims that *som* functions in all respects as a conjunction. Accepting his point of view would imply a different analysis than the one proposed here, as conjunctions are not usually assumed to bear an index.

In conclusion, we subscribe to the view that *der* as used in relative clauses and clefts[16], should not be considered a relative pronoun, but rather a resumptive pronoun following either a gap or a relative pronoun.

On the other hand, it is less clear that *at* in conjunction with *som* should be considered a relative pronoun (or *wh*-word) as suggested by Nølke[17]. We shall adopt a different view here, and consider it a complementiser since its main function is to mark a type of nominal clause, as shown below:

(74) Jeg vidste ikke, at du var kommet.
(I didn't know that you had arrived.)

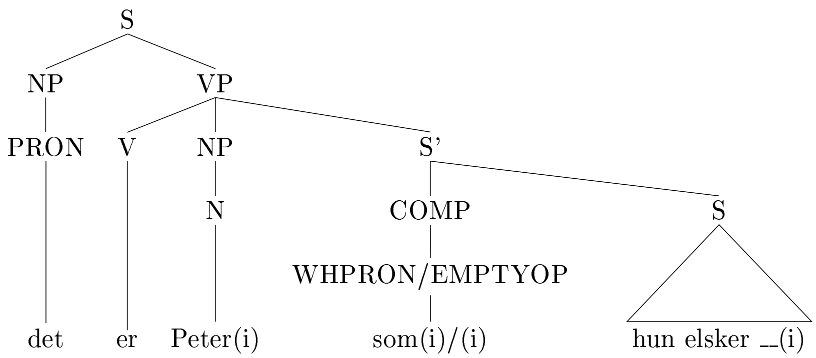
Turning now to clefts of the weak dependency type, Nølke proposes that the trace in the c-clause should be bound by an empty operator in COMP position, while the same position is occupied by the relative pronoun if the dependency is of the strong type. According to this view, the two clefts below receive the same structural representation:

(75) a. Det er Peter, som hun elsker.
(lit: It's Peter who she loves.)

b. Det er Peter, hun elsker.
(It's Peter she loves.)

Thus, in the phrase structure marker produced by syntactic analysis, the empty operator can alternate with a wh-pronoun under COMP:

(76)



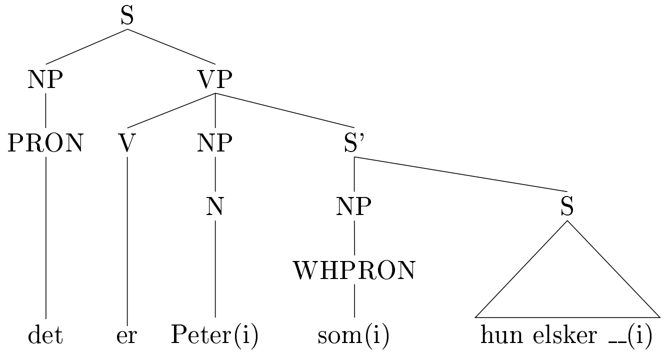
This analysis is coherent with the GB view of UDCs, according to which the dependency between filler and gap is explained in terms of cyclic movement to COMP. COMP is a kind of escape position, which must be available for movement to be possible. Given this constraint, together with the subjacency principle saying that movement is restricted to the domain of two adjacent cyclic nodes (NP and S for English), well-known limitations to wh-movement in English such as the Complex NP Constraint can be predicted. In GB, subjacency is considered a principle of universal grammar. However, Scandinavian unbounded dependency data (cf. Andersson 1982) challenge this view, and in our opinion makes the existence of a COMP node seem less well-motivated.

Therefore, we shall propose a different approach, inspired by non-transformational grammar theories such as GPSG (Gazdar, Klein, Pullum & Sag 1985) and HPSG. First of all, we abandon the idea of a COMP position which the wh-pronoun or an empty operator are moved onto. This implies i.a. that clefts displaying strong and weak dependencies, are assigned different phrase markers, as shown below:

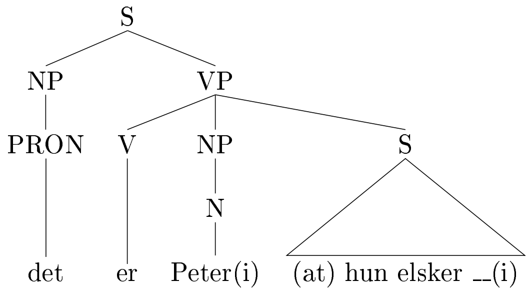
[16] This use of *der* should not be confused with the pleonastic pronoun occurring in sentences such as *Der var en gang en konge* (lit: There was once a king). See Section 5.3.2.

[17] He uses the term *wh*-complementiser.

(77)



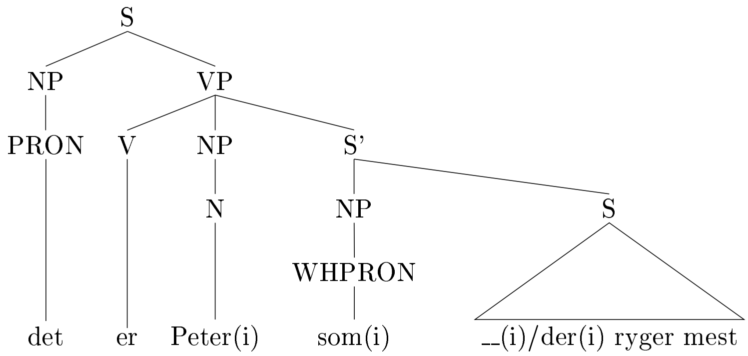
(78)



Note also that in both representations, the c-clause does not modify the NP which is the head of the clause. Thus, the structure clearly shows that this is not a relative clause.

If the extracted constituent is the subject, we distinguish two cases. In the first one a relative pronoun is used, possibly followed by the resumptive *der*, as shown below:

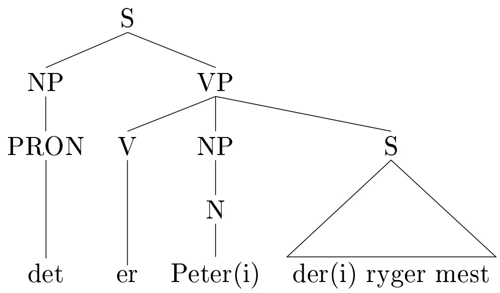
(79)



Note that *der* alternates with a subject gap.

In the second case, no relative pronoun is used, and *der* is then obligatory. The relevant structure is the following one:

(80)



In a non-transformational framework, the relation between filler and gap is not viewed as movement, but as a static relation between nodes in a tree, or features in a feature structure. In particular, the following non-local features are used in HPSG to handle non-local dependencies of various kind:

(81)
$$\left[\begin{array}{ll}\text{NONLOCAL} & \left[\begin{array}{ll}\text{INHERITED} & \left[\begin{array}{ll}\text{QUE} & (\text{set of } \textit{npros})\\ \text{REL} & (\text{set of } \textit{ref} \text{ indices})\\ \text{SLASH} & (\text{set of } \textit{local} \text{ structures})\end{array}\right]\\ \text{TO-BIND} & \left[\begin{array}{ll}\text{QUE} & (\text{set of } \textit{npros})\\ \text{REL} & (\text{set of } \textit{ref} \text{ indices})\\ \text{SLASH} & (\text{set of } \textit{local} \text{ structures})\end{array}\right]\end{array}\right]\end{array}\right]$$

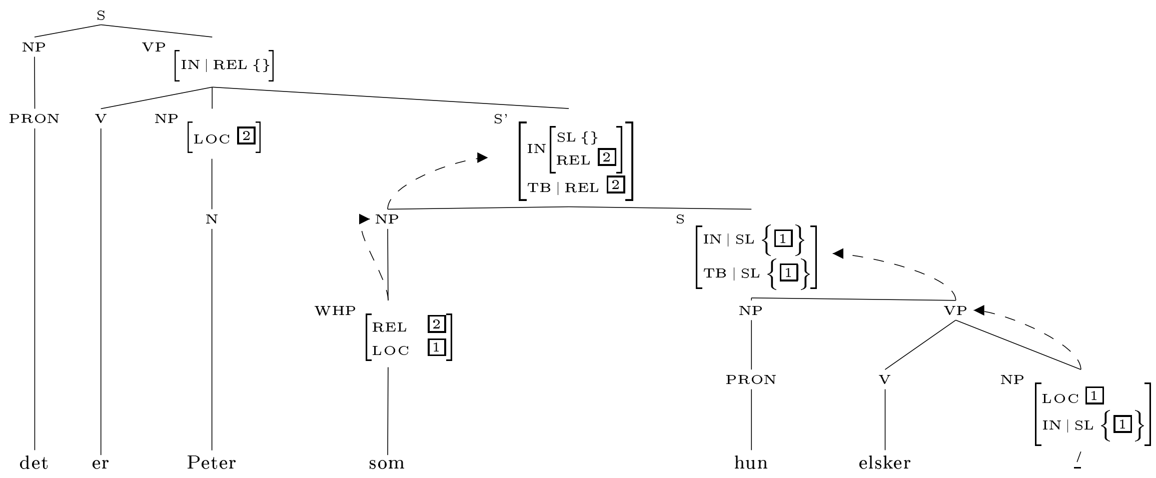
The QUE feature, which takes a set a non-pronominal nominal objects as its value, is used for interrogatives and is not relevant in the present context. REL takes a set of referential indices as its value, and is used for relative pronouns. SLASH takes a set of local structures, and is used for traces.

Thus, the gap in the c-clause of a cleft is represented by a feature structure where all the values for local as well as non-local features are empty, but where the feature LOCAL shares an index with INHERITED|SLASH, namely:

(82)
$$\left[\begin{array}{ll}\text{PHON} & \langle\rangle\\ \text{SYNSEM} & \left[\begin{array}{ll}\text{LOCAL} & \boxed{1}\\ \text{NONLOCAL} & \left[\begin{array}{ll}\text{INHERITED} & \left[\begin{array}{ll}\text{QUE} & \{\}\\ \text{REL} & \{\}\\ \text{SLASH} & \left\{\boxed{1}\right\}\end{array}\right]\\ \text{TO-BIND} & \left[\begin{array}{ll}\text{QUE} & \{\}\\ \text{REL} & \{\}\\ \text{SLASH} & \{\}\end{array}\right]\end{array}\right]\end{array}\right]\end{array}\right]$$

Some information, i.e. syntactic category and semantic restrictions, will then be filled in in this structure by unification with the relevant features on the head of the phrase the trace occurs in. Thus, in the representation below, which is an HPSG-like analysis of sentence (65a), the verb imposes the restriction that the missing complement must be a nominal phrase or clause. Through the index sharing between LOCAL and INHERITED|SLASH in the trace node, the information that the missing node must be an NP is passed up the tree. In HPSG, (cf. Sag & Pollard 1994:164 for more details) a grammar rule called the Head-Filler Rule (or Schema 6) will then make sure that the trace is bound off when the filler is found. Thus, the TO-BIND|SLASH feature is structure-shared with INHERITED|SLASH once the filler has been found, and the value corresponding to the trace is not passed up in the tree. This can be seen in the representation below:

(83) *Cleft representation*



(Feature names have been shortened for lack of space.)

The representation above also shows how the dependency between the relative pronoun and its antecedent is encoded through the REL feature. Somewhat simplifying the HPSG treatment of relative dependencies, the INHERITED|REL feature on the relative pronoun, which is passed up onto the dominating S' node, is made to share an index with the LOCAL feature of the antecedent[18].

A slightly different way of handling the sharing of features between fillers and gaps is the so-called gap threading technique (cf. Alshawi et al. 1991). Gap threading is a simple and computationally tractable list manipulation technique which is relatively easy to implement in a formalism like ALEP. It has therefore been chosen here instead of the SLASH approach used in HPSG. The following NONLOCAL|GAPS features will be used:

(84) $$\begin{bmatrix} \text{NONLOCAL} & \begin{bmatrix} \text{GAPS} & \begin{bmatrix} \text{GAPS\_IN} & \text{(feature list)} \\ \text{GAPS\_OUT} & \text{(feature list)} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Alshawi et al. explain the use of the GAPS features in the following way:

> "Although notions of directionality do not matter in a unification formalism, it is intuitively helpful to think of the first member of the value of 'gaps' as a list of fillers going into a category from the left and out at the right. If a gap of the appropriate type is found within that category, the 'gaps' out will be a list with one fewer members than went in." (p.237)

In other words, the trace that has to be bound is represented as an item on the GAPS_IN list of the constituent ('category' in the quote above) containing the gap. This constituent is in our case the c-clause in a cleft. The value of GAPS_IN is threaded through the whole subtree representing the constituent until the gap is actually found. Then, by unification with the lexical entry corresponding to the gap, the value of the GAPS_OUT list is set to be equal to the GAPS_IN minus the item which is structure-shared with the gap itself. The lexical entry of the gap is as follows:

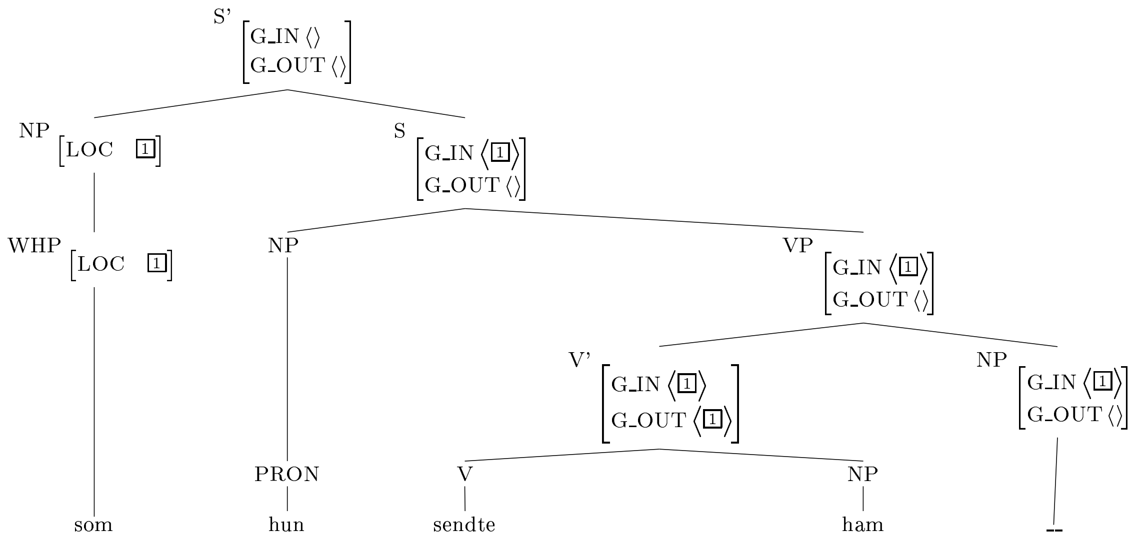
[18]The treatment of relative dependencies presented in Pollard & Sag (1994) postulates a null complementiser as the head of the relative clause. Pollard and Sag need the null complementiser to compute the contribution made by the relative clause to the semantic content of the modified NP. However, there is something unconvincing about a null element that is never realised as an overt element.

(85) $$\begin{bmatrix} \text{PHON} & \langle\rangle \\ \text{SYNSEM} & \begin{bmatrix} \text{LOCAL} & \boxed{1} \\ \text{NONLOCAL} & \begin{bmatrix} \text{GAPS} & \begin{bmatrix} \text{GAPS\_IN} & \left\langle \boxed{1} | \boxed{2} \right\rangle \\ \text{GAPS\_OUT} & \left\langle \boxed{2} \right\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

As can be seen, the GAPS_OUT value is a list which is equal to the GAPS_IN list minus the element which is structure-shared with the LOCAL feature of the trace itself.

The way in which the GAPS features are instantiated in a concrete cleft example is shown in the diagram below, displaying the representation of the c-clause in the cleft *Det er det brev, som hun sendte ham* (It's the letter she sent him):

(86)



Note the way in which the GAPS features are threaded through the VP node until the gap is found and deleted from the GAPS_OUT list.

Note also that the fact that each member of the GAPS tuple is a feature list allows for the treatment of multiple dependencies as found e.g. in Scandinavian languages. We shall return to this point in Section 5.3.3, as well as Chapter 6, where some of the rules implemented in ALEP to handle the GAPS features are discussed in some detail[19].

Finally, we observed earlier that if the resumptive pronoun *der* is used, the dependency created between the antecedent and the pronoun is not unbounded. To account for this restriction, no GAP feature is introduced in the representation by *der*.

Following HPSG, *der* will be treated as a noun of type *ppro* to distinguish it from other pronouns on the one hand, and nouns on the other. Although *der* is not a relative pronoun, we have chosen to let the coindexation with its antecedent occur through the REL feature. Hence, the CONTENT|INDEX feature is structure-shared with a non-local REL feature. The rule that builds an S with *der* as the subject must ensure that the REL feature is passed on to the S node to be structure-shared with the index of the antecedent noun.

The SYNSEM feature in the lexical entry for *der* will be:

[19]A discussion of the way in which gap threading can be used to treat Danish unbounded dependencies in general can be found in Povlsen, Paggio & Underwood (1995).

(87) *der* $\left[\text{SYNSEM}\left[\begin{array}{ll}\text{LOCAL} & \left[\begin{array}{l}\text{CAT}\left[\text{HEAD}\ {}_{noun}\left[\text{nom}\right]\right] \\ {}_{ppro}\ \text{CONTENT}\left[\text{INDEX}\ \boxed{1}\right]\end{array}\right] \\ \text{NONLOCAL} & \left[\text{REL}\left\langle\boxed{1}\right\rangle\right]\end{array}\right]\right]$

The presence of *at* in the right contexts is ensured by the fact that S is or is not marked for an *at* complementiser in the relevant cases.

Compared to Nølke's approach, the treatment proposed here loses in generality, in the sense that a variety of different phrase markers are necessary to cover the data. However, this lack of generality is simply a reflection of the fact that in a non-transformational approach, the syntactic representation reflects surface form rather directly, and thus appears less abstract. The uniformity of the various types of cleft construction is captured, however, through the systematic use of the REL and GAPS features, and, as will be discussed in the next section, through the semantic representation.

To complete our account of the syntax of Danish clefts, we indicate here the representations proposed for wh-clefts and reverse wh-clefts, in (88a) and (88b) respectively:

(88) a. (S (S' (NP [LOCAL $_{[1]}$] S [GAPS_IN $_{<[1]>}$])) VP (V (*være*) XP))
b. (S (XP VP(V (*være*) S'(NP [LOCAL $_{[1]}$] S[GAPS_IN $_{<[1]>}$]))))

In both cases we partly follow Delin's suggestion that the c-clause should indeed be represented as a clause – an S' in fact – rather than as an NP. However, Delin also proposes a flat representation where the predicate of the copula is attached directly under the S node. We do not adopt this suggestion, but maintain the fundamental subject-VP structure also used in non-cleft sentences.

So far, we have not dealt with the syntactic form of the cleft head. In all the cleft types, we assume that the category of the head is an XP, i.e. a maximal projection of any category. In fact, heads that are not maximal projections are disallowed:

(89) a. * Det er hans, de kørte med bil.
(lit: It's his they drove with car.)

b. * Det er grædt, hun har.
(It's cried she has.)

(Nølke 1984:79)

Not all maximal projections, however, seem possible as heads. Thus, impossible heads in it-clefts are finite VPs and nominal clauses. However, if the nominal clause is preceded by the pronoun *det*, clefting becomes possible:

(90) a. * Det var kaldte på mig, Peter.
(It was called me Peter.)

b. * Det er, at hun er så rolig, jeg kan lide ved hende.
(It's that she is so calm I like about her.)

c. Det er det, at hun er så rolig, jeg kan lide ved hende.
(lit: It is it that she is so calm I like about her.)

Note that this constraint does not hold for wh-clefts:

(91) Det jeg kan lide ved hende er, at hun er så rolig.
(What I like about her is that she is so calm.)

Another general constraint on the instantiation of the cleft, which holds for English and for Danish, concerns sentential adverbials, which cannot occur as cleft heads although they are maximal projections:

(92) a. * Det er ærligt talt, jeg ikke forstår dig.
(It's frankly I do not understand you.)

b. * Det er, hvis du kommer, at jeg går med.
(It's if you come that I go.)

Finally, a few other restrictions are due to the semantic content of clefts, as described in the preceding section. The head cannot be a universally quantified phrase:

(93) ? Det er alle børnene, hun leger med.
(It's all the children she plays with.)

Nor can it be a negative polarity item:

(94) ? Det var ikke nogen, der kom.
(It wasn't anybody that came.)

In both cases, the head does not provide a unique instantiation of the variable in the presupposition and the resulting sentence is therefore semantically ill-formed.

The semantic representation of clefts

In the semantic representation of clefts, we want to capture their propositional content as well as their presuppositional structure.

To express the presupposition inherent in the cleft structure, Nølke (1984) proposes that the variable in the presupposition expressed by the c-clause is bound by the λ-operator, and the λ-expression thus obtained is predicated of the referent expressed by the head of the cleft. Applied to three examples, this yields the following representations:

(95) a. Det er Peter, hun elsker.

b. λy (love(she,y)) (Peter)

(96) a. Hvad jeg ikke forstår er, at du allerede går.

b. λy not(understand(I,y)) (already_go(you))

(97) a. Sofia er den, der snakker mest.

b. λx (talk_most(x)) (Sofia)

By λ-conversion, the expression in (95b) above is reduced to:

(98) (love(she,Peter))

which is the semantic representation of the non-cleft:

(99) Hun elsker Peter.
(She loves Peter)

Hence, the truth-conditional equivalence of a cleft and its non-cleft counterpart is reflected in the equivalence between the semantic representations of the two sentences.

In Chapter 4, the focus of a sentence was also treated as a presupposition triggering factor, and it was discussed whether focus could be represented by λ-abstraction. If such a representation is assumed, it is useful to distinguish two different cases. In the first case, i.e. in Danish for it-clefts and wh-clefts, the head of the cleft also carries the focus. This means that the abstractions introduced by focus interpretation and by the presuppositional structure inherent in the cleft's meaning coincide. Thus, the semantic interpretation of

(100) Det er $[\text{PETER}]_F$, hun elsker.

where the focus has been made explicit, will still be (95b) above.

The second and more problematic case is that of reverse wh-clefts, where the focus falls on the c-clause rather than the head. Syntactic structure and focus introduce two different presuppositions. For sentence (97a) above, these can be paraphrased by the two sentences below:

(101) a. Sofia does something.
b. Somebody talks most.

Representing both by λ-abstraction, would yield the following complex expression:

(102) $\lambda x \lambda e$ (e(x)) (talk_most) (Sofia)

The first reduction step is to apply λ-conversion to the inner argument, thus producing:

(103) λx (talk_most(x)) (Sofia)

The second step is to replace (x) with (Sofia) to get:

(104) (talk_most(Sofia))

This is again the semantic content of the non-cleft counterpart to sentence (97a).

However, following the conclusions drawn in Chapter 4, here focus will not be represented by λ-abstraction, but rather by means of a separate feature attached to the focussed constituent(s). Thus, this feature will simply be attached to the constituent constituting the head of the cleft in the first case, and to the c-clause in the second. It remains to be clarified, however, how the presuppositional content of clefts can be conveniently modelled in a feature structure representation.

One possibility would be to represent the presupposition as an incomplete copy of the content of the sentence, where an uninstatiated variable replaces the argument which is bound in the cleft. However, it is not clear how this can be implemented, as it would mean blocking the unification of the content feature for the argument under consideration. Therefore, we shall opt for a simpler approach. We propose to introduce a feature CLEFTPRESUP, the arguments of which will be the content of the main verb and the variable corresponding to the argument which is unbound in the presupposition. The presupposition feature will be part of the CONTEXT feature. Thus, the semantics and presuppositional content of the sentence *Det er Peter, hun elsker* (It's Peter she loves) can be represented in the following way:

(105)
$$\left[\begin{array}{ll} \text{CONT} & \left[\begin{array}{ll} \text{RELN} & \text{elske} \\ \text{ARG1} & \left[\text{CONT} \; \left[\text{RESTR} \; \left[\text{RELN} \; \text{hun}\right]\right]\right] \\ \text{ARG2} & \left[\text{CONT} \; \left[\begin{array}{ll} \text{RESTR} & \left[\text{RELN} \; \text{Peter}\right] \\ \text{INDEX} & \boxed{1} \end{array}\right]\right] \end{array}\right] \\ \text{CONTEXT} & \left[\text{CLEFTPRESUP} \; \left[\begin{array}{ll} \text{RELN} & \text{elske} \\ \text{UNBOUND} & \left[\text{INDEX} \; \boxed{1}\right] \end{array}\right]\right] \end{array}\right]$$

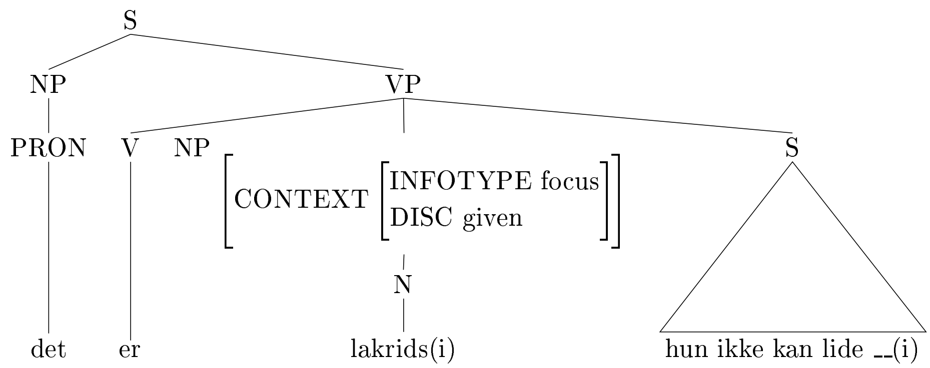
This is clearly not an exhaustive representation of the cleft presupposition, as it does not mention the arguments of the verb which are also part of this presupposition. However, it will suffice in the generation component of an MT system to determine which of the arguments should be the cleft head. On the other hand, the same representation does not tell us anything about how to generate the right focus. This is the subject of the next section.

Analysis and generation of Danish clefts

To conclude our discussion of which syntactic and semantic representations Danish clefts should be assigned, we show here how the phrase markers proposed in the section on the grammar of clefts, would be enriched by the INFOTYPE feature, as well as by a feature signalling whether information is given, inferrable or new.

In it-clefts, we have seen that the head is focus-marked. Additionally, it preferably refers to given information. This is reflected in the contents of the CONTEXT feature in the tree below, which represents the sentence *Det er lakrids hun ikke kan lide* (It's liquorice she doesn't like).

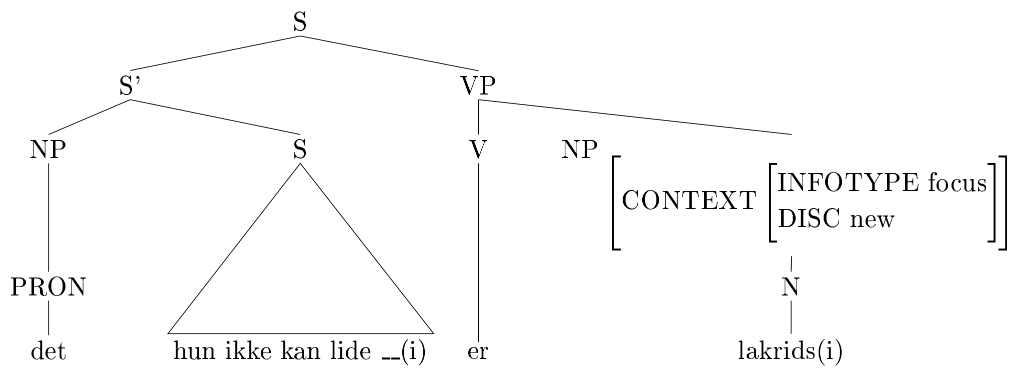
(106)



The possibility that the c-clause should also bear focal prominence is not excluded, as the c-clause is underspecified with respect to the INFOTYPE feature.

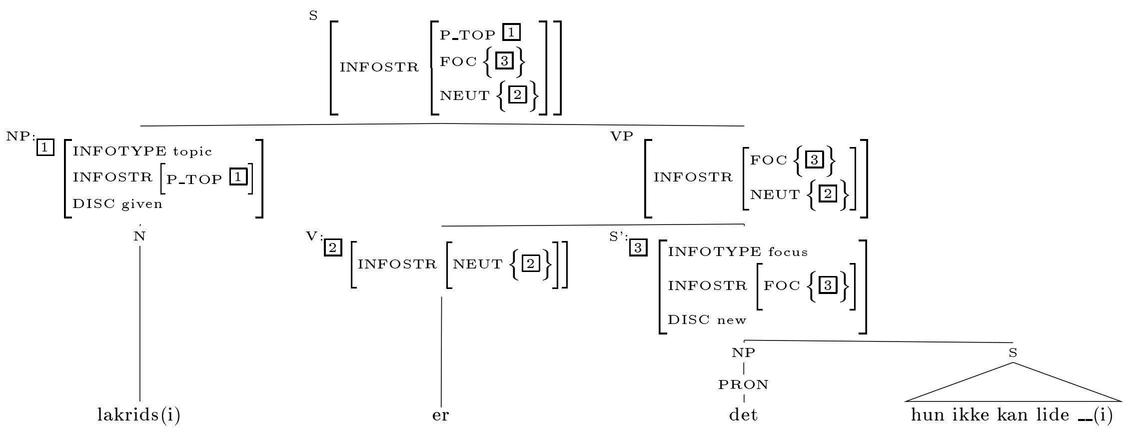
In Danish wh-clefts, the head is again the dominant focus, but this time the information expressed by the head is preferably new. The tree below provides the syntactic representation of the wh-cleft *Det hun ikke kan lide er lakrids* (What she doesn't like is liquorice), enriched with the CONTEXT feature.

(107)



Finally, for reverse wh-clefts, the head will be analysed as the primary topic, and therefore as given, while the c-clause will be focus-marked, and will be considered new information. This is shown below, in the representation of the sentence *Lakrids er det, hun ikke kan lide* (Liquorice is what she doesn't like). This time, the instantiation of the CONTEXT|INFOSTR feature on the top S node as well as on the daughter nodes is made explicit.

(108) *Gaps in clefts*



The generation of Danish clefts will proceed from representations like those above, complete with a presupposition representation as discussed in the preceding section. There may be cases, however, where a feature expressing whether the focus in the cleft is given or new information is not available. In this work at least, we do not provide an algorithm for the computation of a discourse related 'givenness' feature since we have focussed on how the information structure of the single sentence can be predicted from the syntactic shape of the sentence itself. Therefore, in practice we cannot use the INFOSTR|FOCUS feature alone to distinguish between an it-cleft and a wh-cleft, as the head is the focus of the sentence in both cases. This means that either both structures are generated only on the basis of the semantic content and INFOSTR|FOCUS feature, or that a general preference for one of the two construction types is adopted.

5.3.2 Existential sentences

Existential sentences (ES) in Danish are sentences introduced by an unstressed *der* pronoun functioning as a pleonastic subject NP:

(109) Der kom en mand løbende.
(lit: There came a man running.)

The ES above can be parallelled with the sentence below, which has arguably the same semantic content:

(110) En mand kom løbende.
(A man came running.)

The constituent functioning as the grammatical subject in this sentence, namely *en mand* (a man), seems to behave like a grammatical object in the corresponding ES. From here on, although we do not presuppose a transformational treatment of ES, we shall use the term *postponed NP* to refer to the NP following the main verb in ES[20].

[20] Note that we shall not discuss *der* constructions such as the following ones:

Der lugtede meget.
Der blev godt varmt.
Der skrives og tales for meget i vore dage.
(lit: There smelled much.
There got quite warm.
There is written and talked too much i our days.)

In these examples, there is no postponed NP that *der* can be said to be introducing. However, it could be argued that our claim that *der* constructions are topicless constructions, is not contradicted by such cases.

ES are interesting in our context because they are marked with respect to their information structure. In particular, we shall argue that an important pragmatic function of ES is that of detopicalising the subject of the sentence and of marking the postponed NP as (a part of) the focus. As a consequence, the postponed NP in ES cannot carry existential presuppositions of the kind normally associated with a strongly quantified NP functioning as the primary topic of the sentence.

In the following sections, first we deal with the syntax of Danish ES, then we discuss various theories that have been put forward to explain the constraints which the postponed NP in ES is subject to, and finally we conclude that ES can be regarded as topicless constructions. On the basis of this conclusion as well as the syntactic properties of Danish ES, we then formulate a number of criteria for the automatic analysis and generation of ES.

The syntax of Danish existential sentences

The syntax of English ES has been object of much study, especially within generative frameworks. The point of departure for our discussion will be the typology of English ES introduced by Milsark (1974). Milsark distinguishes the following types:

(111) ES
- *be* ES
- Verbal ES
  - IV ES
  - OV ES

IV: inside the verbal unit
OV: outside the verbal unit

Examples:

be ES:	There are rabbits. There was a man arrested. There is rain in the South.
IV ES:	There appeared an angry crowd.
OV ES:	There dwelt in that house an old man.

(quoted in Lumsden 1988:3)

This typology also fits the Danish data. However, the group of verbs which can occur in ES is larger in Danish than in English. Tentatively, we would say that this group comprises at least unaccusative verbs and motion verbs:

(112) a. Der løb en kat igennem døren.
(lit: There ran a cat through the door.
A cat ran through the door.)

b. Der sker ikke ret meget for tiden.
(lit: There happens not much these days.
Not much happens these days.)

But not:

(113) a. * Der græder en pige.
(lit: There weeps a girl)

b. * Der bløder en såret mand på skadestuen.
(lit: There bleeds a wounded man at emergency.)

c. * Der synger en mand på bænken.
(lit: There sings a man on the bench.)

(Nimb 1933:8)

However, intransitive as well as transitive verbs can occur if passivised:

(114) Der synges meget i børnehaven.
(lit: There sing(PASS) a lot in the kindergarten.
They sing a lot in the kindergarten.)

In a recent study, Vikner (1995) proposes that the variation in expletive constructions involving *there* exhibited by the Germanic languages (with Danish as one of them) can be explained according to whether they allow what in GB is called V2 or V-to-I movement. V2 movement is a movement of the finite verb to the second position in the sentence, as in wh-questions in all the Germanic languages; V-to-I movement is the movement of the finite verb from its position in the deep structure to the position immediately following the subject. German, Yiddish and Icelandic have both kinds of movement; Faroese and Danish have V2 but not V-to-I; English and French do not have V2. The following table shows the distribution of V2 and V-to-I movement as well as different verb types in the various Germanic languages:

	Ge.	Ic.	Da.	En.	Fr.
Ergatives	+	+	+	+	+
Transitives	+	+	-	-	-
Passive Transitives	+	+	+	+	+
Intransitives	+	+	+	-	-
Passive Intransitives	+	+	+	-	-
V2	+	+	+	-	-
V-to-I movement	+	+	-	-/+	+

Although we do not intend to discuss Vikner's thesis here, we acknowledge his typology. The category 'intransitives', however, seems to call for further distinctions, since as noted above not all intransitive verbs seem to allow the construction with *der*.

OV ES seem only possible in Danish if the postponed NP is considerably heavy. Thus, the English example of OV ES quoted above can be rendered in one of the following two ways in Danish:

(115) a. I huset boede der en gammel mand.
(In the house there dwelt an old man.)

b. Der boede en gammel mand i huset.
(There dwelt an old man in the house.)

A literal translation, on the other hand, is dubious:

(116) ? Der boede i huset en gammel mand.
(There dwelt in the house an old man.)

The NP can, however, be moved to end-focus position if it is heavy:

(117) Der er kommet løbende en Hund, som bærer Deres Navn på Pladen.
(lit: There is come running a dog, which bears your name on the tag.
A dog came running, wearing a tag with your name on.)
(Hansen 1967:167)

Nevertheless, we shall not consider OV ES any further here. Thus, if we exclude OV ES, Danish ES have a rather uniform syntactic shape, which can be described as follows:

der V* (ADVP) NP (XP)*

The verb can be a simple, a complex or a periphrastic form, either active or passive:

(118) a. Der findes ikke hekse.
(There exist no witches.)

b. Der er blevet udført et stort arbejde i dag.
(lit: There has been done a lot of work today.
There has been a lot of work done today.)

(Nimb 1993:5)

c. Jeg håber ikke, at der er mange børn syge om efteråret.
(lit: I hope not that there are many children ill in autumn.
I hope there aren't many children ill in autumn.)

d. Der accepteres ikke Dankort her.
(lit: There accept(PASS) not Dankort here.
Dankort is not accepted here.)

(Nimb 1993:5)

Note that the postponed NP always follows the whole verbal complex in Danish. This is in contrast with English ES, which obey the so-called left-most *be* condition, according to which the postponed NP must be placed following the first occurrence of the verb *be*. Compare, thus, the following two examples:

(119) a. Der er blevet gjort meget mens jeg var væk.
b. * Der er meget blevet gjort mens jeg var væk.

(Lit: There is been done much while I was away.
There is much been done while I was away.

A lot has been done while I was away.)

(120) a. There was a man being arrested.
b. * There was being a man arrested.

(Lumsden 1988:56)

The XP* following the postponed NP may be a prepositional or adjectival phrase:

(121) a. Der var en del mennesker i byen i går.
(There were a lot of people in town yesterday.)

b. Der var en del mennesker syge.
(There were a lot of people ill.)

(Nimb 1993:5)

c. Der er flere børn med influenza.
(There are several children with the flu.)

In connection with certain verbs, the NP may also be followed by a gerundive form:

(122) Der kom en mand løbende.
(lit: There came a man running.
A man came running.)

When discussing the syntactic structure of ES, at least two issues must be dealt with. The first one regards the relationship between the pleonastic *der* and the postponed NP. As noted in Pollard & Sag (1994), the corresponding English pronoun *there* shares number but not person agreement features with the NP[21]:

[21] This is, however, not true for all varieties of English. In some dialects, in fact, number agreement is not always required.

(123) a. I never heard them at all till there was_{sing} you_{sing}.
b. * I never heard them at all till there $were_{plu}$ you_{sing}.
c. * Everyone else has left and now there am only I/me.

In Danish, there is no subject verb agreement. The only agreement pattern which seems relevant to ES is that between the postponed NP and a predicative AP, e.g.:

(124) Der var flere $mennesker_{plu}$ $syge_{plu}$.
(There were several people ill.)

Therefore, it does not seem necessary to posit any sharing of agreement features between *der* and the postponed NP.

The other issue deserving some discussion is the representation of the predication following the logical subject. Lumsden (1988) among others argues that two different structures are relevant, depending on whether the verb is *be* or not. Thus, for *be* ES, he proposes that the combination of NP and (XP) following the verb should be represented as a small clause:

(125) a. There was [$_{SC}$a girl employed].
b. There was [$_{SC}$a girl singing].
c. There was [$_{SC}$a girl drunk].
d. There was [$_{SC}$a girl in the room].
(Lumsden 1988:50)

Syntactically, Lumsden defines a small clause as a structure consisting of an NP followed by a PP, an AP, or a participle.

For Verbal ES, on the other hand, he proposes a different structure, where the (optional) final PP does not form a clausal constituent with the preceding NP, but is either a daughter or a sister of the VP node, as shown below:

(126) a. There [$_{VP}$occurred a riot] [$_{PP}$at midday].
b. There [$_{VP}$occurred a riot [$_{PP}$at midday]].
(Lumsden 1988:55)

Lumsden argues that the small clause analysis is borne out in the case of *be* ES by the fact that the predicative[22] element following the NP cannot be regarded as a complement of the NP itself. The following data seem to confirm this view:

(127) a. Der er to stearinlys tændte på bordet.
b. * Hvor mange stearinlys tændte er der?
c. Hvor mange stearinlys er der tændte?
(There are two candles lit on the table.
How many candles lit are there?
How many candles are there lit?)

[22]The term *predicative* indicates the fact that a constituent can function as the predicate of a copula verb. However, we also consider PPs such as *med mørkt hår* (with dark hair) predicative. They cannot function as subject complements, but the object of the preposition can typically occur as the object of the verb *have* (have), e.g.:

Hun har mørkt hår.
(She has dark hair.)

Note, however, that in many cases a synonymous predicative adjective can be used in Danish, e.g. *mørkthåret* (lit: dark-haired).

On the other hand, the fact that the predicative element cannot be topicalised also shows that it is more closely connected to the postponed NP than an ordinary sentence adjunct or verbal complement would be:

(128) a. ? Tændte (på bordet) var der to stearinlys.
b. På bordet var der to stearinlys.

(Lit (on the table) there were two candles.
On the table there were two candles.)

However, regarding the whole tail of the ES as a clausal structure makes it impossible to state subcategorisation constraints on the syntactic form of the element following the NP. This is a problem since not all verbs can combine freely with all kinds of predicative elements, as shown by the following examples:

(129) a. Der lå flere børn syge.
b. * Der kom flere børn syge.

(lit: There lay several children ill.
There came several children ill.

(130) a. Der kom en mand løbende.
b. * Der var en mand løbende.

(lit: There came a man running.
There was a man running.)

Whether the restriction at stake is a syntactic or a semantic one is open to question. However, it seems clear to us that it is a restriction coming from the main verb. Therefore, we agree with Pollard and Sag that the internal structure of the ES tail is best accounted for by letting the verb subcategorise for two independent complements rather than a small clause[23].

By treating ES as lexically determined constructions, then, the different constraints imposed by different verbs that can participate in ES, can be described easily. For example, adapting the HPSG proposal to Danish, the two Danish verbs *ligge* (lie) and *komme* (come) will have the two lexical entries below for the readings relevant to the examples just mentioned:

(131) *ligge*

$$\begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD } \textit{verb} \\ \text{SUBCAT} \left\langle \text{NP}_{der}, \boxed{1}\text{NP}\left[\text{AGR}\boxed{2}\right], \text{AP}\left[\text{PRD} +, \text{AGR}\boxed{2}, \text{SUBCAT}\left\langle\boxed{1}\right\rangle\right]:\boxed{3} \right\rangle \end{bmatrix} \\ \text{CONTENT } \boxed{3} & \end{bmatrix}$$

(132) *komme*

$$\begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD } \textit{verb} \\ \text{SUBCAT} \left\langle \text{NP}_{der}, \boxed{1}\text{NP}, \text{GERP}\left[\text{PRD} +, \text{SUBCAT}\left\langle\boxed{1}\right\rangle\right]:\boxed{3} \right\rangle \end{bmatrix} \\ \text{CONTENT } \boxed{3} & \end{bmatrix}$$

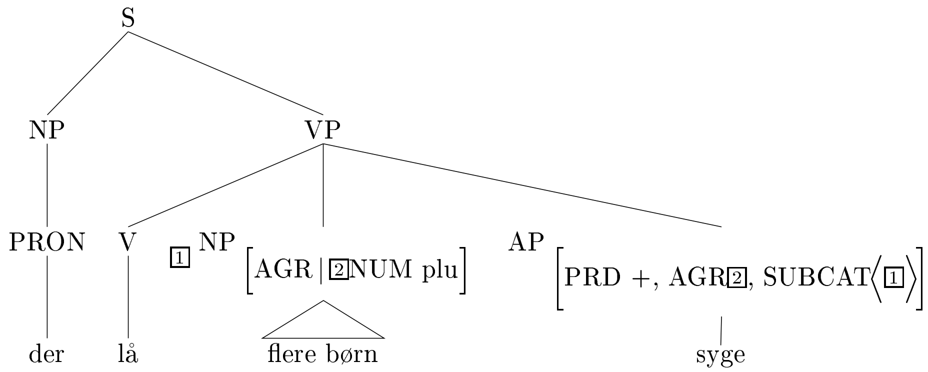
Note that the predicative element, an adjectival phrase in the first case and a gerundive in the second, subcategorises for the postponed NP, and that the content of the whole ES is identical with the content of the predicative element combined with the postponed NP it subcategorises for. In other words, the

[23] For a more general refutation of the concept of small clause, see Chapter 3 in Pollard & Sag (1994).

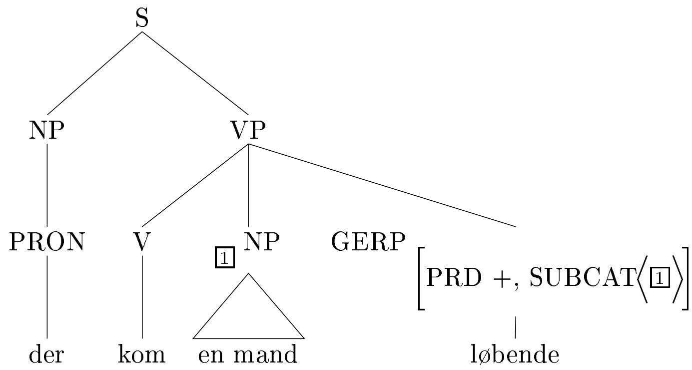
pleonastic *der* does not contribute to the content of the sentence. Furthermore, agreement features[24] are structure-shared between the postponed NP and the predicative AP in the entry for *ligge*.

The two sentences quoted above would then receive the following representations:

(133) a. Der lå flere børn syge.



b. Der kom en mand løbende.



The difference between the cases subsumed by Lumsden under the small clause analysis, and those where the tail contains an element which seems more loosely attached to the postponed NP, can then be captured in functional rather than configurational terms. Consider thus the following two Danish examples:

(134) a. Der var en del børn med snavsede fingre.
b. * Med snavsede fingre var der en del børn.
(lit: There were a lot of children with dirty fingers.
With dirty fingers were there a lot of children.)

(135) a. Der sad ikke en eneste person på caféen.
b. På caféen sad der ikke en eneste person.
(lit: There sat not one single person at the café.
At the café sat there not one single person.)

Med snavsede fingre (with dirty fingers) in the first of the two sets of examples is a predicative PP that cannot be topicalised. On the analysis proposed here, it will be treated as a complement of the verb and subcategorise for the postponed NP just as the AP in example (136a). *På caféen*, on the other hand, is not a predicative PP, and can be topicalised freely. It will be analysed as a sentential adjunct. As a consequence, it will not subcategorise for the postponed NP.

[24] Pollard and Sag advocate a semantic treatment of agreement. Here, however, agreement is treated as a syntactic feature.

The postponed NP

It has been claimed by a variety of grammarians and linguists that the postponed NP in ES must be indefinite. A relatively recent example is that of Rochemont & Culicover (1990), who claim that English ES are subject to the so-called Definiteness Effect, as shown by the following examples:

(136) a. * There was John standing by the window.
b. There was a man standing by the window.

They also note, however, that such a constraint does not apply to what they call presentational *There* constructions – OV ES in Milsark's terminology:

(137) There walked into the room, the man she had no desire to see.
(Rochemont & Culicover 1990:29)

Rochemont and Culicover's main focus is in fact on presentational *There* constructions, which they show to be different from other types of ES in several respects. As we shall see, however, the constraint at play cannot be stated in terms of morphosyntactic definiteness. In fact, we shall be speaking of Quantification Effect (QE), rather than Definiteness Effect (DE).

In the Danish grammatical tradition, it has also been claimed that the postponed NP in ES must be indefinite, although examples where it is syntactically definite but semantically indefinite are observed, e.g.:

(138) Der voksede de smukkeste blomster i haven.
(There grew the most beautiful flowers in the garden.)
(Heltoft 1986:2)

In fact, it is not only the definite article, but a whole set of determiners that do not seem able to occur in ES, as shown by the following examples:

(139) a. Der er mange/nogle/få børn på legepladsen.
(There are many/some/few children on the playground.)
b. ? Der er alle børnene/de fleste børn/børnene på legepladsen.
(There are all the children/most children/the children on the playground.)

Based on whether they can occur in ES or not, a distinction is made by Milsark (1974) between strong and weak quantifiers. Thus, according to this distinction, strongly-quantified determiners are not allowed in ES: this is the QE mentioned earlier. Milsark also points out that the same distinction is useful in explaining a constraint on the occurrence of different types of predicates. In fact, a weak quantifier cannot occur freely in an NP functioning as the subject of a predicate denoting a property, whereas state predicates do not place any restriction on the quantification of the subject:

(140) a. ? A man is intelligent/has red hair.
b. A man is injured/is drunk.
(Lumsden 1988:112)

Strong quantifiers are not subject to the same restriction:

(141) a. Every man is intelligent/has red hair.
b. Every man is injured/is drunk.
(Lumsden 1988:112)

Lumsden argues that the strong-weak distinction can be rephrased in terms of a distinction between presuppositional and non-presuppositional quantifiers. This can be shown by embedding the two quantifier types in a context where ordinary entailments disappear, whilst presuppositions survive:

(142) a. It isn't true that I took a picture of a/some/three/many/a lot of unicorn(s).

b. It isn't true that I took a picture of both/the/all the/most/every/all unicorn(s).

(Lumsden 1988:142)

Thus in the (a) examples, where the final NP is weakly quantified, the existence of the unicorn(s) is not presupposed. It is, on the other hand, in the (b) examples, where the NP is strongly quantified. The same opposition can be found in Danish:

(143) a. Det passer ikke, at du har set en/nogen/tre/mange/en del heks(e).
(It isn't true that you have seen a/some/three/many/a lot of witch(es).)

b. Det passer ikke, at du har set begge hekse/heksen/alle heksene.
(It isn't true that you have seen both witches/the witch/all the witches.)

The existence of the witch or witches is not presupposed in the (a) examples, whilst it is in all the sentences in (b). Lumsden notes that some weak quantifiers are given a strong interpretation when stressed. In the following sentence, for example, the stressed quantifier indicates that the NP carries an existential presupposition, and denotes a subset of the set of unicorns the existence of which is presupposed in the discourse:

(144) If you try to take a picture of SOME unicorns, they get frightened.

(Lumsden 1988:145)

He also notes that the weak interpretation is always favoured in ES. Thus, if *some children* is the postponed NP in an ES, it cannot be stressed:

(145) a. There were some children at the party.

b. ? There were SOME children at the party.

The example above is only acceptable if *some children* is understood as a certain number of children, without reference to a presupposed set. Thus, the general thesis that only weak quantifiers are allowed in ES seems confirmed.

A convincing explanation of why this is the case has been offered, in our opinion, by Dahl (1974), who claims that the inability of the postponed NP in ES to carry existential presupposition, is consistent with an analysis of ES as detopicalising constructions. As we saw in Chapter 4, the primary topic of a sentence is linked with an existential presupposition, and both in English and Danish a definite grammatical subject is perceived as the primary topic. If the subject is an indefinite NP, on the other hand, it does not carry an existential presupposition, and cannot easily be identified with the primary topic of the sentence. In fact, Dahl claims that in this case the sentence is topicless.

ES constitute another example of topicless sentences, i.e. they describe a situation rather than predicating some properties of individuals. The postponed NP has undergone a process of *detopicalisation*. This is reflected by the fact that it cannot carry a presupposition.

Lumsden shows, however, that a strongly quantified NP can occur as the postponed NP in an ES:

(146) There is this car at every dealer's in the country.

(Lumsden 1988:167)

He argues that this is possible when the NP does not denote an individual, or a token reference, but rather a type. Thus, it does not presuppose the existence of a concrete individual (or car in this case) but of a specific type of car.

There are other cases, however, where this explanation does not seem to fit. As noted i.a. in Nimb (1993), strong quantifiers can occur in ES when they are preceded by a focus adverbial:

(147) a. ? Der er vinen tilbage.
(There is the wine left.)

b. Der er kun vinen tilbage.
(There is only the wine left.)

On Nimb's theory, however, focus adverbials make a strong quantifier become weak. Consequently, such examples do not contradict the general pattern according to which ES obey the QE. More crucially from our perspective, focus adverbials clearly mark the NP they precede as being the focus of the sentence. Our claim is, in fact, that the postponed NP in ES cannot be the primary topic of the sentence, but belongs instead to its focus. Therefore, the fact that the postponed NP in ES can be preceded by a focus adverbial, is consistent with our theory.

A clear exception to the QE is the use of definite NPs in the so-called enumerative ES (Milsark 1974). A characteristic of this type of ES is that the postponed NP is a strongly quantified NP denoting one or more members of a set e.g.:

(148) What can we have to drink?
There's the beer we brought with us.

(Lumsden 1988:217)

Although they escape the QE, these examples do not challenge Dahl's view of ES as topicless sentences. Clearly, if interpreted literally as predicating the existence of an NP which is already presupposed, such examples do not convey much information. Rather, they express a state of affairs, or a situation in Dahl's terminology, which is relevant to the discourse, and they are informative in that they further restrict a set mentioned previously. Furthermore, as noted by Lumsden, enumerative ES convey an implicature[25]. Let us compare the example above with an apparently synonymous variant:

(149) What can we have to drink?
The beer we brought with us.

In both cases, *the beer* is in fact the focus in the answer, but unlike the more concise response, the enumerative ES version provides an implicature which can be cancelled, namely the inference that 'we can drink the beer':

(150) What can we have to drink?
There's the beer we brought with us, but it's lukewarm.

(151) What can we have to drink?
? The beer we brought with us, but it's lukewarm.

In the second example, the concluding remark seems to contradict the first part of the answer. Thus in enumerative ES, unlike in the corresponding shorter responses, the relevance of the referents brought to the foreground can be questioned or cancelled. Furthermore, the existence of the set they are members of is not presupposed. This is clearly shown by Heltoft (1986), who mentions the following examples:

[25] By implicature we mean an inference characterised i.a. by the property of cancellability (for a discussion see Levinson 1983).

(152) Ligger der nogen sejlskibe i havnen?
a. Ja, der ligger Danmark og Georg Stage.
b. ? Ja, det er Danmark og Georg Stage.

(Are there any ships in the harbour?
a. Yes, there are the Danmark and the Georg Stage.
b. Yes, it is the Danmark and the Georg Stage.)

(Heltoft 1986:14)

The set 'ships in the harbour' is obviously not presupposed in the discourse, since the very existence of such a set is the point of the question. Therefore, (b) is not a proper answer: as we have seen, clefts carry presuppositions. On the other hand, (a) does not carry a presupposition, and is thus a correct answer.

All the data discussed this far seem to confirm the general thesis that the postponed NP in ES cannot be the primary topic of the sentence. Contrary to what is characteristic of the primary topic of a sentence, in fact, it cannot presuppose the existence of a concrete individual, unless it is preceded by a focus adverbial clearly indicating its focus status.

Some interesting additional evidence pertinent to this claim has been provided by Heltoft. He notes that a sentence where a strongly quantified NP occurs as the postponed NP in an ES, is acceptable if the locational adverb is topicalised. Consider thus the following two examples:

(153) a. ? Der er pengene i kassen.
b. I kassen var der pengene og de ubetalte regninger.

(a. There is the money in the box.
b. In the box there were the money and the unpaid bills.)

(Heltoft 1986:20)

Topicalising the adverbial *i kassen*, makes it possible for the logical subject *the money and the unpaid bills* to be placed in end-focus position, where it clearly functions as the focus of the sentence. Thus, as in the case of strongly-quantified nominals preceded by a focussing adverbial, the QE can be ignored.

Now recall that in a question answer sequence, the focus of the answer always concides with the new information. Thus, by embedding a sentence in the appropriate question answer sequence, one can investigate which of the sentence constituents may be in focus. The test can be used to determine where the focus may fall in ES. Consider the following sequences:

(154) Hvad skete der?
Der løb en stor flok journalister ind i forhallen.
(What happened?
There ran a large group of journalists into the lobby.)

(155) Hvem løb hvor?
Der løb en stor flok journalister ind i forhallen.
(Who ran where?
There ran a large group of journalists into the lobby.)

(156) Hvem løb ind i forhallen?
Der løb en stor flok journalister ind i forhallen.
(Who ran into the lobby?
There ran a large group of journalists into the lobby.)

(157) Hvor løb en stor flok journalister?
? Der løb en stor flok journalister ind i forhallen.
(Who ran into the lobby?
There ran a large group of journalists into the lobby.)

The fact that the sentence can occur in the first three contexts above shows that the focus may include the NP following the verb, and may or may not also include the verb and the final adverbial. The last example shows that the focus not only may but must include the NP following the verb. Thus, our claim is not only that the postponed NP is not the primary topic but that it must be the focus of the sentence (or at least part of it).

The particular information structure of ES has a number of interesting consequences. As noted by Sundman (1980), detopicalising the subject has an effect on quantifier scoping. Consider thus the following two examples:

(158) a. To børn deltog i begge aktiviteter.
b. Der deltog to børn i begge aktiviteter.
(lit: a. Two children participated in both activities.
b. There participated two children in both activities.)

In the (a) version, the preferred reading is that the same two children participated in both activities. This is consistent with the interpretation of the subject *to børn* as the primary topic of the sentence, identifying a set of two concrete individual children which the rest of the sentence is predicated about. In the (b) version, on the other hand, the preferred reading is that two children, possibly but not necessarily the same two, participated in both activities. The sentence is topicless, and expresses a situation rather than predicating something of the primary topic. Slightly more formally, we can say that the quantifier preceding the subject has scope over the quantifier preceding the adverbial in (a), while the opposite is true in (b).

Consider now the much quoted examples below:

(159) a. Two languages are spoken by [everybody in this ROOM]$_F$.
b. Everybody in this room speaks [two LANGUAGES]$_F$.

In both sentences, the quantifier preceding the focussed constituent has narrower scope than the quantifier in subject position. Thus, the postponed NP in existential sentences shows a similar behaviour to that of the focus of a sentence with respect to quantifier scoping.

Another observation pointing in the same direction, still due to Sundman, regards an effect on the scope of the negation operator:

(160) a. Der er ikke nogen børn på legepladsen (men der er måske en hund).
b. Børnene er ikke på legepladsen (men de er et andet sted).
(a. There are no children on the playground (but there is perhaps a dog).
b. The children are not on the playground (but they are in some other place).)

In (a), the negation has scope over the subject NP, whereas the subject is outside the scope of the negation in (b). Since the scope of the negation operator generally includes the focus of the sentence, this is again in accordance with our theory that the postponed NP in ES should belong to the focus of the sentence.

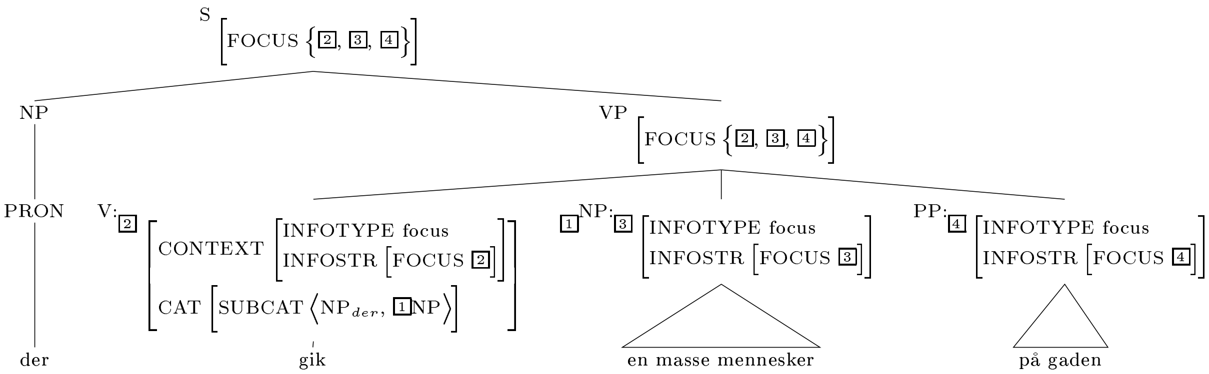
Automatic analysis and generation of Danish existential sentences

Automatic analysis of Danish ES will yield the syntactic structure proposed above. Furthermore, the feature [INFOTYPE focus] will be assigned according to the following principle:

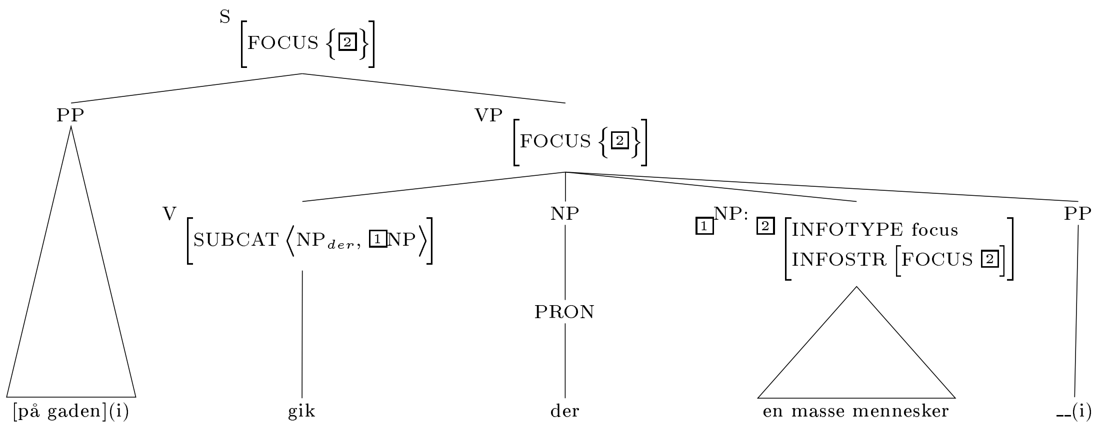
- all the constituents following the pleonastic *der* will be assigned [INFOTYPE focus]

This will yield for example the following two analyses:

(161) Der [gik en masse mennesker på gaden]$_F$.
(lit: There walked a lot of people on the street.)



(162) [På gaden]$_{PT}$ gik der [en masse mennesker]$_F$.
(lit: On the street walked there a lot of people.)



Thus, we take the default focus assignment in the non-topicalised sentence to be such that the whole VP is in focus, whereas in the topicalised sentence the verb, which is placed to the left of the pleonastic pronoun, is not part of the focus. For more details about the coindexation between the topicalised adjunct and the PP trace, cf. Section 5.3.4.

As far as generation is concerned, for an ES to be generated in Danish the following constraints must be satisfied:

- the postponed NP must be focussed (i.e. marked by the feature [INFOTYPE focus]);
- alternatively, if information about focality is not available, the postponed NP must be weakly quantified;
- the main verb must be either an active form of an unaccusative or motion verb, or a passive form.

In addition, the correct syntactic structure must be available, which means that the constituent following the postponed NP in a Danish ES must be either a PP, or an AP, or null. If the main verb is a motion verb, a gerundive is allowed as well. The precise subcategorisation pattern is stated, as mentioned earlier, in the lexical entry of the relevant verb.

These criteria will allow us, for example, to generate the following Danish ES (all quoted previously):

(163) a. Der kom en mand løbende.
b. Der sker ikke ret meget for tiden.
c. Der er blevet udført et stort arbejde i dag.
d. Der var en del mennesker i byen i går.

(lit: There came a man running.
There happens not much these days.
There has been done a lot of work today.
There were a lot of people in town yesterday.)

All of them sound better than their counterparts without *der*, listed below[26]:

(164) a. En mand kom løbende.
b. * Ikke ret meget sker for tiden.
c. Et stort arbejde er blevet udført i dag.
d. En del mennesker var i byen i går

(lit: A man came running.
Not much happens these days.
A lot of work has been done today.
A lot of people were in town yesterday.

5.3.3 Topicalisation

By *topicalisation* in Danish we mean the fact that a constituent other than the grammatical subject – and ignoring complementisers and conjunctions – is placed in theme position. When this happens, the neutral subject-verb order is inverted. The term topicalisation implies that the fronted constituent would be placed in a different position in a non-topicalised sentence. In fact, in the representation of a topicalised construction, a link will be created between the topicalised constituent and the position the same constituent would occupy in the corresponding unmarked sentence. We call this position the *trace* of the topicalised constituent. In the following example, the trace is indicated by an underscore, and the link between the *topicalised* (also called *fronted*, or *extracted*) constituent is indicated by way of an index:

(165) Ham$_i$ vil jeg i hvert fald ikke have noget med $\_\__i$ at gøre.

(Lit: Him$_i$ want I really not have anything with $\_\__i$ to do.
Him I really don't want to have anything to do with.)

Thus, complementisers and conjunctions, which can only be placed in sentence-initial position, and after which the normal subject-verb order is maintained, are not affected by topicalisation.

All the following sentences are examples of topicalisation:

(166) a. Ida har vi glemt at invitere.
b. Maria kan jeg ikke lide at lege med.
c. Smuk er hun.
d. I går spiste vi kylling.

(Ida we have forgotten to invite.
Maria I don't like to play with.
Beautiful she is.
Yesterday we had chicken.)

[26] In some text types, however, the preference may be for non-existential sentences. For example *Dankort accepteres ikke her* (Dankort is not accepted here) is more idiomatic on a sign than its existential counterpart *Der accepteres ikke Dankort her*.

Topicalisation corresponds thus to marked theme in declarative clauses in Halliday's terminology (cf. Section 3.2).

Topicalisation plays at least three different pragmatic functions in Danish. One is that of achieving a contrastive effect. As discussed in Section 4.1.3, this contrastive effect is present when the topicalised constituent is a valency-bound complement, and a full NP. Any complement can be topicalised: the sentences in (166a-c) display topicalisation of the object, of a prepositional object and of the subject complement. Another pragmatic effect can be described as the creation of a setting against which the rest of the sentence should be understood. This is exemplified in (d) above, which is a case of adjunct topicalisation. Adjunct topicalisation exhibits slightly different syntactic properties than complement topicalisation, and it is in fact not obvious that this construction is best described in terms of topicalisation. We shall discuss this in detail in the next section. A third effect can be seen at work in examples where the focus of the preceding sentence is taken up as the primary topic by topicalising a weak personal pronoun, as shown below:

(167) For måske var det Ivan. Ham frygtede vi.
(Because maybe it was Ivan. Him we feared.)
(*DK87*)

In all cases, we consider the topicalised constituent the primary topic of the sentence.

Complement and adjunct topicalisation have different distributions. Adjunct topicalisation is in fact very common both in spoken and written language. Therefore, it should clearly be included in the coverage of any analysis or generation algorithm aiming at dealing with realistic texts. Complement topicalisation, on the other hand, is more restricted to spoken language[27], so that a system having to deal with written texts probably needn't handle it.

In the following sections, first we discuss the grammar of topicalisation in Danish, then we deal with the pragmatic function of topicalisation, and finally we discuss how this construction should be treated in automatic analysis and generation.

The grammar of topicalisation in Danish

Like clefts, which were discussed in Section 5.3.1, topicalisation is a UDC, more specifically a UDC of the strong type, also called a filler-gap construction, where the topicalised element fills a gap (or trace) in the rest of the sentence.

The following examples show that the dependency between the topicalised constituent and the trace can extend over a theoretically unlimited number of sentence boundaries:

(168) a. Ida$_i$ har vi glemt $\_\__i$ at invitere.
b. Ida$_i$ er jeg bange for, vi har glemt $\_\__i$ at invitere.
c. Ida$_i$ er jeg bange for, far har sagt, at vi har glemt $\_\__i$ at invitere.
(Ida we have forgotten to invite.
Ida I'm afraid we have forgotten to invite.
Ida I'm afraid daddy said we have forgotten to invite.)

As mentioned above, it is not clear that the same characterisation can be applied to adjunct fronting. Consider thus the following sentence:

(169) I går gik vi en tur i parken.
(Yesterday we went for a walk in the park.)

[27] However, as pointed out to me by Elisabeth Engdahl, fronting of weak pronouns is not uncommon in children's literature.

It seems counterintuitive to claim that there should be a gap in the sentence, as the initial place is perceived as the default one for the temporal adjunct[28]. However, temporal adjuncts – and indeed adjuncts in general – can be extracted from an embedded sentence:

(170) I morgen$_i$ tror jeg ikke jeg kommer på arbejde $\_\_i$.
(Tomorrow I don't think I'm coming to work.)

Clearly, here the initial adverbial belongs to the nominal clause, so it seems reasonable to posit a gap in sentence-final position. The issue is dealt with at some length in Pollard & Sag (1994). They note that in English, the possibility of adjunct extraction has been claimed to follow from properties of a particular set of verbs, called bridge verbs. Thus, in the following examples the fronted adjunct is clearly extracted from the embedded clause:

(171) a. How long ago do you believe that John left?
b. Why do you suppose that John burned down the building?
c. When do you think your husband is planning to leave you?
(Pollard & Sag 1994:386)

One of the properties of bridge verbs which would make it possible for them to participate in such constructions would be their ability to function in parenthetical tags:

(172) a. How long ago did John leave, do you believe?
b. Why did John burn down the building, do you suppose?
c. When is your husband planning to leave you, do you think?
(Pollard & Sag 1994:386)

On the other hand, adjunct extraction seems to be excluded with other verbs:

(173) a. * How long ago will you deny that John left?
b. Why do you doubt that John burned down the building?
? To collect the insurance.
c. When do you deny that your husband is planning to leave you?
? When he gets enough money to manage on his own.
(Pollard & Sag 1994:385)

However, at least (173a) and (173c) are semantically ill-formed not so much because an interpretation of the adjunct as extracted from the embedded sentence is impossible per se, but because the semantics of the verb and those of the sentence from which the adjunct is extracted, contradict one another. Thus, whilst it makes sense to deny that a certain event took place or is going to take place, it is nonsense to deny when it did or when it is going to. In fact, Pollard and Sag note that examples where adjunct extraction works well together with non-bridge verbs can be constructed:

(174) a. While their parents are in town next week, I doubt that the twins will be able to attend any lectures.
b. During my term as president of this university, I categorically deny that there were any improper appropriations of any federal funds.
c. With no tools but a crowbar and a ballpeen hammer, I'm not very confident that Butch is going to be able to fix that disk drive.
(Pollard & Sag 1994:386)

Turning to Danish data, we find that examples of adjunct extraction from an embedded sentence are possible with bridge as well as non-bridge verbs:

[28] Recall, however, that Halliday treats preposed adjuncts in English as marked themes (cf. Section 3.2).

(175) a. I morgen tror jeg ikke, vi kan komme.
b. I morgen er jeg meget i tvivl, om vi kan komme.
(Tomorrow I don't think we can come.
Tomorrow I doubt very much that we can come.)

Therefore, we can only agree with Pollard and Sag that fronted adjuncts may be extracted from an embedded sentence, and that adjunct extraction is subject to complex semantic and pragmatic constraints.

Recall that our initial objection against considering adjunct fronting a case of topicalisation was that in cases where there is no embedded sentence, there does not seem to be a trace in the sentence. This point deserves a little more discussion.

In fact, not all adverbial types behave alike in this respect. If the adjunct is understood as modifying only the VP instead of the whole sentence, for example, its being placed in theme position is more clearly perceived as a marked option. This is for instance the case with manner adverbials, as shown in the following sentence:

(176) Smukt kan hun i hvert fald ikke synge[29].
(Lit: Beautifully can she at any rate not sing.)

Here, topicalisation is only possible in a contrastive reading. Thus, from a pragmatic point of view, such examples are similar to cases of complement topicalisation.

Also in the case of sentence adjuncts, however, fronting can be perceived as marked:

(177) Aldrig før har jeg set så dårlig en forestilling.
(Never before have I seen such a bad performance.)

In the Danish grammatical tradition, adverbials are usually grouped in three different types, depending on the sentence position in which they can occur. The three types are:

- **central adverbials** (*centraladverbialer*), which in a non-topicalised, declarative main sentence occur right after the finite verb, e.g.:

  altså (then), *da* (then), *jo* (certainly), *nemlig* (in fact), *ikke* (not), *aldrig* (never), *altid* (always);

- **final adverbials** (*slutadverbialer*), which in a non-topicalised, declarative main sentence occur in sentence-final position, e.g.:

  rundt (round), *retur* (back), *ud* (out), *væk* (away), *op* (up);

- **free adverbials** (*frie adverbialer*), which in a non-topicalised, declarative main sentence can occur after the finite verb or in sentence-final position, e.g.:

  alligevel (yet), *allerede* (already), *bagefter* (later), *igen* (again), *i går* (yesterday), *i overmorgen* (the day after tomorrow).

Adverbs belonging to all three types can generally also occur in theme position. Those belonging to the third type, however, are probably also the ones for which fronting is perceived as a non-marked position. This, however, is only a hypothesis which should be validated by further research.

[29]The fact that the adjunct is part of the VP constituent is shown i.a. by the fact that verb and adverb can be topicalised together:

Synge smukt kan hun i hvert fald ikke.
(Lit: Sing beautifully can she at any rate not.)

.

In conclusion, since adjunct extraction from an embedded sentence is possible provided that all the semantic and pragmatic constraints are satisfied, and since, at least for some adjunct types, fronting is perceived as a marked option in opposition to a 'default' word order, we shall consider adjunct fronting a case of topicalisation in general. What this implies in terms of grammar rules and representations will be discussed in brief. But first, let us look at the syntactic representation of complement topicalisations.

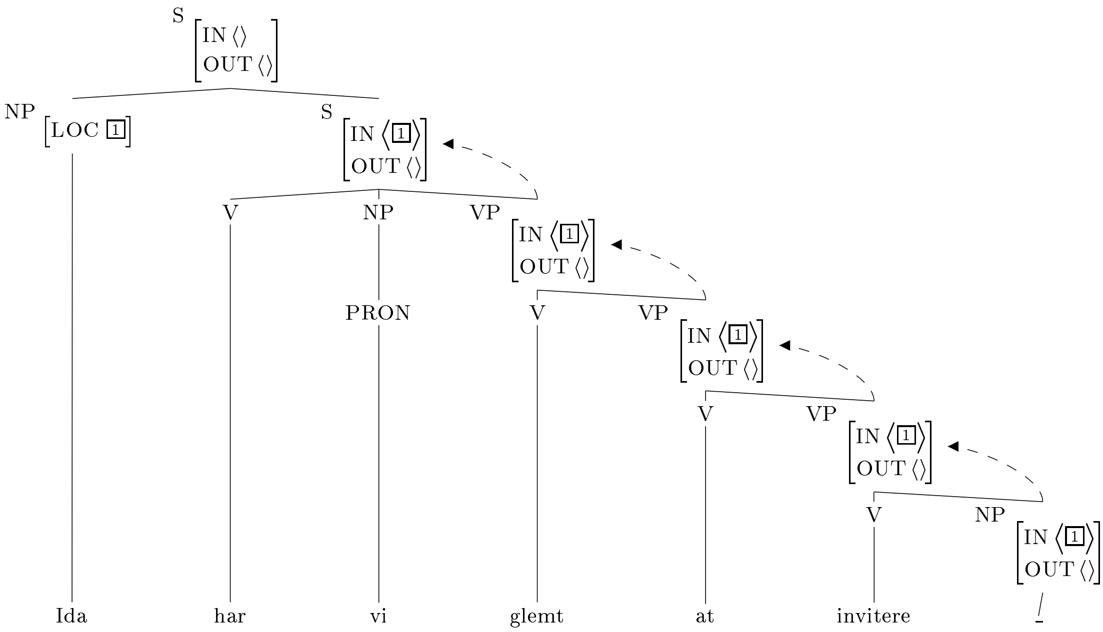
As with cleft sentences (see Section 5.3.1), the non-local feature GAPS will be used to represent the dependency between the extracted element and the trace. At the bottom of the dependency, in the place where the complement is missing, we introduce a trace, which will be recalled is a feature structure where all the values for local and non-local features are empty, but where the feature LOCAL shares an index with the GAPS_IN feature:

(178) $$\begin{bmatrix} \text{PHON} & \langle\rangle \\ \text{SYNSEM} & \begin{bmatrix} \text{LOCAL} & \boxed{1} \\ \text{NONLOCAL} & \begin{bmatrix} \text{GAPS} & \begin{bmatrix} \text{GAPS\_IN} & \left\langle \boxed{1} \mid \boxed{2} \right\rangle \\ \text{GAPS\_OUT} & \left\langle \boxed{2} \right\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The LOCAL features of the trace will be instantiated according to the subcategorisation restrictions imposed by the governor and passed upwards in the structure through the GAPS_IN feature. This will be bound off when a filler is found. In HPSG, the Head-Filler Rule is responsible for bounding off the SLASH feature. In a different framework, the same effect can be achieved by a similar grammar rule. (Cf. chapter 6, where ALEP rules dealing with gapping are shown.)

The mechanism can be seen at work in the following representation of one of the topicalisation examples mentioned previously:

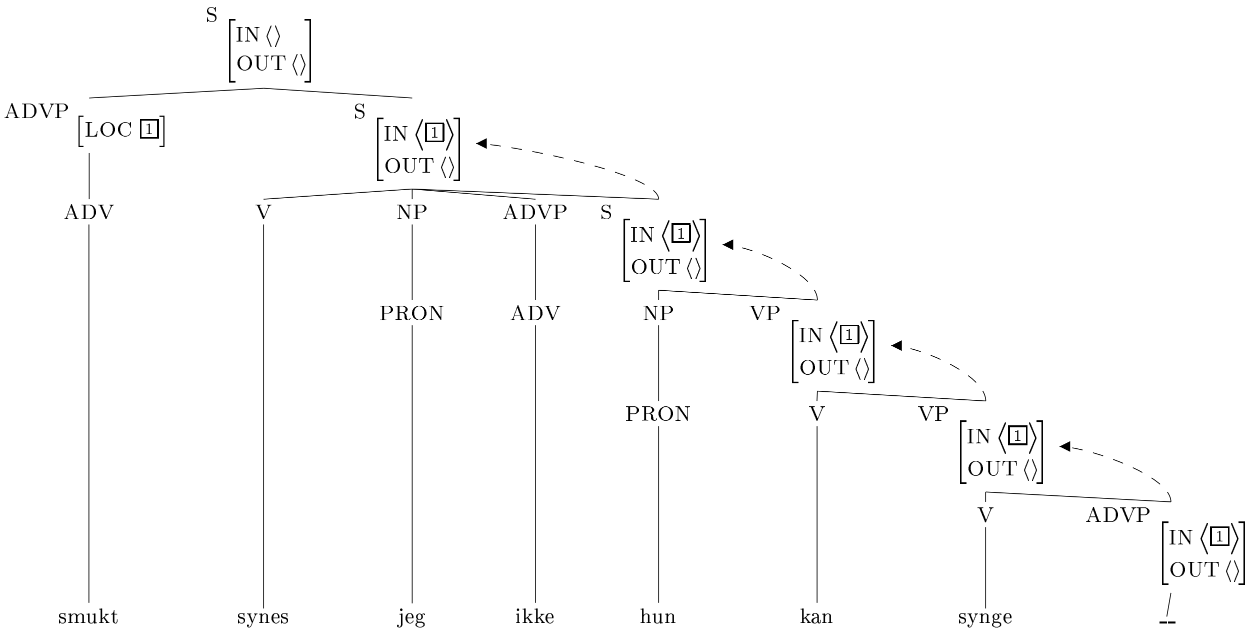
(179) Ida har vi glemt at invitere.
(Ida we have forgotten to invite.)



(NB: For reasons of space, LOCAL has been replaced by LOC, while IN and OUT stand for GAPS_IN and GAPS_OUT.)

Let us now see what the representation of an adjunct topicalisation can look like:

(180) Smukt synes jeg ikke hun kan synge.
(Beautifully I don't think she can sing.)



The mechanism seems at first sight to be functioning in the same way in the two examples. In the case of the adverbial however, the trace is not subcategorised for. This is of course a direct consequence of the fact that an adjunct by definition is not a valency-bound element. Thus, the insertion of a trace in adjunct position (which for Danish will be either in sentence-final position or soon after the finite verb) risks making the grammar very inefficient: there is in fact a risk that adjunct traces will be inserted in every sentence to be then discarded if an adequate filler is not found. Therefore, although the treatment of adjunct topicalisation in terms of extraction seems theoretically well-founded, it may turn out to be impossible in a concrete implementation.

To conclude our discussion of the grammar of topicalisation in Danish, the issue of topicalisation out of relative clauses deserves some attention. It is in fact a peculiarity of topicalisation in the Scandinavian languages in general, and Danish in particular, that the topicalised element can be extracted out of a relative clause, as shown in the examples below (Erteschik-Shir 1982:176):

(181) a. Det$_i$ er der mange$_j$ der$_j$ kan lide $\_\__i$.
(That there are many who like.)

b. Det$_i$ kender jeg mange$_j$ der$_j$ har gjort $\_\__i$.
(That I know many who have done.)

Various proposals have been made in the literature concerning syntactic, semantic and pragmatic constraints on extraction out of complex NPs (i.e. NPs modified by a relative clause or a sentential complement) in the Scandinavian languages. Of these, only one syntactic constraint seems to hold, namely the Complex NP Subject Constraint, described in Allwood (1982). Alwood claims that there is a clear difference between sentences where the relative pronoun in the embedded clause is the subject, and those where it isn't[30]. If the relative pronoun is a subject, in fact, extraction is acceptable. Otherwise, it is ungrammatical (or at least very awkward). Compare thus two examples quoted by Allwood: in the former,

[30] Note that this restriction is not the same as the Subject Condition described e.g. in Pollard & Sag (1994). The Subject Condition is in fact a constraint on extraction which holds for English, stating that a subject can only be slashed if another slashed complement is present, to explain the ungrammaticality of e.g. **That was the rebel leader who rivals of shot the British consul* contraposed to the correct *That was the rebel leader who rivals of shot* (p.183). This constraint does not hold for Scandinavian languages, cf. the Swedish sentence: *Vem står numret till inte i katalogen?* (lit: Who is the number of not in the book?) quoted in Engdahl (1982:165), where the interrogative pronoun "vem" fills a gap in the subject *numret till __*.

the relativised constituent is the subject, and multiple extraction is possible; in the latter, the relativised constituent is the direct object, and extraction makes the sentence quite difficult to understand:

(182) a. En pojke$_i$ känder jag till en flicka$_j$ som$_j$ __$_j$ gav __$_i$ en kyss på Röda Torget.
(lit: A boy know I of a girl who gave a kiss in Red Square.
I know a girl who gave a boy a kiss in Red Square.)

b. ? En flicka$_i$ känder jag till en bok$_j$ som$_j$ __$_i$ gav en pojke __$_j$.
(lit: A girl know I of a book which gave a boy.
I know a book which a girl gave a boy.)

(Allwood 1982:18,22)[31]

The same difference can be observed in the following two Danish examples, which parallel Allwood's Swedish ones:

(183) a. Børn$_i$ er der mange forretninger$_j$ der$_j$ sælger is til __$_i$.
(lit: Children are there many stores there sell ice-cream to.
There are many stores that sell ice-cream to children.)

b. *Børn$_i$ er der mange slags is$_j$ som$_j$ forretningen sælger __$_j$ til __$_i$
(lit: Children are there many types of icecream which the store sells to.
There are many types of icecream which the store sells to children.)

(Erteschik-Shir 1982:187)

However, extraction out of a relative clause where the relativised element is not a subject is not totally impossible, as shown by the following example (the Danish version of an example quoted for Swedish by Andersson 1982):

(184) Selv småbørn$_i$ er der utrolig meget$_j$ man kan lære __$_i$ at gøre __$_j$.
(lit: Even small children is there an awful lot one can teach to do.
There is an awful lot you can teach even small children.)

Note also that this is an example of crossing dependencies.

In conclusion, then, extraction out of relative clauses is allowed by the grammar of Danish with the possible restriction that the relative pronoun must be the subject. However, as we shall see in the next section, complement topicalisation is only possible in the right context, namely if the sentence can be given a contrastive interpretation. This constraint is particularly evident in cases of multiple extractions like the ones just discussed. Out of context, in fact, such examples can be quite difficult to understand.

The pragmatic functions of topicalisation

As mentioned earlier, we distinguish two different pragmatic functions of topicalisation in Danish. One is that of expressing a contrast. As noted by Togeby (1992), and as discussed in Section 4.1.3, this contrast can usually be described as a contrast between alternative pairs of contrastive topic and focus. Let us thus consider a few naturally occurring examples from the Bergenholtz *DK87* corpus for Danish, all showing topicalisation of a valency-bound complement.

(185) Offentligheden fik dengang kendskab til, at han bl. a. blev dømt for at have fået den 29-årige pige til at udsmugle dokumenter og pjecer fra Udenrigsministeriet. [Nogle af papirerne]$_{PT}$ gemte hun i sine sko.

(The public was then made acquainted with the fact that, among other things, he was found guilty of getting a 29 year old girl to smuggle documents and leaflets from the Foreign Ministry. Some of the papers she hid in her shoe.)

[31] Indices and gaps are ours.

(186) Vi bliver faktisk meget bedre, hvis vi får lidt tilbage. Særlig under de morsomme scener. [Bare et lille pip]$_{PT}$ ville vi være glade for.

(We actually become a lot better if we get a little response. Especially in the funny scenes. Even a little peep we would appreciate.)

(187) Michael Carøe som Jørgen Herming ser godt ud, [synge]$_{PT}$ kan han også omend lidt anstrengt til tider, men ...

(Michael Carøe looks good in the role of Jørgen Herming, sing he can too, even if [he is] sometimes a little tense, but ...)

In the first example, the set of alternative pairs could comprise sentences such as:

(188) Andre $_{PT}$ [gemte hun under kjolen]$_F$.
Andre$_{PT}$ [smed hun væk]$_F$.
...
(Others she hid under her dress.
Others she threw away.)

In the second it could include:

(189) Latter$_{PT}$ vil vi [foretrække]$_F$.
En klapsalve$_{PT}$ vil vi [elske]$_F$.
...
(Laughter we would prefer.
An applause we would love.)

And the third:

(190) [Bevæge sig på scene]$_{PT}$ kan han [ikke]$_F$.
(Move on the scene he can't.)

In all the examples above, we have marked the topicalised constituent as the primary topic of the sentence. This is consistent with the fact that topicalised nominals cannot express the new information in a question answer sequence:

(191) Hvad spiste I i går?
a. ? Kylling fik vi.
b. Vi fik kylling.

(What did you eat yesterday?
a. Chicken we had.
b. We had chicken.)

Thus, for the sequence above to be well-formed, the new information provided by the object *kylling* (chicken) cannot be placed in theme position.

The fact that the contrast implicit in a topicalised sentence involves a pair of elements is also noted by Erteschik-Shir (1982), who is concerned with the pragmatic constraints governing extraction out of relative clauses. She quotes the following example:

(192) Is kender JØRGEN mange der kan lide.

(Lit: Ice-cream knows Jørgen many who like.
Jørgen knows many who like icecream.)

(Erteschik-Shir 1982:179)

Her claim is that extraction is acceptable here if *Jørgen* bears contrastive stress, so that the ordered pair <*is Jørgen*> can form what he calls a dominant pair.

We have also seen, however, that such a contrastive interpretation is usually not available if the topicalised element is an adjunct of a certain type. For instance time, place and viewpoint adjuncts[32] often occur in theme position in Danish without triggering a contrastive reading. As discussed in Section 5.2.2, this seems especially to be the case of spatio-temporal adjuncts with general meaning, as opposed to more specific adjuncts that tend to occur in end-focus position. The pragmatic function they play is rather that of establishing a setting against which the sentence must be understood. Likewise, when the topicalised complement is a weak personal pronoun, no contrast is implied.

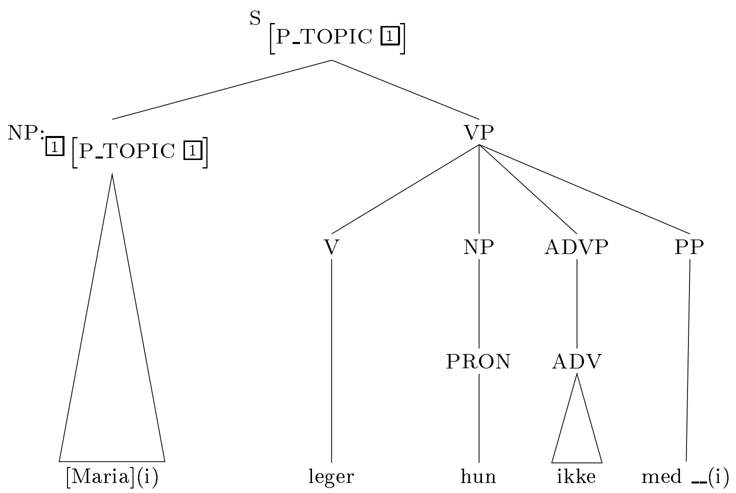
Analysis and generation of topicalised constructions

Our analysis of the information structure of topicalised constructions in Danish is very simple. The topicalised element is in fact always marked as the primary topic of the sentence, as shown below:

(193) Maria$_{PT}$ leger hun ikke med.
(lit: Maria plays she not with.
Maria she doesn't play with)

This would yield the following representation:

(194)



The focus and neutral parts of the sentence will be calculated by means of the heuristics proposed in Section 5.2.4.

As for generation, the following criteria seem relevant:

- if a complement other than the subject is marked as being the primary topic of the sentence, it should be topicalised;
- otherwise, if a general spatio-temporal and viewpoint adjunct (e.g. *i år* (this year), *i Danmark* (in Denmark), *efter min mening* (in my opinion), *med hensyn til ...* (with respect to ...)) is present, it should be generated in theme position, unless focussed in the source text.

Thus, the second criterion is a language-specific one, which does not relate the position of the adverbial in the target text to that of the corresponding adverbial in the source text. An alternative approach would be to record the information that a certain adjunct is placed in theme position in the input, and keep it in the same position in the output unless stronger preferences override this choice.

[32]See Section 5.2.3 for a semantic classification of adjunct types.

The first criterion, on the other hand, makes generation of a topicalised complement depend on the analysis of the same complement in the source text. This means that if the text to be translated into Danish is written in language like e.g. English, where in many cases a contrastive interpretation may be implicit, or where topicalisation of weak personal pronouns is not as used as in Danish, there will be no topicalisation in the Danish text. If translating from other languages – e.g. Italian, where contrast can be indicated syntactically by topicalisation, or Japanese, where contrastive topics are morphosyntactically marked – the generation of complement topicalisation constructions becomes a more realistic option. How this may work can be seen in two examples of translation from Italian into Danish. Both are taken from Jytte Lollesgaard (1989)'s translation of Elsa Morante's *L'isola di Arturo*. For each example, the (a) is the Italian original, and the (b) the Danish translation. Each of them is glossed in English, in a semi-literal way.

(195) a. Almeno dalle altre femmine, uno può salvarsi [...]; ma dalla madre, chi ti salva? (p.143)
(At least from other females, one can escape [...]; but from your mother, who can rescue you?)

b. Andre kvinder kan man i det mindste sende ad helvede til [...]; men din mor, hvordan slipper du af med hende? (p.126)
(Other women one can at least tell to go to hell [...]; but your mother, how do you escape from her?)

(196) a. Arturo [...] è rimasto a Procida senza di me mille volte, e non ha mai fatto storie. (p.139)
(Arturo has stayed in Procida without me a thousand times, and has never made a fuss.)

b. Arturo er jeg rejst fra de første tusind gange, [...] og han har i hvert fald aldrig lavet vrøvl. (p.123)
(Arturo I have left to travel a thousand times, and he has never made a fuss.)

In the former example, the prepositional object of the first clause, *dalle altre femmine* (from other females), is topicalised in Italian. The effect is clearly one of contrast between what one does with other females as opposed to what one does with one's mother, which is the contrastive topic in the second clause – *dalla madre* (from your mother). In the latter example, on the other hand, the primary topic of the first clause is the subject in the Italian original, *Arturo*. In the translation, a different predicate has been chosen by the translator, so that syntactically *Arturo* is now the object of the valency-bound preposition *fra*. However, in order to maintain it as the primary topic of the sentence, it is topicalised. Also in this case, the sentence has a contrastive interpretation, which is more explicit in the Danish version than in the Italian text.

5.3.4 Extraposition

The last syntactic construction we shall be discussing – extraposition of clausal complements – differs from the other phenomena dealt with so far. Although at first blush it seems to qualify as a focus marking construction, extraposition is in fact best explained in terms of syntactic complexity, probably in relation to processing preferences.

Extraposition is a term used to refer to constructions where a constituent usually associated with a different sentence position is placed at the end of the sentence. In this section, we are only concerned with examples involving the pleonastic pronoun *det*[33]:

(197) a. Det er rart at se dig.
(It's nice to see you.)

[33] We thus leave out extraposition out of complex NPs as in:

Jeg synes den bog var for dyr, som jeg en gang hørte dig tale så godt om.
(I think that book was too expensive, which I heard you speak so well about once.)

b. Det forekommer mig dumt at blive ved med at tale om det.
(Lit: It seems me stupid to go on to talk about it.
It seems stupid to me to go on talking about it.)

c. Det er mig ligemeget, hvad De kan og hvad De ikke kan.
(Lit: It is me the same what you can and what you cannot.
It is the same to me what you can and what you cannot do.)

(Hansen, A. 1967:164)

d. Det er slemt som han er faldet i søvn.
(Lit: It is terrible how he has fallen asleep.
It is terrible the way he has fallen asleep.)

(Hansen, A. 1967:164)

e. Det véd jeg nu forvist, at han vil lade den danske folkesang sejre!
(Lit: It know I now for sure that he will let the Danish folk song win!
I know now for sure that he will let the Danish folk song win!)

(Hansen, A. 1967:164)

All the examples above are characterised by the fact that either the grammatical subject (a-d) or the object (e) of the sentence is constituted by the pleonastic pronoun *det* (it), while the "real" subject or object – either a finite or an infinitival clause – is placed in end-focus position[34]. This construction is in fact the default one with clausal subjects. Thus, the non-extraposed variants listed below are very awkward:

(198) a. At se dig er rart.
(To see you is nice.)

b. At blive ved med at tale om det forekommer mig dumt.
(To go on talking about it seems stupid to me.)

Clausal objects, on the other hand, do not have to be extraposed in the same way:

(199) a. Jeg ved ikke, om han kommer i dag.
(Lit: I know not if he's coming today.)

b. ? Jeg ved det ikke, om han kommer i dag.
(Lit: I know it not if he's coming today.)

The example quoted by Hansen (197e above) is rather special as the pleonastic pronoun is topicalised[35]. Thus, extraposition has the effect of avoiding topicalisation of a clausal constituent:

[34] Thus we do not treat examples where the extraposed constituent is an NP, as in:

Det var snak det hele.
(It was nonsense all of it)

also quoted by A. Hansen. It is doubtful, in fact, whether the subject pronoun can be considered a pleonastic element in such cases. Since it exhibits gender agreement with the nominal in end-focus position, in fact, it could be argued that it has referential – more exactly cataphoric – properties:

Den$_{(commongender)}$ var dum den bemærkning$_{(commongender)}$.
(It was stupid that observation.)

[35] Note in passing that the possibility of topicalising a pleonastic pronoun distinguishes Danish from e.g. English, where as noted by Postal & Pullum (1988), "expletive phrases can never be topicalised" (p.649). Thus, they note that the following sentence is ungrammatical:

*Upon it, you may depend that we won't abandon him.

.

(200) At han vil lade den danske folkesang sejre véd jeg nu forvist!
(Lit: That he will let the Danish folk song win I know now for sure!)

Our claim is in fact that extraposition of the kind discussed here, generally serves the purpose of avoiding a clausal constituent in theme position.

At first sight, it would seem that since extraposed subjects are placed in end-focus position, the information structure in such constructions should be clearly marked, in the sense that the extraposed subject should always be the focus of the sentence. However, in many cases it is the predicate that seems to provide the focus of the sentence, while the extraposed constituent is the primary topic. Consider thus the example below:

(201) Hvad synes du om ham?
Det irriterer mig enormt, at han altid kommer for sent.

(What do you think of him?
It annoys me enormously that he is always late.)

Furthermore, the fact that the pleonastic *det* can be topicalised as in the example of object extraposition quoted above, *Det véd jeg nu forvist, at han vil lade den danske folkesang sejre*, is additional evidence of the fact that the object in this case is not the focus of the sentence.

However, the extraposed clause can indeed function as the focus in a different context:

(202) Hvorfor irriterer han dig så meget?
Det irriterer mig enormt, at han altid kommer for sent.

Thus, it seems that no clear account of extraposition is possible in terms of focus marking. However, if a default focus assignment had to be decided on, at least for subject extraposition constructions it would probably have to be one where the predicate and not the extraposed clause, bore the focus.

Rather than being caused by the need to mark the focus in a structural way, subject extraposition seems to be the result of an independent effect, called *heaviness effect*, according to which syntactically complex constituents – clausal ones in this case – tend to be placed in end-focus position. Normally, the heaviness effect reinforces the general pattern according to which new topics occur after given ones. Given information, in fact, tends to be short, often pronominalised, in other words *light*, whilst new topics tend to be long, more complex, and phonetically prominent, i.e. *heavy*. Extraposition of the kind discussed here, however, seems to be a case where the 'given before new' rule and the heaviness effect contradict each other, and where the heaviness effect in fact has priority over the 'given before new' rule. In this case, in fact, the extraposed subject, although possibly expressing given information, is heavy and therefore must be placed in end-focus position.

Similarly, in the example below, where the extraposed constituent is the complement of a preposition, the effect achieved by preposing a pleonastic *det* to the clausal complement is that of moving the latter further to the right, thus allowing a sentence adjunct to be placed between the pronoun and the heavy constituent:

(203) Jeg er ligeglad med det i dag, at han snød mig den gang.

(Lit: I don't care about it today, that he cheated me at that time.)

Even here, however, the default focus assignment seems to be on the predicate rather than on the object.

The term 'heavy' was used above as a metaphor for syntactic complexity. This view is consistent with Hawkins (1994)'s claim that extraposition is motivated by the difficulty of processing a sentence where the lengthy sentential subject delays access to the second constituent of the top S node. Hawkins claims, in fact, that languages in general display a strong preference for word orderings where the number of nodes that have to be built by the parser to access all the immediate sentence constituents, is as small as possible.

In conclusion, extraposition is an interesting construction from the point of view of information structure even though it cannot be considered a focus marking construction. Its peculiarity lies in the fact that it contravenes the general rule that the end-focus position should be the site of the focus of the sentence. The extraposed clausal constituent, in fact, is placed in end-focus position in spite of the fact that it may well refer to presupposed information, and that the focus of the sentence may fall to the left of it. We have in fact suggested that the focus of the sentence should be assigned by default to the predicate rather than the extraposed clause, while the latter should be assumed to be the primary topic. However, this is only a suggestion that certainly needs to be validated by more research.

Thus, rather than being based on focus factors, we propose that the generation of subject extraposition in Danish should be based on the simple syntactic principle that a clausal subject is always extraposed. This will be discussed in more detail in the next section.

The grammar of extraposition

From a syntactic perspective, different views of extraposition have been defended in the literature. We shall start by discussing two opposite views: one is the GB-oriented approach, according to which the pleonastic element cannot occur in strictly subcategorised positions; the other is the opposite view, defended by Postal & Pullum (1988) and recently taken up by Pollard & Sag (1994) that the pleonastic element, although it contributes no semantic value to the sentence, is indeed subcategorised for.

Postal and Pullum argue that the GB claim that expletive NPs (our pleonastic pronouns) do not occur in strictly subcategorised positions, seems founded on a wrong chain of assumptions, namely:

- If α subcategorises for the position β, then α θ-marks β.
- Expletive NPs do not occur in θ-marked positions.
- Expletive NPs do not occur in strictly subcategorised positions.

They provide, in fact, a variety of data which at least for English refute this conclusion. Since, however, it is true that expletive NPs have no thematic role, the axiom that only θ-marked positions should be subcategorised for is disproved.

The data include i.a. examples of object extraposition, of expletive NPs as prepositional objects, and of extraposition of so-called *irrealis clauses*, shown in (204a), (b) and (c) respectively:

(204) a. I regret it very much that we could not hire Mosconi.
b. What do you make of it, then, that his mother is so evasive?
c. I would prefer it if Kim were not informed.
(Postal & Pullum 1988:642, 649)

Postal and Pullum show that an analysis of such examples where the expletive NP is not a complement of the verb, but rather belongs to the clause following it (which would then be analysed as a small clause) is untenable. The possibility of interposing adverbs and parentheticals between *it* and the extraposed constituent (as shown in (b) above), is one of the counterarguments against a small clause analysis. We shall not repeat the entire argument here, but simply agree with Postal and Pullum that at least in English, the pleonastic pronoun *it* can occur in strictly subcategorised positions, for example as the direct object of a verb, or as a prepositional object. Nevertheless, it does not contribute to the meaning of the sentence, and will therefore not appear in the semantic representation. We have already quoted examples in which the Danish *det* has object function. Others are discussed further below. Therefore, Postal and Pullum's argument is valid for Danish as well.

A third approach is that taken in Vikner (1995). Vikner distinguishes between *there* (Da. *der*, Du. *er*, and Ge. *pro*), and *it* (DA. *det*, Du. *het*, and Ge. *es*). The former is treated as an expletive, the second one as a quasi-argument. On his theory, *it* is base-generated in a theta-position and then moved to subject position, where it receives nominative case. The extraposed clause behaves like an apposition that receives neither theta-role, nor case. Structurally, the clause is adjoined to the sentence (or perhaps to the VP: Vikner is not clear on this point).

The intuition that the extraposed clause resembles an apposition seems confirmed by the fact that just like an apposition, it cannot be topicalised:

(205) a. Det irriterer mig, at du er kommet.
(It annoys me that you have come.)

b. * At du er kommet irriterer det mig.
(Lit: That you have come annoys it me.)

c. I dag mødte jeg Peter, min gamle ven.
(Lit: Today met I Peter, my old friend.)

d. * Min gamle ven mødte jeg Peter i dag.
(My old friend met I Peter today.)

(Vikner 1995:239)

However, it is not easy in Vikner's model to account for the fact that in some cases, *det* is obligatory as shown in the Danish examples below:

(206) a. Det sker, at vi går i teatret.
(It happens that we go to the theatre.)

b. Han finder det let at lære hvad som helst.
(He finds it easy to learn anything.)

Compare the sentences above with the non-extraposed variants below, all of which are ungrammatical:

(207) a. * At vi går i teatret sker.
(That we go to the theatre happens.)

b. * Han finder at lære hvad som helst let.
(He finds to learn anything easy.)

Vikner discusses cases similar to (a) above, where it is not obvious how to explain the fact that the clause cannot itself move to subject position. The solution he suggests is based on the hypothesis that in such cases the pronoun and the extraposed clause are base-generated as a small clause. This would yield the following analysis of the sentence under discussion:

(208) Det$_i$ sker [$_{SC}$ t$_j$ t$_i$] [at vi går i teatret]$_j$.

We do not find this solution very convincing, however, as the small clause never seems to surface as such[36]. We also find it unconvincing that the expletive *det* should be considered a thematic element, i.e. an element contributing to the semantic content of the sentence.

A more appealing solution in our opinion, is Pollard and Sag's. They suggest in fact that extraposition should be treated as a lexical phenomenon. According to this proposal, which can easily be adapted to cover the Danish data, the entry of the pleonastic reading of the pronoun *det*[37] would be as shown below:

[36]It may be argued, however, that a small clause analysis is suitable for examples where *det* has object function and is followed by an object complement, as in *Han anser det for usandsynligt, at alle skulle komme* (He considers it unlikely that everybody should come). We leave this point open.

[37]The same form can also be used as a personal or demonstrative pronoun.

(209) *det*

$$\begin{bmatrix} \text{PHON} \left\langle \text{det} \right\rangle \\ \text{SYNSEM} \mid \text{LOCAL} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD noun} \\ \text{SUBCAT} \quad \langle\rangle \end{bmatrix} \\ \text{CONTENT} \mid \text{INDEX} \; {}_{det}[\,] \end{bmatrix} \end{bmatrix}$$

As can be seen, the index associated with the entry is of a specific *det* sort. This is in order to distinguish it from referential pronouns as well as from the other pleonastic pronoun, *der*, which bears an index of sort *der*. This sort can then be used in the lexical entry of a verb that subcategorises for a clausal subject, e.g. *ske* (happen) below:

(210) *ske*

$$\begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD verb} \\ \text{SUBCAT} \left\langle \text{NP}_{det}, \text{S} \left[\text{at, fin}\right]{:}\boxed{1} \right\rangle \end{bmatrix} \\ \text{CONTENT} \begin{bmatrix} \text{RELN ske} \\ \text{ARG1} \; \boxed{1} \end{bmatrix} \end{bmatrix}$$

Thus, the first element on the SUBCAT list of the verb, which in our notation corresponds to the grammatical subject, must be an NP containing an index of the *det* sort, while the second element (the object) must be a finite sentential complement introduced by the complementiser *at*. In the CONTENT feature, there is only one argument, which is structure shared with the sentential complement in the SUBCAT list[38].

In the case of *ske* (happen), this is the only lexical entry needed, as subject extraposition is obligatory. In other cases, in addition to an entry accounting for the extraposition construction, there will also be one taking care of the non-extraposed variant. For example, the verb *glæde* (please) will have (at least) the following two entries:

(211) *glæde*

$$\begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD verb} \\ \text{SUBCAT} \left\langle \text{NP}_{det}, \text{NP}{:}\boxed{2}, \text{S} \left[\text{at, fin}\right]{:}\boxed{1} \right\rangle \end{bmatrix} \\ \text{CONTENT} \begin{bmatrix} \text{RELN glæde} \\ \text{ARG1} \; \boxed{1} \\ \text{ARG2} \; \boxed{2} \end{bmatrix} \end{bmatrix}$$

(212) *glæde*

$$\begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD verb} \\ \text{SUBCAT} \left\langle \text{S} \left[\text{at, fin}\right]{:}\boxed{1}, \text{NP}{:}\boxed{2} \right\rangle \end{bmatrix} \\ \text{CONTENT} \begin{bmatrix} \text{RELN glæde} \\ \text{ARG1} \; \boxed{1} \\ \text{ARG2} \; \boxed{2} \end{bmatrix} \end{bmatrix}$$

[38] It will be noted that our CONTENT feature differs slightly from the HPSG one. The strategy adopted here is taken over from the Eurotra linguistic theory. In Eurotra, arguments are numbered according to their deep syntactic role, so that arg1 would correspond to deep subject, arg2 to deep object and so on (cf. Allegranza et al. 1991).

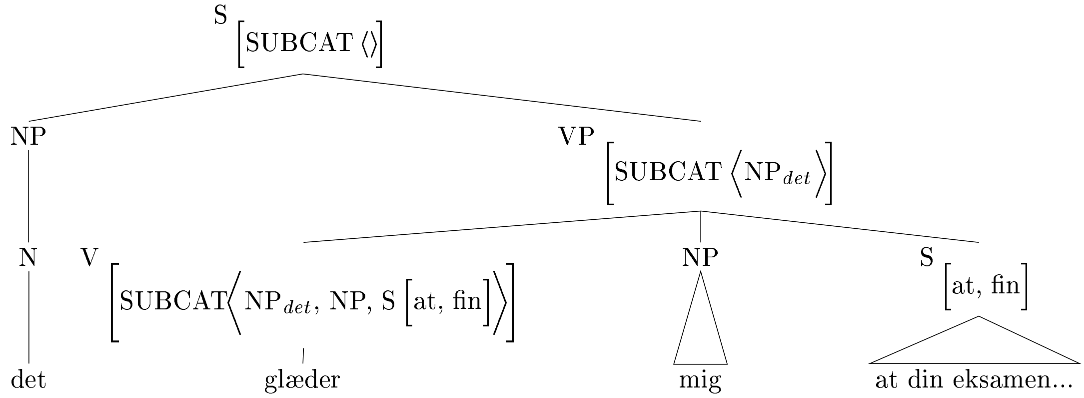
The two entries would account for (a) and (b) below:

(213) a. Det glæder mig, at din eksamen er gået så godt.
(It pleases me that your examination has gone so well.)

b. At din eksamen er gået så godt glæder mig.
(Lit: That your examination has gone so well pleases me.)

The syntactic representation of (a) would look as follows:

(214)

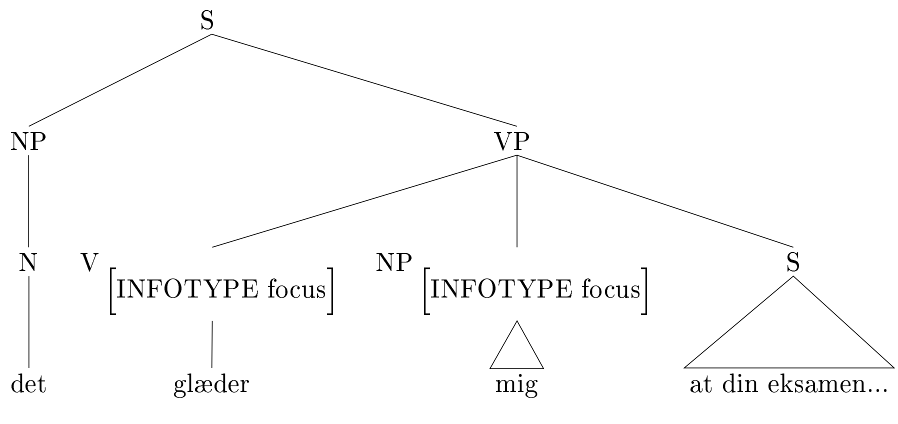


Pollard and Sag suggest that a lexical rule should relate an entry accounting for a non-extraposed structure to one accounting for an extraposed one in those cases where both structures are grammatically correct. The rule should simply move a clausal subject or object to the end of the SUBCAT list, and insert a pleonastic pronoun of the right sort in the position from which the clause has been moved. The structure sharing between SUBCAT elements and arguments in the CONTENT feature is left untouched. Such a rule would also account for the Danish data. However, it should be stressed again that the extraposed construction is indeed the normal structure when the clausal constituent is the subject of the sentence.

Thus, the criteria for generating an extraposed construction in Danish would be of a syntactic nature, and partly depend on the specific lexical head under consideration. In general, every time a clausal deep subject is present, an extraposed construction will be preferred. Thus, whilst in analysis a verb like *glæde* (please), would have at least two lexical entries, only the entry requiring an expletive subject would be used in generation. For verbs such as *ske* (happen), or *have (det) med* (tend), only the entry requiring an expletive subject or object respectively, would be available in analysis as well as in generation.

However, the HPSG-like treatment sketched out here presents a problem when it comes to representing the focus of the sentence. The default focus assignment we have proposed, in fact, foresees that the focus will be placed on the predicate preceding the extraposed clause. In the case of the example just discussed, *Det glæder mig, at din eksamen er gået så godt*, this would yield the representation below:

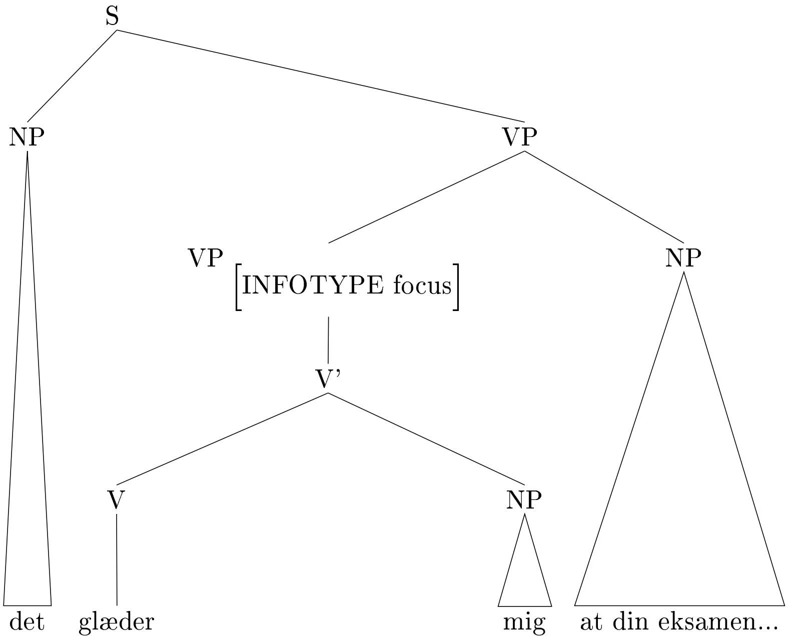
(215)



The problem with this representation is that the two words which constitute the focus of the sentence do not form a single constituent. In Section 3.3.2, it was noted that not all analysts claim a complete isomorphism between focality and constituency. Furthermore, we have looked at examples where the focus was represented as the union of various constituents (see example 161 in this chapter). Thus, technically it is not a problem in (215) above to represent the focus of the sentence as a union of the foci of the verb and the NP following it. However, this representation does not account explicitly for the fact that in the case of extraposition the predicate, i.e. the VP minus the extraposed clause, is either focalised or neutral as a whole[39]. If this is true, it seems an infelicitous consequence of the HPSG representation that there is no single constituent to which focus can be assigned.

The representation above can be compared with a GB-like treatment, where the extraposed clause is adjoined to the final VP:

(216)



In this model, the fact that the verb and the direct object form a constituent makes it easier to state that they should also form a focal entity in one reading of the sentence. The GB model, however, was rejected earlier because of its failure to account for the lexical nature of extraposition, as well for the fact that the pleonastic pronoun *det* can occur in subcategorised positions.

Therefore, we suggest an alternative treatment which tries to combine the HPSG lexical treatment with the GB intuition that the extraposed clause should be adjoined to the VP rather than be directly dominated by the VP node. This could be done by adding a NONLOCAL|EXTRAP feature to the standard HPSG NONLOCAL features, to keep track of extraposed constituents. The EXTRAP feature would take as its value a (unary) list of *local* structures. Thus, the lexical entry of *glæde* (please) would look as follows:

[39]Additional evidence of the fact that it makes sense to analyse the predicate as a constituent comes from the fact that it can be topicalised:

Glæde mig gør det, at din eksamen...
(lit: Please me does it, that your exam...)

Compare this with:

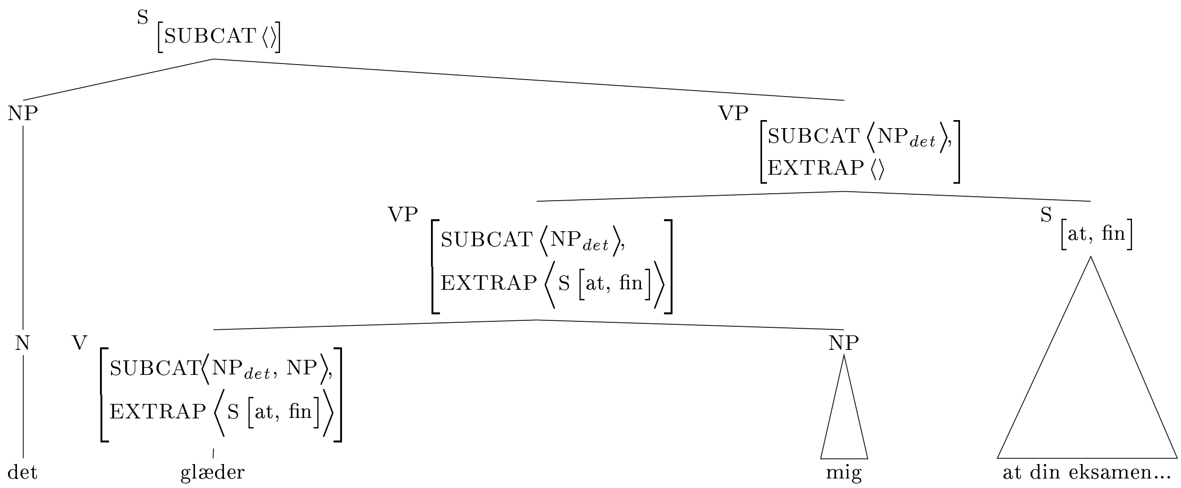
*Glæde gør det mig, at din eksamen...
(lit: Please does it me that your exam...)

(217) *glæde*, revised entry:

$$\left[\text{SYNSEM}\ \left[\begin{array}{l}\text{LOCAL}\left[\begin{array}{ll}\text{CAT} & \left[\begin{array}{l}\text{HEAD verb}\\ \text{SUBCAT}\ \left\langle \text{NP}_{det},\ \text{NP:}\boxed{2}\right\rangle\end{array}\right]\\ \text{CONTENT} & \left[\begin{array}{l}\text{RELN glæde}\\ \text{ARG1}\ \boxed{1}\\ \text{ARG2}\ \boxed{2}\end{array}\right]\end{array}\right]\\ \text{NONLOCAL}\,|\,\text{EXTRAP}\ \left\langle \text{S}\left[\text{at, fin}\right]\text{:}\boxed{1}\right\rangle\end{array}\right]\right]$$

The first argument of the verb is not to be found in the SUBCAT feature any longer, but is the value of the EXTRAP feature. A grammar rule (or a convenient principle compatible with the rest of HPSG principles) should then see to it that the value of EXTRAP is passed upwards in the tree until the extraposed clause is found. On the adjoined VP, the EXTRAP list should be empty, as shown in the representation:

(218)



Note that local and non-local features have been collapsed into a simplified single feature structure for lack of space. A more precise version of the feature structure representing the lower VP, for example, would be the following attribute value matrix:

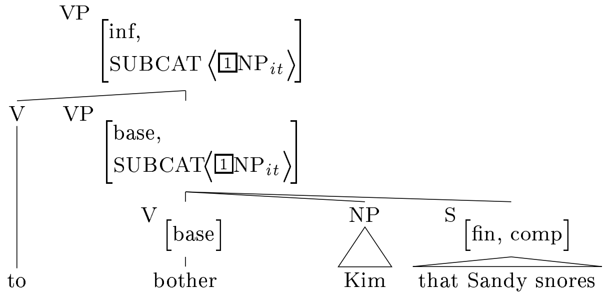
(219)
$$\left[\text{SYNSEM}\ \left[\begin{array}{l}\text{LOCAL}\left[\begin{array}{ll}\text{CAT} & \left[\begin{array}{l}\text{HEAD verb}\\ \text{SUBCAT}\ \left\langle \text{NP}_{det}\right\rangle\end{array}\right]\\ \text{CONTENT} & \left[\begin{array}{l}\text{RELN glæde}\\ \text{ARG1}\ \boxed{1}\\ \text{ARG2}\ \boxed{2}\end{array}\right]\end{array}\right]\\ \text{NONLOCAL}\,|\,\text{EXTRAP}\ \left\langle \text{S}\ \left[\text{at, fin}\right]\text{:}\boxed{1}\right\rangle\end{array}\right]\right]$$

Pollard and Sag argue that the treatment they propose has the advantage of allowing an easy account of raised expletives, e.g. in a sentence such as the following one:

(220) It continues to bother Kim that Sandy snores.

Their theory yields the following representation of the final infinitival clause:

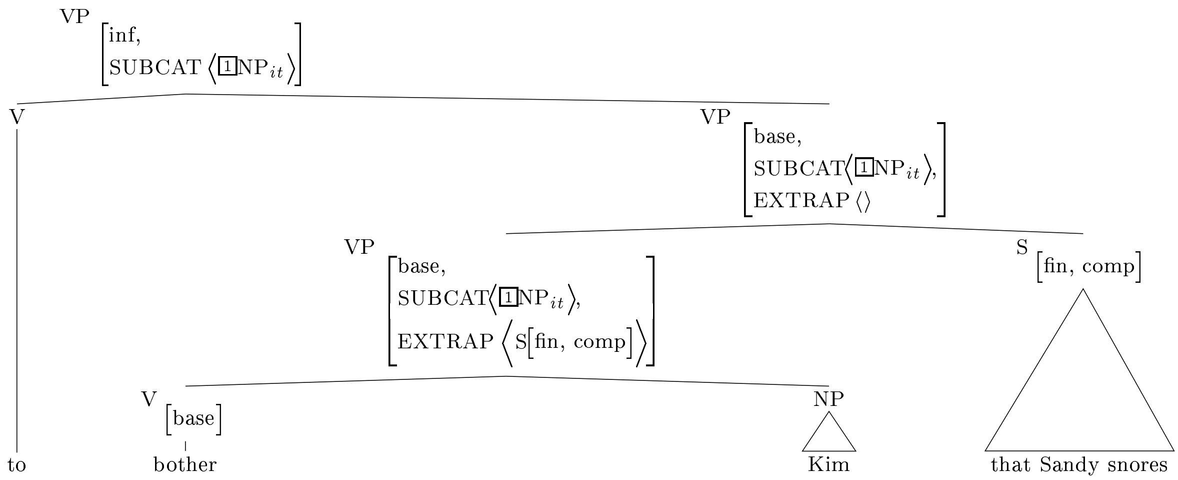
(221)



(Pollard & Sag 1994:152)

This structure may appear as the complement of the raising verb *continue*, the lexical entry of which specifies that the subject of the verb in question must be structure-shared with the unexpressed subject of the unsaturated complement (*it* in this case). The reader is referred to Pollard & Sag (1994) for a more detailed technical explanation. What we want to note here, however, is that our alternative representation of extraposed clauses does not seem to interfere with this mechanism. According to our proposal, in fact, the infinitival clause *to bother Kim that Sandy snores* would receive a different representation, namely:

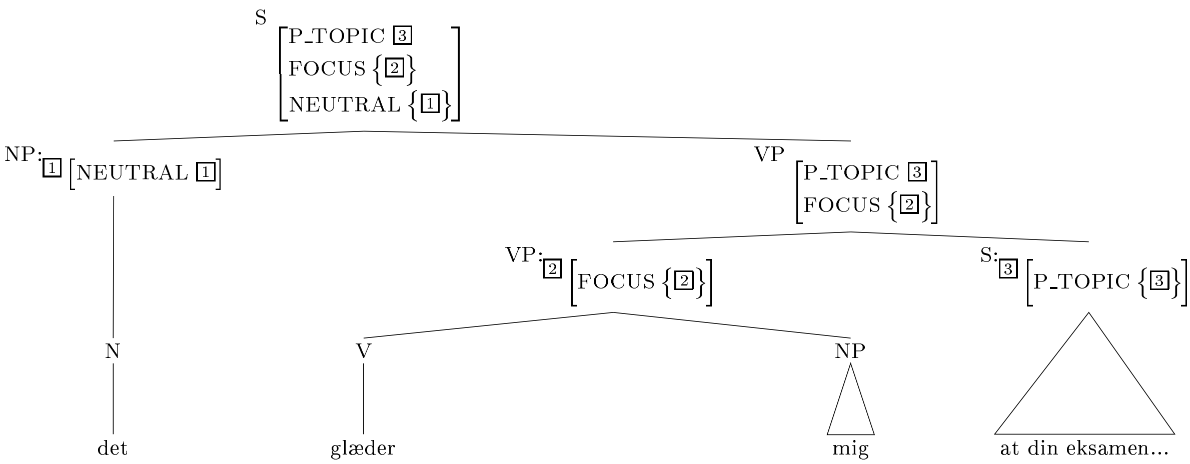
(222)



However, the top node of the structure is exactly the same as in Pollard and Sag's representation, and can therefore form the complement of a raising verb in exactly the same way. Compared to Pollard and Sag's, however, the representation we propose has the additional advantage of making it possible to state the information structure of the construction in a more appropriate way.

We conclude by showing what the information structure representation of the example under discussion would look like in connection with the new structure proposed:

(223) *Information structure in extraposition constructions*



5.4 Summary

In this chapter we have discussed the relationship between information structure and syntax on the basis of Danish data, and dealt with the way in which this relation can be exploited in the analysis and generation of Danish texts with a view to an MT application. We have been working on the assumption that, although a written sentence is often ambiguous with respect to information structure, in most cases a default reading can be assumed. In technical texts in particular, this default reading is hardly ever overridden, and syntax and intonation reinforce one another in creating the information structure of the various sentences.

We have seen that in general, the theme position in Danish tends to be the locus of the primary topic. Two of the syntactic constructions we have analysed, namely existential constructions and topicalisation, can be explained with reference to this characteristic. It was claimed that existential constructions constitute a way of moving a non-topical nominal phrase away from the theme position: the fact that these constructions are quite frequent in Danish, and syntactically less constrained than, say, English existential constructions, can be seen as a consequence of the fact that there is a stronger tendency in Danish to avoid non-topical themes. Topicalisation, on the other hand, serves the function of placing the primary topic in theme position, although in the default grammatical order it would be placed somewhere else in the sentence. The fact that topicalisation (at least in spoken language) is such a common phenomenon is again consistent with our view of the theme as the privileged position of the primary topic in Danish sentences.

The focus of the sentence, on the other hand, was claimed to fall preferably on the final constituent. A consequence of this general tendency in a variety of languages – with Danish as one of them – was shown to be the fact that a focussed complement may be placed in end-focus position even if its default grammatical position is a different one. Such a deviation from the so-called Systemic Ordering was taken as an unambiguous indication of where the focus of a sentence is to be found.

A number of exceptions to the default final focus principle were noted and discussed. The first one is the fact that general temporal and locative adverbials occurring in end-focus positions tend not to be focal. It was shown that this can be due to the very generality of these adverbials, which makes it difficult to identify the elements of the alternative focus set. The second exception is constituted by clefts, where the focus was shown to be structurally marked. It was noted that it-clefts in particular, besides providing a way of marking the focus, are characterised by the fact that they often present a given topic as the focus. Finally, extraposed clausal constituents, although they are placed in focus position, needn't be and often are not focal. A plausible explanation in this case is to be found in the difficulty of processing a sentence the first constituent of which is a clause.

For each of the syntactic constructions studied, including unmarked declarative sentences, implementation strategies for the computation of the information structure were proposed. These were presented as semi-formal statements of the type 'a temporal or locative complementation to the right of the verb is F+ if it has specific meaning, F- otherwise', or 'the head of an it-cleft is focus-marked'. The result of the application of these statements was shown as an instantiation of the INFOSTR and INFOTYPE features in the relevant representations. As for generation, principles aiming at guiding syntactic choice on the basis of the information structure of the input sentence were proposed, again in the form of semi-formal statements such as 'if a complement other than the subject is the primary topic, it should be topicalised'.

An important issue which has only been touched upon is the way in which the various principles for the computation of information structure – some of them based on heuristics rather than rules – interact with each other. In the next chapter, we look at this issue more closely, and see how the problem can be addressed in a concrete implementation of the strategies presented so far.

Chapter 6

Automatic analysis, transfer and generation

After having discussed the way in which the information structure of unmarked declarative sentences and focussing constructions can be represented in a general unification-based framework for the purposes of Machine Translation, in this chapter we look at a concrete implementation of the specifications proposed in the ALEP framework. The implementation covers the language direction English-Danish, and a subset of the syntactic constructions described in the preceding chapters. The purpose of the implementation is thus to show that the approach to information structure taken in this work is in general a sound and feasible one rather than to test all the specifications provided.

We shall be focussing on two main issues, namely the interaction between the various heuristics for the computation of information structure and the rest of the grammar on the one hand, and the way in which information structure information can be made to guide the generation process on the other.

As we shall see, ALEP provides an elegant solution to both issues. In analysis, the computation of information structure has been added on top of what will be called the *core grammar* as an independent refinement component, which adds features to the structures produced by the parser. Generation, on the other hand, proceeds from an interface structure mainly consisting of semantic and information structure information. The ALEP generation algorithm is able to construct a syntactic representation in principle almost completely independent of the syntactic structure of the source sentence.

The structure of this chapter is as follows: in Section 6.1, we introduce some key concepts of the ALEP formalism and give a brief description of the architecture of the system; Section 6.2 describes the way in which focus heuristics have been implemented; Section 6.3 deals with the treatment of focus-marking constructions, in particular clefts and existential constructions; finally, Section 6.4 contains a summary of the chapter.

6.1 Implementing an MT system in ALEP

As it was mentioned in Section 2.4, ALEP is a general purpose language engineering environment supporting a typed feature structure formalism, and one of the applications it is intended for is MT. In this section, we first give a general overview of the formalism, and then describe the architecture of the implemented system.

6.1.1 The formalism

The basic concept of the formalism supported in ALEP is that of typed feature structures. The legal features of a particular implementation are declared in a declaration file. As an example, we show below the CONTEXT features discussed earlier as they are declared in our type declarations for English analysis and Danish generation:

(1) *CONTEXT features*

```
type(synsem:{
           syn => type({syn:{}}),
           content => type({cont:{}}),
           context => type({context:{}})
            }).

type(context:{
            infotype => boolean([{p_topic,focus,neutral}]),
            infostr => type({info:{}}),
            disc =>  boolean([{given,new}]),
            head_disc =>  boolean([{given,new}]),
            presup => functor => none(_)
             }).

type(info:{
         p_topic => list(type({cont:{}})),
         focus => list(type({cont:{}})),
         neutral => list(type({cont:{}}))
          }).
```

Each feature is of a certain type, either a predefined type, such as in this case boolean, list or functor, or a user-defined one. Thus, CONTEXT (we shall continue using capitals to refer to features although feature names are not capitalised in ALEP) is a feature of type 'context', which is defined as subsuming five different features, namely INFOTYPE, INFOSTR, DISC, HEAD_DISC and PRESUP. Let us look at them one by one. INFOTYPE is a boolean feature, i.e. in ALEP a feature type allowing for the use of the negation operator. Its three possible values are those discussed in the previous chapters. INFOSTR is a complex feature, subsuming a set of three features. Each of them takes a list of CONT(ENT) features as its value, in accordance with our model. The DISC feature, which was tentatively proposed as a way of indicating the 'givenness' of discourse referents, takes either 'given' or 'new' as its value. HEAD_DISC has not been discussed previously: it is used to record the givenness status of a cleft head on the topmost S node. Finally, presupposition – which in our approach only applies to clefts – has been implemented as a functor, the default value of which is set to 'none'. A feature of type functor is a feature that takes a logical term as its value.

Thus, the type declaration above can give rise, for example, to the following linguistic description:

(2) *ALEP linguistic description*

$$\left[\text{context} => \left[\begin{array}{ll}\text{context} & \\ \text{infostr} => & \left[\begin{array}{ll}\text{info} & \\ \text{focus} => & \text{(list of contents)} \\ \text{neutral} => & \text{(list of contents)} \\ \text{p\_topic} => & \text{(list of contents)}\end{array}\right]\end{array}\right]\right]$$

This corresponds in the HPSG notation to the following feature structure:

(3) $$\left[\text{CONTEXT}\ {}_{\textit{context}}\left[\text{INFOSTR}\ {}_{\textit{info}}\left[\begin{array}{ll}\text{FOCUS} & \text{(list of contents)} \\ \text{NEUTRAL} & \text{(list of contents)} \\ \text{P\_TOPIC} & \text{(list of contents)}\end{array}\right]\right]\right]$$

Since the two representation types correspond to one another, we shall continue to use the more familiar HPSG one in the example representations we provide in this chapter. ALEP output parse trees can be found in Appendix II.

It must be noted that in this implementation, the organisation of the various syntactic, semantic and pragmatic features in the SIGN is more reminiscent of the approach proposed in Pollard & Sag (1987) than in more recent HPSG theory. The reason is simply a historical one, as the present implementation builds on a prototype implemented in 1993. The main difference between the approach taken here and the sample feature structures shown in the preceding chapters, is the fact that here syntactic and semantic features are kept separate as distinct SYN and CONTENT features subsumed by SYNSEM rather than CAT. Furthermore, the LOCAL/NONLOCAL distinction applies under SYN here, rather than under SYNSEM. In what follows, the basic structure of the SIGN will be:

(4)
$$\begin{bmatrix} \text{SIGN} & \begin{bmatrix} \text{PHON} & \begin{bmatrix}\text{STRING}\end{bmatrix} \\ \text{SYNSEM} & \begin{bmatrix} \text{SYN} & \begin{bmatrix} \text{LOC} & \begin{bmatrix}\text{HEAD}\\ \text{SUBCAT}\\ \text{CAT}\end{bmatrix} \\ \text{BIND} & \begin{bmatrix} \text{GAPS} & \begin{bmatrix}\text{GAPS\_IN}\\ \text{GAPS\_OUT}\end{bmatrix}\end{bmatrix} \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix}\text{RESTR}\end{bmatrix} \\ \text{CONTEXT} & \begin{bmatrix} \text{INFOSTR} & \begin{bmatrix}\text{P\_TOPIC}\\ \text{FOCUS}\\ \text{NEUTRAL}\end{bmatrix}\end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Another difference between the feature structures shown so far and those implemented in ALEP concerns the SUBCAT list, and in particular the order of the complements constituting this list. In fact, due to the way in which ALEP processes this list, the subject must follow the internal complements. In the specifications provided earlier, on the other hand, we followed the more intuitive 'subject $<$ complements' order.

However, these differences do not impinge on the results of this work since the CONTEXT features introduced in the preceding chapters are kept unchanged in the implementation.

In addition to a type declaration, analysis and generation in ALEP make use of grammar rules defining the syntactic backbone of the linguistic representations. A grammar rule is a linguistic description representing the local dependency between a mother node and its daughters as well as the mutual order of the daughters. There is no difference between analysis and generation rules, although they are processed differently. The example below is a simplified version of the rule describing a VBAR[1] consisting of a verbal head and a focussed complement in the Danish generation grammar:

(5) *VBAR rule*

```
sign:{phon => phon:{string => S_v, rest => R_xp},
      synsem => synsem:{
        syn => syn:{type => full,
                    loc => loc:{head => HEAD,
                              subcat => REST,
                              cat => vbar
                             },
```

[1]In the present implementation, any projection of a verbal head is called a VBAR: thus, the VP category is not used.

```
                    str => phrase:{constr => h_compl}
                 },
        content => V,
        context => context:{presup => PRES,
                             infostr => info:{
                                   focus => [F2|F1],
                                   neutral => N},
                             head_disc => Disc}
                    }
     }
<
[
sign:{phon => phon:{string => S_v, rest => S_xp},
      synsem => synsem:{
        syn => syn:{type => full,
                    loc => loc:{head => HEAD,
                              subcat => [C|REST],
                              cat => (v;vbar)
                             }
                 },
        content => V,
        context => context:{presup => PRES,
                        infostr => info:{
                                       neutral => N,
                             focus => F1},
                        head_disc => Disc}
                    }
     },
sign:{phon => phon:{string => S_xp, rest => R_xp},
      synsem => C => synsem:{
             syn => syn:{
                    type => full,
             content => F2,
             context => context:{
                    infotype => focus,
                    infostr => info:{
                             focus => [F2]}}
                           }}
].
```

The linguistic description preceding the '<' operator corresponds to the mother node, which is a feature structure of type SIGN, subsuming the two features PHON and SYNSEM. PHON is used in analysis to keep track of the string being analysed, and in generation to generate the strings corresponding to the structures that are being built. SYNSEM contains the syntactic, semantic and pragmatic information relative to the sign. Note that small letters indicate constant values. Thus, the syntactic category of the mother node is a 'vbar', the construction is of the head-complement type, and is not a gap (as indicated by the 'type => full' specification). All the other features in the description of the mother node have variable values, as indicated by the fact that they start with a capital letter.

To the right of the '<' operator is a list of two SIGN feature structures, defining the daughters. The leftmost daughter is either a 'v' or a 'vbar', allowing for recursion, and the rightmost one is a complement of unspecified category. The SUBCAT feature ensures that the complement generated by the rule has the right syntactic and semantic characteristics, according to the valency frame specified in the lexicon entry of the individual verb forms. The specification 'subcat => [C|REST]' indicates that the leftmost element of the SUBCAT list, which is structure-shared with the SYNSEM value of the complement, should be deleted from the list, while the remaining elements (the REST) are passed on to the mother node. The content of the mother node is set to be the content of the head verb via unification through the variable 'V'.

The rule also shows an example of the way in which information structure features are computed. The mechanism used is also based on list operations. However, instead of consuming elements on a list, as in the case of the SUBCAT feature, elements are appended to the FOCUS list while building up the structure. Thus the value of the FOCUS list of the mother node is obtained by appending the FOCUS value of the verbal complement, which is equal to the semantic content of the complement itself, to the FOCUS list of the head, which may contain zero, one or more elements. The value of the NEUTRAL list, on the other hand, is simply copied from head daughter to mother node.

Transfer rules define mappings between linguistic descriptions for source and target languages by pairing linguistic sub-expressions, provided that the conditions associated with the mapping hold. As an example, we show below a slightly simplified version of the rule transferring the topmost feature structure of the source language representation in our implementation:

(6) *Transfer rule*

```
trule(en,da,<=>,
sign:{synsem => synsem:{
         syn => syn:{
                   loc => loc:{
                   unit => Unit,
                   bar=>max,
                   cat=>s}
                   },
         content=>SEM1,
         context=>Context1}
      } < _,
sign:{synsem => synsem:{
         syn => syn:{
                  loc => loc:{
                  unit => Unit,
                  bar=>max,
                  stype=>main,
                  cat=>s}
                    },
         content=>SEM2,
         context=>Context2}},
[ SEM1 = SEM2,
  Context1 = Context2]).
```

The rule says that an English linguistic description unifying with the left-hand feature structure can be transferred to a Danish linguistic description unifying with the right-hand one (and viceversa, as indicated by the '<=>' operator), provided that the conditions in square brackets hold. The conditions say that the two descriptions must have equal content and context features. The underscore following the '<' operator means that the daughters of the linguistic description to be transferred are ignored. In other words, only the topmost feature structure is transferred, and the generation component must create an appropriate syntactic structure for the target language on the basis of the content and context features carried over in transfer. Note that the PHON feature, which contains the strings corresponding to the words in the source text, is not carried over or mapped directly onto a target language PHON feature. In this implementation we have used an interlingua approach at the lexical level. We use in fact the RELN (relation name) values in the semantic representation to choose the correct lexical entries in generation.

This brief description of the formalism would not be complete without an example of a lexicon entry. ALEP provides a macro facility which makes it possible to define general lexical patterns that are used to expand lexical entries during compilation. Thus, a lexical entry can be as simple as the following:

(7) *Entry for "kræves"*

kræves ~ der_passv['kræves', pres, require].

The string to the left of the tilde is the name of the entry, while the name following the tilde is a call to a macro. The names in square brackets are the arguments of the macro, providing the string value of the verb form, the tense and the relation name, which in this implementation corresponds to the lexical value of the corresponding English lexical entry.

Macros have been defined in the present implementation to generalise over nominal and verbal head features, content features, and subcategorisation patterns. For instance, the macro used in the entry just shown looks as follows:

(8) *Macro for der-passives*

```
macro(der_passv[String,Vform,R],
            sign:{
               lex[String],
               named_spec=>spec:{specinfo=>specinfo:{synth=>y}},
               synsem=>synsem:{
                  syn=>syn:{
                     bind=>bind:{
                        gaps=>gaps:{
                            gap_in=>[],
                            gap_out=>[]}},
                        loc=>loc:{
                            headpassv[Vform],
                            bar => zero,
                            der_o_pass[Arg2],
                            cat=>v
                                  }
                             },
                   content2[R,Arg2],
                   context=>context:{
                         presup=>none(_),
                         infotype=>focus,
                             infostr=>info:{
                                focus=>[cont:{restr=>restr:{reln=>R}}]}}}
                     } ).
```

The macro, which is used for generation as indicated by the value of NAMED_SPEC, defines a passive verb type taking the pleonastic *der* as its subject. GAPS are set to null as the gap feature is not threaded through verbal entries, BAR is zero as for all lexical units, CAT is 'v', PRESUP is set to the default 'none', and the FOCUS list is initialised by setting it to be equal to the RELN of the verb itself. 'Lex', 'headpassv', 'der_o_pass' and 'content2' are again macro calls. At compile time, they are expanded out to generate a complete lexical entry. The expanded entry is shown below. Note that the variable names used in the macros have been replaced by others generated automatically by the system.

(9) *Expanded entry for 'kræves'*

```
sign:{
   phon =>
   phon:{
```

```
   string =>  [kræves | B],
   rest => B },
synsem =>
synsem:{
   syn =>
   syn:{
      loc =>
      loc:{
         head =>                   /* head features */
         vhead:{
            maj => *,
            vform => pres,
            msaspect => simple,
            trans => yes,
            diath => pass },
         subcat =>  [
                 synsem:{          /* object */
                    syn =>
                    syn:{
                       loc =>
                       loc:{
                          head =>
                          nhead:{
                             maj => *,
                             case => acc },
                          subcat => [],
                          so => n_a,
                          cat => (perspron;(relpron;(apn;(np;(nbar;n)))))}},
                    content => C1,
                    context =>
                    context:{
                       infotype => focus } },
                 synsem:{           /* subject */
                    syn =>
                    syn:{
                       loc =>
                       loc:{
                          head =>
                          nhead:{
                             maj => *,
                             case => nom,
                             nform => der },
                          subcat => [],
                          so => n_a } } }],
         bar => zero,
         cat => v },
      bind =>
      bind:{
         gaps =>
         gaps:{
            gap_in => [],
            gap_out => [] } } },
   content =>                      /* content */
   cont:{
      restr =>
      restr:{
         reln => require,
         arg1 =>
         nullcont:{ },
         arg2 => C1,
         arg3 =>
         nullcont:{ },
         mods => [] } },
   context =>                      /* context features */
   context:{
      infotype => focus,
      infostr =>
      info:{
```

```
          focus =>  [
                   cont:{
                      restr =>
                      restr:{
                         reln => require } }] },
       presup =>
       none(   C3 ) } },
 named_spec =>
 spec:{                          /* specifier */
    specinfo =>
    specinfo:{
       synth => y } } }
```

In addition to types, rules and lexical entries, ALEP also makes use of processing declarations of various types to make rule processing and lexicon look-up more efficient. An example is the SPEC(IFIER) feature in the entry above, which indicates that the entry is to be used in generation. These processing declarations, however, are not very important in this context and will therefore not be described any further here. The interested reader is referred to the ALEP literature as well as to the Appendices.

6.1.2 System architecture

The three main components of the implementation are analysis, transfer and generation. ALEP also provides facilities for text handling and morphological analysis. Morphological analysis has not been integrated, so that both the analysis and generation lexica consist of fullform words. Text handling in this implementation is very simple, and only consists in segmenting the input into words and punctuation signs. The output of text handling is a set of chunks. These are mapped onto so-called Linguistic Descriptions (LDs) by the lifting operation. LDs are trees the nodes of which consist of feature structures. In generation, the LDs output by the generation grammar are lowered into chunks, from which the text handler generates the final output document.

As already mentioned, the analysis process has been split up into two phases: parsing proper and refinement. During parsing, the input sentences are analysed on the basis of the core grammar and assigned a syntactic and semantic representation. To these representations are then added information structure features as defined in the refinement grammar. Formally speaking, refinement rules do not differ from other analysis rules. However, it is an advantage to be able to express the focus heuristics in a separate module for two reasons: firstly, the core grammar can be kept simpler and the parsing process more efficient; secondly, heuristics that may have to be modified in a different domain or text type can be changed more easily when kept in a separate module. In our opinion these advantages outweigh the redundancy coming from having to express the legal structures of the grammar twice, i.e. both in the core and in the refinement grammar.

Transfer, as we saw in the preceding section, is a relatively simple operation. In fact, the strategy we have chosen to follow consists of transferring only the topmost feature structure of the representations output by the analysis module. This structure contains the following pieces of information:

- a CONTENT feature describing the content of the whole sentence (this includes the relation names used to select target lexicon entries);
- a CONTEXT feature representing the information structure of the sentence and its presupposition;
- limited syntactic information (e.g. tense and morphosyntactic aspect), which in a more complete implementation would have to be re-expressed as semantic features.

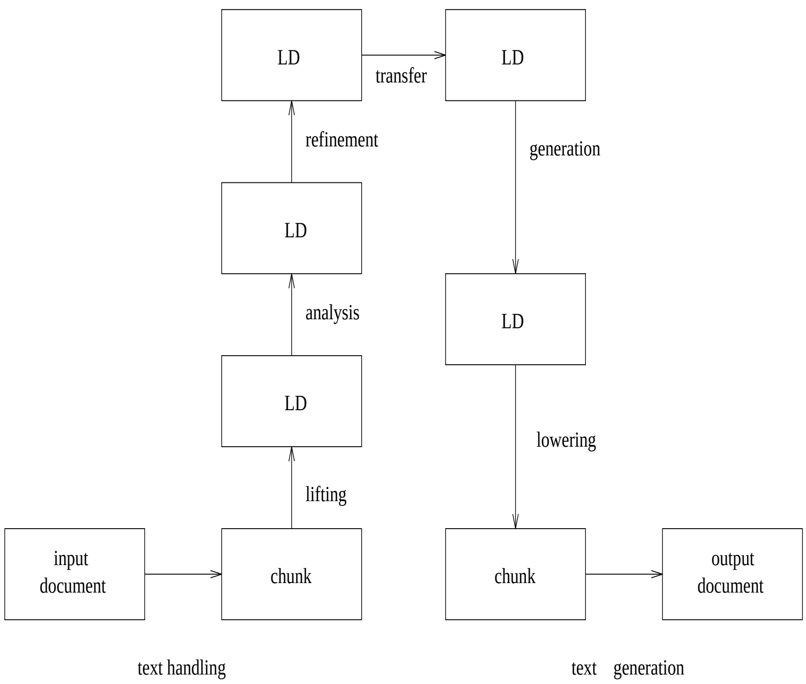
For example, the topmost feature structure in the LD corresponding to the sentence "Amines combat the sludge" will look as follows:

(10) $$\left[\text{SYNSEM}\begin{bmatrix}\text{SYN}\left[\text{LOC}\begin{bmatrix}\text{HEAD}\begin{bmatrix}\text{VFORM pres}\\ \text{MSASPECT simple}\\ \text{DIATH act}\end{bmatrix}\\ \text{CAT s}\end{bmatrix}\right]\\ \text{CONTENT}\left[\text{RESTR}\begin{bmatrix}\text{RELN combat}\\ \text{ARG1 }\boxed{3}\ _{\textit{npcont}}\begin{bmatrix}\text{RESTR}\left[\text{RELN amine}\right]\\ \text{DET none}(\boxed{1})\\ \text{INDEX }\boxed{1}\end{bmatrix}\\ \text{ARG2 }\boxed{4}\ _{\textit{npcont}}\begin{bmatrix}\text{RESTR}\left[\text{RELN sludge}\right]\\ \text{DET def}(\boxed{2})\\ \text{INDEX }\boxed{2}\end{bmatrix}\end{bmatrix}\right]\\ \text{CONTEXT}\left[\text{INFOSTR}\begin{bmatrix}\text{P\_TOPIC}\left\langle\boxed{3}\right\rangle\\ \text{FOCUS}\left\langle\boxed{4},\left[\text{RESTR}\left[\text{RELN combat}\right]\right]\right\rangle\\ \text{NEUTRAL}\langle\rangle\end{bmatrix}\right]\end{bmatrix}\right]$$

We shall not comment on the specific features in the representation yet, but only note that the task of the generation component is to find the best Danish syntactic realisation of the content and information structure of the input English sentence. This is done by way of a head-driven generation algorithm (based on Shieber, Pereira, van Noord & Moore 1990) which in our case is driven by semantic and contextual information.

To summarise, the diagram below shows the various steps that the translation process consists of:

(11) *The translation process in ALEP*



In the following sections, we focus on the strategies used in analysis and generation to handle information structure heuristics in unmarked declarative sentences and focussing constructions.

6.2 Implementing focus heuristics

Recall that information structure is viewed along the two dimensions of topicality and focality. The primitive concerning topicality is the primary topic, which was defined as the topic which the sentence is about, while the primitive concerning focality is the focus, through which one or more elements are emphasised to indicate a change in the informatin built by the discourse as a whole. Information which neither constitutes the primary topic nor the focus of the sentence, is called neutral information. To compute the primary topic, focus and neutral information of an unmarked declarative sentence, a number of heuristic principles were formulated in the preceding chapters.

These heuristics have been implemented in the English analysis component as a set of refinement rules which, on the basis of various factors, compute the value of the INFOSTR feature of the main sentence. In Section 5.2 we drew attention to the fact that some of the heuristics may contradict one another, and thus call for the use of a preference mechanism. Here however, the rules have been formulated in such a way as to be mutually exclusive, in order to rule out cases of overgeneration.

A general issue that had to be resolved was the level at which information structure features had to be assigned in the phrase structure tree. In the model proposed in Chapter 4, the INFOSTR and INFOTYPE features are described as features pertaining to phrasal signs. However, should all phrasal signs have them? The approach taken here is to assign INFOSTR and INFOTYPE to verbal complements, verbal and sentence adjuncts and to the main verb itself in the main sentence. In other words, we neither compute the information structure within subordinate clauses nor within maximal non-verbal projections. The latter choice is consistent with one of the basic assumptions we made, namely that categories like focus or neutral relate to syntactic phrases and not to parts of syntactic phrases. The former, however, is a less obvious choice, as it is open to question whether subordinate clauses have an independent information structure, which somehow gets combined with the information structure of the main sentence.

Another general point we want to make here relates to the relationship between information structure and content features. The proposal we made in Chapter 4 was that P_TOPIC, FOCUS and NEUTRAL should be sets (or lists depending on the implementation) of CONTENT features. The present implementation supports this approach. However, it must be noted that only a part of the verbal content can be taken into consideration when computing the contribution made by the verb to the building up of information structure. To illustrate the point, let us consider again the instantiation of INFOSTR in the simple sentence “Amines combat the sludge”, which we repeat here for clarity’s sake:

$$
(12)\quad \begin{bmatrix} \text{SYNSEM} & \begin{bmatrix} \text{CONTENT} & \begin{bmatrix} \text{RESTR} & \begin{bmatrix} \text{RELN combat} & \\ \text{ARG1 } \boxed{3} & {}_{\textit{npcont}}\begin{bmatrix} \text{RESTR} \begin{bmatrix}\text{RELN amine}\end{bmatrix} \\ \text{DET none}(\boxed{1}) \\ \text{INDEX } \boxed{1} \\ \text{NUMBER plur} \end{bmatrix} \\ \text{ARG2 } \boxed{4} & {}_{\textit{npcont}}\begin{bmatrix} \text{RESTR} \begin{bmatrix}\text{RELN sludge}\end{bmatrix} \\ \text{DET def}(\boxed{2}) \\ \text{INDEX } \boxed{2} \\ \text{NUMBER sing} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{CONTEXT} & \begin{bmatrix} \text{INFOSTR} & \begin{bmatrix} \text{P\_TOPIC } \left\langle \boxed{3} \right\rangle \\ \text{FOCUS } \left\langle \boxed{4}, \begin{bmatrix}\text{RESTR} \begin{bmatrix}\text{RELN combat}\end{bmatrix}\end{bmatrix} \right\rangle \\ \text{NEUTRAL } \langle\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}
$$

We want the analysis of the sentence to be such that the subject is regarded as the primary topic, while the verb and the direct object are both elements in the FOCUS list. Thus, P_TOPIC is correctly computed to be a singleton list, the element of which is structure-shared with the first argument of the verb, as shown by the shared index $\boxed{3}$ in the representation above. The FOCUS list, on the other hand, is a list of two elements. The first element[2] is structure-shared with the second argument of the verb, as indicated by the index $\boxed{4}$. The second element, on the other hand, cannot be structure-shared with the content of the verb itself, as this is in fact the content of the whole sentence. Therefore, only the RELN part of the verbal content, is copied onto the FOCUS list.

Having clarified these general issues, let us now look more closely at some of the rules that perform the actual computation. They can be grouped into three sets of rules, according to the particular assignment they carry out:

- Assignment of fronted adverbials to the P_TOPIC list.
- Assignment of INFOSTR values to subject and verb.
- Assignment of INFOSTR values to verbal complements and adjuncts.

Let us discuss them one by one.

6.2.1 Fronted adverbials

According to our model, a fronted adverbial is always treated as the primary topic of the sentence. This is implemented by copying the content of a fronted adverbial onto the P_TOPIC list, while the values for NEUTRAL and FOCUS are passed on from the VBAR node onto the main S node. The rule that does the job is shown below:

(13) *Rule for fronted adverbials*

```
sign:{phon => phon:{string => S_adv, rest => R_punct},
```

[2]The fact that the order between verb and object is inverted in the FOCUS list is not important: ALEP only allows elements to be appended to the front of a list.

```
     named_spec=>spec:{language=>en, specinfo=>specinfo:{ref=>y}},
     synsem => synsem:{
            syn => syn:{loc => loc:{
                              head => HEAD,
                              subcat => [],
                              cat => s},
                         seg => seg:{first=>y,last=>y}},
            context => context:{
                           presup => PRES,
                           infostr => info:{
                                  p_topic => [PT],  /* computation of
                                   focus => F,         the p_topic */
                                   neutral => N}}
                       }
    }
<

[sign:{phon => phon:{string => S_adv, rest => R_adv},
       synsem => synsem:{
             syn => syn:{loc => loc:{
                              cat => adv;pp},
                          seg => seg:{first=>y}},
             content => PT,                      /* p_topic content */
             context => context:{
                          infostr => info:{
                                  p_topic => [PT]}} /* computation of
                        }                              the p_topic */
    },
sign:{phon => phon:{string => R_adv, rest => R_s},
      synsem => synsem:{
            syn => syn:{loc => loc:{
                              head => HEAD,
                              subcat => [],
                              cat => s},
                         str => phrase:{parsehead => y}},
            context => context:{
                          presup => PRES,
                          infostr => info:{
                                   focus => F,
                                   neutral => N}}
                     }
    },
sign:{phon => phon:{string => R_vp, rest => R_punct},
      synsem => synsem:{
        syn => syn:{type => full,
                  loc=>loc:{cat=>punct}}}
    }
].
```

This rule applies to a tree consisting of a sentential mother node and of three daughters: an initial adverbial or prepositional phrase, a sentence and a full stop.

Note that a feature SEG, subsuming the two boolean features FIRST and LAST, is used to keep track of the position of the various constituents in the sentence. An initial constituent will be typed [SEG|FIRST y] and [SEG|LAST n] while the values will be interchanged for a final constituent, and a constituent in the middle will have the value 'n' ('no') for both features. Thus, the fronted adjunct in the rule above is marked [SEG|FIRST y][3].

The fact that the primary topic of the sentence is an adjunct can be exploited in generation to place the corresponding target language constituent in initial position.

6.2.2 Assignment of INFOSTR values to subject and verb

A number of different rules handle the contribution made by the subject and the main verb to the information structure of the sentence. The rules must take the following constraints into account:

- whether the subject is initial or not: if it is not, it is considered neutral;
- what determination the subject comes with: indefinite subjects are considered focussed, while definite subjects or subjects without determination are considered primary topics when initial and neutral otherwise;
- whether the subject is a pleonastic pronoun: in this case, it makes no contribution to the information structure of the sentence;
- whether the subject is the sbar in a wh-cleft: in this case, it is considered neutral;
- whether the verb is final or not: if it is final, it is considered neutral when the subject is indefinite, focussed otherwise; if it is not final, its INFOTYPE depends on the order of the verbal complements (as explained below).

These heuristics were already discussed at length in Section 5.2, and will therefore not be commented on any further here. To illustrate how they are implemented in practice, however, we show below the rule that computes the information structure in a sentence where the subject is indefinite and the verb is not placed in final position:

(14) *Rule for indefinite subjects*

```
sign:{phon => phon:{string => S_np, rest => R_punct},
      named_spec=>spec:{language=>en, specinfo=>specinfo:{ref=>y}},
      synsem => synsem:{                         /* topmost S node */
             syn => syn:{loc => loc:{
                                cat => s},
                          seg => seg:{first=>y, last=>y}},
             context => context:{
                           presup => PRES,
                           infostr => info:{
                                  p_topic => [],
                                  focus => [F|F2],
                                  neutral => N}}
                        }
     }
<

[sign:{phon => phon:{string => S_np, rest => S_vp},
         synsem => C => synsem:{                /* indefinite subject */
               syn => syn:{loc => loc:{
                                 head => nhead:{},
                                 cat => ~perspron},
```

[3]We are grateful to Bradley Music for having suggested this solution to us.

```
                              seg => seg:{first=>y}},
                 content => F => npcont:{index=>X, det=>exists(X)},
                 context => context:{
                                 infotype => focus,
                                 infostr => info:{
                                         focus => [F]}}
                               }
         },
sign:{phon => phon:{string => S_vp, rest => R_vp},
      synsem => synsem:{                            /* verbal phrase */
             syn => syn:{loc => loc:{subcat => [C],
                                cat => vbar},
                         str => phrase:{parsehead => y}},
             context => context:{
                            presup => PRES,
                            infostr => info:{
                                        focus => F2,
                                        neutral => N}}
                        }
     },
sign:{phon => phon:{string => R_vp, rest => R_punct},
      synsem => synsem:{                            /* full stop */
        syn => syn:{type => full,
                    loc=>loc:{cat=>punct}}}
     }
].
```

Note that in order to unify with the first daughter in the rule, the subject must be in initial position (marked [SEG|FIRST y]), and it must be indefinite, as indicated by the fact that the value of the determination feature (DET) binding the nominal index in the subject CONTENT must be 'exists'. As a result of the application of the rule, P_TOPIC is set to the empty list on the mother S node, while the content of the subject is appended to the FOCUS list. The rest of the FOCUS list and the whole NEUTRAL list are percolated up from the VBAR node.

The assignment of the right INFOTYPE to the subject is quite important when translating into Danish, where as we have seen, focussed subjects are preferably avoided. It is therefore a requirement in the Danish generation grammar that the grammatical subject be either the primary topic of the sentence or a pleonastic pronoun. We return to the issue in the section on focussing constructions, where the strategy used for the generation of existential constructions is discussed in some detail.

6.2.3 Assignment of INFOSTR values to verbal complements and adjuncts

The rules that refine VBARs are constrained by a number of different factors. A VBAR is syntactically speaking a phrase consisting of a mother node which is either a V or a VBAR followed by either a complement or an adjunct. The computation of the contribution made by adjuncts to the information structure of the VBAR is quite simple. All adjuncts are assigned to the FOCUS list by default. Thus, the distinction between general and specific adjuncts discussed earlier is not part of the present implementation.

As for complements, focus-marking constructions are distinguished from unmarked ones. In the former case, INFOTYPE features are assigned in the lexicon. We return to this point in the next section. In the latter case, the assignment of a complement to either the FOCUS or the NEUTRAL list is made to depend on the mutual order of the complements. Thus, if complements are ordered according to SO, they are all assigned to the FOCUS list, and so is the verb. This choice corresponds to the default heuristic assumed in Section 5.2. Otherwise, if the complements do not follow SO, the criterion implemented is again the one already discussed in the same section, namely:

> *If A is the left-hand argument in an adjacent argument pair which does not obey SO, then A is F-, and so are all the arguments placed between the verb and A. The argument following A is F+. The verb is assigned F-.*

Since the order of the complements on the SUBCAT list is a linear one, to be able to analyse sentences where SO is not the default one, the various possible cases of SO deviation have to be enumerated explicitly in the lexicon. For example, the following two (simplified) lexicon entries for the English verb *give* allow us to account for the two possible surface orders IOBJ > OBJ and OBJ > IOBJ:

(15) *Entry for give_1*

```
gives1 ~
sign:{phon => phon:{string =>[gives|R], rest => R},
      named_spec=>spec:{specinfo=>specinfo:{ana=>lex}},
      synsem => synsem:{
        syn => syn:{
          loc => loc:{
             head => vhead:{                      /* head features */
                       maj => v,
                       agr => sing&'3',
                       vform => pres
                            },
             bar => zero,
             subcat => [synsem:{syn => syn:{            /* indirect
                                 loc => loc:{              object */
                                    so => default
                                               }
                                            },
                               content => IO
                                },
                         synsem:{syn => syn:{            /* object */
                                  loc => loc:{
                                     so => default
                                                }
                                             },
                               content => O
                                 },
                         synsem:{syn => syn:{            /* subject */
                                  loc => loc:{
                                     case => nom
                                              }
                                           },
                               content => S
                                }
                          ],
                cat => v
                       }
                },
        content => cont:{restr =>
                  restr:{reln=>give,arg1=>S,arg2=>O,arg3=>IO}},
        context=>context:{}}
     }.
```

(16) *Entry for give_2*

```
gives ~
sign:{phon => phon:{string =>[gives|R], rest => R},
      named_spec=>spec:{specinfo=>specinfo:{ana=>lex}},
      synsem => synsem:{
        syn => syn:{
          loc => loc:{
             head => vhead:{                      /* head features */
                      maj => v,
                      agr => sing&'3',
                      vform => pres
                            },
             bar => zero,
             subcat => [synsem:{syn => syn:{             /* object */
                                 loc => loc:{
                                     so => ndef_left
                                           }
                                         },
                                content => O
                                },
                        synsem:{syn => syn:{             /* indirect
                                  loc => loc:{              object */
                                     head => phead:{
                                          maj=>p,
                                          pform=>Pform
                                                   },
                                     so => ndef_right
                                          }
                                        },
                                 content=>cont:{
                                       restr=>restr:{arg1=>IO}}
                                          },
                         synsem:{syn => syn:{            /* subject */
                                   loc => loc:{
                                      case => nom
                                            }
                                          },
                                  content => S
                                 }
                        ],
              cat => v
                    }
              },
        content => cont:{restr =>
                 restr:{reln=>give, arg1=>S, arg2=>O, arg3=>IO}},
        context=>context:{}}
     }.
```

The two entries differ in the syntactic shape and mutual order of the complements on the SUBCAT list. In the first entry, which accounts for the default IOBJ > OBJ order, the indirect object precedes the direct object on the SUBCAT list, while the last element is the subject. The two complements are respectively mapped onto the third and the second argument in the CONTENT representation. Note that each of the internal arguments is typed ‘so => default’. This feature is used in the grammar to trigger the right focus assignment.

The second entry, on the other hand, accounts for a particular case of SO disruption, namely dative shift. Here the second complement on the SUBCAT list is in fact a prepositional phrase, the complement of which is mapped onto the third argument of the verb, while the first complement on the SUBCAT list is mapped onto the second argument. Note that the value of SO here is not ‘default’, but ‘ndef_left’ for the left-hand complement in the pair disrupting SO, and ‘ndef_right’ for the right-hand one.

Let us now see how the SO feature is used in the grammar to derive the right information structure features for the VBAR. We show below one of the rules applying to complements deviating from SO.

(17) *Rule for SO deviation*

```
sign:{phon => phon:{string => S_v, rest => R_xp},
      named_spec=>spec:{language=>en, specinfo=>specinfo:{ref=>y}},
      synsem => synsem:{                        /* topmost vbar node */
        syn => syn:{loc => loc:{cat => vbar},
                    str => phrase:{constr => h_compl}
                   },
        context => context:{
             presup => PRES,
             infostr => info:{
                focus => [],
                neutral => [N2,
                            cont:{restr=>restr:{reln=>R}}]}}
                }
     }
<
[
sign:{phon => phon:{string => S_v, rest => S_xp},
      synsem => synsem:{                            /* verbal head */
        syn => syn:{loc => loc:{
                       subcat => [C|REST],
                       cat => v
                             },
                    str => phrase:{parsehead => y}
                   },
        content => cont:{restr=>restr:{reln=>R}},
        context => context:{presup => PRES}
                }
     },
sign:{phon => phon:{string => S_xp, rest => R_xp},
      synsem => C => synsem:{                     /* left-hand complement
        syn => syn:{loc => loc:{                     in a pair deviating
                      so => ndef_left,               from SO */
                      cat => ~sbar}},
        content => N2,
        context => context:{
                    infostr => info:{
                             neutral => [N2]}}}
     }].
```

The rule refines a VBAR consisting of a V and a complement typed with the feature [SO ndef_left], i.e. the left-hand complement in a pair deviating from SO. As a result of refinement, the NEUTRAL list on the mother node will be a list consisting of the RELN of the verb and of the content of the complement. The FOCUS list is still empty, as the focussed complement(s) to the right of the neutral one are refined by a different rule.

After refinement, the CONTENT and CONTEXT features relevant to the topmost S node in the representation of the sentence "Information technology gives better communication possibilities to the research community" will look as follows[4]:

(18)
$$
\left[\begin{array}{ll}
\text{SYNSEM} & \left[\begin{array}{ll}
\text{CONTENT} & \left[\begin{array}{ll}
\text{RESTR} & \left[\begin{array}{ll}
\text{RELN give} & \\
\text{ARG1 } \boxed{3} & {}_{\textit{npcont}}\left[\begin{array}{l}
\text{RESTR } \left[\text{RELN information\_technology}\right] \\
\text{DET none}(\boxed{1}) \\
\text{INDEX } \boxed{1} \\
\text{NUMBER sing}
\end{array}\right] \\
\text{ARG2 } \boxed{4} & {}_{\textit{npcont}}\left[\begin{array}{l}
\text{RESTR } \left[\begin{array}{l}
\text{RELN communication\_possibilities} \\
\text{MODS } \left\langle \text{RESTR } \left[\text{RELN better}\right] \right\rangle
\end{array}\right] \\
\text{DET none}(\boxed{2}) \\
\text{INDEX } \boxed{2} \\
\text{NUMBER plur}
\end{array}\right] \\
\text{ARG3 } \boxed{5} & {}_{\textit{npcont}}\left[\begin{array}{l}
\text{RESTR } \left[\text{RELN research\_community}\right] \\
\text{DET def}(\boxed{6}) \\
\text{INDEX } \boxed{6} \\
\text{NUMBER sing}
\end{array}\right]
\end{array}\right]
\end{array}\right] \\
\text{CONTEXT} & \left[\begin{array}{ll}
\text{INFOSTR} & \left[\begin{array}{l}
\text{P\_TOPIC } \left\langle \boxed{3} \right\rangle \\
\text{FOCUS } \left\langle \left[\text{RESTR } \left[\begin{array}{l} \text{REL to} \\ \text{ARG1 } \boxed{5} \end{array}\right]\right] \right\rangle \\
\text{NEUTRAL } \left\langle \boxed{4}, \left[\text{RESTR } \left[\text{RELN give}\right]\right] \right\rangle
\end{array}\right]
\end{array}\right]
\end{array}\right]
\end{array}\right]
$$

As desired, the NEUTRAL list consists of the RELN of the verb and the CONTENT of the object, while the FOCUS list consists of the CONTENT of the prepositional phrase expressing the focussed indirect object. Note that there is a slight mismatch between the semantics and the information structure representation: the content of the indirect object in the CONTENT feature is in fact equivalent to the content of the NP which is the argument of the preposition *to*; the content of the whole prepositional phrase, on the other hand, has been copied onto the FOCUS list. This seems reasonable, however, as in our approach the whole PP is in focus.

The refinement module implemented for English analysis consists of 26 rules, accounting for the various information structure constraints we have identified. The application of the rules is quite fast, and the overall strategy seems in general to deliver the desired results. In generation, however, a different approach had to be taken, as such a high number of rules would constitute too large a search space for the generation algorithm – which, it is important to remember, has to create a syntactic structure which is not there to start with.

[4] Compound nominals are not treated in the present implementation. Thus, in order to get the correct Danish equivalents in the translation, compounds are written as single words in the input sentences.

Thus, INFOTYPE values are assigned in the generation lexicon. This is a more efficient solution as it makes it possible to restrict the number of rules considerably. On the other hand, it makes the lexicon less flexible, as default assignments such as [INFOTYPE focus] to complements following a verb in the default SO, due to their heuristic nature, conceptually seem to belong to the grammar rather than to the lexicon. To make the difference clearer, we show below one of the lexicon entries for the Danish verb *give*, namely the one accounting for dative shift:

(19) *Entry for Danish "give"*

```
giver ~
sign:{phon => phon:{string =>[giver|R], rest => R},
      named_spec=>spec:{specinfo=>specinfo:{synth=>y}},
      synsem => synsem:{
        syn => syn:{
           loc => loc:{
              head => vhead:{           /* head features */
                          maj => v,
                          vform => pres,
                          diath => act
                            },
              bar => zero,
              subcat => [synsem:{             /* object */
                            syn => syn:{
                               loc => loc:{
                                   head => nhead:{
                                             maj => n,
                                             case => acc},
                                   so => ndef_left,
                                   subcat => []}},
                             content => O,
                             context => context:{
                                 infotype => neutral}
                                 },
                          synsem:{             /* indirect object */
                            syn => syn:{
                               loc => loc:{
                                   head => phead:{
                                             maj => p,
                                             pform => for},
                                   so => ndef_right,
                                   subcat=>[]}},
                            content=> PP => cont:{
                                 restr=>restr:{arg1=>IO}},
                            context => context:{
                                infotype => focus}
                                   },
                          synsem:{             /* subject */
                            syn => syn:{
                               loc => loc:{
                                   head => nhead:{
                                            maj => n,
                                            case => nom
                                                 },
                                   subcat => []}},
```

```
                            content => S
                                  }
                        ],
                cat => v}},
          content => cont:{restr =>                          /* content */
                   restr:{reln=>X=>give, arg1=>S, arg2=>O, arg3=>IO}},
          context=>context:{                         /* context features */
                      infotype=>neutral,
                      infostr=>info:{
                            neutral => [cont:{
                            restr => restr:{reln => X}}]}}
                   }
       }.
```

The main difference with respect to the corresponding lexical entry for the English *give*, apart from a number of features relative to the syntactic shape of the various complements, is the fact that the INFOTYPE feature is already assigned to both internal complements, and that the RELN of the main verb is also already assigned to the NEUTRAL list (as desired in case of SO deviation; it is assigned to the FOCUS list otherwise). Thus, only three VBAR rules are necessary in generation to account for the various possibilities deriving from different information structure values: one for focussed complements, one for neutral ones, and one for adjuncts, which in the present implementation are always focussed[5]. As an example, we show below the rule generating neutral complements:

(20) *Generation rule for neutral complements*

```
sign:{phon => phon:{string => S_v, rest => R_xp},
      named_spec=>spec:{language=>da, specinfo=>specinfo:{synth=>y}},
      synsem => synsem:{                          /* topmost vbar node */
        syn => syn:{type => full,
                    loc => loc:{head => HEAD,
                              subcat => REST,
                              cat => vbar
                             },
                    bind => bind:{
                    gaps => gaps:{
                            gap_in => GAP4,
                            gap_out => GAP6}
                               },
                    seg => seg:{first=>F,last=>L},
                    str => phrase:{constr => h_compl}
                 },
        content => V,
        context => context:{presup => PRES,
                            infostr => info:{
                                  focus => F1,
                                  neutral => [N2|N]}, /* neutral
                            head_disc => Disc}           complement
                    }                                    taken off the
     }                                                   NEUTRAL list */
<
[
sign:{phon => phon:{string => S_v, rest => S_xp},
      synsem => synsem:{                              /* verbal head */
```

[5]To these must be added a rule allowing for the generation of empty complements to account for complement gapping.

```
          syn => syn:{type => full,
                      loc => loc:{head => HEAD,
                                  subcat => [C|REST],
                                  cat => (v;vbar)
                                 },
                      bind => bind:{
                        gaps => gaps:{
                              gap_in => GAP4,
                              gap_out => GAP5}
                                 },
                      seg => seg:{first=>F,last=>n},
                      str => phrase:{parsehead => y}
                   },
          content => V,
          context => context:{presup => PRES,
                              infostr => info:{
                                        neutral => N,
                                        focus => F1},
                              head_disc => Disc}
                      }
       },
    sign:{phon => phon:{string => S_xp, rest => R_xp},
          synsem => C => synsem:{syn => syn:{       /* neutral complement */
                      type => full,
                      seg => seg:{first=>n,last=>L},
                      bind => bind:{
                        gaps => gaps:{
                              gap_in => GAP5,
                              gap_out => GAP6}}},
          content => N2,
          context => context:{
                      infotype => neutral,
                      infostr => info:{
                               neutral => [N2]}}
                     }}
    ].
```

As expected, the content of the neutral complement is deleted from the NEUTRAL list of the mother node, and the rest of the NEUTRAL list is passed down to the verbal head together with the entire FOCUS list.

This strategy makes it possible to generate Danish sentences displaying either default SO or SO deviation. In particular, the translation of the English sentence *Information technology gives better communication possibilities to the research community* will be as expected:

(21) Informationsteknologi giver bedre kommunikationsmuligheder for forskningsverdenen.

(Information technology gives better communication possibilities to the research community.)

In conclusion, the implementation has shown that the heuristics proposed for the treatment of information structure in Danish unmarked declarative sentences can be formalised and applied in the generation component of an MT system to help the system generate a more accurate and varied output.

6.3 Implementing focussing constructions

To test the specifications provided in the preceding chapters, two of the focussing constructions discussed have been implemented, namely clefts and existential constructions.

6.3.1 Clefts

We have seen that Danish does not differ much from English as far as cleft constructions are concerned. Thus, the strategies used in analysis and generation to deal with these constructions are quite similar. Nevertheless, it must be noted that the Danish generation component does not rely on the structure built by the analysis grammar: Danish clefts are generated on the basis of the semantic and contextual information associated with the input sentence. Thus, the input sentence does not have to be a cleft for the system to generate a cleft as the Danish translation. One could imagine, for example, that a German sentence like *Gerade der Schlamm traegt zur Rostbildung bei* (Precisely the sludge contributes to rust formation) could be translated in Danish as *Det er slammet der bidrager til rustdannelse* (It is the sludge that contributes to rust formation). In this section, we focus on the strategy used in generation.

Let us consider the representation built by the English grammar for the cleft example *It is the sludge which contributes to rust formation*:

(22)
$$
\begin{bmatrix}
\text{SIGN} & \begin{bmatrix}
\text{PHON} & \begin{bmatrix}\text{STRING it is the sludge which \_ contributes to rust\_formation .}\end{bmatrix} \\
\text{SYNSEM} & \begin{bmatrix}
\text{CONTENT } \boxed{5} & \begin{bmatrix}
\text{RESTR} & \begin{bmatrix}
\text{RELN contribute} & \\
\text{ARG1 } \boxed{3} & \begin{bmatrix}
\text{RESTR} & \begin{bmatrix}\text{RELN sludge}\end{bmatrix} \\
\text{DET def}(\boxed{1}) & \\
\text{INDEX } \boxed{1} & \\
\text{NUMBER sing} &
\end{bmatrix}_{npcont} \\
\text{ARG2 } \boxed{4} & \begin{bmatrix}
\text{RESTR} & \begin{bmatrix}\text{RELN rust\_formation}\end{bmatrix} \\
\text{DET none}(\boxed{2}) & \\
\text{INDEX } \boxed{2} & \\
\text{NUMBER sing} &
\end{bmatrix}_{npcont}
\end{bmatrix}
\end{bmatrix} \\
\text{CONTEXT} & \begin{bmatrix}
\text{INFOSTR} & \begin{bmatrix}
\text{FOCUS} & \left\langle \boxed{3} \right\rangle \\
\text{NEUTRAL} & \left\langle \boxed{5}, \begin{bmatrix}\text{RESTR} & \begin{bmatrix}\text{RELN be}\end{bmatrix}\end{bmatrix} \right\rangle
\end{bmatrix} \\
\text{PRESUP} & \begin{bmatrix}\text{contribute} & \begin{bmatrix}\text{UNBOUND } \boxed{1}\end{bmatrix}\end{bmatrix} \\
\text{HEAD\_DISC given} &
\end{bmatrix}
\end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

As can be seen, the PRESUP feature is a term consisting of a predicate and an UNBOUND argument. The predicate corresponds to the RELN of the main verb in the cleft clause. The argument corresponds to a nominal index, which is structure-shared with the nominal index in the CONTENT feature of the cleft head *the sludge*. The UNBOUND argument refers thus to the variable in the presupposition of the sentence which is associated with a semantic restriction in the semantic representation.

Turning now to the information structure of the sentence, we see that the FOCUS list corresponds to the content of the cleft head, while the cleft clause and the copula verb are placed on the NEUTRAL list. Finally, the feature [HEAD_DISC given] indicates that the cleft head is a given piece of information in the discourse at hand.

Before turning to the generation of the equivalent Danish cleft, let us discuss the content of the INFOSTR list in more detail. What one would like to express is the fact that the content of the cleft clause *excluding* the cleft head is neutral information. However, due to the way in which lists are used in ALEP, we can only state the fact that the whole CONTENT of the cleft clause should be part of the NEUTRAL list. The problem is that it is not possible in ALEP (at least in the current version) to add more than one element at a time to a list. In other words, the following operation (where both A and B are lists consisting of zero, one or more elements) is disallowed:

(23)
$$\begin{array}{ccc} & [A|B] & \\ \diagup & & \diagdown \\ A & & B \end{array}$$

The left-hand value of the list in the specification of the mother node must correspond to a single element:

(24)
$$\begin{array}{ccc} & [B|A] & \\ \diagup & & \diagdown \\ A & & [B] \end{array}$$

This means that the formalism enforces a certain isomorphism between information structure and constituency: for each complement or adjunct that is added to the parse tree, only one item can be added to the information structure of the mother node in the tree. Hence, even if the complement is a clause – as in the example at hand – it can only contribute one element to the information structure of the main sentence. In other words, the information structure value of the SBAR in our example cannot be:

(25)
$$\begin{bmatrix} \text{NEUTRAL} \left\langle \begin{bmatrix} \text{RESTR} \begin{bmatrix} \text{RELN contribute} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{RESTR} \begin{bmatrix} \text{RELN rust\_formation} \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}$$

It has to be the whole CONTENT feature of the clause, namely:

(26)
$$\begin{bmatrix} \text{NEUTRAL} \left\langle \begin{bmatrix} \text{RESTR} \begin{bmatrix} \text{RELN contribute} \end{bmatrix} \\ \text{ARG1} \quad {}_{\textit{npcont}}\begin{bmatrix} \text{RESTR} \begin{bmatrix} \text{RELN sludge} \end{bmatrix} \\ \text{DET def}(\boxed{1}) \\ \text{INDEX } \boxed{1} \\ \text{NUMBER sing} \end{bmatrix} \\ \text{ARG2} \quad {}_{\textit{npcont}}\begin{bmatrix} \text{RESTR} \begin{bmatrix} \text{RELN rust\_formation} \end{bmatrix} \\ \text{DET none}(\boxed{2}) \\ \text{INDEX } \boxed{2} \\ \text{NUMBER sing} \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}$$

Thus, the head of the cleft (*sludge*) is represented both on the FOCUS list and on the NEUTRAL list as part of the content of the cleft clause. This is not optimal, however it does not impair the generation of the correct construction in Danish.

With the exception of the STRING feature, the representation shown in (22), complete with some syntactic information which has been omitted here, is used by the generation grammar to generate a cleft.

As mentioned earlier, the generation algorithm is a semantically-driven one. This means that it tries to downwards-expand the input representation (the main sentence node) until it finds the *pivot*. In Schieber et al. the pivot is the lowest node in a tree sharing the same semantic information as the root. In our case, however, the pivot also partly shares the CONTEXT features of the root. In particular, it shares the PRESUP feature. Thus, the algorithm downwards-expands the S node until it finds a verb node the content of which unifies with the content of the whole sentence, and which carries the same presupposition. This is the reading of *være* (be) occurring in cleft constructions. In the Danish lexicon, there are four entries corresponding to this use of *være*, accounting for it-clefts and wh-clefts with subject and complement gapping, namely:

(27) *Entries for cleft readings of "være"*

er ∼ itcleftv1[er, pres, be].
er ∼ itcleftv2[er, pres, be].
er ∼ whcleftv1[er, pres, be].
er ∼ whcleftv2[er, pres, be].

The expanded entry relevant to our example is the following one, where the head of the cleft is structure-shared with the gapped subject in the cleft clause:

(28) *Expanded entry for it-cleft reading of "være"*

```
sign:{
   phon =>
   phon:{
      string =>  [er | B],
      rest => B },
   synsem =>
   synsem:{
      syn =>
      syn:{
         loc =>
         loc:{
            head =>
            vhead:{
               maj => *,
               vform => pres,
               msaspect => simple,
               trans => yes,
               diath => act },
            subcat =>  [
                   synsem:{            /* the cleft head */
                      syn =>
                      syn:{
                         loc =>
                         loc:{
                            head =>
                            nhead:{
                               maj => *,
                               case => acc },
                            subcat => [],
                            so => n_a,
                            cat => n },
                         type => full },
                      content => B1 =>
                      npcont:{
                         restr => C1,
                         det => D1,
                         index => E1,    /* cleft head index */
                         number => F1 },
                      context =>
                      context:{
                         infotype => focus,
```

```
                 infostr =>
                 info:{
                    focus =>  [
                            npcont:{
                               restr => C1,
                               det => D1,
                               index => E1,
                               number => F1 }] },
                 disc => given } },

           synsem:{            /* the cleft clause */
              syn =>
              syn:{
                 loc =>
                 loc:{
                    head =>
                    vhead:{
                       msaspect => simple },
                    so => n_a,
                    cat => sbar },
                 type => full,
                 bind =>
                 bind:{
                    whhead => som,
                    rel =>  [
                            npcont:{
                               restr => C1,
                               det => D1,
/* rel index =                 index => E1,
   cleft head index */         number => F1 }] } },
              content => I2 =>
              cont:{
                 restr => J2 =>
                 restr:{
                    reln => K2,
                    arg1 => B1 } },
              context =>
              context:{
                 infotype => neutral,
                 infostr =>
                 info:{
                    neutral =>  [
                            cont:{
                               restr => J2 }] },
                 presup =>
                 none(   U2 ) } },

           synsem:{         /* the subject of the cleft clause */
              syn =>
              syn:{
                 loc =>
                 loc:{
                    head =>
                    nhead:{
                       maj => *,
                       case => nom,
                       nform => det },
                    so => n_a,
                    cat => perspron },
                 type => full },
              content =>
              npcont:{
                 index => E1 } }],  /* subject index =
     bar => zero,                      rel index */
     cat => v },
  bind =>
  bind:{
     gaps =>
```

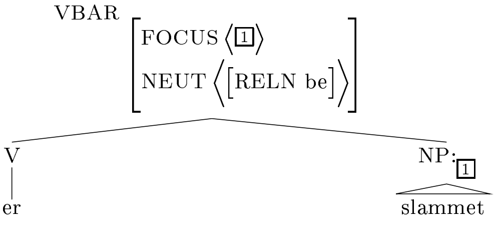
```
          gaps:{
             gap_in => [],
             gap_out => [] } } },
      content => I2,
      context =>                           /* cleft context */
      context:{
         infotype => neutral,
         infostr =>
         info:{
            neutral =>  [
                    cont:{
                       restr =>
                       restr:{
                          reln => be } }] },
         head_disc => given,
         presup =>
         cleftpresup(K2,
                 unbound(E1 ) ) } },
   named_spec =>
   spec:{
      specinfo =>
      specinfo:{
         synth => y } } }
```

Let us consider how the semantics of the cleft construction is computed. In order to achieve the desired effect, namely that the verb in the cleft clause rather than *være* (be) should be the semantic head of the cleft as a whole, the CONTENT feature of *være* is structure-shared with the CONTENT of the cleft clause, which corresponds to the SBAR subcategorised for by *være*. The value of the CONTENT feature will eventually consist of a RELN corresponding to the RELN of the verb in the SBAR, and of a number of arguments. The index of one of these arguments (in our case the subject of the SBAR) is structure-shared with the REL index, and through this with the index of the cleft head. This coindexation between the REL index and the extracted subject (or complement) in the SBAR is ensured as desired through the GAPS features. We return to this point below.

As for the CONTEXT feature, we see that it contains a cleft presupposition the predicate of which is structure-shared with the CONTENT of the verb in the cleft clause, while the argument corresponds as described earlier to the INDEX of the cleft head. Furthermore, the entry also contains the constraint that the discourse value of the cleft head should be 'given'. This feature distinguishes it-clefts from wh-clefts.

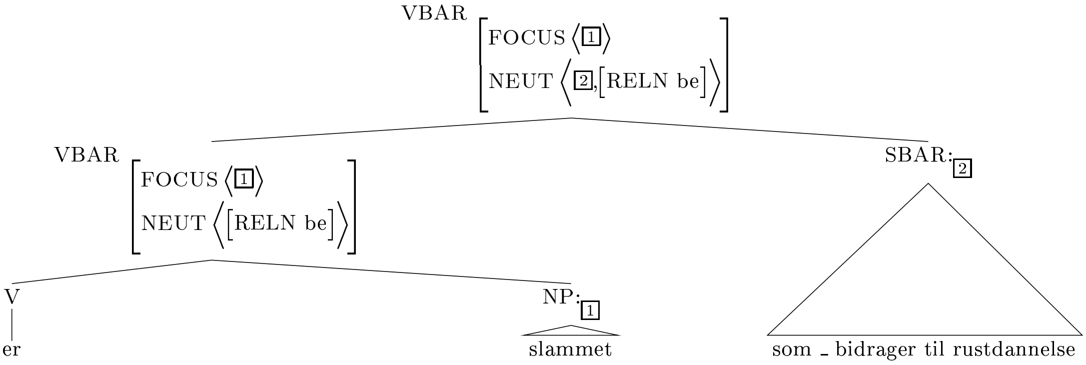
The feature INFOSTR is also already assigned in the lexicon entry so that the cleft head is focussed, while the verb and the cleft clause are neutral.

Having found the pivot and unified it with the lexicon entry shown above, the algorithm proceeds both top-down and bottom-up from it to generate the various complements and adjuncts. This is done by applying grammar rules that have the pivot as their semantic head, and that are part of a chain of rules connecting it back to the root of the tree which is being built. In this case, the generation algorithm will build a VBAR consisting of the verb *er* (is) and a focussed NP complement, namely:

(29)


On top of it, it will then built another VBAR attaching a neutral SBAR complement to the representation which is being built:

(30)



The SBAR itself is generated by downwards-expanding the semantic representation down to the verbal head *bidrager* (contributes) and then generating the prepositional complement *til rustdannelse* (to rust formation) and the empty subject. A rule taking a relative pronoun and a clause containing a gap as daughters of an SBAR concludes the chain. We show the rule below:

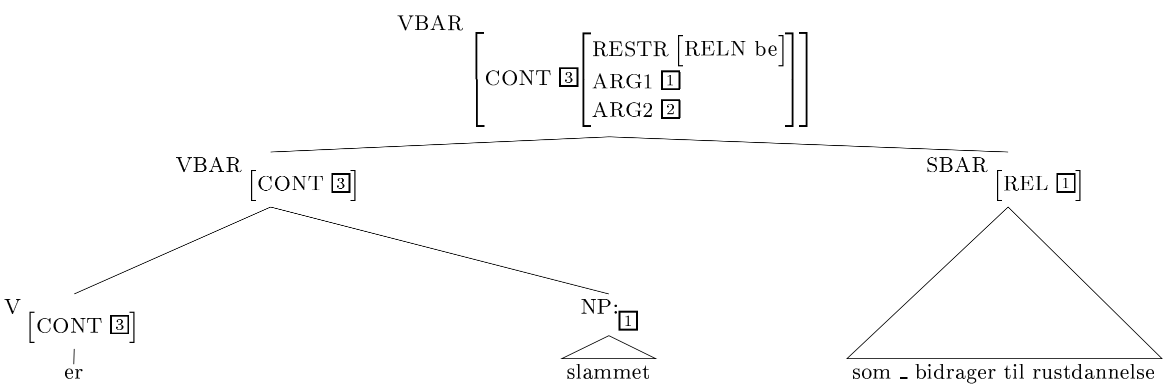
(31) *Rule generating an SBAR*

```
sign:{phon => phon:{string => S_np, rest => R_s},
      named_spec=>spec:{language=>da, specinfo=>specinfo:{synth=>y}},
      synsem => synsem:{                 /* topmost sbar node */
        syn => syn:{type => full,
                    loc => loc:{head => vhead:{
                                        maj=>Maj,
                                        vform=>pres,
                                        msaspect=>Aspect,
                                        trans=>T,
                                        diath=>D},
                              subcat => [],
                              cat => sbar
                             },
                    seg => seg:{first=>F,last=>L},
                    str => phrase:{constr => h_compl},
                    bind => bind:{
                                 gaps => gaps:{
                                      gap_in => IN,
                                      gap_out => OUT
                                              },
                                whhead => W,
                                rel => Rel
                               }
                 },
        content => Vp,
        context => context:{presup => none(_)}
                  }
     }
<
[
sign:{phon => phon:{string => S_np, rest => S_s},
       synsem => GAP => synsem:{                 /* relative pronoun
         syn => syn:{type => full,                  structure-shared
                     loc:{                          with the gap */
                          cat=>relpron
                                  },
                     seg => seg:{first=>F,last=>n},
                     bind=>bind:{
```

```
                                whhead => W,
                                rel => Rel
                               }
                    }
                            }
         },
sign:{phon => phon:{string => S_s, rest => R_s},
      synsem => synsem:{                          /* clause containing
        syn => syn:{type => full,                    subject gap */
                    loc => loc:{head => vhead:{
                                       maj=>Maj,
                                       vform=>pres,
                                       msaspect=>Aspect,
                                       trans=>T,
                                       diath=>D},
                              subcat => [],
                              stype => sub,
                              cat => s
                             },
                    seg => seg:{first=>n,last=>L},
                    str => phrase:{parsehead => y},
                    bind => bind:{
                                 gaps => gaps:{
 /* the gap bound off by              gap_in => [GAP|IN],
    the rel pron is taken             gap_out => OUT
    off the GAP_IN list */                    }
                                }
                 },
        content => Vp,
        context => context:{presup => none(_)}
                     }
     }
].
```

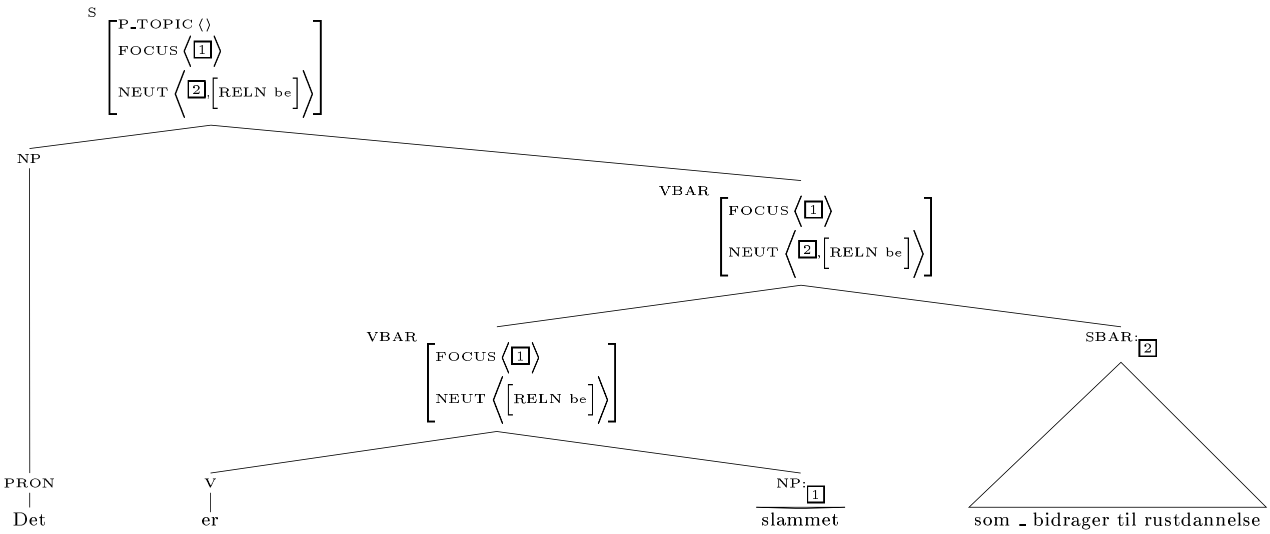
Note especially the way in which the GAPS feature is used: the SYNSEM value of the relative pronoun is structure-shared with the first element on the GAP_IN list of the clause following the pronoun. This element is deleted from the GAP_IN list of the mother SBAR node: as discussed earlier, this means that the gap has been bound off by the filler. The REL feature, which contains the content of the relative pronoun, is passed on upwards in the structure, to be unified with the content of the NP constituting the cleft head, as shown below:

(32) *The REL feature*



Finally, the subject pronoun *det* (it) is generated by a rule consisting of a mother S node, and two daughters, the first of which is a pleonastic pronoun and the second one a VBAR having the same pleonastic pronoun on its SUBCAT list. The final structure will look as follows:

(33) *It-cleft representation*



Thus, the implementation shows that the CONTENT and CONTEXT features provide a sufficient basis for the generation algorithm to output the correct cleft construction in Danish, so that as desired, the generation of a cleft is made independent of the syntactic form of the source sentence.

6.3.2 Existential constructions

The last topic we are going to deal with in this chapter is the strategy used in the Danish generation grammar to account for existential constructions. In Chapter 5, Danish existential constructions were claimed to be detopicalising constructions, and a lexical treatment of these constructions was proposed. According to this approach, verbs that can occur in existential sentences are coded accordingly in the dictionary. This strategy has been followed in the implementation. Consider thus the following two entries from the Danish generation lexicon, corresponding to the verbs *fremkommer* (occur/occurs) and *kræves* (are/is required):

(34) *Entries relative to existential constructions*

fremkommer ~ der_o_v['fremkommer', pres, emerge].
kræves ~ der_passv['kræves', pres, require].

Macros have been defined to expand both active and passive verbs that can take a pleonastic *der* (there) subject. The macro relevant to *kræves* was shown earlier (cf. example 7 in this chapter), together with the corresponding expanded entry (cf. example 8). Here we show the macro that describes the SUBCAT feature of an active verb like *fremkomme*:

(35) *Macro for active der-constructions*

```
macro(der_o[Arg1],
        subcat=>[
          synsem:{syn=>syn:{            /* postponed NP */
                    loc=>loc:{
                      head=>nhead:{
```

```
                            maj=>n,
                            case=>acc
                               },
                  so=>n_a,
                  cat=>(np;nbar;apn;n;perspron;relpron),
                  subcat=>[]
                          }
                       },
            content=>Arg1,
            context=>context:{
                        infotype=>focus}
            },
    synsem:{syn=>syn:{           /* expletive subject */
               loc=>loc:{
                  head=>nhead:{
                            maj=>n,
                            nform=>der,
                            case=>nom
                               },
                  so=>n_a,
                  subcat=>[]
                        }
                     }
            } ]  ).
```

The macro defines a SUBCAT list consisting of two elements: the first one is the only internal complement – corresponding to what we have called the postponed NP in our account of existential constructions; the second element is the pleonastic *der*. Note that the postponed NP is marked [INFOTYPE focus]: recall that our claim was that existential constructions are focussing constructions, and that the postponed NP always constitutes the focus. Note also that the same constituent contributes to the CONTENT of the main sentence (in the expanded entry, the variable 'Arg1' is structure-shared with the value of ARG1 in the CONTENT of the verb), as opposed to the pleonastic subject. All that is needed to make the treatment complete is to ensure that the P_TOPIC list of the main sentence is empty. This is done in a grammar rule generating main sentences with a pleonastic subject, as shown below:

(36) *Sentence rule taking pleonastic subjects*

```
sign:{phon => phon:{string => S_np, rest => R_punct},
      named_spec=>spec:{language=>da, specinfo=>specinfo:{synth=>y}},
      synsem => synsem:{                             /* topmost S node */
        syn => syn:{type => full,
                    loc => loc:{head => vhead:{
                                         maj=>Maj,
                                         vform=>pres,
                                         msaspect=>Aspect,
                                         trans=>T,
                                         diath=>D},
                                bar => max,
                                stype => main,
                                subcat => [],
                                cat => s
                               },
                    seg => seg:{first=>y,last=>y},
                    str => phrase:{constr => h_compl},
```

```
                    bind => bind:{gaps => gaps:{
                                   gap_in => [],
                                   gap_out => []},
                                  rel => []}
                 },
        content => Vp,
        context => context:{presup => none(_),
                             infostr => info:{
                                      p_topic => [],
                                      focus => Focus,
                                      neutral => Neutral}}
                    }
     }
<
[
sign:{phon => phon:{string => S_np, rest => S_vp},
      synsem => C => synsem:{           /* pleonastic der subject */
        syn => syn:{type => full,
                    loc=>loc:{cat=>perspron,
                              head=>nhead:{
                                    nform=>der}},
                    seg => seg:{first=>y,last=>n}
                   }
             }},

sign:{phon => phon:{string => S_vp, rest => R_vp},
      synsem => synsem:{
        syn => syn:{type => full,                /* verbal head */
                    loc => loc:{head => vhead:{
                                        maj=>Maj,
                                        vform=>pres,
                                        msaspect=>Aspect,
                                        trans=>T,
                                        diath=>D},
                              subcat => [C],
                              cat => (v;vbar)
                             },
                    seg => seg:{first=>n,last=>y},
                    str => phrase:{parsehead => y},
                    bind => bind:{gaps => gaps:{
                                    gap_in => [],
                                    gap_out => []}}
                 },
        content => Vp,
        context => context:{presup => none(_),
                             infostr => info:{
                                      focus => Focus,
                                      neutral => Neutral}}
                     }
     },

sign:{phon => phon:{string => R_vp, rest => R_punct},
      synsem => synsem:{                             /* full stop */
        syn => syn:{type => full,
```

```
                        loc=>loc:{cat=>punct}}}
          }
      ].
```

The rule generates an S node the daughters of which are a *der* subject, a V or a VBAR and a punctuation sign. The FOCUS and NEUTRAL lists are shared between mother node and head daughter, whereas the P_TOPIC list of the mother is simply set to be the empty list.

The mechanism used to generate existential constructions is thus quite simple: all that is required is for the argument which in a non-existential sentence would be the grammatical subject, to be focussed. However, it remains to be shown how we avoid generating both an existential and a non-existential version of the same input sentence. For example, one could imagine that the input English sentence below would yield both the Danish translations shown below:

(37) Another antioxidant is required.

a. En anden antioxidant kræves.
(Another antioxidant is required.)
b. Der kræves en anden antioxidant.
(There is required another antioxidant.)

As a matter of fact, assuming that the indefinite subject in the English sentence is assigned to the FOCUS list (as is the case in our implementation) the generation grammar can only produce (b) as a correct translation. The reason is simply that a general constraint that the grammatical subject cannot be focussed, has been enforced in all the sentence rules in the Danish generation grammar. Thus, although both readings of the verb *kræves* are coded in the dictionary, only the one taking a pleonastic *der* subject is used in this case.

The topmost feature structure of the representation built in generation (the CONTENT and CONTEXT features of which are the same in the corresponding analysis representation) is shown below:

(38)
$$\left[\text{SIGN}\ \left[\begin{array}{ll} \text{PHON} & [\text{STRING der kræves en\_anden antioxidant.}] \\ \text{SYNSEM} & \left[\begin{array}{ll} \text{CONTENT} & \left[\text{RESTR} \left[\begin{array}{ll} \text{RELN} & \text{require}\\ \text{ARG1} & {}_{\textit{nullcont}}[\,] \\ \text{ARG2} & \boxed{2}\ {}_{\textit{npcont}}\left[\begin{array}{ll}\text{RESTR} & [\text{RELN antioxidant}]\\ \text{DET} & \text{also\_exists}(\boxed{1})\\ \text{INDEX} & \boxed{1}\\ \text{NUMBER} & \text{sing}\end{array}\right] \end{array}\right]\right] \\ \text{CONTEXT} & \left[\begin{array}{ll}\text{INFOSTR} & \left[\begin{array}{ll} \text{P\_TOPIC} & \langle\rangle \\ \text{FOCUS} & \left\langle \boxed{2}, [\text{RESTR}\ [\text{RELN require}]]\right\rangle\\ \text{NEUTRAL} & \langle\rangle\end{array}\right]\\ \text{PRESUP} & [\text{none}]\\ \text{HEAD\_DISC} & \text{given}\end{array}\right]\end{array}\right]\end{array}\right]\right]$$

To conclude, the specifications provided for the treatment of existential constructions have proved a simple and elegant way to account for this quite common syntactic phenomenon in the Danish generation grammar.

6.4 Summary

In this chapter, a concrete implementation of some of the specifications provided in this thesis for the treatment of information structure in an MT system, has been presented. The implementation consists of a toy English-Danish translation system running on the ALEP language engineering platform.

First a short account of the formalism supported by ALEP was given: ALEP is based on the concept of typed feature structures and thus provides a good environment for writing and testing HPSG-like grammars.

Then the general architecture of the system was described. The two main architectural choices discussed were the division of labour between the core grammar and the refinement module on the one hand, and the generation style adopted in the present implementation on the other. The use of refinement rules to add features to the analyses produced by the parser on the basis of the core grammar, has proved an elegant way of modelling the heuristic principles for the computation of information structure that were proposed earlier. As far as generation is concerned, the choice of letting the semantic and information structure information associated to the input sentences drive the generation process, has also proved a viable approach, allowing for the generation of a more varied and coherent output. This strategy makes the generation of focussing and topicalising contructions independent of the particular syntactic structures associated with them in the source language.

Furthermore, the way in which the particular specifications proposed in the preceding chapters for a number of phenomena has been implemented, was discussed in some detail. Both in the case of the heuristic principles computing the information structure of unmarked declarative sentences, and of the two focussing constructions that have been implemented – clefts and existential constructions, the results of the implementation show that the specification proposed constitute a feasible and elegant way of dealing with these phenomena in an MT system.

Chapter 7

Conclusions

At this point, let us consider whether we have achieved the goals set at the beginning of this work. Our main objective was to propose a method for the treatment of information structure which would contribute to the generation of more coherent and varied output in MT. This general objective included a number of more specific subgoals.

The first of these subgoals was to show that the absence of a proper treatment of information structure impoverishes the quality of MT output. Our discussion of a number of different approaches to MT has shown that this is indeed the case, and that this failure is especially serious in relatively sophisticated systems, where the surface structure of the translation can be quite far from that of the source sentence. To remedy this situation, a number of different theories of information structure were reviewed in order to understand what principles a treatment of information structure in MT should be based on.

The most promising and viable approaches in this context were found to be those focussing on the internal structure of the individual sentences of a text rather than on a theory of discourse, since integrating a full discourse model into a fully operational MT system still seems to constitute an unrealistic task. Thus, a model was proposed – this was our second subgoal – according to which the information structure of a sentence is described along the two dimensions of topicality and focality. Topicality concerns the entities (topics) participating in a discourse. It was claimed that, with the exception of topicless sentences, it is possible to single out one of the topics occurring in a sentence as the most important one, i.e. the primary topic, and that it is important to identify the primary topic of a source sentence to generate it in the right place in the target sentence. Focality, on the other hand, deals with the fact that one or more elements in a sentence are emphasised to indicate a change in the pragmatic information built by the discourse as a whole. As for the primary topic, there is a complex relationship between the focus of a sentence and its syntax, and an explicit marking of the focus in the source can contribute to a more accurate translation.

On the basis of these dimensions, three primitive categories were defined, corresponding to features in a typed feature structure formalism. These are P_TOPIC, FOCUS and NEUTRAL. Together, they provide the necessary information to re-express the information structure of an input sentence using the syntactic means available in the target language. A proposal was made to add the three features to the CONTEXT part of an HPSG-like sentence analysis, and general principles concerning the computation of these three features were formulated.

Most of the literature on information structure deals with Slavic languages or English. An important aim of this work – our third subgoal – was to explore the relationship between information structure and syntax in Danish. It was thus shown that syntactic factors such as definiteness, word order and semantic content of certain constituents (adjuncts in particular), can be used to determine the information structure of a Danish sentence. A number of heuristics were proposed based on a set of such criteria. However, more work would be necessary to establish the validity of these heuristics across different text types. In addition to the fact that certain syntactic and semantic factors seem to indicate the most likely information structure interpretation of a sentence, it was shown that in certain syntactic constructions a more direct mapping between syntax and information structure is possible. This is the case with clefts, existential sentences and topicalisation. This mapping was described and formalised, and specifications for the automatic analysis and generation of all three constructions were provided. In the case of extraposition, on the other hand, it was claimed that, contrary to what one would expect, the disruption of the default grammatical order does not reflect a marked information structure, but is rather due to the difficulty of processing clausal subjects.

Providing formal specifications for the treatment of the phenomena under discussion was our fourth subgoal. A subset of these specifications was successfully implemented in the ALEP language engineering platform, so that at least for this subset the feasibility of the strategies proposed has been demonstrated in practice. The outcome of the implementation is that adding an extra layer of information to the analysis representations built by the parser, namely the features pertaining to information structure, makes it possible to choose between different surface realisations of the same semantic content. For example, the system can be made to choose between different mutual orders of the complements of a verb, or between a cleft and a non-cleft version of the same proposition, or between an existential sentence and its non-existential counterpart.

In conclusion, then, the treatment of information structure proposed, which consists of a general model for the representation of information structure as well as algorithms for the computation of the information structure of Danish sentences, serves its purpose by providing a means of generating more accurate automatic translations.

This work has raised a number of questions that deserve further research. The first one regards the very definition of our primitive categories. While it was possible to give FOCUS a uniform formal semantic characterisation coming from the literature on focus semantics, the category P_TOPIC could only be defined informally as 'what the sentence is about'. A possible solution to the problem may come from centering theory, according to which the centre of a sentence is defined formally in terms of its relation with the most salient topic of the preceding sentence. Perhaps the primary topic could profitably be defined in terms of centering, thus widening the approach to information structure to include intersentential information.

Another issue that was raised but clearly not exhausted in this thesis is the relationship between focus and presupposition: our claim was that it is not necessary to make this relation explicit for the purposes of MT. However, in our treatment of clefts a representation of the presupposition expressed by clefts is part of the representation proposed. Thus, it remains to be investigated to what extent a general theory of presupposition could or should be integrated in an MT system.

Finally, the third issue that needs to be mentioned, and which has not been dealt with at all in this research, is the resolution of pronominal anaphora. This is clearly an important task for a system aiming at the generation of coherent text. We feel again that centering theory, for its relative simplicity and its computational tractability, may provide a promising approach which it would be worthwhile to test in an MT application.

Bibliography

Ahrenberg, L. (1990), A grammar combining phrase structure and field structure, *in* H. Karlgren, ed., 'Proceedings of COLING-90', Vol. II, Helsinki, pp. 1–6.

Allegranza, V., Bennett, P., Durand, J., Eynde, F. V., Humphreys, L., Schmidt, P. & Steiner, E. (1991), Linguistics for Machine Translation: The Eurotra linguistic specifications, *in* C. Copeland, J. Durand, S. Krauwer & B. Maegaard, eds, 'The Eurotra Linguistic Specifications', Vol. 1 of *Studies in MT and NLP*, Luxembourg: Office for Official Publications of the European Communities, pp. 15–124.

Allwood, J. (1982), The Complex NP Constraint in Swedish, *in* E. Engdahl & E. Ejerhed, eds, 'Readings on Unbounded Dependencies in Scandinavian Languages', Stockholm: Almquist & Wiksell, pp. 15–32.

Alshawi, H., Arnold, D. J., Backofen, R., Carter, D. M., Lindop, J., Netter, K., Pulman, S. G., Tsujii, J. & Uszkoreit, H. (1991), Eurotra ET6/1: Rule formalism and virtual machine design study (final report), Technical report, CEC: Luxembourg.

Andersson, L. (1982), What is Swedish an exception to? Extractions and island constraints, *in* E. Engdahl & E. Ejerhed, eds, 'Readings on Unbounded Dependencies in Scandinavian Languages', Stockholm: Almquist & Wiksell, pp. 33–45.

Arnold, D., Balkan, L., Humphreys, R. L., Meijer, S. & Sadler, L. (1994), *Machine Translation - An Introductory Guide*, Manchester: NCC Blackwell.

Bennett, P. & Paggio, P., eds (1993), *Preference in Eurotra*, Vol. 3 of *Studies in MT and NLP*, Luxembourg: Office for Official Publications of the European Communities.

Bergenholtz, H. (1988), 'Dk87: Et korpus med dansk almensprog', *Hermes* **1**, 229–237.

Brennan, S. E., Friedman, M. W. & Pollard, C. J. (1987), A centering approach to pronouns, *in* 'Proceedings of the 25th ACL Meeting', Stanford: Stanford University.

Brown, K. & Miller, J. (1980), *Syntax: A Linguistic Introduction to Sentence Structure*, London: Hutchinson.

Chafe, W. L. (1976), Givenness, contrastiveness, definiteness, subjects, topics, and point of view, *in* C. N. Li, ed., 'Subject and Topic', New York: Academic Press, pp. 25–56.

Chomsky, N. (1971), Deep structure, surface structure and semantic interpretation, *in* D. Steinberg & L. A. Jakobovitz, eds, 'Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology', London: Cambridge University Press, pp. 183–216.

Dahl, Ö. (1974), Topic-comment structure revisited, *in* Ö. Dahl, ed., 'Topic and Comment, Contextual Boundness and Focus', number 6 *in* 'Papiere Zur Textlinguistik', Hamburg: Helmut Buske, pp. 1–24.

Daneš, F. (1966), *Les problèmes du centre et de la périphérie du système de la langue*, number 2 *in* 'Travaux Linguistiques de Prague', Prague-Paris.

Declerck, T. (1996), Dealing with cross-sentential anaphora resolution in alep, *in* 'Proceedings of COLING-96', Vol. I, Copenhagen, pp. 280–285.

Delin, J. L. (1989), Cleft Constructions in Discourse, PhD thesis, University of Edinburgh.

Diderichsen, P. (1946), *Elementær dansk grammatik*, Copenhagen: Gyldendal. 3rd edition 1962.

Dik, S. (1989), *The Theory of Functional Grammar*, Functional Grammar Series, Dordrecht: Foris Publications.

Dyvik, H. (1991), Linguistics and machine translation, *in* 'Proceedings of the Scandinavian Conference on Computational Linguistics', Bergen, pp. 67–78.

Engdahl, E. & Vallduví, E. (1996), Information packaging in HPSG, *in* C. Grover & E. Vallduví, eds, 'Studies in HPSG', Vol. 123 of *Edinburgh Working Papers in Cognitive Science*, Centre for Cognitive Science, University of Edinburgh, pp. 1–32.

Erteschik-Shir, N. (1982), Extractability in Danish and the pragmatic principle of dominance, *in* E. Engdahl & E. Ejerhed, eds, 'Readings on Unbounded Dependencies in Scandinavian Languages', Stockholm: Almquist & Wiksell, pp. 175–190.

Firbas, J. (1992), *Functional Sentence Perspective in Written and Spoken Communication*, Cambridge: Cambridge University Press.

Gazdar, G., Klein, E., Pullum, G. & Sag, I. (1985), *Generalized Phrase Structure Grammar*, Cambridge, Mass: Harvard University Press.

Geluykens, R. (1992), *From Discourse Phenomena to Grammatical Construction*, Amsterdam: John Benjamins Publishers.

Givon, T. (1983), *Topic Continuity in Discourse*, Amsterdam: John Benjamins Publishers.

Groenendijk, M. & Simpkins, N. K. (1995), *ALEP Reference Guide*, Luxembourg: Cray Systems.

Grosz, B. (1977), The representation and use of focus in dialogue understanding, Technical report, Technical Note 151. Stanford Research Institute, Menlo Park, CA.

Grosz, B. J., Joshi, A. K. & Weinstein, S. (1995), 'Centering: A framework for modeling the local coherence of discourse', *Computational Linguistics* **21**(2), 203–225.

Haegeman, L. (1991), *Government & Binding Theory*, Oxford & Cambridge: Blackwell.

Hajičová, E. (1987), Focussing – a meeting point of linguistics and artificial intelligence, *in* P. Jorrand & V. Sgurev, eds, 'ARTIFICIAL INTELLIGENCE II: Methodology, Systems, Applications', Elsevier Science Publishers B.V. (North-Holland), pp. 311–321.

Hajičová, E., Kuboň, P. & Kuboň, V. (1990), Hierarchy of salience and discourse analysis and production, *in* H. Karlgren, ed., 'Proceedings of COLING-90', Vol. III, Helsinki, pp. 144–166.

Hajičová, E., Sgall, P. & Skoumalová, E. (1993), Identifying topic and focus by an automatic procedure, *in* 'Proceedings of the 7th Conference of the EACL', Utrecht, pp. 178–182.

Hajičová, E., Skoumalová, E. & Sgall, P. (1995), 'An automatic procedure for topic-focus identification', *Computational Linguistics* **21**(1), 81–94.

Halliday, M. A. K. (1967), 'Notes on transitivity and theme in English (part 2)', *Journal of Linguistics* **3**, 199–243.

Hansen, E. (1980), 'Fremmedsprogsundervisningens skadelige indflydelse på den danske grammatik og det danske sprog gennem tiderne', *SAML* **6**, 281–293.

Hansen, E. & Lund, J. (1983), *Sæt Tryk På. Syntaktisk Tryk i Dansk*, Copenhagen: Lærerforeningernes Materialeudvalg.

Hauenschild, C. (1988), Discourse structure - some implications for Machine Translation, *in* D. Maxwell, ed., 'New Directions in MT. Distributed Language Translation', Vol. 4, Dordrecht: Foris Publications, pp. 145–155.

Hawkins, J. A. (1994), *A Performance Theory of Order and Constituency*, Cambridge Studies in Linguistics, Cambridge: Cambridge University Press.

Heltoft, L. (1986), 'The pragmatic syntax of Danish *der*-constructions', *ROLIG-papir 40* pp. 1–23.

Hoffman, B. (1996), Translating into free word order languages, *in* 'Proceedings of COLING-96', Vol. I, Copenhagen, pp. 556–561.

Hutchins, W. J. & Somers, H. L. (1992), *An Introduction to Machine Translation*, Academic Press, Harcourt Brace Jovanovich Publishers.

Levinson, S. C. (1983), *Pragmatics*, Cambridge: Cambridge University Press.

Lumsden, M. (1988), *Existential Sentences*, Routledge: London.

Maegaard, B. & Hansen, V. (1995), PaTrans - Machine Translation of patent texts from research to practical application, *in* 'II Language Engineering Convention', London, pp. 1–8.

Mathesius, V. (1975), *A Functional Analysis of Present Day English on a General Linguistic Basis*, The Hague.

McKeown, K. R. (1985), *Text Generation*, Cambridge: Cambridge University Press.

Milsark, G. (1974), *Existential sentences in English*, Cambridge, Mass: MIT Press.

Nimb, S. (1993), En formel semantisk analyse af danske 'Der'-sætninger, Master's thesis, University of Copenhagen.

Nølke, H. (1984), 'Clefting in Danish?', *NyS* **14**, 72–111.

Paggio, P. (1994), Theme and focus in an MT application, *in* P. Bosch & R. van der Sandt, eds, 'Focus & Natural Language Processing', number 6 *in* 'IBM Working Papers', Heidelberg: IBM Deutschland Informationssysteme GmbH, pp. 553–562.

Pollard, C. & Sag, I. A. (1987), *Information-Based Syntax and Semantics*, Stanford: CSLI.

Pollard, C. & Sag, I. A. (1994), *Head-Driven Phrase Structure Grammar*, Chicago: The University of Chicago Press.

Postal, P. P. & Pullum, G. K. (1988), 'Expletive noun phrases in subcategorised positions', *Linguistic Enquiry* **19**(4), 635–670.

Povlsen, C., Paggio, P. & Underwood, N. (1995), LINDA specifications for phrase structure in Danish, Technical report, Centre for Language Technology: Copenhagen.

Prince, E. F. (1979), On the given-new distinction, *in* 'Papers from the 15th Regional Meeting of the Chicago Linguistics Society', Department of Linguistics, University of Chicago, pp. 267–278.

Prince, E. F. (1981), Towards a taxonomy of given-new information, *in* P. Cole, ed., 'Radical Pragmatics', London: Academic Press, pp. 223–255.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1995), *A Comprehensive Grammar of the English Language*, London: Longman.

Rochemont, M. (1989), *Implementing Focus in Machine Translation*, number 9 *in* 'Eurotra-D Working Papers', Saarbrücken: IAI/Eurotra-D.

Rochemont, M. S. & Culicover, P. W. (1990), *English Focus Constructions and the Theory of Grammar*, Cambridge: Cambridge University Press.

Rooth, M. (1992), 'A theory of focus interpretation', *Natural Language Semantics* **1**(1), 75–116.

Schmidt, P., Rieder, S., Theofilidis, A. & Declerc, T. (1996), Lean formalisms, linguistic theory, and applications. grammar development in alep, *in* 'Proceedings of COLING-96', Vol. I, Copenhagen, pp. 286–291.

Selkirk, E. (1984), *Phonology and Syntax: The Relation Between Sound and Structure*, Cambridge, Mass: MIT Press.

Sgall, P., Hajičová, E. & Panenova, J. (1986), *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Dordrecht: Reidel Publishing Company.

Shieber, S. M., Pereira, F. C. N., van Noord, G. & Moore, R. C. (1990), 'Semantic-head-driven generation', *Computational Linguistics* **16**(1), 30–41.

Sidner, C. (1979), Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse, PhD thesis, MIT, Cambridge Mass.

Steinberger, R. (1994), Treating 'free word order' in Machine Translation, *in* 'Proceedings of COLING-94', Vol. I, Kioto, pp. 69–75.

Strawson, P. F. (1964), 'Identifying reference and truth values', *Theoria* **30**, 96–118. Reprinted in Strawson, P. F. (1971), *Logico-Linguistic Papers*, London: Methuen.

Styś, M. E. & Zemke, S. S. (1995), Incorporating discourse aspects in English-Polish MT: Towards robust implementation, *in* 'Recent Advances in NLP 95'.

Sundman, M. (1980), Existentialkonstruktionen i svenskan, PhD thesis, Äbo Akademi.

Togeby, K. (1982), *Grammaire française*, Copenhagen: Akademisk Forlag.

Togeby, O. (1992), *PRAXT. Pragmatisk tekstteori*, Århus: Aarhus Universitetsforlag.

Vallduví, E. & Engdahl, E. (1995), 'The linguistic realisation of information packaging', *Linguistics*. forthcoming.

van Leusen, N. & Kálmán, L. (1994), The interpretation of free focus, unpublished manuscript.

van Noord, G., Dorrepaal, J., van der Eijk, P., Florenza, M., Ruessink, H. & des Tombe, L. (1991), 'An overview of MiMo2', *Machine Translation* **6**(3), 201–214.

Vikner, S. (1995), *Verb Movement and Expletive Subjects in the Germanic Languages*, Oxford: Oxford University Press.

Vlk, T. (1988), Topic/focus articulation and intentional logic, *in* 'Proceedings of COLING-88', Budapest, pp. 720–725.

Dansk resumé

Denne afhandlings overordnede mål er at foreslå en metode til behandling af informationsstruktur i maskinoversættelse (MT) med det formål at skabe mere varieret og sammenhængende outputtekst, end det er tilfældet i de fleste eksisterende systemer.

Som diskussionen af en række forskelligartede MT-systemer har påpeget er eksisterende algoritmer for det meste ikke i stand til at skelne mellem sætningstyper, som inholdsmæssigt svarer til hinanden, men hvor sætningsleddene indgår i forskellige diskursmæssige relationer. Denne mangel har vist sig at være specielt alvorlig i forholdsvis komplicerede systemer, hvor der i oversættelsen ofte ses bort fra inputsætningens ordrækkefølge.

I afhandlingen diskuteres en del forskellige teorier om informationsstruktur, og der argumenteres for, at de mest lovende kandidater til en MT-applikation er de teorier, hvor informationsstruktur betragtes som et aspekt vedrørende de enkelte sætninger snarere end hele teksten. Dette hænger sammen med, at integrering af en fuldt udviklet diskursteori i et MT-system stadig synes at være en urealistisk opgave. I den model, som er brugt i denne afhandling, består en sætnings informationsstruktur af to forskellige aspekter. Det ene kaldes *topikalitet*. Det handler om, hvordan de diskursreferenter (*topics*), som optræder i en sætning, relaterer sig til den foregående tekst. Det hævdes, at en af disse referenter har en særlig status, idet den udgør sætningens hovedemne (*primary topic*). Hovedemnet er det, som sætningen handler om. Det andet aspekt, som indgår i informationsstrukturen, handler om sætningens fokus. Ordet bruges til at betegne det forhold, at der i enhver sætning er en del, der fremhæves, mens sætningens øvrige dele danner baggrund. Fokus defineres desuden ved hjælp af formel semantik som en kontrastmængde. En formel definition af begrebet hovedemne opstilles ikke, men det bliver foreslået, at en sådan definition muligvis kunne opnås inden for *centering*-teori. Både for hovedemne og for fokus gælder det, at der i ethvert sprog findes specifikke regler for, hvordan de kan udtrykkes i den syntaktiske struktur. Derfor er det vigtigt at identificere dem i inputsætningen for korrekt at kunne gengive informationsstrukturen i oversættelsen.

Topikalitet og fokus formaliseres her ved hjælp af tre basiskategorier, som svarer til træk i en formalisme baseret på trækstrukturer. De tre træk er P_TOPIC, NEUTRAL og FOCUS. Tilsammen udtrykker de al den information, som er nødvendig for at kunne gengive kildetekstens informationsstruktur ved hjælp af målsprogets egne syntaktiske midler. Trækkene foreslås tilføjet CONTEXT-delen af en HPSG-inspireret sætningsanalyse og generelle principper til beregning af deres værdi opstilles.

Forholdet mellem informationsstruktur og syntaks er et centralt emne i denne afhandling. Den relevante forskningslitteratur handler om slaviske sprog eller engelsk. Et vigtigt mål ved dette projekt er imidlertid at undersøge, hvilke relationer der er mellem informationsstruktur og syntaks på dansk. Som resultat af denne undersøgelse påvises, at lingvistiske faktorer såsom bestemthed, ordrækkefølge og visse sætningsleds semantiske egenskaber kan bruges til at formulere heuristiske principper til at beregne informationsstruktur i analysefasen og til at vælge den mest adækvate syntaktiske struktur i generering. Det vil dog kræve mere arbejde at fastlægge, hvorvidt de generaliseringer, som er udtrykt i disse principper, gælder bredt i forskellige teksttyper. Disse defaultstrategier kompletteres af egentlige regler til beregning af informationsstruktur i en række syntaktiske konstruktioner. Det påvises således, at sætningens informationsstruktur entydigt er afspejlet i syntaksen i eksistentielle *der*-konstruktioner, kløvninger (*it-clefts* og *wh-clefts*) og topikalisering-konstruktioner. Specifikationer til analyse og generering af disse konstruktioner i HPSG fastlægges.

En delmængde af de analyse- og genereringsstrategier, som er beskrevet i afhandlingen, er blevet implementeret med positive resultater i ALEP-systemet. Implementeringen viser, at generering ud fra en repræsentation af sætningens semantiske indhold og informationsstruktur ikke blot er processeringsmæssigt mulig men også leverer bedre resultater end en syntaktiskbaseret syntese i de tilfælde, hvor kilde- og målsprog benytter sig af forskellige syntaktiske midler til at udtrykke den samme informationsstruktur.

Endelig peges der i afhandlingen på to vigtige emner, som på trods af deres relevans ikke er blevet berørt, og som derfor fortjener yderligere forskning. Det ene er forholdet mellem fokus og præsupposition. Det ville være en interessant opgave at undersøge, hvordan en præsuppositionsteori ville kunne integreres i maskinoversættelse. Det andet er behandlingen af pronominer, som er nødvendig for at opnå sammenhæng i outputteksten. *Centering* synes igen at være et lovende kandidat med hensyn til, hvordan pronominel anafor kunne behandles i et maskinoversættelsessystem.

Appendix I: type declarations and coverage description

Appendix I documents the ALEP implementation. It includes a copy of the type declarations for both English analysis and Danish generation, a description of the grammatical coverage of both modules, and a list of the words coded in the lexica.

For those interested in inspecting the actual grammar and lexicon files, they are accessible through my home page on the web, together with a compressed, uuencoded dvi file containing the thesis itself. The address is: "file://cst.ku.dk/www/patrizia/home.html".

Type declaration files

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%    Type declarations for English analysis    %
%                                              %
%            Patrizia Paggio                   %
%              July 1996                       %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Contents:
% =========
%
% 1. The sign
% 2. Processing features
% 3. Content features
% 4. Context features
% 5. Syntactic features


% 1. The sign

type(sign:{phon => type({phon:{}}),
           synsem => type({synsem:{}}),
           named_spec => type({spec:{}})
           },'').


type(phon:{string => list(atom),
           rest => list(atom)
          }).

type(synsem:{
           syn => type({syn:{}}),
           content => type({cont:{}}),
           context => type({context:{}})
            }).

% 2. Processing features

type(spec:{
    language => atom({ en,da }),
    specinfo => type({specinfo:{}}),
    ruleinfo => type({ruleinfo:{}})
          }).


type(
  ruleinfo:{
    tlm_id => atom,
    lift_id => atom,
    ana_id => atom,
    ref_id => atom,
    synth_id => atom,
    lower_id => atom
```

```
    }, '' ).


type(
  specinfo:{
    lift => atom({ y,n }),
    ana => atom({ lex,ws,ps,n }),
    ref => atom({ y,n }),
    default => atom({ y,n }),
    synth => atom({ y,n }),
    lower => atom({ y,n })
      }, '' ).

% 3. Content features

type(cont:{restr => type({restr:{}})
          }).

type(restr:{reln => atom,
          arg1 => type({cont:{}}),
          arg2 => type({cont:{}}),
          arg3 => type({cont:{}}),
          mods => list(type({cont:{}}))
           }).

type(npcont:{det => functor,
            index => functor,
            number => boolean([{sing,plur}])}).

type(detcont:{det => functor,
             index => functor}).

type(nullcont:{}).

cont > {npcont,
        detcont,
        nullcont}.

% 4. Context features

type(context:{infotype => boolean([{p_topic,focus,neutral}]),
             infostr => type({info:{}}),
             disc =>  boolean([{given,new}]),
             head_disc =>  boolean([{given,new}]),
             presup => functor => none(_)
            }).

type(info:{p_topic => list(type({cont:{}})),
          focus => list(type({cont:{}})),
          neutral => list(type({cont:{}}))
         }).

% 5. Syntactic features
```

```
type(syn:{mkey => atom,
          hkey => atom,
          seg => type({seg:{}}),
          loc => type({loc:{}}),
          type => boolean([{full,empty}]),
          str => type({phrase:{}}),
         bind => type({bind:{}})
        }).

type(seg:{first => boolean([{y,n}]),
          last => boolean([{y,n}])
       }).

type(
  phrase:{
    constr =>
boolean([{h_subj,h_compl,h_subj_compl,h_spec,h_adj,h_mark,h_fill, coord}]),
    parsehead => boolean([{y,n}]),
    headed => boolean([{left,right}]),
    heading => boolean([{left,right}])}).

type(loc : {head => type({head:{}}),
             unit => atom,
             subcat => list(type({synsem:{}})),
             so => boolean([{default,ndef_left,ndef_right,n_a}]),
             bar => atom({max,phrase,zero}),
             cat => boolean([{n,nbar,np,apn,
                             v,vbar,vp,
                             p,pp,a,d,s,adv,
                             relpron,perspron,
                             conj,
                             sbar,advp,
                             punct}])
             }).


type(bind:{gaps => type({gaps:{}}),
          whhead => boolean([{what,which}]),
          rel => list(type({cont:{}})),
          slash => list(type({loc:{}})),
          extrap => list(type({synsem:{}}))
         }).


type(gaps:{gap_in => list(type({synsem:{}})),
          gap_out => list(type({synsem:{}}))
         }).



type(head:{}).

type(nhead:{maj => boolean([{n}]),
          hkey => atom,
```

```
          case => boolean([{nom,acc,gen}]),
          agr => boolean([{plur,sing}, {'1','2','3'}]),
          nform => boolean([{norm,it,there}])
           }).

type(ahead:{maj => boolean([{a}]),
                  hkey => atom,
          agr => boolean([{plur,sing}, {'1','2','3'}])
           }).

type(vhead:{maj => boolean([{v}]),
                  hkey => atom,
          agr => boolean([{plur,sing}, {'1','2','3'}]),
          vform => boolean([{pres,past,bse,psp,prp,inf,ger,pas}]),
          msaspect => boolean([{simple, perf}]) => simple,
          trans => atom({yes,no}),
          diath => atom({act,pass}),
          aux => atom({yes,no})
            }).

type(phead:{maj => boolean([{p}]),
                  hkey => atom,
          pform => boolean([{of,on,to,from,*}]),
          ptype => boolean([{conj,prep,*}]),
          prd => atom({yes,no})
            }).

type(dhead:{maj => boolean([{d}]),
          hkey => atom
            }).

type(advhead:{maj => boolean([{adv}]),
                  hkey => atom
            }).

type(conjhead:{maj => boolean([{conj}]),
                  hkey => atom,
               ptype => boolean ([{conj,prep}])
            }).

type(puncthead:{maj => boolean([{punct}]),
                  hkey => atom
            }).

head > {nhead,
        ahead,
        vhead,
        phead,
        dhead,
        advhead,
        conjhead,
        puncthead
        }.
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%    Type declarations for Danish synthesis    %
%                                              %
%            Patrizia Paggio                   %
%              July 1996                       %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Contents:
% =========
%
% 1. The sign
% 2. Processing features
% 3. Content features
% 4. Context features
% 5. Syntactic features

% 1. The sign

type(sign:{phon => type({phon:{}}),
           synsem => type({synsem:{}}),
           named_spec => type({spec:{}})
           },'').


type(phon:{string => list(atom),
           rest => list(atom)
          }).

type(synsem:{
           syn => type({syn:{}}),
           content => type({cont:{}}),
           context => type({context:{}})
            }).

% 2. Processing features

type(spec:{
    language => atom({ en,da }),
    specinfo => type({specinfo:{}}),
    ruleinfo => type({ruleinfo:{}})
          }).


type(
  ruleinfo:{
    tlm_id => atom,
    lift_id => atom,
    ana_id => atom,
    ref_id => atom,
    synth_id => atom,
    lower_id => atom
      }, '' ).
```

```
type(
  specinfo:{
    lift => atom({ y,n }),
    ana => atom({ lex,ws,ps,n }),
    ref => atom({ y,n }),
    default => atom({ y,n }),
    synth => atom({ y,n }),
    lower => atom({ y,n })
      }, '' ).

% 3. Content features

type(cont:{restr => type({restr:{}})
           }).

type(restr:{reln => atom,
           arg1 => type({cont:{}}),
           arg2 => type({cont:{}}),
           arg3 => type({cont:{}}),
           mods => list(type({cont:{}}))
            }).

type(npcont:{det => functor,
             index => functor,
             number => boolean([{sing,plur}])}).

type(detcont:{det => functor,
              index => functor}).

type(nullcont:{}).

cont > {npcont,
        detcont,
        nullcont}.

% 4. Context features

type(context:{infotype => boolean([{p_topic,focus,neutral}]),
              infostr => type({info:{}}),
              disc =>  boolean([{given,new}]),
              head_disc =>  boolean([{given,new}]),
              presup => functor
             }).

type(info:{p_topic => list(type({cont:{}})),
           focus => list(type({cont:{}})),
           neutral => list(type({cont:{}}))
          }).

% 5. Syntactic features

type(syn:{mkey => atom,
          hkey => atom,
```

```
          seg => type({seg:{}}),
          loc => type({loc:{}}),
          type => boolean([{full,empty}]),
          str => type({phrase:{}}),
         bind => type({bind:{}})
        }).

type(seg:{first => boolean([{y,n}]),
          last => boolean([{y,n}])
       }).

type(
  phrase:{
    constr =>
boolean([{h_subj,h_compl,h_subj_compl,h_spec,h_adj,h_mark,h_fill, coord}]),
    parsehead => boolean([{y,n}]),
    headed => boolean([{left,right}]),
    heading => boolean([{left,right}])}).

type(loc : {head => type({head:{}}),
             unit => atom,
             subcat => list(type({synsem:{}})),
             so => boolean([{default,ndef_left,ndef_right,n_a}]),
             bar => atom({max,phrase,zero}),
             stype =>  boolean([{main,sub}]),
             ms_def => boolean([{indef,def}]),
             cat => boolean([{n,nbar,np,apn,
                              v,vbar,vp,
                              p,pp,a,d,s,adv,
                              relpron,perspron,
                              conj,
                              sbar,advp,
                              punct}])
             }).


type(bind:{gaps => type({gaps:{}}),
          whhead => boolean([{som,det_som}]),
          rel => list(type({cont:{}})),
          slash => list(type({loc:{}})),
          extrap => list(type({synsem:{}}))
         }).


type(gaps:{gap_in => list(type({synsem:{}})),
          gap_out => list(type({synsem:{}}))
         }).



type(head:{}).

type(nhead:{maj => boolean([{n}]),
          hkey => atom,
```

```
          case => boolean([{nom,acc,gen}]),
          agr => boolean([{plur,sing}, {common,neuter}]),
          nform => boolean([{norm,det,der}])
           }).

type(ahead:{maj => boolean([{a}]),
            hkey => atom,
            agr => boolean([{plur,sing}, {common,neuter}])
           }).

type(vhead:{maj => boolean([{v}]),
                  hkey => atom,
          vform => boolean([{pres,past,bse,psp,prp,inf,ger,pas}]),
          msaspect => boolean([{simple, perf}]) => simple,
          trans => atom({yes,no}),
          diath => atom({act,pass}),
          aux => atom({yes,no})
            }).

type(phead:{maj => boolean([{p}]),
                  hkey => atom,
          pform => boolean([{af,paa,til,fra,*}]),
          ptype => boolean([{conj,prep,*}]),
          prd => atom({yes,no})
            }).

type(dhead:{maj => boolean([{d}]),
            agr => boolean([{plur,sing}, {common,neuter}]),
            hkey => atom
            }).

type(advhead:{maj => boolean([{adv}]),
                  hkey => atom
            }).

type(conjhead:{maj => boolean([{conj}]),
                  hkey => atom,
               ptype => boolean ([{conj,prep}])
            }).

type(puncthead:{maj => boolean([{punct}]),
                  hkey => atom
            }).

head > {nhead,
        ahead,
        vhead,
        phead,
        dhead,
        advhead,
        conjhead,
        puncthead
        }.
```

Grammatical coverage

The following grammatical structures are covered by the English core analysis grammar:

1. Coordination
   - Coordination of nouns
   - Coordination of adjectives
2. The clause
   - Purpose clauses
   - Adverb fronting
   - Sentences with subject gap
   - Sentences with a complement gap
   - Relative clauses
   - It-clefts and wh-clefts
3. The NP
   - Modified by determiners
   - Modified by adjectives
   - Modified by relative clauses
   - Modified by reduced relative clauses
   - Modified by PPs
   - Modified by complements
4. The PP
5. The verbal phrase (VBAR)
   - Vbar/v + complement
   - Vbar with a complement gap
   - Vbar followed by an adverbial clause

The English refinement grammar covers the following focus heuristics:

1. Adverb fronting
2. Simple clause
   - NP subject and VERB is the rightmost element
   - NP subject and VERB not the rightmost complement
   - The subject is not an NP
   - The subject is *it* or *there*
   - The subject is an sbar in a wh-cleft construction
3. The VBAR
   - VBAR rules refining vbars in cases of SO disruption
   - VBAR rules refining vbars in cases of default SO
   - VBAR in focus-marking constructions

- VBAR rules for head_adjunct structures

The following features are transferred in the English-Danish transfer grammar:

- diath (diathesis)
- msaspect (morphosyntactic aspect)
- vform (verbal form)
- unit (number of current sentence in the input text)
- content
- context

The Danish generation grammar covers the following phenomena:

1. The sentence
    - Clauses with topical subject
    - Clauses with pleonastic *der* subject
    - It-clefts and wh-clefts
    - Sentences with subject gap
    - Sentences with complement gap (but see caveat below)
2. The NP
3. The PP
4. The VBAR
    - Taking focussed complements
    - Taking neutral complements
    - Taking an empty complement
    - Taking a focussed adjunct

Caveats:

The generation grammar has not been optimised as much as the analysis grammar. This means that for some of the test sentences shown in Appendix II, it takes the system a long time to terminate after having produced a result. Therefore, test sentences should be run in "first only" mode, so that Alep interrupts the process as soon as the first result is produced.

Complement gapping has been implemented both in analysis and generation. However, whereas the rules work correctly in analysis, they do not produce the desired result in generation. More debugging would be needed in order to correct the error.

Lexical coverage

The following entries are included in the English analysis lexica (core and refinement):

Adjectives:

abovementioned
aliphatic
animal
aromatic
better
copperfree
critical
direct
entrapped
existing
fat
finished
geostationary
increasing
natural
other
poor
present
saturated
supplementary
ten
total
typical
unknown
unsaturated
useful
utmost
vast
vegetable
visible

Adverbs:

almost
also
now
furthermore
especially
above
in_addition
in_another_aspect
long
preferably
seriously
sometimes
suitably

Conjunctions:

and
or

Determiners:

a
an
another
the
this

Nouns:

ability
architectural_and_software_considerations
attention
communication
communication_possibilities
context
coverage
diversity
earth
england
factor
form
formation
globe
impact
information_technology
integration
invention
japan
man
operation
orbit
practice
research_community
satellites
surface
technology
types
use
water

Prepositions:

although
between
by
from
in
on

of
to
when
with

Pronouns:

it
which
what

Terms:

additives
alcoholamines
amines
antioxidant
antioxidants
basestock
bisuccinimides
carbon_atoms
chlorination_route
coal
concentrate_composition
crankcase
crankcase_oil
dispersants
distillate_fuels
engine
earth_surface
fuel
internal_combustion_engine
lubricating_composition
lubricating_oil_composition
lubricating_viscosity
monosuccinimides
nonchlorination_route
oil
oils
oxidative_stability
piston_varnish
polyamines
polyoxyalkylene_polyamines
rust_formation
shale
sludge
succinimides
tetraethylene_pentamine
triethylene_tetramine
varnish

Verbs:

accompanied
are

be
been
can
cannot
claimed
combat
combats
comprising
contribute
contributes
derived
diluting
emerge
emerged
emerges
employed
give
gives
has
have
include
included
is
likes
like
lubricate
may
mentioned
obtained
permits
preferred
provides
puts
require
required
restrict
restricts
treated
use
used
using
will

Others:

fullstop

In Danish generation, lexical coverage is substantially more restricted. The lexicon basically covers the word readings needed for the generation of the test sentences shown in Appendix II. It was felt that a large lexicon was not necessary to demostrate the feasibility of the generation strategies proposed in this thesis. Furthermore, in its current version ALEP does not seem very efficient in the way it selects entries in the generation process, and the lexicon has therefore been kept small not to burden the generation process unnecessarily.

The following words are coded (with the readings necessary for the generation of our test sentences):

Adverbs:

alvorligt
derudover
derudover
helst
i_en_anden_sammenhæng
især
længe
nu
næsten
også
ovenover
passende
sommetider

Conjuctions:

eller
og

Determiners:

denne
dette
en
en_anden
et
et_andet

Nouns:

forskningsverden
forskningsverdenen
informationsteknologi
informationsteknologien
kommunikationsmuligheder

Prepositions:

i
til

Pronouns:

der
det
det_som
som

Terms:

aminer
antioxidant
antioxidanten
antioxidanter
antioxidanterne
belægning
belægningen
motor
motoren
rustdannelse
rustdannelsen
slam
slammet
tilsætningsstoffer
tilsætningsstofferne

Verbs:

bekæmper
bidrager
er
fremkomme
fremkommer
fremkommet
give
giver
kan
kræves

Others:

fullstop
trace

Appendix II: test suites and results

Test suites and results

Two different types of test suite were used to test the implementation: a set of sentences that were run interactively to test the whole translation process, and a series of analysis suites that were regularly run as batch processes to test the analysis coverage. The current version of ALEP does not provide the means to test the whole translation in batch mode. In fact not much work has been carried out in the ALEP community involving transfer and generation. Results are shown below for all the suites.

Test sentences

The results obtained for our first test suite are shown by providing the input sentences and the output generated by the system. Sample ALEP representations showing the analysis and generation feature structures built by the system for each of the sentences below are shown at the end of the Appendix.

Input:

additives combat the sludge.
another antioxidant is required.
the sludge contributes to rust_formation[1].
it is the sludge which contributes to rust_formation.
what contributes to rust_formation is the sludge.
the sludge contributes to rust_formation in the engine.
information_technology gives better communication_possibilities to the research_community.
information_technology gives the research_community better communication_possibilities.

Output produced:

tilsætnigsstoffer bekæmper slammet.
der kræves en_anden antioxidant.
slammet bidrager til rustdannelse.
det er slammet som bidrager til rustdannelse.
det_som bidrager til rustdannelse er slammet.
slammet bidrager til rustdannelse i motoren.
informationsteknologi giver bedre kommunikationsmuligheder for forskningsverdenen.
informationsteknologi giver forskningsverdenen bedre kommunikationsmuligheder.

Analysis test suites

The analysis module was regularly tested by means of 6 different test suites. Rather than show the detailed reports generated by ALEP as a result of each test, however, we simply list below the sentences that the system could analyse and those for which analysis failed.

Analysed sentences:

Additives can be used.
Additives combat the sludge and varnish.
Additives combat the sludge.
Amines are used.
Amines include alcoholamines.
Amines restrict the sludge.
Amines which restrict the sludge include alcoholamines.
An additive can be used.
Another antioxidant has emerged.
Another antioxidant is required.

[1] As already mentioned, compound nominals are not treated in this implementation, and are therefore written as single words in the test sentences.

Information_technology gives better communication_possibilities to the research_community.
Information_technology gives the research_community better communication_possibilities.
Invention permits practice.
It is the sludge which contributes to rust_formation.
It is the sludge which restricts the engine.
It is the sludge.
Satellites restrict the practice of the invention.
Satellites restrict the practice.
Sludge can be used.
Sludge contributes to rust_formation in the engine.
Sludge contributes to rust_formation.
Sludge has emerged.
Sludge is used.
The invention permits the practice.
The sludge contributes to rust_formation in the engine.
The sludge contributes to rust_formation.
The sludge is it.
What contributes to rust_formation is the sludge.
What restricts the engine is the sludge.

Sentences the analysis of which failed:

An additive can be used.
Amines can restrict the sludge.
It is the sludge which has emerged.
Sludge can be accompanied by amines.
Sludge is accompanied by amines.
Sludge which amines restrict has emerged.
What has emerged is sludge.