Proceedings of the Scandinavian Conference in Computational Linguistics Bergen 28-30 November 1991

edited by:

Kjell Morland
and
Kari Sørstrømmen

Bergen 1992

Published by: Norwegian Computing Centre for the Humanities - 1992

ISBN 82-7283-066-3
ISSN 0800-5796

No. 56 in the Report Series of the Norwegian Computing Centre for the Humanities

Typeset and layout: Kari Sørstrømmen
Printed by: Grafisk Hus A/S, Bergen / Dupliseringstjenesten, Universitetet i Bergen

Norwegian Computing Centre for the Humanities, Harald Hårfagresgt. 31, N-5007 Bergen, Norway
Tel.: + 47 (0)5 212954

Contents

Preface

The Scandinavian Conference on Computational Linguistics (Nordiske datalingvistikkdager) **1991** was arranged at the Faculty of Arts, University of Bergen 28-30 November.

The conference committee had the following members:

Professor Helge Dyvik, Department of Linguistics and Phonetics
Senior Executive Officer Knut Hofland, Norwegian Computing Centre for the Humanities
Office Manager Kjell Morland, Norwegian Computing Centre for the Humanities
Research Fellow Torbjørn Nordgård, Department of Linguistics and Phonetics
Senior Lecturer Roald Skarsten, Computing Section of the Faculty of Arts
Junior Executive Officer Kari Sørstrømmen, Norwegian Computing Centre for the Humanities
Lecturer Ivar Utne, Department of Scandinavian Languages and Literature

This volume contains 16 articles. Some papers read at the conferece were not later submitted for publication. The articles are printed in alphabetical order by author's name, and no attempt has been made to organize the papers according to topic.

The conference was supported financially by Nordisk Forskarutdannelseakademi, Norwegian Council for Research in the Humanities, and the Faculty of Arts, University of Bergen.

Bergen, 3 June 1992

Kjell Morland and Kari Sørstrømmen

Conceptual text representation for multi-lingual generation and translation

Lars Ahrenberg & Stefan Svenberg
Linköping University

This paper presents some ideas and preliminary results of a project aimed at automatic generation and translation of text from conceptual (interlingual) representations. In the first part we give some arguments for treating translation and generation as closely coupled processes and motivate the need for conceptual text representations to aid these tasks. In the second part we describe an implemented method for multi-lingual generation of sentences.

1. Introduction

During the last decade unification-based grammar formalisms have become standard tools for grammatical description within computational linguistics. A significant advantage of them is their declarativeness, something which implies that they can be used by parsers and generators with the same ease. Two developments in recent years are particularly interesting from the point of view of translation and multi-lingual generation. The first is the idea that descriptions of different languages can be related by virtually the same means as descriptions belonging to different levels of description within a single language (Kaplan et al., 1989; van Noord, 1990; Russell et al., 1990). This latter development is especially suitable for those who favour a transfer model of translation, as well-known description levels such as phrase structure, functional structure and logical form can be reused and set into correspondences.

The second idea is that parsing and generation can be seen as basically the same processes that differ only in their input (Shieber, 1988; Zajac & Emele, 1990; Emele et al., 1990). This would make it possible to handle all processing by the same mechanisms and by means of a single grammar and dictionary for each language. This idea, on the other hand, gives something to those who are interested in interlingua approaches, since, if a common representation language can be found for the description of different languages, we can use it in the different grammars and get translation systems and multi-lingual generation system almost for free.

In this paper we report on a project which is concerned with developing the second idea.[1] In the next section we describe briefly its goals and motivations. Then, in section 3, we provide some arguments in favour of conceptual text-representations and propose a way of characterizing semantic equivalence of generated texts. In section 4, we show how the conceptual representations are used for bi-lingual sentence generation. In the last section, finally, we indicate briefly our plans for the future.

2. The project

The project is explorative with the overall aim to judge the possibilities of using interlingual representations in applications such as document generation, intelligent (on-line) handbooks, and translation. We do this by studying a specific text genre, the service manual, and in particular the expository sections, where an object is described and its function is explained. Our initial corpus comprises 86 sentence pairs from two service manuals issued by Volvo Truck Corporation. More specific goals of the project are to develop an interlingual representation language for this genre and algorithms for multi-lingual generation from the interlingual representations. We also investigate the possibilities of using the interlingua for automatic translation. Two languages are considered, Swedish and English.

The interlingua should cover all relevant aspects of the texts: structure, content and textual coherence, but we are only concerned with the linguistic variation that can be found within the genre. For instance, as all sentences of our corpus are declarative and present-tense, we treat such properties as structural invariants that need not be accounted for on semantic or pragmatic grounds.

A basic assumption of the project is that it is useful to consider translation and multi-lingual generation as closely related tasks. If we view the general generation problem as one of finding a text (or all texts) that satisfies a given set of constraints in a given context, it is easy to see that translation fits this definition. The source text provides one set of constraints while the grammar and genre-specific rules of the target language provide another set. On the other hand, in multilingual generation, we must somehow ensure that the texts that are generated are equivalent in important respects, e.g. as regards content, style and progression. The requirements of such an equivalence relation comes very close to what one demands of a relation between translations.

3. Towards a characterization of equivalence

The usual way to characterize the relation of translation equivalence is perhaps with reference to a number of description levels on which two texts should be identical or corresponding. For instance, Carbonell & Tomita (1987) mentions the following factors to be important for good translations: pragmatic invariance (matters of illocutionary force, style etc.), semantic invariance, structural invariance, lexical invariance and spatial invariance. Ignoring the latter, which is concerned with the external properties of the text such as length and page layout, it is interesting to note how invariance is decribed with respect to the different levels. In the case of pragmatics and semantics the notion is one of "preserving invariant" the relevant properties, while structural invariance is explained as "preserving as far as possible" the syntactic structure, and lexical invariance is explained as "preserving a one-to-one mapping of words or phrases from source to target text". We see that it is easier to imagine a common, language-independent representation for the higher levels of texts, whereas the lower levels, such as syntax and lexis, can only be brought into correspondence with each other.

However, even if two lexemes, or two constructions, of different languages cannot be treated as having the same properties, but set into a correspondence for lack of better alternatives, we can still give them a common description. From the formal point of view there seems to be no relevant distinction between identity and correspondence, since, if two elements correspond, we can introduce a property at the interlingua level that is expressed by these two elements (and no others) in the two languages. The association between the elements and the interlingual descriptor is then made in the grammars of the two languages. Conversely, of course, a perceived identity of meaning of two elements from different languages can be represented as a trivial correspondence between the elements. Note, though, that by using an interlingua representation we can decompose a

correspondence between often, quite complex sets of properties into two simpler relations both of which relate a simple descriptor at the interlingua level to complex language-specific representations, just as a word form is associated with its morpho-syntactic and semantic properties in a dictionary. This is advantegous not least from the point of view of multi-lingual generation.

Now, if we want to get at the bottom of a generation problem, we will need to refer to complex combinations of properties. But this is the case whether we work with interlingua representations or correspondence rules. With the interlingua approach we then have the advantage that all information is accessible at a single level of description. It is often argued that in the case we deal with languages that make the same kind of distinctions and use the same kind of constructions, as is the case with Swedish and English, we need not consider the relevant pragmatic and semantic properties, but merely note the correspondences. However, in our corpus we find several pairs of sentences, such as the following, that seem difficult to describe on the basis of structurally based correspondence rules only:[2]

1. An ellipsis occurs only in one of the languages, but not in the other.
   S: Basväxeldelen manövreras mekaniskt, rangeväxeln pneumatiskt.
   E: The basic section is mechanically operated; the range gear is pneumatically operated.

2. An integrated complement corresponds to the subject head, while the subject corresponds to a subject modifier.
   S: Spärrventilen har till uppgift att förhindra växling av rangeväxeln när ...
   E: The purpose of the inhibitor valve is to prevent inadvertant shifting of the range gear when ...

3. A passive clause corresponds to an active clause with an anaphoric subject.
   S: Spärrventilens kolv trycks upp ur fördjupningen på kolvstången ...
   E: This moves the plunger of the inhibitor valve out of its dimple ...

4. A simple NP corresponds to a coordinated, disjunctive NP.
   S: Den här nedkylningen kallas laddluftkylning.
   E: This cooling process is known as charge air cooling or intercooling.

As a basis for describing the semantic equivalence of two sentences that are translation equivalents we appeal to the notion of topic, or topical question (Carlson, 1983; Ahrenberg, 1987). Speaking informally, we can say that a necessary condition for two sentences being translation equivalents is that they answer the same question by the same standards, where standards refer to such things as truthfulness, completeness, clarity and relevance.[3] This is actually also a condition that can be applied in practice; often it is not a difficult task to decide which question or questions a given text sentence attempts to answer, as evidence both from its form and its context can be used.

If we look at the four sentence pairs above, the topical questions of the first pair can be rendered in English as *How is the basic section operated, and how is the range gear operated?* At the conceptual level we may introduce a concept, Operation, with two arguments, one for the object (gear or gear set) being operated upon and one for the manner in which it is done. A question that relates to this concept, one that makes it a topical concept, can be represented as a propositional structure which is unspecified with respect to one of its arguments:

$$\begin{bmatrix} \text{aspect} & \text{Operation} \\ \text{thing} & \text{r-gear1} \\ \text{value} & [\,] \end{bmatrix}$$

As for the occurrence of the finite verb in the second conjunct of the English sentence and its absence in the corresponding Swedish sentence, it can be handled completely within the grammars of the two languages. We need not formulate a separate correspondence rule saying that a finite verb in one language can correspond to nothing in the other language under certain circumstances.

As for the second pair of sentences we are faced with a correspondence pattern which is even more involved than the splitting/fusing examples discussed by Kaplan et al. (1989) and Sadler & Thompson (1991). We can avoid introducing explicit correspondence rules, if we state the rules in terms of relations between interlingua descriptions and language-specific grammatical descriptions, however. Moreover, they become quite simple because the interlingua description is a simple one. The topical question is *What is the purpose of the inhibitor valve?* with the interlingua description

$$\begin{bmatrix} \text{aspect} & \text{Purpose} \\ \text{thing} & \text{inh-valve2} \\ \text{value} & [] \end{bmatrix}$$

The rules we need are associated with the concept Purpose in the knowledge-base as explained in section 4.4.

The third pair illustrates the importance of co-text. The topical question may be rendered as *What happens in connection with this?*, where *this* refers to an event of shifting the range gear described in the previous sentence. While the event is explicitly referred to in the English sentence, it is not so in the Swedish sentence, illustrating the common property of texts that causal relationships between events are often not given explicit expression. At present we have not defined text-level rules, so this sentence-pair cannot be handled by our generator, but it seems clear that a correspondence rule using only structural information is not sufficient for the purpose.

The fourth pair of sentences, finally, addresses the topical question of what a certain process is called. As it happens two terms are used for it in English while only one is used in Swedish. The result is that a disjunction is used in English – probably in response to some standard of completeness – with no corresponding disjunction in the Swedish sentence. The fact that a simple NP in one language in some cases can correspond to a disjunctive NP in another language is for obvious reasons not something that one would like to express as a general possibility of structural correspondence. However, without access to a semantic/pragmatic level of description one cannot express the appropriate constraints.

4. The unification-based generator

4.1 Overview

In this section we show how sentences can be generated from conceptual representations. We refer to these representations as content descriptions as they mainly contain semantic information. The generation process roughly have the stages illustrated in figure 1.

The first module of the generator constructs language specific grammatical descriptions using relational grammar rules and information in the common knowledge-base (KB). These will then be fed into the surface string generator, which has its own grammar. Between these two main modules there is also an interface whose purpose is to fine-tune the incoming grammatical descriptions so that they conform to the demands of the surface generator.

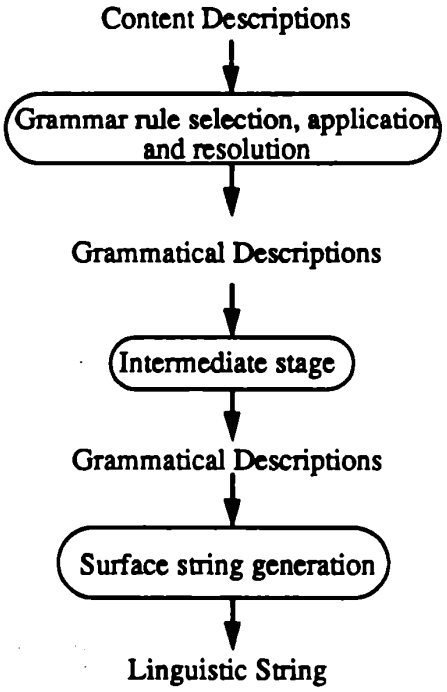


Figure 1: Basic architecture of the generator

An editor is built on top of the KB enabling content descriptions, grammatical descriptions, grammar rules, domain and linguistic knowledge to be interactively and incrementally added or modified. All information is coded as feature-value matrices with unification as the sole method of combining.

The generation process is not language specific. The KB can support knowledge for several languages. Given a content description the system concurrently generates the corresponding strings for all the supported languages. In the next subsections we will describe content descriptions, grammar rules, and KB for the first stage of generation. The surface string generation module is modeled on Shieber (1988) and uses grammars written in the PATR-II-format.

4.2 Content descriptions

In order to achieve multi-lingual generation we want to work with language independent chunks of information. The traditional approach is to use logical propositions, such as a conjunction of facts, which the text is supposed to express. The content descriptions in the system contain logical propositions, but also distinguish them on the basis of pragmatic function. Moreover the content descriptions need not be fully specified as some of the content may be retrieved from the KB.

Propositions can be of a number of different types. They are expressed differently depending on the type. The types can be grouped into two main categories. First those that have a purely logical meaning, currently being the simple type thing-aspect-value (tav) and the conjunction (and). Secondly, we have a number of types that besides having a logical interpretation also have a non-logical function. Below we will discuss instances of both types and give small examples of their use.

- Thing-aspect-value, which is used to express a value of an aspect of a certain thing. The example below shows a simple fact written as an attribute-value matrix which says that a thing called r1000 has the color green.

$$\begin{bmatrix} \text{type} & \text{tav} \\ \text{thing} & \text{r1000} \\ \text{aspect} & \text{color} \\ \text{value} & \text{green} \end{bmatrix}$$

- And, a logical conjunction of two other propositions:

$$\begin{bmatrix} \text{type} & \text{and} \\ \text{a1} & [] \\ \text{a2} & [] \end{bmatrix}$$

Most often the a1 value is of type tav while the a2 is either a tav or an and.

- Reference-scope, which logically is a conjunction. It has two attributes, called ref and scope each containing propositions about a given object (or set). The non-logical aspect of it lies in the way the two propositions have been divided. The first argument specifies a proposition which is used for making reference to the object. The intention is that the object and proposition must be known to the reader at this point, either by having been introduced at an earlier point in the text or included in the reader's background knowledge. The proposition will determine how reference to the object will be achieved at the surface level; e.g. by using its proper name, a pronoun, or a definite description, possibly including a number of adjectives and so on. The scope value contains information that is new in the current context. It identifies the proposition that is asserted about the object.The reference-scope type is used to express the contents of simple clauses. The first argument will form the subject of that clause while the scope value is used to build the predicate, including verb and complements.

Below is a simple example of a content description of the clause *The basic section is mechanically operated.* "The basic section" is known while "mechanically operated" is new information which is asserted in the sentence.

$$\begin{bmatrix} \text{type} & \text{rs} \\ \text{ref} & \begin{bmatrix} \text{type} & \text{tav} \\ \text{thing} & {}^{1}[\text{r1000bs}] \\ \text{aspect} & \text{isa} \\ \text{value} & \text{b-section} \end{bmatrix} \\ \text{scope} & \begin{bmatrix} \text{type} & \text{tav} \\ \text{thing} & {}^{1}[] \\ \text{aspect} & \text{Operation} \\ \text{value} & [] \end{bmatrix} \end{bmatrix}$$

The value r1000bs means the basic section of the gearbox r1000. The aspect *isa* gets the reflexive and transitive closure of the classes r1000 belongs to. From these, the concept b-section has been chosen. It has the name "basic section" attached to it.

The scope value is an unspecified proposition, i.e. one corresponding to a topical question as explained above. The unspecified value will be retrieved from the KB during the generation process.

4.3 Mapping content descriptions onto grammatical descriptions

A content description represents properties that are common to equivalent sentences of different languages. In addition, each sentence in the equivalence class satisfies a language-specific grammatical description. Grammatical descriptions and content descriptions are related roughly as the two sides of the classical Saussurean linguistic sign. Thus, it is natural to express possible relations between grammatical features and content features as a relation between partial structures. In the current grammar we actually use a number of different relations, which encode both the language and the textual function of the structures being related, as in the following examples where cd and gd stand for content and grammatical descriptions respectively:

1. inform-sw(cd, gd), refer-eng(cd, gd), describe-eng(cd, gd)

Another possibility would be to encode the function and language as arguments:

2. signify(cd, languge, function, gd)

A specification of the relation under a more complete generation scheme would have to take many other aspects into account, e.g. as follows:

3. signify(cd, language, function, type, user-model-in, context-in, context-out, user-model-out, gd)

where 'type' is the type of text object; such as clause or sentence. 'user-model-in' is a model of what background knowledge the reader posesses at this point, 'context-in' records relevant objects mentioned in the text so far to aid pronoun generation and ellipsis. The user-model-out will reflect the fact that the reader has accomodated the new information in the content description. This can be used to ensure that the new information indeed is relevant, coherent and consistent with what has been said and with what the reader already knows.

From now on version 1 of the relation specification will be considered since it is the one has been implemented.

Given a content description and a grammatical description the generator will try to prove the relation between them. Either description may be partially specified. The generator will during the proof procedure suggest instantiations to complete the specifications. If we would like to generate a grammatical description that argument should initially be uninstantiated. If the process was successful the generator returns the answers one by one even if there are an infinite number of them. In principle, we could also parse a grammatical description which would lead to a content description. At the time of writing this is not yet practical for efficiency reasons. The generator works according to the principle known as SLD-resolution (see e.g. Nilsson & Maluszynski, 1990). The attribute-value matrices are coded as directed acyclic graphs. The selection function is graph unification. This is similar to, but not necessarily limited to, the way logic programming languages work.

The grammar is a rule base, where each rule generally takes the following form:

$rel_0(cd_0, gd_0) \leftarrow rel_1(cd_1, gd_1), ..., rel_n(cd_n, gd_n).$

We refer to the left-hand side as the head and the right-hand side as the body of the rule. The head consists of a single term, while the body may contain any number of terms. The rules currently in

use actually contain terms with additional arguments, but we ignore these here.

Given a content description for a sentence, we will first construct a term having inform-sw or inform-eng as a functor. The content description will become the first argument and the second argument will be initialized to the null dag, giving, say, inform-sw(cd, []). The generation of the grammatical description will start by picking out those rules which have inform-sw as the functor of their heads. They will then be tried out one by one. The arguments of the term will be unified with the corresponding argument of the rule head. If that succeeds all subgoals appearing on the right hand side of the rule must hold. These are tried out recursively. The arguments of the subgoals share material with each other and particularly with the head arguments. As new material is unified into the structures,the changes become immediately visible everywhere. When all the subgoals have successfully been proved the process suspends in that state. The arguments of the initial term have been fully instantiated and can be picked out. After that the process resumes by backtracking to choice points in the SLD-tree where alternative instantiations can be found.

The rules are written in such a way that a structural depth-first analysis will be performed on the content description. This also means that the grammatical description will be built top-down. The subgoals take care of the substructures. The recursion is stopped, aside from failing unifications, by rules having no right hand side or by built-in rules accessing the KB.

The far most important built-in rule is the the primitive retrieval operation, iget. It has three arguments: carrier, indicator, and value. The carrier denotes an object that has a value stored under a certain indicator. It can, from the generator's point of view, be regarded as a simple property. The iget is used for two different purposes:

1. Checking the validity of the content description. The world modelled in the KB must sanction the information expressed there.
2. Retrieving linguistic information from domain objects. All domain concepts mentioned in the content description contribute fragments of grammatical descriptions. The rules glue these fragments together to form the gd.

We end this subsection with two rules that handles reference-scope descriptions. The first rule says that in order to relate a cd of this type and a corresponding gd, besides the condition that they shall unify with the argument matrices, the reference information must be possible to express as part of a grammatical subject and the scope information must be expressed as part of a grammatical predicate. In addition the rule adds more features to the gd, in this case tense information.

$$\text{inform-sw}\left(\begin{bmatrix}\text{type} & \text{rs}\\ \text{ref} & {}^{1}[]\\ \text{scope} & {}^{3}[]\end{bmatrix}, \begin{bmatrix}\text{subject} & {}^{2}[]\\ \text{predicate} & {}^{4}[\text{vform pres}]\end{bmatrix}\right) \leftarrow$$

$$\text{refer-sw}\,({}^{1}[], {}^{2}[]), \text{describe-sw}\,({}^{3}[], {}^{4}[])$$

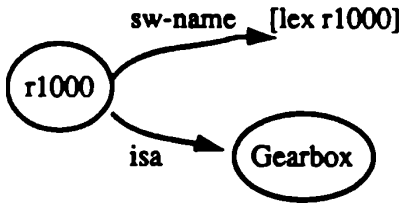
$$\text{describe-sw}\left(\begin{bmatrix}\text{thing} & {}^{1}[\,]\\ \text{aspect} & {}^{2}[\,]\\ \text{value} & {}^{3}[\,]\end{bmatrix}, {}^{4}[\,]\right) \leftarrow$$

$$\text{iget}({}^{1}[\,], {}^{2}[\,], {}^{3}[\,]), \text{iget}\left({}^{2}[\,], \text{sw-epred}, \begin{bmatrix}\text{arg} & {}^{5}[\,]\\ \text{body} & {}^{4}[\,]\end{bmatrix}\right), \text{iget}({}^{3}[\,], \text{sw-name}, {}^{5}[\,])$$

The predicate is obtained by applying a rule for the functor describe-sw. The right-hand side of this rule consists of three calls to the KB, where the first answers the topical question and the other two retrieves linguistic information as explained in the next section.

4.4 The knowledge base

The knowledge necessary for the first stage of the generation process is kept in a single knowledge base. The primary aim of the knowledge base is to store information about the domain objects. It is organized as an inheritance network, allowing instances to inherit properties from their concepts. The properties can be divided in two classes, domain related and language related. In the small example below the r1000, called the carrier, is an instance of the concept Gearbox, and it has the property sw-name, also called the indicator, defined to be the value [lex r1000].



Associated with each indicator, there is a method which knows how the value is to be retrieved. The value can either be cached in the carrier node directly, inherited from a concept through isa-links, or otherwise computed. It is the fact that the knowledge is only accessible through the domain objects that makes generation much easier than parsing. In order to make the system bidirectional we would also have to make domain objects accessible from their property values, especially the linguistic properties.

Linguistic information can be associated with domain concepts in different ways. Nouns are stored under an indicator name which is then differentiated for the two languages as in the example above. Information relevant for predicates is stored under the indicator epred which is differentiated in the same way. In the case of the concept Operation, the Swedish epred-value is a structure, which says that the predicate should contain the verb *manövrera* in the passive voice, and a manner adverbial expressing the manner of operation:

$$\begin{bmatrix} \text{arg} & {}^{1}[\,] \\ \text{body} & \begin{bmatrix} \text{passive} & \text{yes} \\ \text{compl1} & \begin{bmatrix} \text{manner} & {}^{1}\begin{bmatrix} \text{aform} & \text{t} \end{bmatrix} \end{bmatrix} \\ \text{lex} & \text{manövrera} \end{bmatrix} \end{bmatrix}$$

In section 3 we considered the following sentence pair, and introduced a concept Purpose for its description:

S: Spärrventilen har till uppgift att förhindra växling av rangeväxeln när ...
E: The purpose of the inhibitor valve is to prevent inadvertant shifting of the range gear when ...

The Swedish sentence can be handled by the previous rule for reference-scope propositions.[4] The structure for the Swedish predicate information associated with Purpose would be as follows:

$$\begin{bmatrix} \text{arg} & {}^{1}[\,] \\ \text{body} & \begin{bmatrix} \text{compl1} & \begin{bmatrix} \text{pobj} & \begin{bmatrix} \text{prep} & \text{till} \\ \text{obj} & \begin{bmatrix} \text{lex} & \text{uppgift} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{lex} & \text{ha} \\ \text{compl2} & \begin{bmatrix} \text{infcomp} & {}^{1}[\,] \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

For the English sentence we actually have to use a different clause-level rule, introducing relations of the asserted proposition both with the subject and the predicate. The first association will use the English name of the concept, while the second will use the predicate information. In conjunction these rules will produce a partial grammatical description corresponding to the pattern *the purpose of ... is ... .*

$$\text{inform-eng}\left(\begin{bmatrix} \text{type} & \text{rs} \\ \text{ref} & {}^{1}[\,] \\ \text{scope} & {}^{3}[\,] \end{bmatrix}, \begin{bmatrix} \text{subject} & {}^{2}\begin{bmatrix} \text{pmod} & \begin{bmatrix} \text{prep} & \text{of} \\ \text{obj} & {}^{4}[\,] \end{bmatrix} \end{bmatrix} \\ \text{predicate} & {}^{5}\begin{bmatrix} \text{lex} & \text{be} \\ \text{vform} & \text{pres} \end{bmatrix} \end{bmatrix}\right) \leftarrow$$

$$\text{refer-eng}\,({}^{1}[\,], {}^{4}[\,]), \text{describe-eng}\,({}^{3}[\,], {}^{5}[\,]), \text{identify-eng}\,({}^{3}[\,]\,{}^{2}[\,])$$

5. Status and future work

We believe that the generation procedure as far as sentences are concerned is powerful and flexible enough to handle most of the sentence-level phenomena of our corpus. However, this remains to be proved, as the current grammar only covers a fraction of the corpus. Moreover, we will extend the rule base to handle paragraph-level phenomena such as coherence relations and anaphoric dependencies as well.

If anything, the generation procedure is at present rather too powerful and unconstrained, so we want to investigate further what constraints to put on it. As for speed, the bottleneck of the generation process is the surface generator. Ideally we would like to eliminate it completely and work with a single grammar incorporating phrase structure as well as functional grammatical information.

Notes

1. The project, Konceptuell textrepresentation för automatisk generering och översättning, is funded by the Centre for Industrial Information Technology (CENIIT) at the Linköping Institute of Technology.
2. We take the sentence pairs in the corpus as *prima facie* instances of translation equivalents and thus necessary to account for. This may be questioned in a few cases, e.g. where one sentence contains a modifier having no counterpart at all in the other sentence, but all such exceptions need careful motivation.
3. Two paragraphs may be considered equivalent if they answer the same questions in the same order.
4. It needs a different rule for describe-sw, however, as the value of the asserted proposition need not be a simple concept.

References

Ahrenberg, L. (1987): Interrogative structures of Swedish: aspects of the relation between grammar and speech-acts. RUUL 14 (Doct. diss.), Uppsala University, Department of Linguistics.

Carbonell, J.G. and Tomita, M. Knowledge-based machine translation, the CMU approach. In S. Nirenburg (ed.) *Machine Translation.* Cambridge University Press, pp. 68-89.

Carlson, L. (1983): *Dialogue Games.* Dordrecht, Reidel.

Emele, M., Heid, U., Momma, S. and Zajac, R. (1990): Organizing linguistic knowledge for multi-lingual generation. *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, 20-24 August, 1990, pp. 102-107.

Kaplan, R.M., Netter, K., Wedekind, J., Zaenen, A. (1989): Translation by structural correspondences. *Proceedings of the 4th EACL Conference*, Manchester 10-12 April, 1989, pp. 272-281.

Nilsson, U. and Maluszynski, J. (1990): *Logic, Programming and Prolog.* John Wiley & Sons.

Russell, G., Ballim, A., Estival, D., Warwick-Armstrong, S. (1991): A language for the statement of binary relations over feature structures. *Proceedings of the 5th EACL Conference*, Berlin 9-11 April, 1991, pp. 287-292.

Sadler, L. and Thompson, H. (1991): Structural non-correspondence in translation. *Proceedings of the 5th EACL Conference*, Berlin 9-11 April, 1991, pp. 293-298.

Shieber, S.M. (1988): A uniform architecture for parsing and generation. *Proceedings of the 12th*

International Conference on Computational Linguistics, Budapest, 2-27 August 1988, pp. 614-619.

van Noord, G. (1990): Reversible unification based machine translation. *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki 20-24 August, 1990, pp. 299-304.

Zajac, R. and Emele, M.(1990): Typed unification grammars. *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, 20-24 August, 1990, Vol. 3 pp. 293-298.

Lars Ahrenberg & Stefan Svenberg
Department of Computer and Information Science
Linköping University
S - 581 83 Linköping
Email: {lah,ssv}@ida.liu.se

Operajonell maskinomsetjing: Kvar møter vi veggen?

Magnar Brekke
Norges Handelshøyskole

Roald Skarsten
Universitetet i Bergen

1. Introduksjon

1.1. Utgangspunkt

Dette er ein rapport om ei undersøking av produktiviteten ved bruk av maskinomsetjing (heretter MT) som del av eit større kommersielt omsetjingsprosjekt. Ei prosjektgruppe ved Det historisk-filosofiske fakultet ved Universitetet i Bergen gjennomførte i tidsrommet 1988-90 det s.k. Prosjekt PHILENTRA på oppdrag frå Phillips Petroleum Company Norway. Den maskinelle delen av oppdraget vart utført ved bruk av Weidners MacroCAT system for engelsk-til-norsk. Dette programmet vart utvikla ved UiB 1986-87 (Prosjekt ENTRA) på grunnlag av tilsvarande engelsk-tysk versjon (jfr. Brekke og Skarsten 1987). ENTRA-systemet vart utvikla på DEC MicrovaxII i samarbeid med Digital Equipment Corporation, Norge, og vart kjørt på denne maskinen i prosjektperioden.

1.2. Avgrensing

Rapporten er orientert ut frå ønsket om å vinna empirisk kunnskap om den rolle eksisterande programvare for MT kan spela innanfor eit meir omfattande omsetjingsprosjekt med reelle rammefaktorar (så som kostnadseffektivitet, tidspress, kvalitetskontroll m.m.). Undersøkinga er ein konkret konfrontasjon mellom eit operasjonelt og praktisk innretta MT-system og dei språklege, økonomiske og sosiale realitetane som det er nødvendig å kunna takla. Denne rapporten må sjåast som ei tilbakemelding frå eit frontavsnitt som er viktig dersom vi meiner at det vi som språkspesialistar driv med, skal ha relevans for det samfunnet vi lever i. Det som følgjer er altså innretta mot praktisk utforsking heller enn teoretisk forklaring av dei faktorar som spelar inn ved bruk av MT som omsetjingsreidskap.

2. Prosjektdesign

Den overordna arkitekturen for Prosjekt PHILENTRA går fram av flytskjemaet i Appendiks 1 (med mindre justeringar, jfr. nedanfor). Tekstane kom frå oppdragsgivar som trykt original og som ASCII-fil på diskett. Kvart deldokument (kapittel) i papirkopi gjekk til translatør for term-ekser-

pering. Lista over ekserperte termar gjekk så saman med papirkopien til terminolog, som ved søk i Norsk termbanks NOT-base etablerte sannsynlege ekvivalentar, ofte i konsultasjon med oppdragsgivar.

I mellomtida vart ASCII-teksten lasta inn i den elektroniske tekstbanken og førebudd for MT-systemet. Eit vokabularsøk gav ei liste over ukjende ord som saman med termane frå termsøket vart lagde inn i systemordboka. Teksten vart deretter kjørt gjennom ENTRA-programmet og resultatet overført til diskett i WordPerfect-format. Denne ut-teksten vart lasta inn på PC og danna grunnlaget for translatøren si redigering (*post-editing*) og tilbakeføring til opphaveleg tekstformat. Ferdigredigert tekst vart så revidert og kontrollert før siste utskrift og retur til oppdragsgivar.

Tekstmassen besto av omlag 2000 sider driftshandbøker for oljeindustrien, fordelt over 31 ulike subdomene (jfr. Appendiks 2). Desse spenner over svært mange tekniske fagområde, frå prosesskjemi over det meste av mekanisk industri til drikkevatn og telekommunikasjon. Det seier seg sjølv at den filologiske tekstforståing her er sterkt avhengig av fagspesialistens innsikt i funksjon og arbeidsmåte, og av at ulike ekspertar kan etablera ei effektiv kontaktform.

PHILENTRA-teamet besto for det meste av 5 personar på deltid, som seg i mellom representerte følgjande relevante funksjonar: terminolog, translatør, engelsk fagspråkspesialist, edb-spesialist og sekretær/koordinator (jfr. seksjon 3). Alle var rutinerte PC-brukarar, dei fleste med hovudfag eller meir, og med relativt sterk praktisk språkkompetanse. Etter visse innkjøringsvanskar og personskifte i starten fungerte dette teamet bra gjennom heile prosjektperioden.

3. *Forskingsdesign*

Undersøkinga av tidsbruk og produktivitet i totalprosjektet omfattar fire viktige delområde med kvar sine spesifikke funksjonar. Innhaldet i kvart av desse delområda er lista opp nedanfor; forkortingane refererer til kurvene på fig. 2. Ein gitt medarbeidar hadde sjeldan alle eller berre funksjonar under eitt delområde:

3.1. Terminologiarbeid (TRM):

- Termekserpering i papirkopi
- Termkontroll mot Weidner systemordbok (*Vocabulary Search*)
- Termsøk i Norsk termbanks NOT-base
- Termavklaring (telefon, telefaks, brev) med oppdragsgivar
- Oppdatering av systemordbok (*Dictionary Update*)

3.2. Redigering (= post-editing) (EDT):

Med utgangspunkt i tilgjengeleg originaltekst i papirkopi, termliste for deldokument, og maskinprodusert ut-tekst i WP-format på diskett:

- Rekonstituering av originalformat
- Tekstredigering

3.3. Tekstrevisjon (REV) (NB: ved annan person enn redigerar):

- Kritisk gjennomlesing av uttekst
- Konsekvenskontroll, dokument-internt og mot tidlegare dokument
- Oppretting

3.4. Administrasjon (ADM):

- Konvertering og overføring av elektronisk tekst
- Kjøring av maskinsystemet

- Kopiering og utskriving
- Ymse sekretærfunksjonar

Etter denne skissa av prosjektet og undersøkinga går vi no over til å drøfta dei viktigaste problema vi støytte på i den praktiske gjennomføringa.

4. Problemstillingar

4.1. "Virkelighetskonfrontasjon"

Prosjekt PHILENTRA har ei rekke interessante problemstillingar. I det som her er kalla "virkelighetskonfrontasjon" ligg det at vi ville ha ei tilbakemelding på kva problem det er viktigast å få løyst først dersom MT skal kunna bli til større samfunnsmessig nytte. Kva ville utfallet bli av ein konkret konfronasjon mellom eit operasjonelt MT-system og den røyndom ("virkelighet") av språklege, økonomiske og sosiale realitetar som forretningsverda representerer? Frå vår side låg det eit behov for å knyta den lingvistiske refleksjon til eit anna utgangspunkt enn det teoretiske, der det lingvistisk interessante normalt blir definert ut frå ein bestemt teori, og innanfor denne teorien sin problemhorisont.

Lingvistisk interessante problem synte seg ofte å spela ein perifer rolle i denne samanhengen. Det var ofte lingvistisk trivielle problem som vart avgjerande for kor vellukka resultatet vart pga. den kumulative effekten. Eit i og for seg overkomeleg lingvistisk problem med høg tekstfrekvens, men der løysinga ikkje er implementert i systemet, skaper eit stort problem i praksis, medan eit teoretisk komplisert problem som førekjem sjeldan, i praksis er ubetydeleg fordi omsetjinga likevel må redigerast. Problemfrekvens blir med andre ord viktigare enn teoretisk vanskegrad, og vanskegrad i dette praktiske perspektivet er igjen relatert til dei tastetrykka translatøren må utføra.

Ein annen viktig subdimensjon er tidsdimensjonen. I ein konkret arbeidssituasjon med faste leveringsfristar var det uråd å prioritera "akademiske problem". Dette ga oss også som akademikarar ei anna nyttig erfaring. Vi hadde planlagt at det faglege utviklings- og oppgraderingsarbeidet skulle gå parallelt med oppdragsprosjektet, slik at sjølve MT-systemet kunne kontinuerleg forbetrast frå dokument til dokument; tidspresset gjorde dette i praksis umogeleg for oss. Når ein arbeider i en anbodssituasjon, vil det faglege aspektet tapa med ein gong det kjem kryssande interesser inn. Det faglege arbeidet bør difor gå parallelt med og uavhengig av anbodssituasjonen, dersom det let seg gjera, eller skje i etterkant, dersom det er dei same personane som er involvert.

Tidsdimensjonen aktualiserte også ei rekkje andre latente problemstillingar som ikkje let seg kontrollera på same måte som når ein skaper seg si eiga ideelle verd og løyser problema teoretisk. MT vil, med sitt verifiserbare krav til effektivitet, stilla dei engasjerte akademiske miljøa andsynes ein ny type utfordringar, og det er ikkje gitt at MT-forsking primært vil bli ein akademisk disiplin i framtida dersom ein ikkje maktar å takla slike "virkelighetskonfrontasjonar", noko det ikkje er tradisjon for i lingvistiske miljø. Omsetjing har nemleg hittil ikkje hatt ein etablert og veldefinert rolle innan lingvistikkdisiplinen.

4.2. Evalueringsvanskar

Som kjent er det i seg sjølv ein vanske å evaluera kvalitet ved omsetjing ut over det grunnleggjande kriteriet at kunden skal vera nøgd med resultatet. Denne rapporten gir ikkje rom for vidare drøfting av dette men nøyer seg med å påpeika at vi i prosjektrapporteringa vår har støytt på tilhøve som ikkje så lett lar seg eintydig definera. Grensa mellom generelt vokabular og terminologi er ikkje lett å finna, og slett ikkje grunngi; tilsvarande gjeld for grensa mellom termar og kollokasjonar. Vi har følgt den praksis som Norsk Termbank har etablert ut frå erfaring og skjønn.

Vi vil også gjera merksam på at når vi samanliknar med manuell omsetjing, tar vi som gitt at dette

inneber bruk av elektronisk tekstbehandling og termoppslag, på same måte som under redigering av MT-produsert ut-tekst. Vi vil også nemna at evalueringa ikkje byggjer på ein ideell prosedyre, i og med at vi er part i saka (for eit klårt døme på dette sjå Maklovitch 1985) og fordi tidspresset også her kan vera årsak til mindre målefeil.

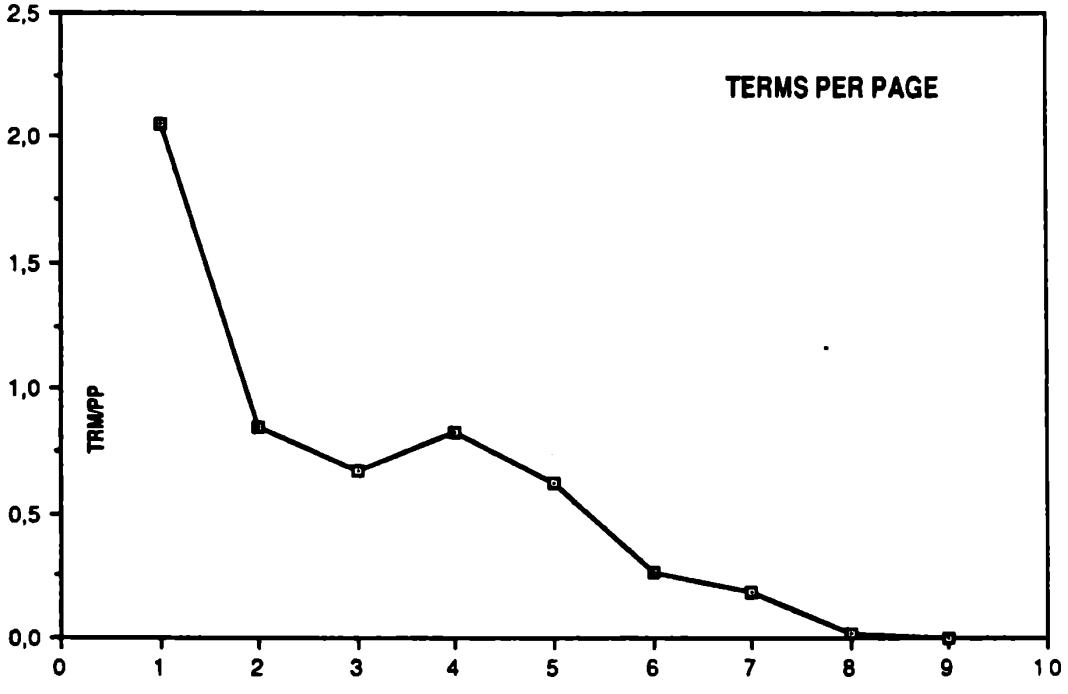
Translatørens psykiske reaksjonar på "samarbeid" med MT er også ein noko variabel faktor. Det er utan vidare klårt at det er meir anstrengande å "reparera" eit dårleg tekstutkast enn å streva med å omsetja direkte frå grunnteksten. Som vi snart skal sjå, skulle dette tilhøvet få ein viktig innverknad på sluttfasen av prosjektet og representera ei feilkjelde som ikkje vart avdekka før det underliggande talmaterialet var gjennomarbeidd.

Desse atterhalda er likevel ikkje av ein slik karakter at det generelle biletet som kjem fram vil kunna endrast vesentleg. Det tyder berre at ein bør vera romsleg med "desimalnivået" i tolkinga av resultata. I hovudsak refererer våre målingar til eit representativt utval som omfattar nær 50% av total tekstmasse.

4.3. Delproblem

4.3.1. Terminologi

Den gjennomsnittlege forekomsten av nye termar pr. side for kvar av dei 9 dokumenta er framstilt i fig. 1. Prosjektkalkylen vart lagt etter ein del stikkprøver under pilotfasen, og desse ga oss grunn til å rekna med at nye termar ville opptre relativt tett i startfasen men etter kvart sjeldnare. Terminologiarbeidet ville i startfasen m.a.o. kunna påføra prosjektet eit større tap som etter ei viss tid ville bli kompensert etter at terskelarbeidet var gjort. Som fig. 1 tydeleg viser, vart denne forventninga innfridd. I det første dokumentet (nr. 1 langs horisontal akse), som var på totalt 420 sider, fann vi i gjennomsnitt litt i overkant av to nye termar pr. side (vertikal akse). Alt etter det



Figur 1: Documents

andre dokumentet, då 30% av totalsidetalet var tilbakelagt, var dette talet for termar pr. side meir enn halvert, og slik vidare med ein gjennomgåande fallande tendens fram mot det siste dokumentet (nr. 9 langs horisontal akse), som inneheldt ingen nye termar. Den grafiske kurven i fig. 1 stadfester

at den terminologiske grunnlagsinvesteringa ved bruk av MT er tung og at lønsemd berre kan oppnåast ved eit visst volum.

I praksis synte det seg likevel svært vanskeleg å styra ein konsekvent og til ei kvar tid oppdatert termbruk berre ut frå systemordboka. Som eit separat delprosjekt produserte vi difor ei trykt termliste på grunnlag av omlag 700 ekserperte termar og faste uttrykk, med både engelsk og norsk oppslag. Termlista fungerte som eit konkret referansepunkt og ei fokusering for den kontinuerlege dialogen vi førte med oppdragsgivar omkring termbruken.

4.3.2.. Ekvivalens

Ekvivalensproblematikken er ikkje løyst berre ved terminologisk registrering. Igjen er det kunden som bestemmer kva som er ei funksjonelt ekvivalent omsetjing. Det synte seg her at petroleumsbransjen hadde innarbeidd i språket sitt den spesielle forma for norsk som i fagleg samanheng er kjend som Norwenglish. *Flash gas* td. skulle i følgje prosessingeniørane heita det same på norsk, heller enn "avdampingsgass", som var Norsk Termbanks ekvivalent. Spørsmålet om kven som skulle vera normgivar, og konflikten rundt det, kom difor inn med ei tyngde som vi ikkje hadde forutsett.

Eit anna problem som vi ikkje hadde forutsett, hadde samanheng med fagbestemte subkulturar og språkkonvensjonar. Det synte seg at etablert norsk ingeniørspråk er tydeleg basert på tysk tradisjon og tyske formuleringer, medan dokumenta reflekterte anglo-amerikansk ingeniørspråk. I slike tilfelle er normgrunnlaget høgst uavklara, og alt etter tradisjonsforankring kunne vi frå tilsette i same selskap få ulike svar på kva som var rett teknisk språkføring.

Ein tredje faktor var fraseologien og den "tause" kunnskapen. Med det meinest at ei omsetjing i og for seg kunne vera grammatisk rett, og isolert sett ein uklanderleg setning, men "det var berre ikkje slik ein sa det", og det visste alle som var i bransjen. Ein kunne akseptera nokre avvik, men når dette vart det vanlege, vart teksten opplevd som sær, sjølv om han ikkje var direkte feil, jfr. døme 1:

(1) *Atmospheric Equipment Drain System* skulle heita

"System for avløp fra atmosfærisk utstyr" og ikkje noko anna, som td.

"Avløpssystem for atmosfærisk utstyr"

Vidare må nemnast problemet med "gjendikting", som for MT er uløysande. Såleis fastslo oppdragsgivar at døme 2:

(2) *Tank Pressure Device* burde bli gjengitt som:

"gassvakt for å varsle dersom det oppstår isolasjonsfeil på transformatorviklingen"

Dette er eit nokså ekstremt døme på gjendikting kopla med implisitt encyklopedisk kunnskap og ligg langt frå det ein vanlegvis reknar som omsetjing, maskinell eller ikkje.

For det fjerde kan nemnast samansette nominaluttrykk, eit velkjent crux for all syntaktisk analyse. Det er ikkje mindre kjent som eit vanskeleg MT-problem, ikkje berre for automatisk analyse men også i høg grad for transfer-reglane som lagar overgang mellom kjeldespråk og målspråk. Mellom typologisk ulike språk som td. engelsk og norsk blir dette ekstra kinkig. Det som på engelsk er ein del av frase-syntaksen, må på norsk ofte fordelast mellom termdanningsreglane og preposisjonsfrasebruken (døme 1 ovanfor illustrerer dette godt).

Her er vi inne på eit felt som synest å liggja svært dårleg til rette for automatisert omsetjing, særleg når opphopinga av element i den engelske nominalfrasen overstig 4-5. Dette er ofte tilfelle i kompakt teknisk dokumentasjon, og døme (3) er førebels ytterpunktet i samlinga vår:

(3) *frame molded case thermal magnetic automatic 3 pole air circuit breaker*

Det er mange problem som kjem til syne her: er dette ein term, eller eit konglomerat av termar (med td. *frame molded, 3 pole* og *air circuit breaker* som kjerneelement)? Eller ser vi her berre det ulykksalige resultatet av forfattaren sitt behov for å korta ned til eit samansett nominaluttrykk det som burde vore sagt med tre-fire setningar? Døme (3) viser ein type problem som ikkje har tradisjon for å vera lingvistisk interessant, men ei løysing må finnast. Etter inngåande konsultasjonar vart vi stående ved (4):

(4) "3-polet kompaktbryter fastmontert med overlast- og kortslutningsvern"

men den kom sjølvsagt ikkje ut av maskinen. Også her er det bygd inn encyklopedisk kunnskap som ikkje var eksplisitt i det engelske uttrykket.

Gjennom heile tekstmassen førekom liknande relativt faste og tilbakevendande sekvensar av nominal-element, gjerne svært like uttrykk men med små variasjonar i ordstilling. Ein konkordans var eit nyttig hjelpemiddel for å finna desse. Sjølv om få var så ekstreme som døme 3, vart det likevel, innanfor prosjektrammene, uråd å avgjera terminologisk status for mange av dei. Vi var dermed avskorne frå å leggja slike sekvensar inn i systemordboka og såleis henta ut nokon rasjonaliseringsgevinst i dette høvet.

Vi kan ikkje forlata ekvivalensproblematikken utan å peika på problemet med det språklege "bevissthetsnivået" som karakteriserer mange bedrifter. Ofte har selskapet ein vag eller uavklart *company policy* m.o.t. språkbruk. Det er vår erfaring at slike spørsmål er lågt prioritert og kanskje ikkje kjem til å få noko anna løysing enn at den som er siste formelle kontrollerande instans også blir språkleg normgivar utan at det skjer etter medviten overveging. Dermed risikerer vi at ein annan dag kan det vera eit anna synspunkt som er "det rette". I tillegg kjem ei allmenn forventning om at same kor dårleg og uklar originalteksten er, skal sluttproduktet vera høgglanspolert og uangripeleg ("taus forventning"?).

4.4 Tekstleverandørtilhøve

Prosjektet vart kalkulert med utgangspunkt i tekst som alt var elektronisk tilgjengeleg. Det seier seg sjølv at MT basert på tekst som ikkje var i elektronisk format ville gitt eit helt anna reknestykke, sjølv med optisk innlesingsutstyr. Dette problemet vil gradvis avta og er vel alt i dag eit marginalt praktisk problem pga. den generelle bruken av tekstbehandling.

Det at teksten har elektronisk format sikrar likevel ikkje alltid rask MT-behandling. Ofte kan forekomsten av visse kontrollkodar (som td. "hard linjeskift" ved slutten av kvar linje i ein ASCII-tekst) skapa uforutsette og tidkrevjande vanskar. Sideformatering kan representera eit vesentleg større problem. Vår erfaring var at utprega teknisk spesifikasjonsstoff i spalter ikkje eignar seg for MT, jfr. døme (5):

(5)	YSLL 5083 Local panel mounted YSLL 5084 low-low UV intensity switch for each pod.	Set at 24 % full intensity.

No er nye system undervegs som kan takla også slike problem, men det krev då ei spesialprosedyre for deformatering og reformatering.

I ein konkret produksjonssamanheng blir det viktig ikkje berre med termbruk og stilnivå men også meir banale aspekt som gjeld typografisk og ortografisk kvalitet på inntekst. I MT-samanheng blir dette spørsmålet endå viktigare fordi ein setningsbasert analyse ikkje gir rom for *educated guessing*. Det same problemet eksisterer også for setningar med trykkfeil som i seg sjølv gir legitime ordformer. Døme (6) illustrerer problemet:

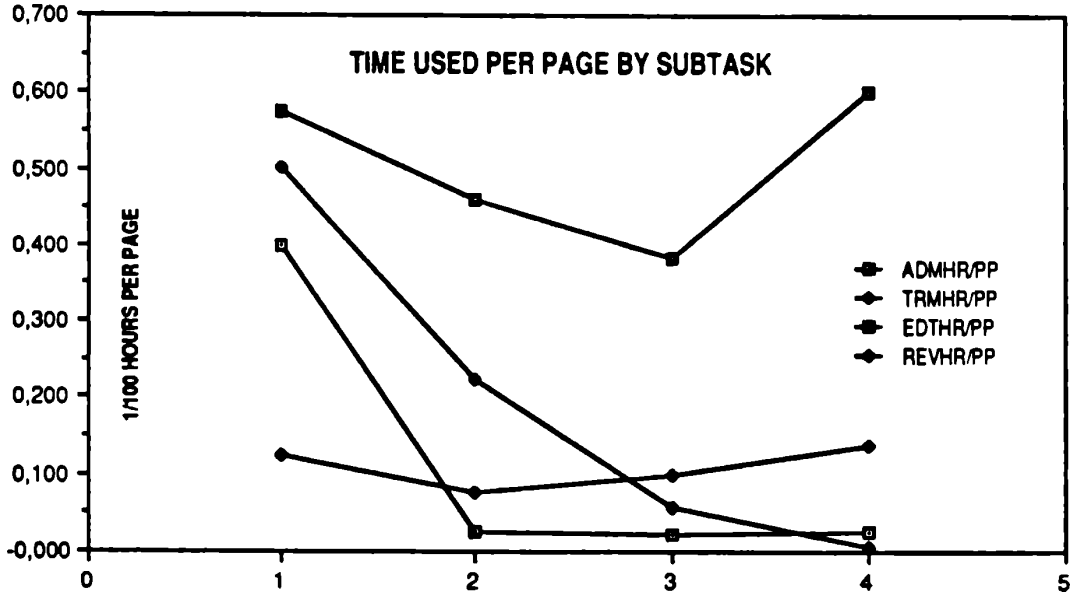
(6) *The ling-up consits of 23 unit of metre-clad circuit breakers.*
(link-up)(consists) (metal)

Det er ikkje vanskeleg å skjøna at for mykje av dette skaper større problem for MT enn for tradisjonell omsetjing.

Til slutt skal nemnast eit delproblem som vi gløymde å ta omsyn til under planleggjinga: med nye reidskap og arbeidsmåtar følgjer det ofte visse psykologiske reaksjonar som påvirkar samarbeidskulturen i eit prosjekt. Vi opplevde i starten ein viss uvilje hos oppdragsgivars dokumentasjonsansvarlege, truleg fordi dei opplevde MT som ein potensiell konkurrent til etablerte produksjonsrutiner. Det tok ei stund før samarbeidskulturen vart slik at dette problemet kunne seiast å vera løyst. Det som frå vår side var testing av eit spennande MT-prosjekt, med sikte på ytterlegere forbetringar, var noko ukjent og dermed kanskje negativt lada og til og med trugande for enkelte medarbeidarar. Dette vart ei påminning om kor viktig det er i forkant av eit prosjektsamarbeid å etablera ei felles grunnleggjande forståing for problem og avgrensingar og slik sikra at oppdragsgivar alt i starten har realistiske forventningar til resultatet.

4.5 Er MT konkurransedyktig?

Med nemnde føresetnader og kompliserande faktorar er vi så komne fram til hovudproblemstillinga: Er MT konkurransedyktig, og i så fall, kor stor er gevinsten? Framstillinga vil i det følgjande vera knytt til fig. 2 (for kodar sjå seksjon 3 ovanfor):



Figur 2: Documents

Fig. 2 er ei samanstilling av den tida (uttrykt som timar pr. side langs vertikal akse) som gjekk med til administrasjon, terminologiarbeid, redigering og revisjon av fire utvalde dokument (langs horisontal akse) mellom i alt ni. I tillegg til første og siste dokument (respektive nr. 1 og 4 på fig. 2) valde vi dei to som representerte 1/2-vegs- og 3/4-dels-merket i prosjektgjennomføringa (respektive nr. 2 og 3 på fig. 2), dvs. etter 1048 og 1530 sider. Utvalet utgjer nær 50% av heile tekstmassen (995 av i alt 2022 sider).

Den markerte nedgangen i nye termar pr. side (synt på fig. 1) er på fig. 2 reflektert som eit sterkt fall i ressursbruken på terminologiarbeid (jfr. nest øvste kurve over dokument 1, som går mot null ved dokument 4). Behovet for administrasjon, særleg tungt i innkjøringsfasen (jfr. nest nedste kurve

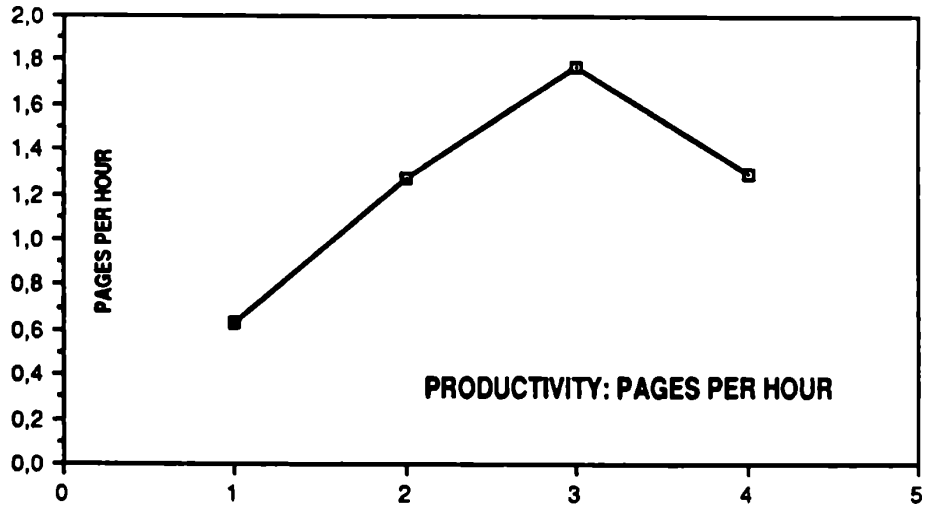
over dokument 1) blir også sterkt redusert i resten av prosjektperioden. Behovet for revisjon av ferdig translatør-redigert tekst (jfr. nedste kurve over dokument 1) er relativt stabilt men svakt aukande mot slutten. At behovet for redigering skulle auka så dramatisk mot slutten som øvste kurve på fig. 2 syner, er eit faktum som krev ei litt grundigare forklaring.

Det hadde seg nemleg slik at våre to translatørar, etter at 3/4 av prosjektet (seks dokument på til saman 1530 sider) var tilbakelagt, kom med framlegg om å omsetja siste 1/4-delen manuelt. Dei var på dette stadiet godt innkjørde i realia og terminologi og veltrente i språkstil og formuleringsmåte for desse dokumenta. Leveringsfristen nærma seg, og translatørane meinte at dei no kunna tasta inn norsk versjon (på skjermen) raskare og meir rasjonelt direkte frå grunnteksten (på papir) enn om dei skulle halda fram med å "reparera" på MT-uttekst med stadig tilbakevendande feil og manglar.

Prosjektleiinga (=forfattarane) opplevde framlegget som litt av eit antiklimaks, fordi vår utgangshypotese tilsa at det var nettopp i denne fasen at dei største fruktene av MT ville kunna haustast. I den pressa tidssituasjonen var det viktigaste likevel å ferdigstilla siste dokument innan den avtalte leveringsfristen. Vi gav dermed grønt lys for direkte manuell omsetjing av dei siste tre dokumenta, mellom desse altså det aller siste, representert som nr. 4 på fig. 2. Resten av arbeidet synte god framdrift og vart avslutta i rett tid, og translatørane var tilfredse med arbeidssituasjonen i innspurten.

Det skulle etter dette ikkje vera uventa at tidsbruken på redigering (=manuell omsetjing med elektronisk termoppslag og tekstbehandling) skyt i veret og er høgare pr. side for dokument 4 enn for nokon av dei andre. At tidsbruken på revisjon (jfr. nest øvste kurve over dokument 4 på fig. 2) ikkje berre blir redusert men som nemnt syner ein svak auke, er vanskelegare å forstå. Men det overordna synspunktet blir: korleis påverka overgangen til manuell drift den totale produktiviteten i prosjektet?

For å kunna svara på dette, er det naturleg å gå over på den vanlege produktivitetsindeksen som translatørar gjerne brukar, nemleg produksjon av ferdige sider pr. time arbeidsinnsats, anten det gjeld administrasjon, termarbeid, redigering eller revisjon. Slår vi saman desse tala for kvart av dei fire utvalde dokumenta, får vi som resultat kurven i fig. 3. Produktivitetskurven syner bra stigning frå 0,62 sider pr. time for dokument 1 (det aller første), gjennom 1,27 sider pr. time for dokument 2 (etter vel 1000 sider) og opp til 1,77 sider pr. time for dokument 3 (etter vel 1500 sider). Her inntreffer så det totalt uventa i dette prosjektet, like overraskande for translatørar som for prosjektleiing: ved overgang til manuell omsetjing i ein fase der behov for administrasjon og termarbeid er ved eit minimum, går den totale sideproduksjonen **ned** til 1,29 pr. time.



Figur 3: Documents

Dette var så pass sensasjonelt at vi, etter at undersøkinga var avslutta, gjekk tilbake i materialet og kontrollerte også dei to andre manuelt omsette dokumenta, altså det sjuande og åttande. Desse tala vart litt høgare men framleis markert under tala for dokument 3. I gjennomsnitt for manuell omsetjing var produktiviteten 1,42 sider pr. time arbeidsinnsats, altså ca. 25% lågare enn det beste av maskinomsetjing. Denne differansen er oppsiktsvekkjande særleg av to grunnar:

1. Den manuelle omsetjinga er utført av høgt kvalifiserte og rutinerte translatørar med solid erfaring med den aktuelle terminologi og språktype.
2. Fallet i produktiviteten frå den maskinelle fasen var stikk i strid med desse translatørane si psykologiske oppleving av arbeidssituasjonen m.o.t. kva metode som var mest produktiv.

Det viktigaste spørsmålet som vi ikkje har fått svar på, er sjølvsagt korleis kurven i fig. 3 ville ha gått etter td. 2000 eller 3000 sider dersom vi hadde halde fram med maskinell omsetjing som del av prosessen, og særleg på kva punkt vi ville ha fått ei utflating. Den oppfatninga som vi etter kvart fekk, er at under dei rådande rammevilkår (kvaliteten i det eksisterande systemet, ressursar til justering og forbetring undervegs o.l.) kunne vi ikkje ha venta særleg meir stigning i produktiviteten. Det er ein slik tentativ konklusjon som ligg til grunn for vår satsing på neste generasjons utviklingsverktøy for maskinomsetjing (jfr. nedanfor) heller enn på raffinering av det eksisterande.

5. Generaliseringar

Denne undersøkinga av eit konkret empirisk materiale hadde til formål å klarleggja om eit eksisterande operasjonelt MT-system av dagens programgenerasjon har vore produktivitetsfremjande i den praktiske gjennomføringa av eit oppdragsprosjekt innan fagtekstomsetjing. Vi har i det føregåande konstatert at svaret er ja, og i tillegg at dette MT-systemet i sin rette samanheng er konkurransedyktig i samanlikning med tilsvarande prosjekt utan slikt MT-innslag. Vi vil no drøfta i kor stor grad desse generaliseringane kan vera gyldige også utanfor ramma av Prosjekt PHILENTRA.

5.1 Weidners system og MT generelt

Weidners MacroCAT var resultatet av forsking og utvikling gjort på slutten av 70-talet, og er representativt for operasjonelle MT-system i store deler av 80-åra. Av dei mest alvorlege avgrensingane kan nemnast at Weidner-systemordboka berre tillet éin ekvivalent i kvar ordklasse. Den nye generasjonen av MT-system som no er under utvikling eller implementering er sterkare lingvistisk funderte (td. LFG) og har langt meir robuste analyse- og transferstrategiar (td. *Chart Parsing, Unification*). I tillegg er det utvikla meir sofistikerte teknikkar for registrering og utveljing av synonym. Ein aktuell representant for denne programgenerasjonen er MT TOOLKIT frå ECS, Provo (sjå Higinbotham, Her and Pentheroudakis 1987).

5.2 Petroleumsspråk og fagspråk generelt

Kan vi slutta frå det materialet vi har testa til fagspråk generelt? Det er ei svært komplisert oppgåve å definera fagspråk, anten det gjeld terminologi eller andre sider ved språket, og det er det ikkje plass for å gjera her. I lys av det som er sagt under seksjon 2. ovanfor er det liten grunn til å sjå på "petroleumsspråk" som noko einskapleg fenomen. Det vi har for oss, er eit generelt ingeniørspråk som tar farge av det subdomenet som er omhandla, td. prosesskjemi, og i same grad som ein får grep om oppbygning, funksjon og terminologi i subdomenet, vil ein også kunna handtera omsetjing av fagtekst innan subdomenet.

Vi er av den oppfatninga at med eit corpus som omfattar 31 subdomene (jfr. Appendiks 2) er det rimeleg å gå ut frå at våre tal ville vera dekkande også for andre bransjar. Det er liten tvil om at med eit corpus som spenner over færre subdomene ville kostnadseffektiviteten vore endå større.

5.3 Kostnadseffektivitet

Termtetthet og spreiing over subdomene er svært viktige variablar i eit kostnadsperspektiv. Ein meir uventa faktor var omfanget og betydningen av den skjulte kunnskapen som ligg i fraseologiske konvensjonar. Lingvistisk interessante problem som td. pronominalisering var i dette perspektivet perifere, fordi det var lite brukt i denne teksttypen, medan eit translatørproblem som val av rett preposisjon vart svært sentralt. Generelt sett er det tekstvolumet som utgjer den kritiske massen for kostnadseffektivitet. I vårt prosjekt låg denne grensa på ca. 1000 tekstsider.

6. Konklusjonar

Samanfattande kan det seiast at MT har gitt lingvistikken ein ny dimensjon som prøvebenk for lingvistisk teori og datamaskinelle metodar. Operasjonell MT skaper konkrete høve for uttesting og verifikasjon av teoribaserte hypoteser og forslag mot det fulle spekteret av autentiske språklege fenomen. Dette virkar definerande og fokuserande for problemstillingane.

På same tid har vi sett at MT avdekker heilt nye utfordringar i den tause kunnskapen som ligg i evna til god omsetjing, og gjennom forsøket på å implementera noko av denne. Omsetjing per se har som kjent ingen særleg tradisjon som vitskapleg fagdisiplin. Dette tilhøvet er no iferd med å endra seg, og MT vil kunna koma til å spela ei viktig rolle i denne samanhengen. For oss frå eit akademisk miljø har det vore positivt å arbeida med litt større grad av resultatorientering, kopla med gjennomførde rutinar for kvalitetssikring. Det er likevel viktig å sikra det faglege utviklingsarbeidet rom ved sida av produksjonsprosessen.

Prosjekt PHILENTRA har vist at Weidner-generasjonen av operasjonelle MT-system var rimeleg brukbar, men han er no eit tilbakelagt stadium. Det er likevel ikkje tilfeldig at det "gamle" Systran-systemet framleis er konkurransedyktig, trass i dei store lingvistiske avgrensingane det har. Forklaringa er truleg at systemet har store og varierte ordbøker med opptil ein kvart million ord. Til samanlikning kan det nemnast at vi hadde omlag 15.000 oppslagsord i vårt Weidner-system.

Dette hindrar sjølvsagt ikkje at det må vera eit primærmål for MT å få eit best mogeleg lingvistisk fundert utgangspunkt for systemet; det er difor vi også no har gått over til det LFG-baserte systemet MT TOOLKIT. Men det er nødvendig at utviklingsarbeidet undervegs kan relaterast til "virkelighetskonfrontasjonar" av den typen som her er rapportert. Den erfaringa som vi i Bergen har med MT TOOLKIT gir grunnlag for å hevda at når Weidners MacroCAT kunne gi ein viss produktivitetsgevinst, vil eit system basert på MT TOOLKIT kunna gjera det i endå høgare grad, når først terskelinvesteringa er gjort, i form av terminologisk detalj-spesifikasjon og generell ordboksoppdatering. Den nye generasjonen av operasjonelle MT-system gir langt betre høve til implementering av varierte kriterium for val av ord og setningskonstruksjonar.

Visse ekvivalensproblem (m.a. fraseologi) representerer formidable utfordringar for parametriske tilnærmingsmåtar og kan kanskje berre løysast ved individuelle tilpasninger frå prosjekt til prosjekt. På same måte som det finst eigne ordbøker for idiomatiske uttrykk kan det lagrast oppslag for ferdige omsetjingar med fraseologiske preferanser. Vansken, i høve til dei fleste idiomatiske uttrykka, er at dei ikkje er allment kjende på førehand, og i bransjen er ein seg heller ikkje alltid medviten at det nettopp er tale om ein spesiell og karakteristisk fraseologi. Mykje ligg ugjort nettop her, men når fraseologien først er identifisert og gjennomarbeidd, ligg MT bra til rettes for konsekvent bruk.

Når vi så har nådd "the moment of truth", spørsmålet om MT er lønsam, utgjer tekstvolum, termtetthet og spreiing over subdomene viktige objektive element i svaret. Men det er også subjektive element, visse psykologiske faktorar i arbeidssituasjonen, som må takast svært alvorleg. Det er viktig å hindra at translatørene må arbeida med altfor mykje halvgod språkføring som må rettast - slik at dei til slutt står i fare for å missa si naturlege språkkjensle fordi dei har vore omgitt

så lenge av unaturleg språkføring. I denne samanhengen er det den grammatisk uklanderlege omsetjinga som er farlegast, ikkje den påviseleg ugrammatiske omsetjinga.

Det kan ikkje understrekast sterkt nok at MT i dag ikkje er og neppe vil bli eit trugsmål mot translatørane som yrkesgruppe. Tvert imot er MT alt i dag, og vil i aukande grad bli, ein effektiv translatørreiskap som i sin rette samanheng vil kunna redusera rutinearbeid, sikra overordna konsekvens og fremja produktivitet i omsetjing, særleg av fagtekst.

Referansar

Brekke, M. og R. Skarsten (1987) "Prosjekt ENTRA: Engelsk til norsk", i A. Hartnack & H. Ruus (red.) *Nordisk Seminar om Maskinoversættelse*. EUROTRA, København.

Brekke, M. og R. Skarsten (1987) "Machine translation – a threat or a promise?", i C. Picken (red.) *Translating and the Computer* 9. Aslib, London.

Brekke, M. (1990) *Prospects and Perspectives in Machine Translation*. Engelsk institutt, Universitetet i Bergen.

Higinbotham, D., O.-S. Her og J. Pentheroudakis (1989) "The LFG-Based Bravice Machine Translation System". Unpublished article, ECS, Provo, Utah, to appear in *Computer Processing of Chinese and Oriental Languages*.

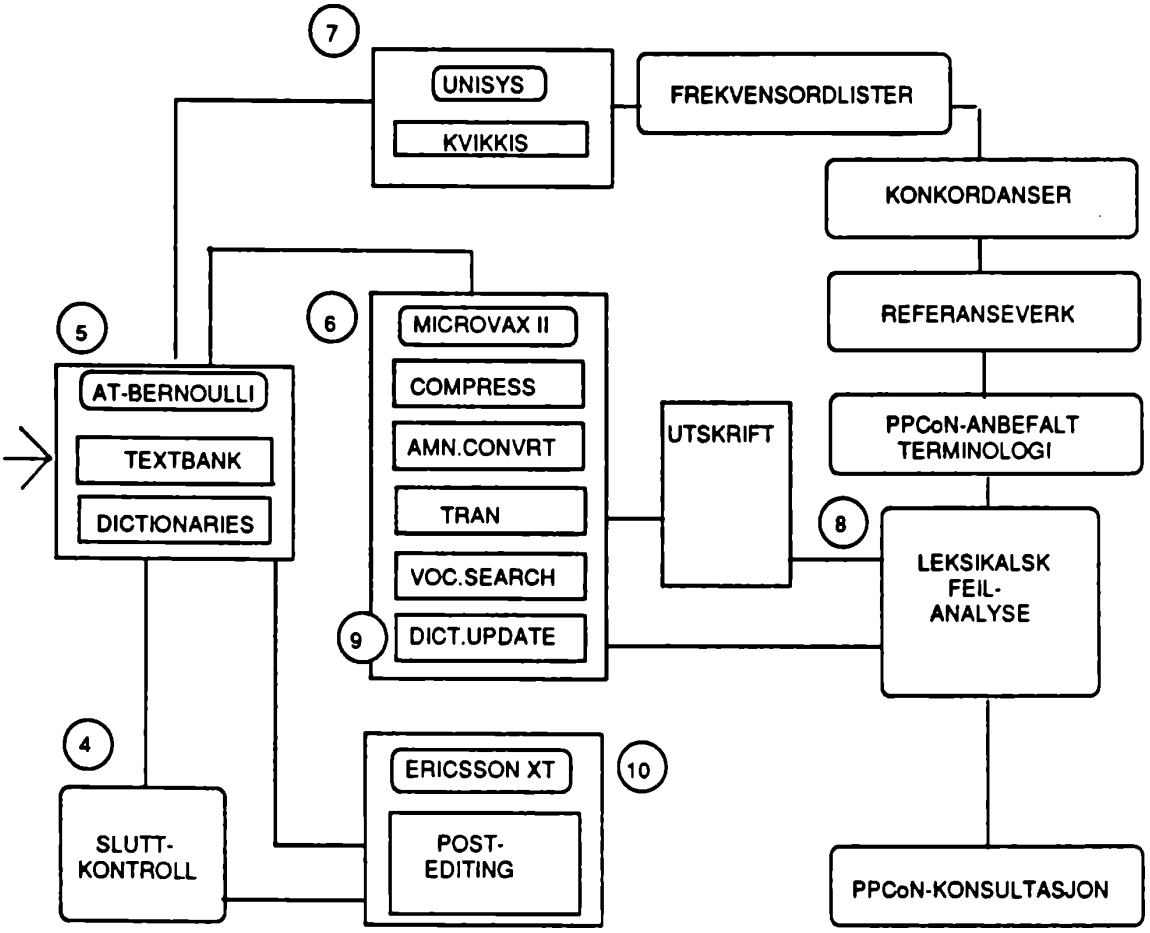
Maklovitch, E. (1986) "Machine Translation: Trial and Errors. Foredrag ved Aston University's Conference on Machine Translation, april 1986.

Skarsten, R. (1991) Developments within Machine Translation", i R. Skarsten, E.J. Kleppe and R.B. Finnestad (red.) *Understanding and History in Arts and Sciences, Vol.1, Acta Humaniora Universitatis Bergensis*. HF-fakultetet, Universitetet i Bergen.

Magnar Brekke
Institutt for språk
Norges Handelshøyskole
Helleveien 30
N-5035 Bergen-Sandviken

Roald Skarsten
Edb-seksjonen v/HF
Universitetet i Bergen
Sydnesplass 9
N-5007 Bergen

Appendix 1: Prosjekt PHILENTRA flytskjema



Appendix 2: Subdomene, petroleumsproduksjon

INSTRUMENT AND UTILITY AIR
CHEMICAL INJECTION
SEPARATION, FLASH GAS AND BOOSTER COMPRESSION
COMMUNICATIONS
[MASTER) CONTROL (STATION)
CRANES AND HOISTS
DEHYDRATION AND VENTILATION
DEW POINT UNITS, PROPANE REFRIGERATION, GAS COMPRESSION
DIGITAL CONTROL
ALARM, SHUTDOWN AND STARTUP FACILITIES
WASHDOWN AND FIREFIGHTING
FLARE
FUEL AND LUBRICATION
GAS/GASLIFT
GLYCOL REGENERATION
HOT OIL
HEATING, VENTILATION, AIR CONDITIONING
HYDRAULIC
METERING AND TELEMETRY
PIG & SPHERE LAUNCHING
POWER GENERATION
PRODUCTION WELLHEADS
OIL RECOVERY
RELIEF
SEAWATER COOLING
SEPARATION
OIL STORAGE
TURBINES
POTABLE WATER/ FRESH WATERCOOLING/ FRESHWATER SUPPLY

Unification Based Transfer

Björn Beskow
Uppsala University

Abstract

This paper concerns formal tools and methods for describing the transfer phase of a machine translation process. The unificational framework of the MulTra formalism is elaborated. The transfer relation representing a relation between translation units is formally defined. A control structure based on specificity is introduced, and shown to provide a way of treating some linguistic phenomena.

1 Introduction

1.1 The MulTra system

The MulTra system is being developed within the project "Multilingual Support for Translators and Writers". The project is sponsored by Digital Equipment Corporation and by NUTEK. It runs in tandem with a sister project at IPPI in Moscow.

The MulTra system is a support system for translating and writing. The system is fully interactive, where the translator or writer always is in command. The informational domain of the system is not just text, but consists of several different kinds of elements, ranging from ordinary text to layout and 3D graphics. The system thus operates in a *Compound Document Environment.*

One functionality of the system is a Machine Translation component. The translator or writer can use this component by selecting an element or a region of the document and querying the system for a translation.

The transfer approach to machine translation has been adopted. Analysis is performed by a chart parser producing feature structures (see [Sågvall-Hein 1987]). Synthesis is performed by a unification-based generator taking feature structures as input. This paper will focus on the transfer phase of the MulTra system.

1.2 Transfer Process

Techniques and tools for analysis within the field of Machine Translation (MT), and during the last years to some extent also generation, are reasonably well understood and developed. The transfer process, on the contrary, is still often described as a black box. Very few descriptions of the transfer component in an MT system has been given. If the transfer process is described at all, the description is almost always procedurally formulated, and sometimes even hardcoded. This leads to several unfortunate consequences. The *translation relations* which the transfer rules are supposed to describe, becomes hidden behind sequentially ordered transformations. The grammar writer is

forced to try to foresee complicated interaction between rules. The maintainability and extendibility of the grammar becomes poor.

During the last years, though, some efforts have been made to describe the transfer process within a unificational framework (see for example the ELU project in Geneva Estival *et.al.* 19901 or Karlsson *et.al.* in Helsinki [Carlsson & Vilkuna 1990]). This is also the approach of the MulTra system. For reasons of transparency, a transfer formalism should be declaratively formulated. The formalism should be transparent, in both an external and internal sense. The possible interaction between rules should be fully controlled by the formalism. Rules should be stated in a linguistically intuitive way.

The unificational paradigm has shown to be fruitful within analysis and generation, and has become a de facto standard in the research community. Formalisms within the paradigm is characterized by their declarativeness and transparency. The MulTra formalism is thus formulated in terms of unification and subsumption.

I will first describe the declarative formulation of the MulTra formalism. Discussing the procedural reading of the formalism will lead me to the issue of control. I will show how a control structure based on the notion of specificity can be defined within the unificational paradigm.

2 Translation Relations

The purpose of a transfer component is to capture the translation relations that may hold between units in two languages (see [Wikholm 1991] or [Ingo 1991]). These units can be of different type. A translation relation can hold between two *lexemes*,, like 'rock' in English and 'klippa' in Swedish. It can hold between two *structural(units*,, phrases (see [Wikholm 1989]), like for example a noun phrase consisting of a head and a prepositional phrase in English, and a phrase with the same structure in Swedish. It may also hold between combinations of lexical and structural units, like core phrases (see [Sågvall-Hein *et.al.* 1990] or [Östling 1991]), or cases where a lexeme in one language is related to a phrase in the other language.

These different cases have in MT traditionally been referred to as a distinction between *lexical* and *structural* transfer. But since the different cases are conceptually equal on the level of translation relations, the notion of translation relations implies that no sharp distinction should be made between structural and lexical transfer. You should be able to describe them with the same type of rules.

The basic building blocks in unification-based formalisms are *feature structures* (see [Shieber 1986]). A feature structure is an unorded set of feature value pairs. Feature structures provides a unifolm framework for representing translation units. A translation unit, be it a lexical or a complex unit, is in the MulTra formalism represented by a feature structure. A transfer relation is thus a relation between two feature structures.

3 Transfer Rules

The relations between translation units are defined by transfer rules. These rules describe the feature structures representing the translation unit being related, and places restrictions on these structures.

3.1 Syntax

Feature structures are usually described by a set of identity equations holding over the feature structure being described. An identity equation states what value should be for a specified feature. The special symbol '*' usually refers to the 'root' of the feature structure being described. In the MulTra formalism, an identity equation may contain variables, prefixed by a questionmark '?'.

These are not variables in the feature structure being described, but in the description itself. A variable thus denotes a substructure. A special variable is 'ANY' which can be viewed as the anonymous variable.

A transfer rule consists of four parts:

- A label
- A set of identity equations which describes the source feature structure.
- A similar set of identity equations describing the target feature structure
- A set of transfer equations relating variables mentioned in the source and target identity equations. A transfer relation states that the transfer relation holds between the subparts of the source and target structure denoted by the variables.

Figure 3-1 below is a simple example of a lexical rule. It describes the relation between two structures representing lexemes in Swedish and German. It says that a structure with the value 'montering' on the *lex* feature stands in the translation relation to a structure having the value 'montiren' on the *lex* feature.

It does not contain any transfer relations. It is thus an 'atomic' rule.

```
Label
     Montering
Source
     <* lex> = montering
Target
     <* lex> = montieren
Transfer
     {}
```

Figure 3-1 Lexical rule

Figure 3-2 below is a little more complex rule, describing the translation relation between two prepositional phrases. It relates a structure having the substructure denoted by ?prep1 as value on the *prep* feature, and the substructure denoted by ?rect1 as the value on the *rect* feature, with a structure that has ?prep2 as value on the *prep* feature and ?rect2 as the value of the *rect* feature. It also says that the substructures denoted by ?prep1 and ?prep2 stands in transfer relation to each other, and ditto for ?rect1 and ?rect2.

```
Label
     PP
Source
     <* prep>    = ?prep1
     <* rect>    = ?rect1
Target
     <* prep>    = ?prep2
     <* rect>    = ?rect2
Transfer
     ?prep1      <=> ?prep2
     ?rect1      <=> ?rect2
```

Figure 3-2 Structural rule

Figure 3-3 below is an example of a hybrid rule relating structures with both structural and lexical information. It describes the relation that holds between a structure representing an NP with the lexeme 'Whisky' as its head, modified by the PP 'on the rocks', and a structure representing an NP with the lexeme 'Whisky', modified by the PP 'med is'.

Since the transfer rules relates feature structures, not lexemes or trees, the grammar writer may take an arbitrarily large context into account.

```
Label
        whisky-on-the-rocks
Source
        <* head lex>          = 'whisky'
        <* prep lex>          = 'on'
        <* rect def>          = def
        <* rect det lex>      = 'the'
        <* rect head lex>     = 'rocks'
Target
        <* head lex>          = 'whisky'
        <* prep lex>          = 'med'
        <* rect head lex>     = 'is'
Transfer
        {}
```

Figure 3-3 Hybrid rule

3.2 Semantics

To this syntax, we have to add some semantics to say what it means that two feature structures stands in the transfer relation. Given two feature structures S and T, we start by defining the set of *applicable* rules wrt. S and T. We say that a rule is applicable iff its source part subsumes S. Let A be the set of applicable rules as defined in 3-1 below:

A = { r: source(r) subsumes S }

Definition 3-1 Applicable rules

The use of subsumption. in the definition will guarantee that the transfer process be sound. No feature can be added to S by any rule.

We then define the set of actually *applied* rules to be a subset of the set of applicable rules A. with the additional demands that

- the target part of the rules be unified with T and
- all the transfer equations in the rules holds.

Let a be the set of applied rules as defined in 3-2 below:

a= { r: source(r) subsumes S &
target(r) unifies with T &
∀ e in equations(r): e holds}

Definition 3-2 Applied rules

Note that the definition of applied rules relies on the transfer relation yet to be defined.

Now we shall define the *most general* feature structure that unifies with the source part of all the applied rules. Let C be a structure as defined in 3-3 below:

$\forall r \in \alpha$: C unifies with r.

Definition 3-3 Completeness

C will help us guarantee that the transfer process is complete, by acting as a 'record' of all applied rules.

We are now ready to formally define the transfer relation. We say that the two feature structures S and T stand in the transfer relations, if and only if there exists a C, such that the source structure S subsumes C:

S <=> T if and only if

$\exists$ C as in Definition 3-3 : S subsumes C.

Definition 3-4 Transfer Relation

The definition implies that, if S is not the empty structure, there exists a empty set of applied rules, and T is the unification of all target parts of the applied rules.

This is a recursive definition of the transfer relation, and the recursion is introduced above in the definition of applied rules. It is recursive over sub-parts of the structures S and T, mentioned in the rules. The recursion eventually ends with atomic rules, which is to say, rules that lack further transfer equations.

The transfer process, or the actual computing of a transfer relation, consists in creating a constructive proof that the transfer relation holds. The procedural reading of the transfer process follows readily:

- Given a source structure S and an underspecified target structure T:
- Compute the set A of applicable rules wrt. S.
- Apply the rules in A in parallel, by unifying the target parts with T and recursively compute the transfer equations, yielding the applied rules a.
- Compute the structure C by unifying the source part of all applied rules in a.
- Verify that S subsumes C.

From the definition follows completeness of the transfer relation. Since C is the unification of the source part, and S subsumes C, all features of the source structure have been considered in the transfer process.

It can also be seen from the definition that the transfer process will terminate, since the recursion introduced in the transfer equations relate substructures.

4 Control

The definition of the transfer relation implies that *(any subset* of the applicable rules which yields an appropriate C could be applied. To this general framework, we will add a control structure specifying which applicable rules to apply. We will base this control structure on the notion of *specificity.*

4.1 Specificity

Specificity is a general heuristics well understood and used in artificial intelligence, and elaborated by for example the ELU project in Geneva ([Estival *et.al.* 1990]). The basic idea is that a more specific rule should precede or block the application of a more general rule.

Specificity can be defined in terms of subsumption between the source parts of two rules. A rule A is more specific than a rule B if the source part of B subsumes the source part of A.

If the rules are equal by subsumption, a rule is more specific if it contains more transfer equations.

A rule A is more specific than a rule B if
- source(A) ≠ source(B) and
  source(B) subsumes source(A), or
- source(A) ≡ source(B) and
  | transfer(A) | > | transfer(B) |

Definition 4-1 Specificity

Specificity thus constitutes a partial order on the transfer rules. This partial order is independent of the source and target feature structures S and T.

We then have to modify the definition of the set of applied rules to reflect this principle of specificity. We constrain the set of applied rules further by saying that if a rule r is in the set of applied rules, no other rule s which is more specific than r is in the set. If a rule is applied, we guarantee that no other more specific rule also is applied. So, let α be the set of applied rules as defined in 4-2 below:

α = { r: source(r) subsumes S &
target(r) unifies with T &
∀ e in equations(r):
e holds &
¬∃s ∈ α : s is more specific than r}

Definition 4-2 Applied rules

This definition can be viewed strict, as actually prohibit the application of less specific rules. It may also be used to define a *preference ordering* on the transfer relations. A preference ordering is thus a partial ordering on the results of the transfer so that a more specific reading is preferred or precedes a more general interpretation.

4.2 Blocking

This control strategy can be used to handle for example blocking phenomena in an elegant way. Blocking occurs in cases where there is a specific translation which one wants to block a more general interpretation. For example, consider the translation relations in figure 4-1 below:

'Whisky on the rocks' <=>

'Whisky med is'

'Whisky' <=> 'Whisky'

'on' <=> 'på'

'rock' <=> 'klippa'

Figure 4-1 Translation relations

'Whisky on the rocks' we prefer to translate into Swedish as 'Whisky med is'. But 'on' in isolation can be translated as 'på', and 'rock' as 'klippa'.

The general interpretation of an English NP consisting of a noun and a PP is described by the rule shown in figure 4-2 below, which just relates the noun and the prepositional phrase by the transfer equations.

```
Label
     General-np
Source
     <* head> = ?head1
     <* prep> = ?prep1
     <* rect> = ?rect1
Target
     <* head> = ?head2
     <* prep> = ?prep2
     <* rect> = ?rect2
Transfer
     ?head1   <=> ?head2
     ?prep1   <=> ?prep2
     ?rect1   <=> ?rect2
```

Figure 4-2 General NP rule

The application of this general rule can is blocked by the rule previously described above in figure 3-3, since that rule is more specific than the general rule in figure 4-2. The source part of the rule in figure 3-3 describes the *lex* value of the head and the preposition, and the *rect*. Thus the source part of the general rule subsumes the source part of the rule in figure 3-3.

4.3 Lexicalized Phrases

As mentioned above, since the transfer relation is a relation between feature structures, the grammar writer can take an arbitrarily large context into account when folmulating a transfer rule. The notion of specificity in conjunction with this possibility of choosing context, also provides for a way of treating a similar class of notorious problems in machine translation, lexicalized phrases and idioms. Consider for example the English phrase 'kick the bucket'. It is a lexicalized phrase, which should be translated to 'dö' in Swedish. The main verb of the phrases may be inflected. This translation relation is described by the rule described in figure 4-3 below.

The rule relates an English verbal phrase whose head verb is 'kick' and whose object is 'the bucket', with a Swedish verb phrase whose head verb is 'dö'. The two phrases share value on the *tense* feature. This rule is more specific than a general reading of the verb phrase, and thus blocks the general reading.

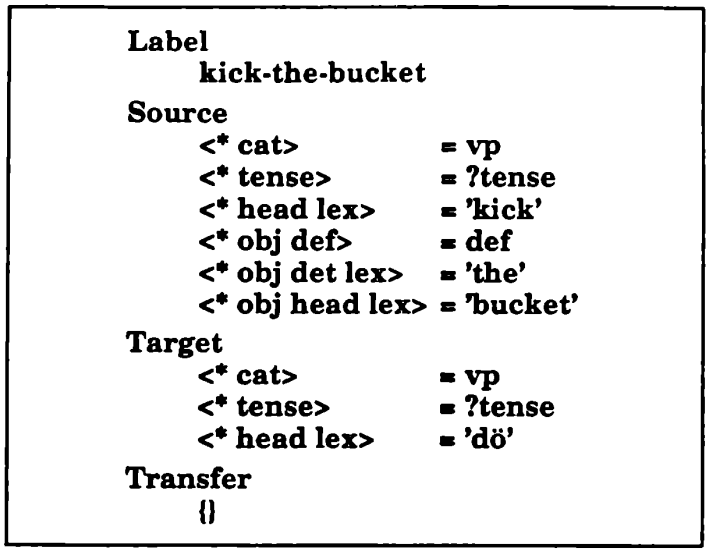


Figure 4-3 Kick the bucket

5 Conclusions

In this paper, the transfer formalism of the MulTra system has been described. Feature structures have been shown to provide a uniform way of representing translation units. A declarative formulation of the transfer relation have been stated in terms of unification and subsumption. The procedural reading of this formulation have been elaborated.

Further, the notion of specificity have been discussed, and a control structure based on this principle have been introduced. It has been shown how this control structure provides a way of treating some interesting linguistic phenomena. This raises further questions about on which level in the translation process these phenomena should be treated that is, in the analysis or in the transfer phase. Considerations like this remains to be investigated.

It may also turn out that the principle of specificity is a too powerful heuristics. But that would in any way lead to further insight into blocking phenomena, and might also reveal other interesting generalizations to be done in this field.

In conclusion, I have claimed that the unificational paradigm provides a very suitable framewolk for describing the transfer process, and shown how such a description can been made.

6 References

Carlsson, L. & Vilkuna, M. *Independent Transfer using Graph Unification* COLING 1990

Estival, D. *et.al.* A *Syntax and Semantics for Feature-Structure Transfer* in Third Conference on Theoretical and Methodological Issues of MT, 1990

Ingo, R. *Från källspråk till målspråk Introduktion i översättningsvetenskap "From* Source language to Target Language", Lund, 1991

Shieber, S. *An Introduction to Unification-based Approaches to Grammar.* Volume 4 of CSLI Lecture Notes. CSLI, Stanford, 1986

Sågvall-Hein, A. *Parsing by means of Uppsala Chart Processor*, in Bolc, L. (ed), Natural Language Parsing Systems, Berlin, 1987

Sågvall-Hein, A., Cederholm, Y ., van Stam Rydehell, M. & Wikholm, E. *Fraser i kärnlexikonet.* "Phrases in the Core Vocabulary", Uppsala University, 1990

Wikholm, E. *Sammansättning som översättningsenhet*, "Compound as Translation Unit", Gothenburg University, 1989

Wikholm, E. *Übersetzungstheorie und maschinelle Übersetzungen* "Translation Theory and Machine Translation", Scandinavian Conference of Computational Linguistics, Bergen, 1991

Östling, A. *A Swedish Core Vocabulary for Machine Translation* Scandinavian Conference of Computational Linguistics, Bergen, 1991

Björn Beskow
Uppsala University
Department of Linguistics
Box 513
S-751 20 Uppsala
E-mail: beskow@ling.gu.se

A Unification-based Grammar of Serial Verb Constructions

Adams B. Bodomo
University of Trondheim

0. Introduction

In many languages of West Africa and also in the African-Carribean Creoles there exists a unique and productive grammatical phenomenon involving an intricate interplay of a series of verbs and their arguments within the borders of what seems to be a monoclausal construction. Various names have been used to designate this phenomenon. Among them are serial verbs, verb serialisation, consecutive verbs, sequential verbs and serial verb constructions We, here, adopt the term serial verb constructions (hereafter, SVCs).

SVCs present a number of problems with regard to information categorisation and are therefore the subject of intense debate in current grammatical theories. The major issues are summarised in section 1.2. In section 1.3, I give a brief presentation of representative approaches currently being suggested within the various grammatical theories and formalisms.

My hypothesis in this paper lies within a computational linguistic framework where I regard SVCs as complex predicates derived from the conceptual or argument structure of two or more verbs, first by the formal process of unification and then by PS rules of the language in question. I take it that unification operates at all levels of the grammar, including the lexicon, the morphosyntactic level and the *gestalt* level (to be defined in section 2.4). In section 2.0, we briefly introduce unification, a concept popularly employed within computer science and linguistics as an information combining operation on feature structures. In subsequent sections we then begin to demonstrate the unification formalism with SVCs as they occur in Dagaare and other Voltaic languages at the various grammatical levels.

1.0 Serial Verb Constructions

1.1. Data: The structure of Voltaic SVCs

The following data (1 – 6) illustrate the structure of SVCs as they occur in some major Voltaic languages (Dagaare, Kusaal, Dagbane). Data from some Kwa languages (Akan, Ewe, Yoruba) are also added for comparative purposes.

(1) **Dagaare** (a Voltaic language in Ghana, Burkina Faso and Côte d'Ivoire):

(i) *Bayuo daaɛ Ayuo lɔɔ*
push+p.c Ayuo fall
Bayuo pushed Ayuo down

(ii) *Bayɔɔ ŋmɛ la Ayɔɔ kʊ*
beat p.c kill
Bayor beat Ayor to death

(iii) *Bayɔɔ na dʊgɛɛ nɛn kʊɔr*
FUT boil+p.c meat sell
Bayor will boil meat and sell it

(iv) *Bayɔɔ na dʊgɛɛ nɛn a kʊɔr* (non svc)
FUT boil+p.c meat and sell
Bayor will boil meat and then sell it

(v) * *Bayɔɔ na dʊgɛɛ nɛn na kʊɔr*
FUT boil+p.c meat FUT sell
Bayor will boil meat and sell it

(2) **Kusaal** (a Voltaic language in Ghana, Burkina Faso and Togo):

(i) *O bʊɔɪ nɛ kɔ̀ɔ̀m nu*
S/he pour+PAST p.c water drink+PAST
S/he poured water and drank it

(ii) **O bʊɔɪ nɛ kɔ̀ɔ̀m o nu*
S/he pour+PAST p.c water s/he drink+PAST
S/he poured water and drank it

(iii) **O bʊɔɪ nɛ kɔ̀ɔ̀m nu o*
S/he pour+PAST p.c water drink+PAST it
S/he poured water and drank it

(3) **Dagbane** (a Voltaic language in Ghana and Togo):

(i) *O zaŋ la kpargʊ yɪ*
s/he took p.c shirt wore
S/he took a shirt and wore it

(ii) * *O zaŋ la kpargʊ yɪ la*
s/he took p.c shirt wore p.c
S/he took a shirt and wore it

(iii) *Dzemi kpaa la kpam naŋ mwali nɪ*
poured p.c oil put river inside
Dzemi poured oil into the river

(iv) **Dzemi kpaa la kpam naŋ kpam mwali nɪ*
poured p.c oil put oil river inside
Dzemi poured oil into the river

(4) **Akan** (a Kwa language in Ghana):

(i) *Kofi tɔɔ nsuo nʊmʊɪ*
buy water drink
Kofi bought water and drank it

(ii) *Kofi tɔɔ nsuo nvmvi nv
buy water drink it
Kofi bought water and drank it

(5) **Ewe** (a Kwa language in Ghana and Togo and Benin):

(i) Kɔmi fo Ami dze anyi
beat fall ground
Kɔmi knocked Ami down

(ii) Kofi fo Ama wui
beat kill+pron
Kofi beat Ama to death

(iii) ?Kɔmi fo Ama wu
beat kill
Kofi beat Ama to death

(iv) * Kofi fo Ama wu Ama
beat kill+pron
Kofi beat Ama to death

(6) **Yoruba** (a Kwa language in Togo, Benin and Nigeria):

(i) Olu gbe aso wo
took dress wear
Olu put on some clothes

These characteristics are captured by the following well-formedness conditions in (7) below:

(7)
i. All the verbs in an SVC must share a single structural subject.
ii. In an SVC there is only a single tense node
iii. SVCs have a single polarity clitic. (p.c)
iv. There is an absence of connectors or complementizers within the string of verbs.
v. Dydadic verbs must share internal arguments.

The well-formedness condition in (7i) will account for the ungrammaticality of (2ii) since there is an undesirable copy of the subject pronoun 's/he'. Similarly, the extra occurences of the future tense marker and the polarity clitic in (1v) and (3ii) respectively violate conditions (7ii) and (7iii), thereby rendering them ungrammatical. The data in (1iv) illustrate (7iv) which is actually a semantic well-formedness condition because, eventhough the construction is syntactically correct, it does not have the normal semantic reading of SVCs where the actions are intuitively more tightly related. Condition (7v) is very important in SVC constructions in Voltaic. It accounts for the fact that V2s and subsequent verbs in the SVC don't need internal arguments whether in the pronominal forms (2iii) or as copies (3iv). Notice, however, that there seems to be a difference between Voltaic and Kwa serialisation with respect to the extra occurrence of internal arguments in their pronominal forms. This difference explains why (5ii) is grammatical and (2iii) is not.

1.2 The issues

The above characteristics of SVCs pose a number of problems for the theory of grammar and the major issues being discussed include the following:

1. What are the syntactic and semantic processes involved in these complex predicates that are SVCs and at what level of the grammar do they occur?
2. How can syntactic and semantic information be categorised in these constructions?
3. Do these constructions express a single event or a series of events and, if so, is verb order in these constructions crucial to the understanding of these phenomena?
4. And why do certain languages serialise at all while others don't? In other words, can we establish a serialisation parameter?

Issue no.1 might, for instance, involve outlining the syntactic processes that are able to bring so many verbs together in one clause: but for memory failure, an unlimited number of verbs can occur in one clause. Would these be cases of complementation, coordination, adjunction or some other processes? And if so at what level of the grammar are these likely to occur: the lexicon, the d_structure, the s_structure ?

Issue no.2 is related to no.1 but, in addition, it might involve accounting for how the information for each verb and its arguments are distributed in the whole complex. What, for example, is it that enables a verb to share some of its arguments with others? This is what we are directly concerned with here and although the other issues are interesting in their own right it might not be possible to suggest solutions for them in this paper. In the next section we show how some earlier analysis have attempted to solve the problems.

1.3 Possible solutions

As a reaction to the above problems many solutions have been proposed within various grammatical theories and formalisms. There are no tight compartments between these approaches but it is possible to group them into what I will call the lexical-conceptual, the syntactico-semantic and the cognitive approaches.

1.3.1 The lexical - conceptual approach

This approach to the analysis of SVCs is currently pursued by researchers such as Déchaine (1987, 1988) and Lefebvre (1986, 1987, 1991), the main idea being that SVCs 'are derived complex predicates which are formed prior to D-structure by means of operations on the lexical conceptual structure (LCS) of verbs.'

The representative analysis from this group is Lefebvre (1991). With data from causative SVCs (what she terms 'take' serial verb constructions) in the Fon language, Lefebvre claims that the process of serialisation originates from the lexicon. The resulting complex predicate is then projected onto the syntax as a bi-headed VP. The example sentence she uses is shown below:

(8)

Kɔ̀kú	*sɔ́*	*àsɔ̃́*	*yì̀wá*	*àxì*
Koku	take	crab	go/come	market

Koku brought (direction away/ towards speaker) the crab to the market.

The LCS of (8) will then be represented as follows in (9):

(9)

a. *sɔ́*: [x cause [y undergo change of location]]

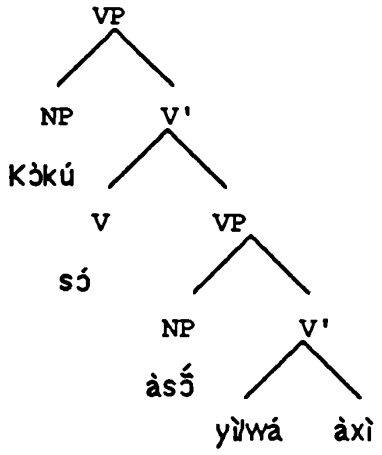
b. *yì̀wá*: [y undergo change of location
away from/towards speaker to location z]

c. *sɔ́-yì̀wá*: [x cause [y undergo change of location
away from/towards speaker to location z]]

The verbs *sɔ́* and *yì̀wá* will receive the LCS as shown in (9a) and (9b) respectively. To Lefebvre, a certain process (which she fails to mention) 'conflates' the LCS to form the complex

lexical predicate in (9c). This complex predicate then projects into the syntax as a bi-headed VP in an essentially complementation configuration (as shown in (10))

(10)



by the following X-bar theory (11)

(11)

a. XP → SpecX' X'

b. X' → X YP

As Larson (1991) points out, such a theory allows heads to have at most a single complement from a maximal projection, forcing a binary branching structure in which the two heads of the complex predicate are inserted into two available V positions.

1.3.2 The syntactico-semantic approach

While the foregoing group of researchers consider SVCs to be a product of the lexicon, others such such as Baker (1989, 1991) Ayewole (1988) and Hale (1991) consider it to be a post-lexical phenomenon, taking place at the level of syntax (and possibly, other post-lexical levels). Baker (1989) is the representative analysis for this group.

The structural characteristics of the SVC that we saw earlier in (7) (section 1.2) threaten the entirety of the theta-theory, especially the Projection Principle (PP) and the theta criterion. Naturally, therefore, most analyses within the GB framework are concerned with analysing SVCs in the light of the problems posed by the threat to theta theory. One recent analysis in this direction is Baker (1989). Most of the SVCs we have seen so far seem to violate the Projection Principle as stated by Chomsky (1981) as in (13) below:

(12) *Kofi naki Amba kiri* (taken from Baker (ibid)).

hit kill

Kofi struck Amba dead

(13) The projection principle

Suppose α is a lexical category and β is a position of argument type.

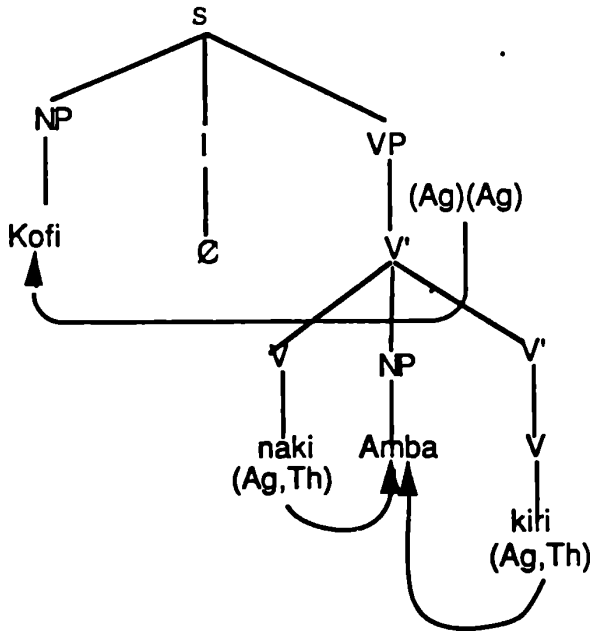
a. If β is an immediate constituent of a one-bar level projection of α at some syntactic level, then α θ-marks β in α'

b. If α θ-marks β as a lexical property, then α θ-marks β at all syntactic levels

All the V2s e.g *lɔɔ, yi, nʊmʊi, kiri, wo* in the above examples are transitive verbs but there is no argument following them in the surface structure. The object seems to be 'deleted' at this level by identity with the object of V1, from the point of view of old-fashioned transformational grammar. This means then that the complementation properties of such verbs are not represented at all the levels of the syntax of SVCs, thereby violating the PP as stated above.

To solve this problem, Baker proposes that double-headed VPs be allowed in serialising languages, thereby making it possible for both verbs to 'share' the single object NPs in each of the constructions. Figure (14) below (also taken from Baker (ibid)) illustrates the principles for this approach with the Sranan sentence in (12). According to him, to get a double-headed VP one of the verbs must project immediately to the V' level. The VP and higher V' would then be projections of both verbs, 'naki' and 'kiri'. The arrows indicate theta-role assignments. According to the standard conditions on theta-role assignment (stated in Chomsky (1986)) 'naki' directly theta-marks 'Amba' while 'kiri' indirectly or predicationally theta-marks 'Amba'. Quoting Williams (1984), Baker again claims that the two verbs can theta-mark 'Kofi' by the fact that the external (agent) theta-roles of the verbs percolate to their maximal projections, which is the VP, thereby being assigned to the subject. In this way, according to Baker, the lexical theta-role-assigning properties of both verbs are satisfied in this structure, and the PP would then be obeyed.

(14)



The conclusions behind such an approach are that X' theory is extended in such a way that serialising languages allow V's to embed within V' to form a double-headed construction. Finally, he outlines the kind of consequences that these conclusions may have on what kind of verbs may combine in an SVC, their linear order and the structural positions of their argument NPs as follows in (15):

(15)
a. Each verb may or may not θ-mark the subject of the whole serial VP.
b. For each other argument α in the SVC:
   i. α must be θ-marked by all the verbs that follow it.
   ii. α must be θ-marked by the verb that immediately precedes it.
   iii. α is not θ-marked by any verb that precedes it other than as in (ii).
c. For each verb in the SVC, the arguments of that verb must appear in the following hierarchical order: Agent>Instrument>Patient/Theme>Goal/Location.

Baker's proposals certainly constitute an important contribution to the discussion on SVCs as long as we limit ourselves to the kind of V NP V structures that he mainly employs as the motivation for his analysis. A closer look at longer strings would reveal a number of objections to some of Baker's conclusions. As an illustration, consider (16) visa vis (15bi) repeated below:

(15bi). α must be θ-marked by all the verbs that follow it.

(16) *Zo gaa wuo haani wa ku ma*
run go collect blackberries come give me
"Go and collect some blackberries for me"

While the verb 'give' would θ-mark blackberries it cannot be said that 'come' does the same with blackberries, thereby violating (bi). In other words (16) is a counter example to Baker's characterisation of SVCs. Further still, this whole idea of abstracting double-headedness in serialising languages might introduce more problems than solutions, for it is difficult to see how three, four or five verbs may have their argument structure crunched together without getting in to conflict with some GB- theoretic issues like the θ-criterion and certain principles of word order. For example, as Awoyale (1988: p.6) points out, it might not be possible on (GB?) theoretical grounds for one verb to form part of the argument structure of another verb. As another example to one of the above reservations, the theta criterion, at least, in its classical sense would not be obeyed in Baker's proposed analysis, since in effect each of the two NPs in the configuration receives two theta roles.

However, some of these objections would only seem to reflect GB internal problems created by Baker's 'double-headed VP' approach. The contribution offers us a lot of good premises to build upon.

1.3.3 The cognitive approach

While the approaches discussed above usually concern themselves with mainly grammatical issues, the cognitive approaches would concern themselves with the *relationships* that exist between grammatical categorisation and mental categorisation i.e the 'grammatical packaging' vis-a-vis the 'cognitive packaging'. They are therefore mostly interested in issues like the clausehood and eventhood of SVCs. An example of such approaches is Givón (1991).

This study is most interested in the sense in which the series of verbs in an SVC jointly form a single event. While grammatical approaches would use structural diagnoses like the distribution of inflection and agreement, Givón employs a different method involving an elicitation of serial and non-serial constructions in discourse to see the way in which native speakers of serial and non-serial languages structure their information. The underlying principle for this investigation is the Distance Principle, an iconicity principle which relates grammatical organisation to conceptual organisation. This is stated below in (17):

(17) The temporal-physical distance between chunks of linguistically-coded information correlates directly with the conceptual distance between them.

One interpretation of this principle is that pause separations dividing finite clauses (single event domains) in non-serial languages should be comparable to those separating verb sequences in serialising languages if SVCs really define a single event.

Givón's methodolgy involved presenting speakers of serialising and non-serialising languages with a short movie which they were asked to describe orally later. Pause measurements were taken and the probabilities that were computed showed that there were no significant differences between pauses separating finite clauses in non-serialising languages and those separating SVCs in serialising languages.

2.0 Unification grammar

In this section of the study we propose our own solutions to some of the problems posed by attempts to represent the structure of SVCs. We do this from a grammar formalism that has been variously referred to as *unification grammar, unification-based grammars, informational grammars* or *information-based grammars*. A representation of grammar from this viewpoint is in consonance with the structure of SVCs outlined in (7) as highly information-sharing, the argument being that this is possible because of a unification operation which can occur at any level of the grammar.

Before embarking on the formal representation of SVCs from this information processing perspective at our three proposed levels, we present a brief characterisation of unification grammar in section 2.1. For the purpose of achieving a concise analysis we do not envisage giving a full expository account of unification and unification grammar. An adequate number of introductory references exist in the literature for the purpose e.g Shieber (1985), Carlson and Linden (1987) and Uszkoreit (1990).

2.1 The Concept

The general idea behind unification is that it is a computational (i.e. a formal) operation that merges the information of two or more feature structures if this information is consistent. In other words, it is an operation which enables two feature structures to share the same consistent information. Uszkoreit (1990) gives a much more formal definition of the concept of unification as follows in (18) and (19):

(18) A type t_0 is the unification of two types t_1 and t_2, iff t_0 is the least informative type that is subsumed by both t_1 and t_2.

or

(19) A type t_0 is the unification of two types t_1 and t_2, iff t_0 is subsumed by both t_1 and t_2 and t_0 subsumes all other types t_i that are also subsumed by t_1 and t_2.

Below in (20) is an example of unification (from Shieber (ibid)) as an operation which combines the information from two feature structures to obtain a feature structure that includes all the information of both.

(20) [cat: np]
unifies with [agreement:[number:singular]]

$$= \begin{bmatrix} \text{cat: np} \\ \text{agreement: } \begin{bmatrix} \text{number:singular} \end{bmatrix} \end{bmatrix}$$

In (21) we have an example of an attempt at unification which fails because of inconsistent information in both feature structures:

(21) [agr[per:3]] unifies [agr[per:2]] = failure

Unification is often said to be *commutative, associative* and *indempotent.*

Any grammar formalism a part of which contains unification as described above would then be termed a unification grammar or a unification-based grammar formalism. Such a formalism does not necessarily have to conform to any particular grammar theory; it could also be built just as a tool for formulating and testing theories of natural language. In this sense we would say there are basically two groups of unification-based grammar formalisms. Lexical Functional Grammar (LFG) and Generalised Phrase Structure Grammar (GPSG) are those recognised in the literature to be based on grammatical theories while the second consists of Functional Unification Grammar (FUG) and PATR which were developed as tools for evaluating grammatical theories.

What is, however, more important is that these formalisms, irrespective of whether they were developed as tools or as grammatical theories in their own right, are built on the same principles. All the current formalisms are characterised by a combination of a unification framework for processing grammatical information within complex feature structures and a context free phrase structure part.

2.2 The lexicon: LCS unification

In Lefebvre's diagrammatical representation (figure repeated below) of how complex predicates are formed in the lexicon she fails to mention the process which she says 'conflates' the predicates of the various verbs. Developing that idea further, we claim that that process is indeed a unification operation: the LCSs of the various verbs can be represented as feature structures and merged together as long as the information in both structure do not conflict. We illustrate this with the following Dagaare sentence (22) which has an almost equivalent translation as Lefebvre's example sentence (8) from Fon.

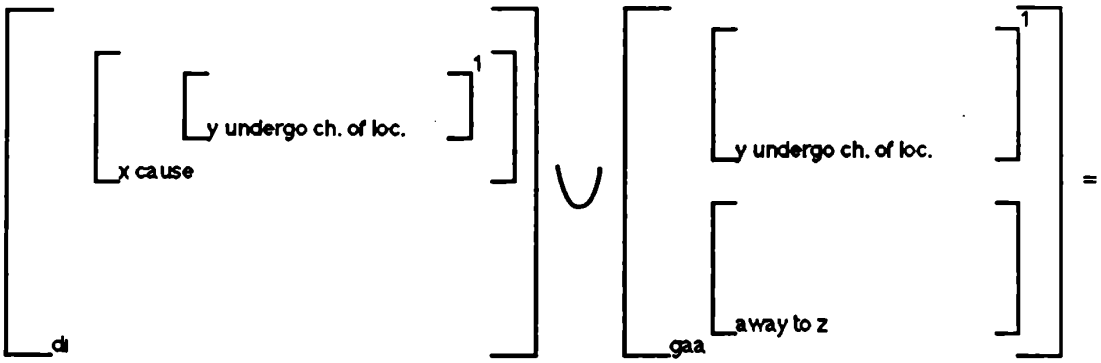
(22) *Badɛr dɪ la kyɪ gaa daa*
take p.c millet go market
Bader took some millet to the market

The verbs 'take', 'go' and their compound will have an identical LCS as in (9), repeated below:

(9)
a. *sɔ́:* [x cause [y undergo change of location]]
b. *yɪ̀ɲwá:* [y undergo change of location
away from/towards speaker to location z]
c. *sɔ́-yɪ̀ɲwá:* [x cause [y undergo change of location
away from/towards speaker to location z]]

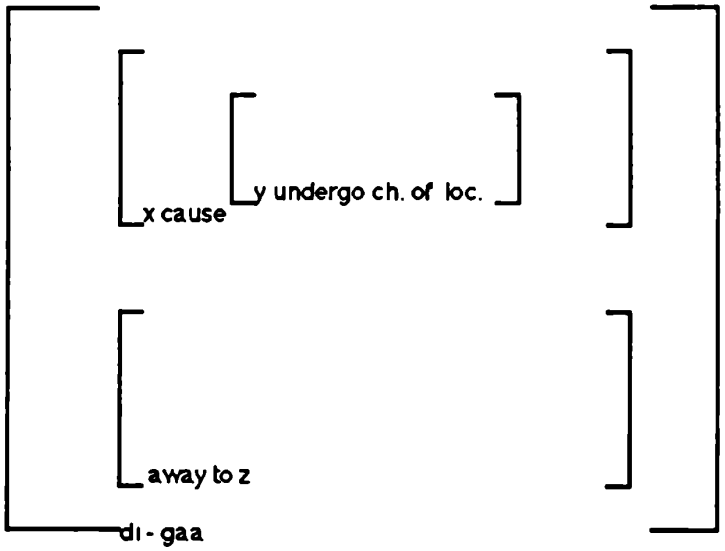
The reanalysis of this representation in a unification formalism is shown below in (23). The LCSs such as (9) are seen in terms of feature structures containing bundles of conceptual information. The conceptual feature structure of 'take' is then merged with the conceptual feature structure of 'go' through a process of unification marked by the 'U' in (23a).

(23a)



The resulting complex LCS is represented in (23b) as a feature structure which is more complex that the previous two. As will be noticed parts of the structure with similar information are marked with similar numbers (here 1). It is these parts which are merged together because they contain *consistent* information. This is the sense in which we say unification can take place also inside the lexicon.

(23b)



In the next section, we see how the unification operation would combine consistent syntactic information about arguments of two predicates together to form complex syntactic feature structures. We will also see that the result of this conceptual unification shows up in the syntax, first, as discontinuous predicates in the c-structure, and then as predchains in the f_structure.

2.3 The syntactic level: LFG unification

In section 1.3.2, we saw that Baker's analysis of SVCs is essentially at the syntactic level where he regards the two verbs in the series as sharing the internal argument. We take a much more radical approach and claim that not just only the internal argument but other units like the external

argument, tense and polarity clitics are also shared. This we claim is possible through the unification operation being discussed.

We adopt essentially Baker's approach whereby V's can be embeded within V's to form double-headed constructions. By this extension, we are able to account for the recursive nature of VPs in SVCs. (24) can be regarded as part of the PS rules of an SVO serialising language like Dagaare:

(24) S → NP, AUX, VP.
VP → V'.
V' → V, NP, V'.
V' → V.
V' → V, NP.
V' → V, PP.
NP → DET, N.
AUX → TENSE.

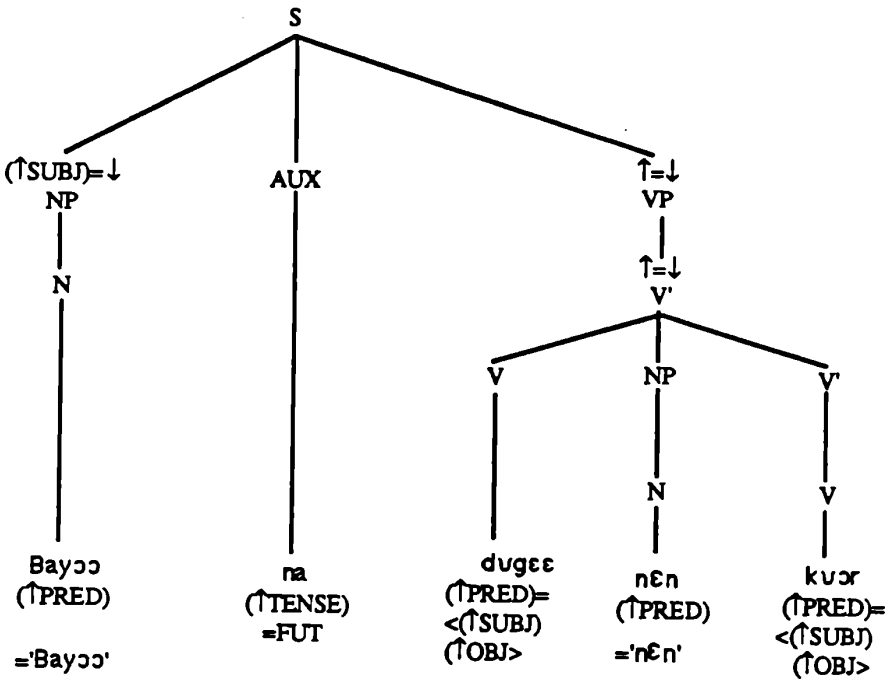
These rules can then generate the data (1iii), repeated below as (25):

(25) *Bayɔɔ na dugɛɛ nɛn kuɔr*
FUT boil+p.c meat sell
Bayor will boil meat and sell

This structure can be represented in LFG style c-structure and f-structure as shown in the diagrams below in (26a) and (26b) respectively. To indicate how unification operates within the c-structure, we use the metavariable notation of LFG in which up and down arrows are used on the NP and VP nodes, which are regarded as feature structures, to show the direction in which grammatical information flows from one feature structure to another. In the diagram below, then, the up and down arrows on the NP and VP nodes are read respectively as "ups subject is down" and "up is down". The "up" refers to the feature structure of the mother node, which in this case is S and the "down" refers to the feature structure of the node itself. So this would mean that, with respect to (↑SUBJ)=↓ all the functional information carried by the NP in (26) goes into the subject part of the mother's (i.e S's) feature structure while with respect to ↑=↓ all the functional information carried by the VP (i.e the VP's feature structure) is also direct information about the mother's feature structure. This is exactly the notion of unification we have been alluding to so far: the information of the NP is *unified* with that of S and the information of the VP is also *unified* with that of S. The = in each of these representations expresses this idea of unification much clearer: the information of the 'daughter' node indicated by ↓ is identified or unified with the information of the 'mother' node, indicated by ↑. The end result is that both mother and daughter *share* the same information. Put in a simpler way, NP is the subject of S and VP the functional head of S.

From the above explanation we may then say that the two verbs *dugɛ* and *kuɔr* share the same information, not just only about the internal argument *nɛn* but also about the future tense *na* and the polarity clitic -*e*[1]

(26a)



(26b)

SUBJ	[PRED 'Bayɔɔ']		
OBJ	[PRED 'nɛn']		
AUX	TENSE	FUT	[PRED 'na']
PREDCHAIN	PRED1	'dugɛɛ <(↑SUBJ)(↑OBJ)>'	
REST	PREDCHAIN	PRED2	'kʊɔr <(↑SUBJ)(↑OBJ)>'

As was hinted at the end of section 2.3, the predchain seen in (26b) is the result of the conceptual unification described in that section. Now, the unification at this syntactico-semantic level of the grammar involves the arguments of this complex predicate. In the next section we will see that the conceptual unification again creates a complex predicate or a 'sharing gestalt' whose 'points' are then unified due to the unification operating at that level too.

2.4 The gestalt level: Gestalt unification

The field of linguistic semantics is in a continuous and rapid state of development. Quite apart from more established theories like Montague's (logical) model theoretic semantics, Longacker's theory of cognitive grammar, Lakoff's cognitive semantics and Jackendoff's conceptual semantics, among many others, newer models are coming up quite often.

One of such theories is the Gestalt theory currently being developed at the University of Trondheim, Norway (cf Hellan and Dimitrova-Vulchanova (1991, forthcoming)). One of the main aims of the theory is to pay much closer attention to the relationship between syntactic and conceptual levels than other theories do. By the term gestalt of a sentence is meant among a couple of usages, as the idealised model of reality that any 'piece' of reality has to match in the relevant respects in order to be realised as the interpretation of that sentence. From this we realise that there is no one-to-one relationship between gestalts and the situations-in-the-world they model: there may be several ways of modeling a reality as shown below in (27):

(27)

1. He knocked John on the head
2. He knocked the head of John

The sentences 1 and 2 may be said to be two gestalts representing what is probably one situation-in-the-world. Different languages have different ways of modeling these gestalts and it is one of the aims of the theory to show how this is done and also the number of gestalts in languages of the world (gestalt language universals?) One of such attempts at extracting language universals is the establishing of the completedness parameter, which is one of the cardinal notions of the theory (Dimitrova-Vulchanova and Hellan (ibid)). According to this parameter, languages of the world model gestalts, which are essentially conceptual notions, on the basis of this parameter and they have mechanisms for adjusting constructions nearer or away from this parameter in both dimensions. As a result, we can have completed and uncompleted gestalts or better still more or less complete gestalts.

Like the Jackendofian model, this model also has three modules: the conceptual module/level, the gestalt module/level and the morphosyntactic module/level. However, in this work we will be concerned with just one level - the gestalt level, where we hope to show how the unification model we have been developing may be used to represent structures at this level.

At this level the basic unit is the gestalt, whose superordinate form is the super gestalt. A super gestalt divides into a root gestalt (RG) and a dependent gestalt (DP). The topology of the RG is made up of relations and points, the points ,in turn, dividing into centre point and limitation point. But we also have a predicate. The sentences 1 - 4 below in (28) illustrate this basic topology:

(28)

1. He walked :	Centrepoint - relation
2. He is a teacher :	Centrepoint - relation - predicate
3. He painted the house :	Centrepoint - relation - limitation point
4. He painted the house red:	Ctre.pt. - relation - limitn.pt. - pred.

There are many types of gestalts at this level but we will be most interested in a special type called *sharing gestalts*. These are instances of non-iconic gestalts in which different relations would share points. Here the world's languages divide into two main types of gestalts: dependent gestalts and serial gestalts and the claim, to be elaborated in subsequent work (Bodomo (forthcoming)), is that in the expression of sharing gestalts in causative constructions of the format (29):

(29) S → NP + VP [NP XP]

the XP of non-serialising languages like Norwegian, English and French is essentially -V while that of serialising languages like Dagaare, Ewe and Yoruba is essentially +V (serialisation parameter for gestalt formation). The two cases are exemplified by the following Norwegian and Dagaare sentences (30).

(30a)

i. *Sigurd slo Sigrid ned*
hit+PAST down
Sigurd knocked Sigrid down

ii. *Gøran sparket ballen flat*
kick+PAST ball+DEF flat
Gøran kicked the ball flat

iii. *Ingebjørg spiste kjøleskapet tomt*
eat+PAST fridge+DEF empty
Ingebjørg ate up (the contents of) the refrigerator

(30b)

i. *Bayuo ŋmɛ la Ayuo lɔɔ*
beat p.c down
Bayuo knocked Ayuo down

ii. *Dɔɔdaa ŋmɛ la a bɔɔl pur*
beat p.c DEF ball explode
Dordaa kicked the ball flat

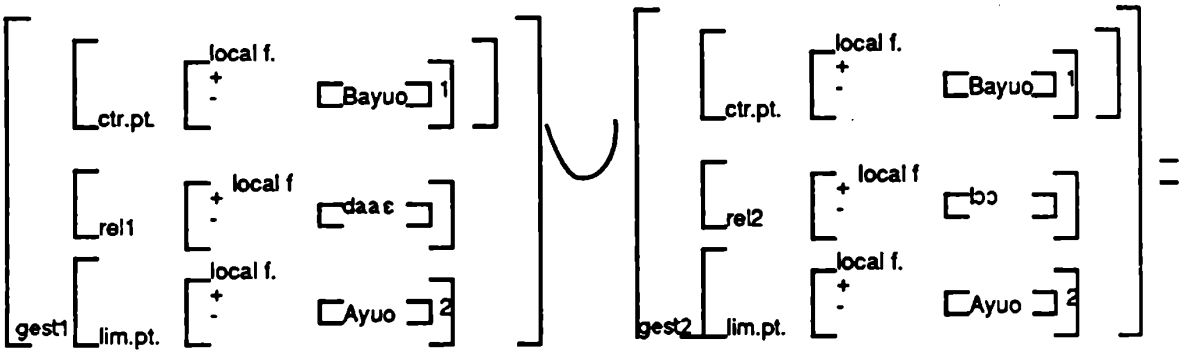
iii. *Boŋlakyɛr dɩ la a frigyi baar*
eat p.c DEF fridge finish
Bonlacher ate up (the contents of) the refrigerator

Having said this, we proceed to analyse how an SVC can be represented within the framework of what we call a gestalt unification grammar (GUG). The representative sentence here is (1i) repeated below:

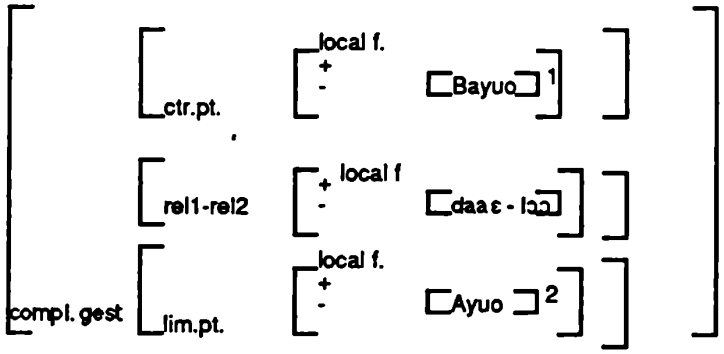
(1i) *Bayuo daaɛ Ayuo lɔɔ*
push+p.c Ayuo fall
Bayuo pushed Ayuo down

The verb 'push' together with its arguments expresses gest1 while 'fall' with its arguments expresses gest2. In what way then or by what mechanism do we say that the two gestalts 'share' points? Again, as was done in previous sections, we reinterpret the gestalt topology as a feature structure of the attribute value type. This situation is depicted in (31a) where structures which contain similar information about points are marked with thesame number (here 1 and 2). A unification operation, symbolised by 'U' runs over these structures, merging their information together.

(31a)



(31b)



The result of this information is (31b) above. Briefly, we say that the various local features (attributes of the component parts of the gestalt), especially those of the relations, would then combine together to specify the global features (attributes of the whole complex). As a result of the merging together of the points through the unification operation at this level, we say that the various gestalts (which have now become a complex gestalt) share points. This is how we may arrive at a complex or a 'sharing gestalt' within a unification grammar framework.

3.0 Conclusion

In conclusion, the main concerns of this paper have been to offer an alternative approach to the representation of SVCs from a computational linguistic perspective. After a brief discussion of the characteristics of SVCs within major Voltaic languages of West Africa we gave a brief exposition of some earlier treatments of SVCs. The rest of the paper then consisted of an account of SVCs using a unification grammar framework.

The unification operation, as has been shown, is a powerful tool which can be used to analyse linguistic data independent of theory. In this paper it has been shown to bring together analyses from diverse theories, including lexicalist, syntactico-semanticist and conceptualist ones. From this perspective we claim the unification grammar approach to be a clean, unifying way of representing SVCs.

The paper, which is mainly a demonstration of how information is processed between verbs and their arguments within an SVC, has, however, not accounted for many other issues concerning SVCs such as eventhood, clausehood, verb order and parametrisation and should constitute topics for future research.

Note

1 In this and many of the feature structures in this paper it has been necessary to do some representational underspecifications in order archieve simpler feature structures and thereby archieve useful generalisations. Here, it would have been necessary, for instance, to account for how the polarity clitic -ɛ, like the future tense particle, is also shared by the two verbs in the construction.

References

Awoyale, Y. (1988) 'Complex Predicates and Verb Serialisation,' ms., University of Ilorin and MIT, Cambridge, Massachusetts.

Baker, Mark (1989): 'Object Sharing and Projection in Serial Verb Constructions,' *Linguistic Inquiry* 20, 513-553.

Bodomo, A.B. (1990) 'A Unification Grammar of Serial Verb Constructions' term paper, Department of Linguistics, University of Trondheim.

Bodomo, A.B. (forthcoming) 'Sharing gestalts among Scandinavian and West African: A case of parametric variation?

Carlson, L. & Linden, K. (1987) 'Unification as a grammatical tool,' *Nordic Journal of Linguistics* 10, 111-136.

Chomsky, Noam (1981): *Lectures on Government and Binding*, Foris, Dordrecht.

Chomsky, Noam (1986): *Barriers*, MIT Press, Cambridge, Massachusetts.

Déchaine, R.-M. (1986) *Opérations sur les structures d'argument: Le cas des constructions sérielles en haitien*, Master's thesis, Université du Québec à Montréal.

Dimitrova-Vulchanova, M. and Hellan, L. (1991) 'Clitics and the Completedness Parameter.' In Nordic Journal of Linguistics, 14, 1-39.

Givon, T. (1991) 'Some Substantive Issues Concerning Verb Serialisation: Grammatical vs Cognitive Packaging.' In *Serial Verbs: Grammatical, Comparative and Cognitive Approaches.* Studies in The Sciences of Language Series 8. ed. by C. Lefebvre, Amsterdam/Philadelphia: John Benjamins.

Hale, K. (1991) 'Misumalpan Verb Sequencing Constructions' In *Serial Verbs: Grammatical, Comparative and Cognitive Approaches.* Studies in The Sciences of Language Series 8. ed. by C. Lefebvre, Amsterdan/Philadelphia: John Benjamins.

Hellan, L. and Dimitrova-Vulchanova, M. (forthcoming): Propositional Gestalts and Grammar

Kaplan, R. & Bresnan, J. (1982) Lexical-Functional Grammar: A formal System for Grammatical Representation. In Bresnan, J. (ed), *The Mental Representation of Grammatical Relations.* Cambridge, Mass.: MIT Press, 173-281.

Larson, R.K. (1991) 'Some Issues in Verb Serialisation' In *Serial Verbs: Grammatical, Comparative and Cognitive Approaches.* Studies in The Sciences of Language Series 8. ed. by C. Lefebvre, Amsterdam/ Philadelphia.

Lefebvre, C. (1991) 'Take Serial Verb Constructions: Please,' In *Serial Verbs: Grammatical, Comparative and Cognitive Approaches.* Studies in The Sciences of Language Series 8. ed. by C. Lefebvre, Amsterdan/Philadelphia: John Benjamins.

Lefebvre, Claire ed. (1991) *Serial Verbs: Grammatical, Comparative and Cognitive Approaches.* Studies in The Sciences of Language Series 8. Amsterdam/Philadelphia.

Shieber, S. (1985). An Introduction to Unification-Based Approaches to Grammar. Paper presented as a Tutorial Session at the 23rd Annual Meeting of the Association for Computational Linguistics. Chicago, Illinois: University of Chicago

Uszkoreit, H. (1990) 'Unification in Linguistics' lecture notes SSLLI, Leuven, Belgium, 1990.

Williams, E. (1984) 'Grammatical Relations', *Linguistic Inquiry* 15, 639-673

Adam B. Bodomo
Moholt Alle 34-12
7035 Trondheim
E-mail: adams.bodomo@avh.unit.no

Maskinoversættelse af tyske NPer

Ellen Christoffersen og Margrethe H. Møller
Handelshøjskole Syd, Kolding

Ved Institut for Erhvervsforskning i Kolding har en projektgruppe fra Handelshøjskole Syd siden sommeren 1989 arbejdet med udviklingen af det tysk-danske modul til maskinoversættelsessystemet METAL.

METAL er et af de "gamle" maskinoversættelsessystemer i den forstand, at man har arbejdet på udviklingen af det i 20-30 år. Først kun i USA, ved University of Texas, og siden midten af firserne også i Europa, idet den tyske Siemens-koncern købte systemet. I dag er der METAL-arbejdsgrupper i USA, Tyskland, Belgien, Spanien og Danmark. Siemens sælger systemet med sprogparrene tysk-engelsk, engelsk-tysk og tysk-spansk, og andre sprogpar er under udvikling. I den danske arbejdsgruppe udvikler vi moduler med dansk, i første omgang det tysk-danske genereringsmodul.

METAL arbejder efter transfermodellen:

Den danske generering bygger på et analyseresultat, der dels er fremkommet vha. morfologisk og syntaktisk analyse af den tyske tekst, dels omfatter de semantiske oplysninger, de enkelte leksemer er forsynet med i den tyske ordbog.

METAL skal først og fremmest oversætte fagsproglige tekster. I det lingvistiske arbejde har den danske arbejdsgruppe derfor taget udgangspunkt i fagsproglige tekster. Vi ønskede at begynde vort arbejde med enkle tekniske tekster som f.eks. styklister og indholdsfortegnelser til brugervejledninger, og derfor var det også naturligt for os i første projektfase at begynde med at oversætte NPer.

Da vi kunne oversætte indholdsfortegnelser, satte vi en foreløbig slutstreg under NP-genereringen og gik over til at oversætte sætninger.

I dette indlæg vil vi beskæftige os med nogle principielle problemer i maskinoversættelse, eksemplificeret ved to problemer, vi ofte er stødt på i vore fagsproglige NPer. Det drejer sig dels om et *leksikaliseringsproblem*: Hvilke oplysninger skal være tilgængelige i ordbøgerne og hvilke oplysninger kan man beregne sig til? – og dels om et *strukturproblem*: En struktur på udgangssproget kræver somme tider en strukturændring på målsproget, og i andre tilfælde kan den gengives ved samme struktur på målsproget. Her må man søge at finde lingvistiske variabler og værdier, der både kan beregnes og også er tilstrækkelige til at træffe valget mellem de to muligheder.

Et strukturproblem: Tyske genitivkonstruktioner

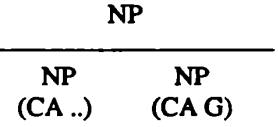
(Ellen Christoffersen)

Først vil jeg gøre rede for det problem, der opstår, når en bestemt struktur i udgangssproget modsvares af flere forskellige strukturer i målsproget. Et sådant problem møder vi, når der i den tyske tekst optræder et nominalsyntagme som attribut til et andet nominalsyntagme. Det attributive nominalsyntagme står i genitiv, mens det overordnede nominalsyntagmes kasus er bestemt af dets syntaktiske status i sætningen.

Tysk struktur

Det tyske analyseresultat af denne struktur ser ud som vist på figur 1.

TYSK STRUKTUR



Figur 1

Denne struktur finder vi i følgende eksempler hentet fra de fagsproglige tekster, vi har arbejdet med i forbindelse med oversættelse af tyske nominalsyntagmer til dansk:

1. Eine Anzahl mehrsprachiger Lexikoneinträge
   Et antal flersprogede leksikonartikler
2. Die Vorschriften des Herstellers
   Producentens forskrifter
3. Der Austausch nichttragender Teile
   Udvekslingen af ikke-bærende dele
4. Die Verwindungsstabilität des Fahrzeuges und die Sicherheit der Insassen
   Køretøjets vridningsstabilitet og passagerernes sikkerhed
5. Folgende Verformungen des Rahmenbodenverzuges
   Følgende deformationer af rammebunden
6. Bremsdruckkreise der Antriebsachse
   Trækakselens bremsekredse
7. Ausbeulen leichter Blechschäden
   Opretning af lette pladeskader
8. Vor Beginn der Reparaturarbeiten
   Inden reparationsarbejdernes begyndelse
9. Ein Dutzend aktueller Probleme
   Et dusin aktuelle problemer
10. Rahmenkontrollmaße und Blechstärken der BMW-Modellreihen
    BMW-modelrækkernes rammekontrolmål og pladetykkelser

11. Der Anfang des Benutzerhandbuches
    Brugervejledningens begyndelse
12. Eine Reihe der Übersetzungsvorschläge
    En række af oversættelsesforslagene

Forsøger man at gruppere eksemplerne efter de kriterier, der anvendes i de gængse grammatikker, opstår følgende fire grupper:

Objektiv genitiv:

Som objektiv genitiv betegnes det "logiske objekt" for den handling, der er udtrykt i det verbalsubstantiv, som er kerne i den overordnede NP.

1. Ausbeulen leichter Blechschäden
   Opretning af lette pladeskader
2. Der Austausch nichttragender Teile
   Udvekslingen af ikke-bærende dele
3. Folgende Verformungen des Rahmenbodenverzuges
   Følgende deformationer af rammebunden

Possessiv genitiv:

Den possessive genitiv udtrykker et tilhørsforhold eller et direkte ejendomsforhold.

4. Die Verwindungsstabilität des Fahrzeuges und die Sicherheit der Insassen
   Køretøjets vridningsstabilitet og passagerernes sikkerhed
5. Rahmenkontrollmaße und Blechstärken der BMW-Modellreihen
   BMW-modelrækkernes rammekontrolmål og pladetykkelser
6. Bremsdruckkreise der Antriebsachse
   Trækakselens bremsekredse

Subjektiv genitiv:

Som subjektiv genitiv betegnes det "logiske subjekt" for den handling, der er udtrykt i det verbalsubstantiv, som er kerne i den overordnede NP.

7. Vor Beginn der Reparaturarbeiten
   Inden reparationsarbejdernes begyndelse
8. Der Anfang des Benutzerhandbuches
   Brugervejledningens begyndelse
9. Die Vorschriften des Herstellers
   Producentens forskrifter

Paritiv genitiv:

Den partitive genitiv er et genitivled, som betegner en mængde, der består af ensartede dele, eller en ensartet masse, medens den overordnede NP betegner en del af denne mængde eller masse.

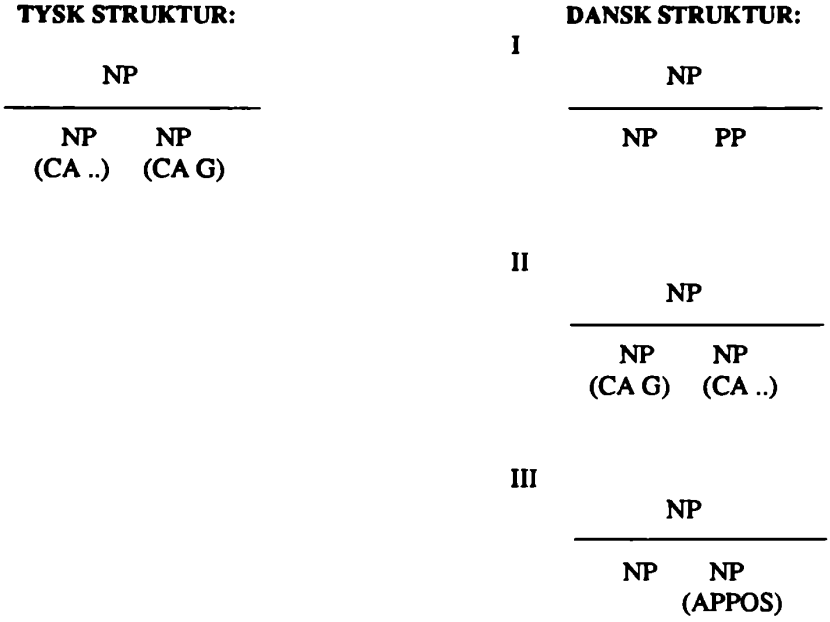
10. Eine Anzahl mehrsprachiger Lexikoneinträge
    Et antal flersprogede leksikonartikler
11. Ein Dutzend aktueller Probleme
    Et dusin aktuelle problemer

12. Eine Reihe der besten Übersetzungsvorschläge
    En række af de bedste oversættelsesforslag

I de tyske eksempler er der strukturel overensstemmelse, men semantisk er der tale om en stor variationsbredde.

Danske strukturer

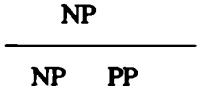
Til ovennævnte tyske struktur svarer tre strukturer på dansk, der ser ud som vist i figur 2. Den af strukturerne, der forekommer hyppigst i de fagsproglige tekster, vi har arbejdet med, er nævnt først, og den mindst hyppige sidst.



Figur 2

Dansk struktur I:

Det tyske attribut i genitiv skal på dansk erstattes af en præpositionsforbindelse. Denne struktur anvendes på dansk ved transfer af objektive genitiver og visse partitive genitiver.

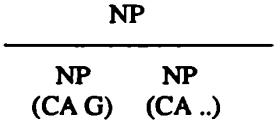


1. En række af oversættelsesforslagene (partitiv genitiv)
   Eine Reihe der Übersetzungsvorschläge
2. Opretning af lette pladeskader (objektiv genitiv)
   Ausbeulen leichter Blechschäden
3. Udvekslingen af ikke-bærende dele (objektiv genitiv)
   Der Austausch nichttragender Teile

4. Følgende deformationer af rammebunden (objektiv genitiv)
   Folgende Deformationen des Rahmenboden-verzuges

Dansk struktur II:

Det tyske attribut i genitiv skal på dansk være en NP i genitiv, der står foran den overordnede NP. Denne struktur anvendes på dansk ved transfer af subjektive og possessive genitiver.

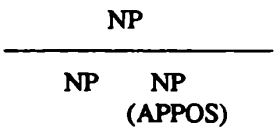


5. Inden reparationsarbejdernes begyndelse (subjektiv genitiv)
   Vor Beginn der Reparaturarbeiten
6. Brugervejledningens begyndelse (subjektiv genitiv)
   Der Anfang des Benutzerhandbuches
7. Producentens forskrifter (subjektiv genitiv)
   Die Vorschriften des Herstellers
8. Køretøjets vridningsstabilitet og passagerernes sikkerhed (possessiv genitiv)
   Die Verwindungsstabilität des Fahrzeuges und die
   Sicherheit der Insassen
9. BMW-modelrækkernes rammekontrolmål og pladetykkelser (possessiv genitiv)
   Die Rahmenkontrollmaße und Blechstärken der BMW-Modellreihen
10. Trækakselens bremsekredse (possessiv genitiv)
    Bremsdruckkreise der Antriebsachse

Dansk struktur III:

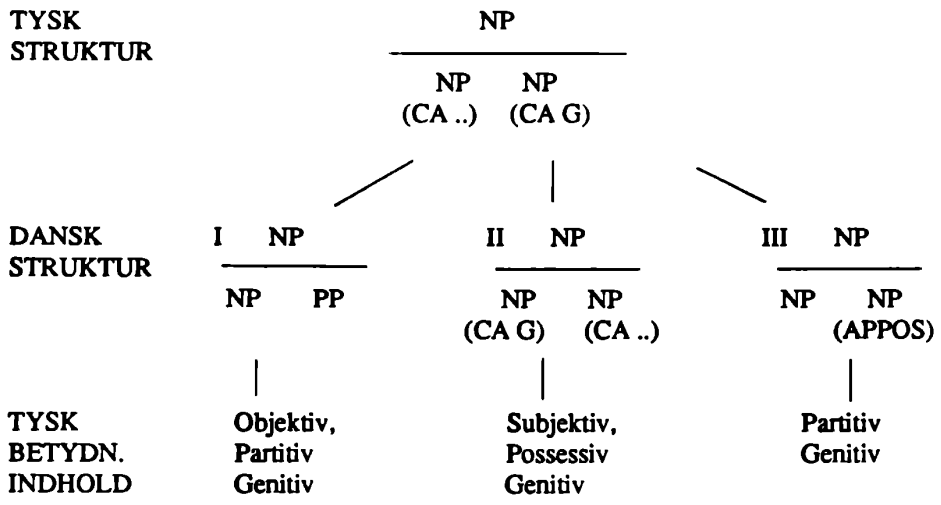
Det tyske attribut i genitiv skal på dansk være en apposition med samme syntaktiske status som den overordnede NP.

Denne struktur anvendes på dansk ved transfer af visse partitive genitiver, nærmere betegnet artens genitiv.



11. Et antal flersprogede leksikonartikler (partitiv genitiv)
    Eine Anzahl mehrsprachiger Lexikoneinträge
12. Et dusin aktuelle problemer (partitiv genitiv)
    Ein Dutzend aktueller Probleme

Et betydningsindhold på tysk behøver ikke at være knyttet til en bestemt struktur på dansk. Således omfatter både dansk struktur I og III visse partitive genitiver (se figur 3).



Figur 3

Inden for den ene tyske struktur kan vi altså konstatere en semantisk differentiering i 4 betydningsvarianter, nemlig objektiv, subjektiv, possessiv og partitiv genitiv. Dette forhold er imidlertid kun en begrænset hjælp til den strukturelle disambiguering, der skal foretages fra tysk til dansk.

Hvordan kommer vi frem til den rigtige danske struktur?

Det ville være ideelt, om de differentieringskriter, som vi har brug for, blev leveret af den tyske analyse. Dette er imidlertid ikke tilfældet. For tiden analyseres en tysk genitivkonstruktion som en NP, der består af to NPer, hvoraf den sidste står i genitiv (se figur 1). Det er der en historisk forklaring på, idet det første tyske analysemodul blev udviklet til sprogparret tysk-engelsk, hvor det ikke har været nødvendigt at levere disse informationer. Der er derfor behov for, at den tyske gruppe udvikler analysen på dette punkt.

Da vi ikke har kunnet vente på at få de manglende oplysninger leveret af den tyske analyse, har vi selv måttet finde løsninger for at kunne danne den rigtige struktur på dansk.

Vi har derfor udviklet en procedure, der beregner, under hvilke betingelser en tysk genitiv skal gengives med den danske struktur I, II og III:

I	tysk genitiv - dansk PP	(objektiv og partitiv genitiv)
II	tysk genitiv - dansk s-genitiv	(possessiv og subjektiv genitiv)
III	tysk genitiv - dansk apposition	(partitiv genitiv)

Selve proceduren bearbejder genitiverne i den omvendte rækkefølge, således at den mest specielle og lettest identificerbare struktur behandles først.

Tysk genitiv → *dansk apposition*

Denne danske struktur skal anvendes, når den overordnede tyske NP angiver et mål, en mængde o.l. (substantiver som *Anzahl, Meter, Kiste, Glas*), og når den overordnede NP er uden determiner.

Den praktiske løsning har bestået i, at vi i den tysk-danske ordbog har forsynet de tilsvarende danske substantiver (*antal, meter, kasse, glas*) med et træk, der viser, at det tyske genitivattribut modsvares af en dansk apposition (trækket Marker of Genitive med værdien APPOSition).

Tysk genitiv → dansk s-genitiv

De tyske genitiver, der på dansk skal gengives som s-genitiv, kan identificeres ved hjælp af tre betingelseskomplekser.

Det første betingelseskompleks anvender som kriterium, om der på tysk som determiner for overleddet anvendes enten den bestemte artikel eller 0-artikel.

Her er der anvendt et slags "kollokationkriterium".

I andet betingelseskompleks anvendes som hovedkriterium den regel, at alle danske subjektive genitiver optræder som s-genitiv. Opgaven består så i at identificere den overordnede tyske NP som subjektiv genitiv. Det er den, hvis kernen er et "intransitivt" verbalsubstantiv, eller kernen er et "transitivt" verbalsubstantiv og underleddet kan fungere som agens, dvs. i den tyske ensprogede ordbog er forsynet med det semantiske træk **HUM**an eller **POT**ent.

Den praktiske løsning består i, at vi slår op i den danske ordbog for at kunne hente oplysningen om, hvorvidt den danske ækvivalent til kernen i den overordnede tyske NP er et verbalsubstantiv, og for at få oplysning om dets transitivitet.

Er det danske substantiv et verbalsubstantiv, har vi for at kunne hente denne oplysning forsynet det med trækket **DEVERBAL**. Kan verbalsubstantivet forbindes med en objektiv genitiv er det i ordbogen forsynet med trækket **TRANS**itivity med værdien True.

I tredie betingelseskompleks identificerer vi en possessiv genitiv ved udelukkelsesmetoden dvs. ved at konstatere, at det overordnede substantiv ikke er forsynet med trækket **DEVERBAL** i den danske ordbog og ikke har fået tildelt et træk i den tysk-danske ordbog, der viser, at det tyske genitivattribut skal oversættes ved en dansk præpositionsforbindelse.

Tysk genitiv → dansk PP

De resterende tyske genitiver (objektiv, partitiv genitiv) oversættes ved danske præpositionsforbindelser. Den danske præposition er enten den, der måtte være angivet på den overodnede NPs kernesubstantiv ved trækket Marker og Genitive, ellers er det præpositionen "af".

Et leksikalsk problem: Oversættelse af tyske komposita

(Margrethe H. Møller)

I et maskinoversættelsessystem er der principielt to måder at oversætte komposita på: man kan enten leksikalisere dem eller oversætte dem kompositionelt.

Ved leksikalisering optager man kompositumet i ordbøgerne som ét ord. Ved kompositionel oversættelse optager man de led, kompositumet er sammensat af, i ordbogen – medmindre de findes der allerede – og lader systemet oversætte dem ét for ét og sammenføje dem til et kompositum på målsproget.

Sammensætning er en meget anvendt orddannelsesmekanisme i tysk fagsprog, og man kan ikke leksikalisere alle komposita. I praksis vil man oversætte kompositionelt, når det overhovedet kan lade sig gøre, og kun ty til leksikalisering, når der ikke er andre muligheder. Det sparer kodningsarbejde, idet leksikalisering indebærer, at der oprettes artikler i alle tre ordbøger, og det giver mulighed for at behandle sammensætninger, der ikke har været registreret tidligere.

Det, jeg gerne vil vise, er, hvornår der ikke er andre muligheder end leksikalisering, og hvilke ulemper leksikaliseringen på den anden side har.

Først og fremmest er kompositionel oversættelse kun mulig, når det tyske og det danske kompositum har samme struktur. Hvis det tyske kompositum skal oversættes ved et dansk simpleksord eller et flerleddet udtryk, som vi har mange eksempler på i de fagtekster, vi beskæftiger os med, er vi nødt til at leksikalisere kompositumet.

Vi skelner mellem kontekstuafhængig og kontekstafhængig kompositionel oversættelse.

(1) Serien | produktion – serie | produktion
(2) Dienst | programm – hjælpe | program

I eksempel (1) kan begge kompositumets led oversættes direkte, uafhængigt af konteksten. Det tyske *Serie* oversættes til dansk *serie*, og det tyske *Produktion* oversættes til dansk *produktion*. I eksempel (2) er oversættelsen af førsteleddet kontekstafhængig: normalt oversættes det tyske substantiv *Dienst* ved det danske substantiv *tjeneste*, men når det står sammen med *Programm*, oversættes det ved verbet *hjælpe*. Dette kan angives i transferordbogen.

Danske komposita kan have fugebogstaver, og som hovedregel gælder det, at et givet ord altid anvender samme fugebogstav, når det indgår som første sammensætningsled i et toleddet kompositum. Det danske substantiv *ændring* anvender f.eks. fugebogstavet s som i *ændringsforslag*, *ændringsønske*. Vi angiver, hvilket fugebogstav der skal anvendes, i den danske ordbog.

Nogle ord anvender desuden en speciel allomorf, når de indgår i et kompositum. Det danske substantiv *maskine* anvender f.eks. allomorfen *maskin* som i *maskinfejl*. Også dette angiver vi i den danske ordbog.

Der kan være disambigueringsproblemer, f.eks. når analysen vælger substantivet *die Steuer* i stedet for verbet *steuern* i det tyske kompositum *Steuerzeile*, så vi får oversættelsen **skattelinie* i stedet for den rigtige styrelinie. Her er den eneste løsning leksikalisering af det tyske kompositum og dets danske oversættelse.

Men bortset fra det fungerer kompositionel oversættelse fint, når der er tale om toleddede komposita. Det bliver straks vanskeligere, når kompositumet har mere end to led. For det første er det da ikke muligt i den danske generering at beregne, hvor der skal indsættes fugebogstaver, for det andet kan vi ikke stole på at få den rigtige oversættelse af et led, hvis dette leds oversættelse er kontekstafhængig. Begge dele skyldes, at systemet ikke er i stand til at analysere sig frem til kompositumets interne struktur.

I eksemplerne:

(3) [[Wörter | buch]eintrag]
ordbogsartikel
(4) [Fach[buch | handlung]]
specialbog_handel

ser vi, at substantivet *bog* skal have fuge-s, når det optræder som andet sammensætningsled som i *ordbog* i eksempel (3), men ikke, når det er første sammensætningsled som i *boghandel* i eksempel (4).

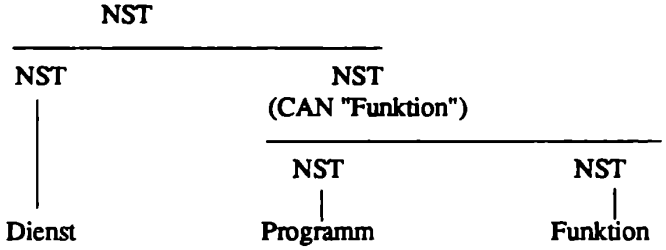
Da systemet ikke har kriterier til at afgøre hvordan komposita med flere end to led er opbygget, har man i den tyske analyse truffet et valg: man analyserer altid højrerekursivt. Alternativet ville være at levere flere analyser. Det må imidlertid være legitimt at træffe et sådant valg i et operationelt maskinoversættelsessystem.

I den danske generering gør vi ikke noget forsøg på at vælge specielle allomorfer eller indsætte fugebogstaver i disse komposita, vi sammenføjer simpelt hen de kanoniske former, så vi får oversættelserne (5) *ordbogartikel* og (6) *specialboghandel*. Vi overlader det til posteditoren at indsætte fuge-s. Hvis et sådant kompositum optræder gang på gang og posteditoren ikke ønsker hver gang at skulle indføje fuge-s, må han leksikalisere kompositumet.

Det, at komposita altid analyseres højrerekursivt, får også konsekvenser for den kontekstafhængige kompositionelle oversættelse i transferordbogen:

(5) Datenbankdienstprogramm
databasehjælpeprogram
(6) Dienstprogrammfunktion
* tjenesteprogramfunktion

Eksempel (5) bliver oversat korrekt, fordi det tyske substantiv *Dienst* med den højrerekursive analyse bliver genkendt som førsteled i kompositumet *Dienstprogramm*. Det gør det derimod ikke i eksempel (6), jf. nedenstående analysetræ, og den kompositionelle oversættelse bliver derfor forkert.



Hvis man vil være sikker på at få oversat *Dienstprogramm* rigtigt i alle sammenhænge, må man leksikalisere det.

I praksis vil en bruger af systemet vælge leksikalisering, når der er tale om fagsproglige termer. Det er vigtigt, at de bliver oversat korrekt og på samme måde hver gang.

Samtidig med, at man løser et problem ved at leksikalisere, skaber man imidlertid et andet. Man ødelægger nemlig muligheden for at oversætte sideordnede komposita med ellipse.

På tysk udelader man ofte andet sammensætningled, hvis det er fælles for to sideordnede komposita:

(7) Änderungs- oder Umstellungswunsch
ændrings- eller omstillingsønske
(8) Maschinen- oder Bedienungsfehler
maskin- eller betjeningsfejl
(9) Kopf- und Fußzeile
top- og bundlinie

I eksempel (7) – (9) er andetleddene *Wunsch*, *Fehler* og *Zeile* fælles for de to sideordnede komposita. I eksempel (7) skal det danske førsteled *ændring* have fuge-s. I eksempel (8) anvendes den specielle allomorf *maskin* af det danske førsteled *maskine*. I eksempel (9) skal det tyske substantiv *Kopf*, da det indgår i et kompositum med andetleddet *Zeile*, have den kontekstafhængige oversættelse *top* i stedet for den normale, direkte oversættelse *hoved*.

En forudsætning for, at vi kan behandle sideordnede komposita med ellipse er, at den tyske analyse får mulighed for at analysere konstruktionen kompositionelt. Kompositaenes bestanddele skal altså findes i ordbøgerne, og kompositaene må ikke være leksikaliserede. Dette skyldes, at analysen følger princippet om "the longest possible match".

Hvis man f.eks. har leksikaliseret det tyske kompositum *Fußzeile*, får man ikke en kompositionel analyse. Det betyder, at *Fußzeile* i eksempel (9) ikke bliver genkendt som et kompositum. *Kopf* bliver dermed heller ikke genkendt som førsteled i et kompositum med *Zeile* som andetled. Vi får

altså ikke den rigtige kontekstafhængige oversættelse af *Kopf*, nemlig *top*, men derimod den normale oversættelse *hoved*, altså **hoved- og bundlinie*.

Så kan man naturligvis vælge at leksikalisere sideordningen *Kopf- und Fußzeile*, men en sådan leksikalisering kan gribe om sig. De to komposita kan jo også optræde i andre kombinationer og med andre konjunktioner end *und*:

Kopf- und Fußzeile
Kopf- oder Fußzeile
Kopf- bzw. Fußzeile
Kopf- Fußzeile
Fuß- und Kopfzeile
etc.

Komposita optræder hyppigt i tyske fagsproglige tekster, både alene og i sideordninger med ellipse. I vort arbejde med fagsproglige korpora har vi været nødt til at leksikalisere mere, end vi fra begyndelsen havde forestillet os.

Ellen Christoffersen og Margrethe H. Møller
Institut for erhvervsforskning
Handelshøjskole Syd
Engstien 3
DK-6000 Kolding

Linguistics and Machine Translation

Helge Dyvik
University of Bergen

Abstract

This talk discusses some of the challenges posed for theoretical linguistics by translational problems. It briefly discusses different aspects of the concept of translational equivalence and the degree to which this relation may be captured by machine translation (MT) systems. It is argued that the formal requirements and multilingual perspective of MT provide new criteria of adequacy for semantic theories. This motivates further development of linguistic semantics as a prerequisite for a semantically based theory of machine translation. In particular it seems that the semantics of translational equivalence needs to be able to refer to a typology of linguistic objects as part of its "ontology". Some of the points are illustrated by a short presentation of an experimental MT project. Finally it is briefly indicated how the treatment of linguistic objects as part of the semantic model fits into a situation theoretic account.

1. Translational Equivalence as Empirical Evidence

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'. To-day, 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.

Within this area of increased interaction between the two fields the flow of information – or inspiration – is usually seen as unidirectional: linguistic theories of syntax and semantics provide representational tools for the developers of MT systems, but remain rather unperturbed by the activity themselves. However, the increased contact between linguistics and the activity of translation should not be seen as a pure case of applied science, in which results from linguistics are simply used to solve practical problems. Translation should also be recognized as an important testing ground for linguistics, reinforcing a much-needed multilingual perspective in the development of more or less formal linguistic theories. Specifically, the "pre-theoretic" concept of translational equivalence may provide valuable criteria of adequacy for linguistic theories. This is especially evident within semantics. Within linguistic semantics we try to characterize the meaning of a linguistic expression by translating it into some meaning representation. This meaning representation is, basically, just another linguistic expression, and the natural question is how this alternative expression brings us closer to the characterization of meaning than did the original expression. There are several relationships a meaning representation may enter into which may

motivate it as an informative characterization of meaning. For one thing we may relate it to different expressions within the same language, for instance by means of a model theoretic interpretation, thereby characterizing various semantic relations like synonymy, entailment etc. between expressions; for another the meaning representation can be used to mediate between language and other contacts with reality, like sight and action; for yet another the representation can be related to expressions in different languages, and hence characterize semantic relations across language borders. The relation of translational equivalence involves just this last-mentioned type of semantic relations and hence constitutes an empirical domain with respect to which semantic representations need to be adequate. The relation of translational equivalence is not easy to pin down, but a corpus of actual texts paired with their translations, sorted according to quality by bilingual informants, would be a starting point, giving us part of the extension of the relation, so to speak. The translational relation is manifested in such a collection of actual translations. To the extent that semantic representations *are* adequate with respect to such phenomena – that is, to the extent that they classify equivalent expressions together – they are also motivated as something more than just arbitrary alternative encodings of some semantic content.

It is an age-old insight that translational equivalence involves much more than denotational equivalence. (I am leaving aside now the important question of whether the translation relation should be conceived as an equivalence relation at all; we may note that good translation is frequently assumed to be irreversible.) The literature on translation discusses several sub-species of the relation; the following categories are suggested by Werner Koller (1983:186ff.):

1. *Denotational equivalence*
2. *Connotative equivalence*
3. *Text-normative equivalence*
4. *Pragmatic equivalence*
5. *Formal equivalence*

Denotational equivalence means equivalence with respect to properties of a described situation (conceived as a section of "objective reality"). However, languages usually allow alternative verbalizing strategies, which may induce different *perspectives* on what may still be conceived as the same described situations, highlighting different aspects of their temporal structure or of the relationships between their participants. (Example: "I can only stay until 5 PM" vs. "I have to leave at 5 PM".) Similarly, alternative verbalizations may belong to different sub-languages or levels of style. Equivalence with respect to such properties is *connotative* equivalence. *Text-normative* equivalence is equivalence with respect to properties characteristic of certain text types. A French business letter may have a different structure from a Norwegian business letter; a French/ Norwegian pair of letters are text-normatively equivalent only if they each satisfy the respective French and Norwegian conventions for such texts. *Pragmatically* equivalent texts or utterances are equivalent with respect to the communicative acts performed with them, and their effects on the receiver (such as the degree of politeness accompanying a request, irony, etc.). Cultural differences frequently make the achievement of pragmatic equivalence difficult. Finally, *formal* equivalence is equivalence with respect to properties of assonance, rhyme, rhythm and word-play.

Obviously, all these types are not always of equal importance, but they all belong in a principled discussion of translational equivalence. We may note that denotational equivalence primarily concerns the described situation, pragmatic equivalence the discourse situation, whereas connotative, text-normative and formal equivalence primarily concern, in an irreducible way, *the linguistic signs themselves*. If two texts are equivalent with respect to chosen perspectives, levels of style, text-structural norms and metrical properties, they can be said to be equivalent with respect to the *types of linguistic devices* used in them. This suggests that a precise characterization of translational equivalence presupposes not only adequate denotational semantics and pragmatics enabling us to

refer to such properties of described situations and discourse situations which translationally equivalent expressions have to bear a common relation to. It seems that we also need a universal typology of linguistic devices: linguistic expression types themselves are an irreducible part of the 'ontology' of things which translationally equivalent expressions have to have in common. Thus, formal aspects of a text may create types of meaning that we want to re-create in a translation. If we take it as axiomatic that it is the task of semantics to account for all aspects of meaning, these translational phenomena bring new problems within the scope of semantic theories.

2. *Limits of MT: Equivalence from Pre-Established Correspondences*

Translational equivalence is basically a relation between *texts* or *utterances* rather than between lexical and grammatical elements in language descriptions. In other words, translational equivalence concerns *parole* rather than *langue*. It is an analytic task to reduce translational equivalence between texts, as far as possible, to a function of correspondence relations between elements of language descriptions – to relations of *langue* rather than *parole*. However, it seems very likely that there are types of translational equivalence that cannot be so reduced. Thus, while connotative properties of individual lexemes might be given some representation in the lexicon, it is difficult to see how information about connotative and stylistic properties of complex phrases can be derived compositionally, in the way information about denotative properties are. In such cases the only solution seems to be to list such complex phrases with their properties as idioms, but extensive use of this solution is at least impractical because it leads to an explosion in the inventory of idioms. Furthermore, global stylistic properties of texts that concern, for instance, the frequency of certain constructions cannot even in principle be reduced to simple correspondence relations between linguistic descriptions. There are also aspects of formal translational equivalence (for instance, in translations of poetry) and culture-bound pragmatic equivalence that seem to presuppose genuine creativity on the part of the translator, and hence to be something that cannot be reduced to pre-established correspondences.

These limitations are at the same time limitations on the possibility of machine translation. A machine translation system is possible to the extent that we have been able to reduce aspects of translational equivalence to functions of pre-established correspondence relations between finite linguistic descriptions. The linguistic analysis in the descriptions may show any degree of sophistication, and we may grant the possibility that the translation algorithm is able to exploit contextual information of various kinds to choose among alternatives – the fact remains that any equivalence between texts established by the system must be compositionally derivable from the elements in the descriptions and the pre-established correspondence relations between them. This holds true whether we implement the relations as simple pointers or as representations of some kind, representing what the corresponding entities have in common. Such representations, if we use them, can reasonably be seen as semantic representations. Since it seems plausible to claim that all aspects of translational equivalence are meaning related, it should be the task of semantics to account for the correspondence relations between languages on which translational equivalence is based. Semantic representations, then, should capture whatever a translational equivalent has to be equivalent with respect to, to the extent that this is derivable from the linguistic descriptions. Such representations are at the same time potential *interlingua expressions* – it should in principle be possible to implement a translation algorithm that translates a source text to such a representation, and generates a target text from it.

3. *Interlingua as a Theoretical Tool*

This does not mean that it would be a good idea to do so in practice. The majority view is that the transfer technique is to be preferred over the interlingua technique, and there are often good arguments to support this view. The goal of constructing a true interlingua is often considered as unrealistic, or at least impractical. Such a universal intermediate representation is easily seen at least as a detour; in many cases it seems simpler to define direct transitions from source representations to target representations. Still, the idea of an interlingua should not be discarded out of hand. In the first place, the possibility of interlingua-based translation does not imply the possibility of a truly universal interlingua, common to all conceivable language pairs. An interlingua can be specific to a given language pair and still be an interlingua. In the second place, we should distinguish between the theoretical *possibility* of an interlingua and the *practical utility* of having such an interlingua actually implemented in a system. Even if we decide against the latter, working out an interlingua might be a useful part of a principled study of translational equivalence. In fact, on reflection it is easily seen that the distinction between interlingua and transfer is not a deep distinction of principle, but rather a fairly superficial one of implementation. It is never the case that transfer must be preferred over interlingua because transfer is the only *possible* option: interlingua is possible whenever transfer is. Machine translation actually implies the possibility of interlingua-based machine translation. This is because machine translation, as I have already pointed out, is only possible to the extent that translational equivalence can be reduced to a function of pre-established correspondences between elements of two language descriptions. Since the language descriptions are finite, there will be a finite number of such pre-established correspondences. The correspondences can be implemented as simple pointers, but it is of course also possible to label each correspondence with a unique name, and to write rules for combining such labels compositionally in tandem with the rules for combining the corresponding expressions syntactically. In short, the correspondence relations, being finite in number, can evidently be described in a metalanguage. Such a metalanguage would be a theoretically possible interlingua between the two languages, and even if we don't implement it as an interlingua the metalanguage might be a useful theoretical tool for keeping our ideas straight while writing the programs.

4. *The PONS Project: Exploiting similarity between languages*

Now, I haven't worked out such a theoretically interesting interlingua yet. I have, on the other hand, been working on an experimental translation system, and I should like to give a brief sketch of it in order to illustrate some of the points I have made. The project is called PONS – acronymic for "Partiell Oversettelse mellom Nærstående Språk" (Partial Translation between Closely Related Languages). The starting point was an idea about studying some aspects of the relationship between linguistic motivation and computational tractability in MT systems. A common objection to linguistically sophisticated language descriptions as modules in translation systems is that they will inevitably make analysis and synthesis hopelessly redundant and inefficient, compared to quick ad-hoc shortcuts from source construction to target construction, without regard for their full set of grammatical and semantic properties. This is evidently a valid consideration – it would be irrational to spend time finding a lot of grammatical and semantic information if you really don't need it. On the other hand, linguistic motivation and computational tractability could be combined if we could achieve a linguistically well-informed system able to refrain from using all its knowledge all the time. If the system had some means of evaluating the complexity of a given translational task in advance, it could infer the amount of analysis required and adjust its mode of operation accordingly. The possibility of taking shortcuts would obviously be most frequent during translation between closely related languages; hence it was natural to try out these ideas in relation to translation between

a pair of Scandinavian languages like Norwegian and Swedish. The formal tools of unification grammar seemed well suited to the task, since feature structures can be more or less underspecified: information can be removed from them as the need arises.

It is not only considerations of efficiency that motivate such an attempt to achieve a system that uses its knowledge in a considered way. It seems quite plausible to assume that human translators behave in a similar manner. We seem able to adapt the amount of information we bring to bear on a problem to its complexity. Thus it is obviously far easier for us to translate between closely related languages than between languages that are genetically and typologically further apart. The Scandinavian languages are almost limiting cases in this respect. In translations between Danish, Swedish and Norwegian, or between Bokmål and Nynorsk, there will frequently (but far from always, of course) be a word-by-word type correspondence between source and target texts. A natural procedure in such cases would be first to inspect the sentence to be translated sufficiently closely to determine that it contains no constructions departing from the unmarked word-by-word case, and then translate word-by-word, making morphological adjustments along the way. The human translator would not take the trouble to reflect carefully and at length on the content and connotations of the source sentence, in abstraction from its syntactic form, and then try to encode this content "from scratch" in the target language with no regard for the way it was expressed in the source text. Only to the extent that the constructions do not allow fairly simple formal mappings is such a closer consideration of semantic properties necessary. The translator will use as much of the structure of the source sentence as possible – as long as she can trust that a similar sentence structure in the target language is translationally equivalent. And this is not a case of laziness; it is because this method is a precondition for a good translation, that is, a translation which renders the properties of the source text as reliably as possibly, *including its way of using language*. By "a way of using language" I have in mind production of the type of meaning which is captured by connotative equivalence (such as equivalence with respect to the chosen perspective on a situation), pragmatic equivalence and formal equivalence. Languages have different resources, different device inventories, for creating such meaning – that is what makes translation difficult – but the more closely two languages are related, the greater the overlap between their inventories of linguistic devices will be. It is not a facile shortcut, but the satisfaction of an independent purpose of translation which takes place when we choose equivalent linguistic devices in a translation – for instance word-by-word translation whenever that is possible. ('Possible' here means 'preserving translational equivalence' – including connotative equivalence etc. In other words, it is not a counter-argument to the claim made here that word-for-word translations frequently are bad – when they are bad, they presumably are not the closest possible equivalent.) In such cases we pass directly from a device in one language to an equivalent device in the other; careful consideration of the actual job done by these devices is superfluous. This does not mean that meaning suddenly is unimportant, it simply means that equivalence of meaning has been established once and for all, given our knowledge of the relationship between the languages, and therefore we need not worry about it every time.

Hence there are both practical and theoretical reasons to try to develop a system able to take "shortcuts" past an involved semantic analysis whenever this is possible because of a certain degree of structural correspondence between the two languages. This is a basic idea in the PONS project. The system is implemented (in Interlisp), except for a few modules that are not fully integrated yet. The linguistic descriptions are developed within an extended and modified version of Lauri Karttunen's D-PATR, a framework for developing unification-based grammars. The idea of shortcuts is captured by a distinction between three different modes of operation, corresponding to three different degrees of correspondence between source and target constructions. We will look at the most elaborate mode, Mode 3, first.

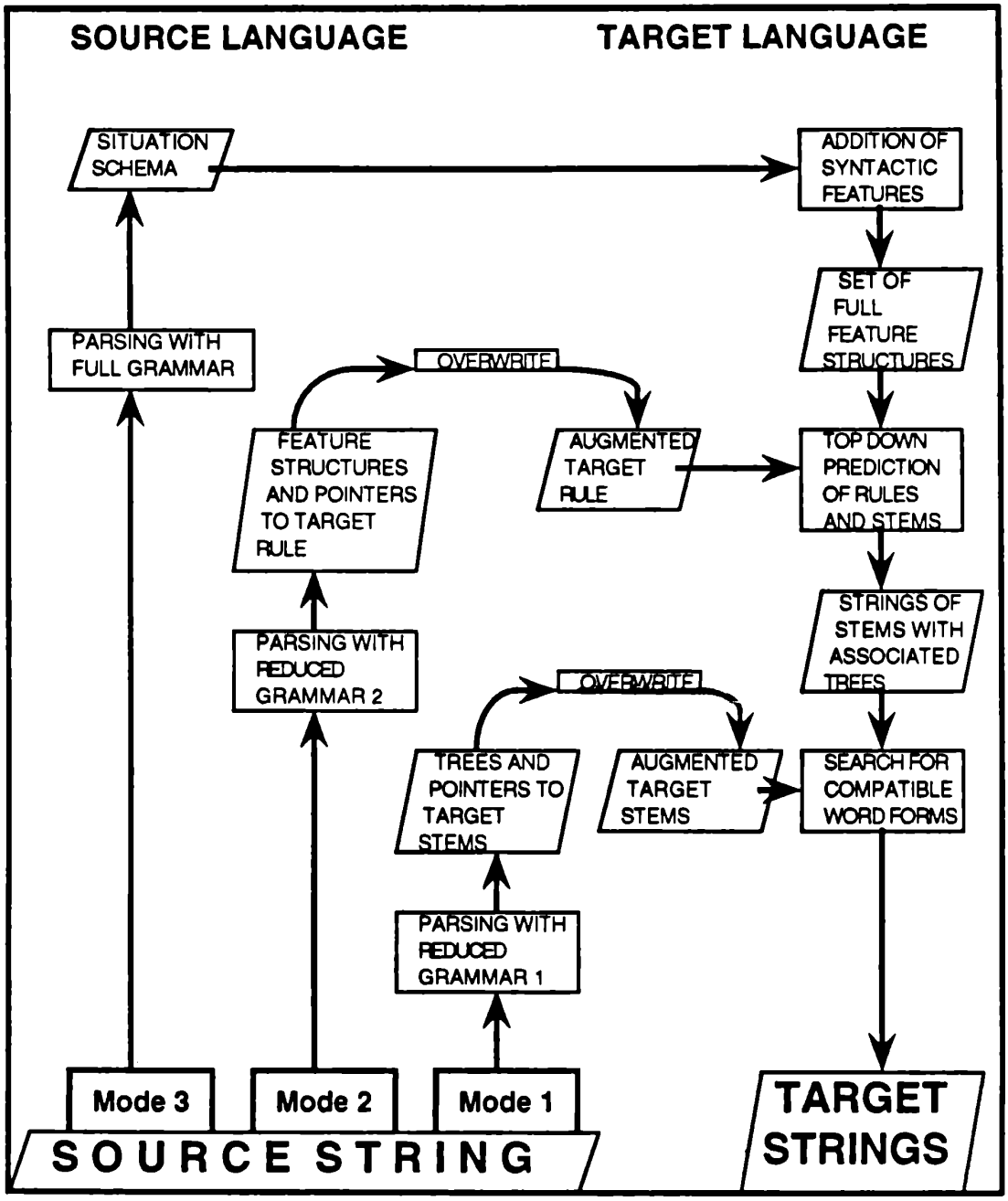


Fig. 1. The three modes of PONS

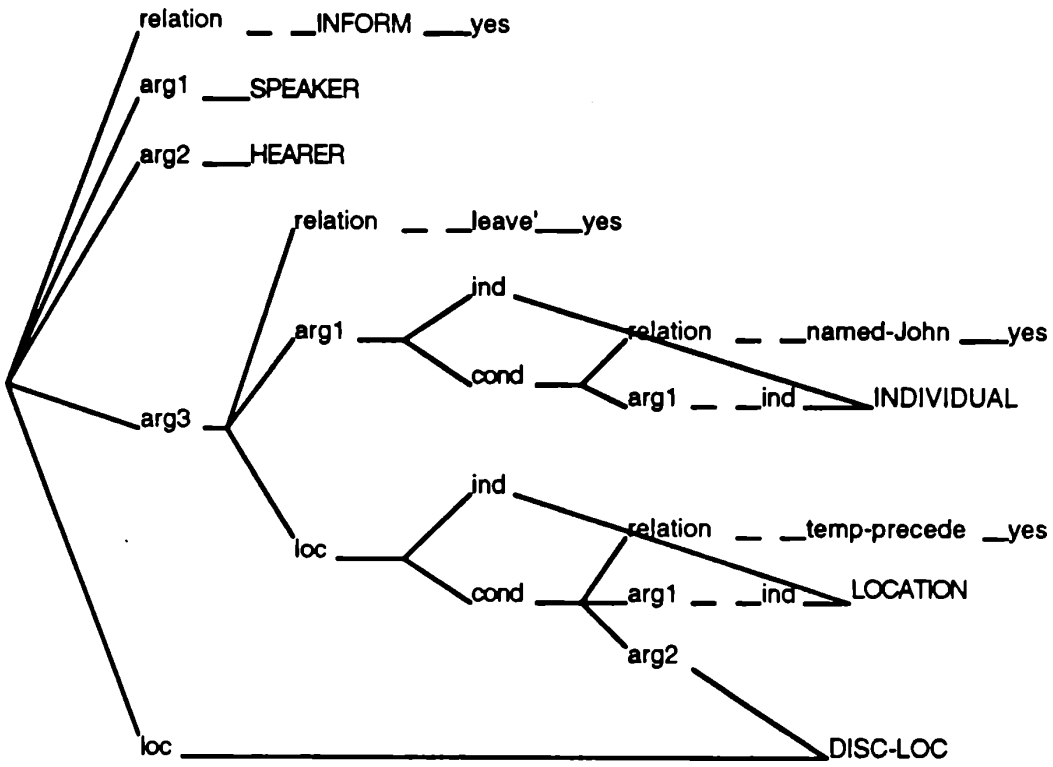


Fig. 2 Situation schema, PONS version

5. Mode 3: Semantic Representations as Interlingua

The Mode 3 path from source string to target strings is described in the outer circle of fig. 1. The language description on which parsing is based consists of a set of annotated phrase structure rules, a set of stems, in which each stem contains all information which is common to all forms of that stem, and a set of word forms containing further information. All entities – rules, stems and word forms – are represented as feature structures, or directed graphs, which are unified during parsing. The result of a parse is a phrase structure tree and a feature structure. The feature structure consists of substructures, one of which is a semantic representation in the form of a *situation schema* (fig. 2).

The situation schemata I am using are somewhat modified versions of the situation schemata introduced by Fenstad, Halvorsen, Langholm and van Benthem in 1987. A situation schema is a representational format suited to be interpreted by situation theory. In Situation Semantics the meaning of a sentence is conceived as a relation between types of discourse situations and types of described situations: the meaning constrains these situations to be of certain types. The example in fig. 2 represents the sentence "John left". In the schema attributes represent parameters of situations (a situation is constituted of a *relation*, a set of *arg*uments to it and a *loc*ation, among other things),

while values represent the corresponding entities in the situations represented. The "outermost layer" represents aspects of the discourse situation, while the described situation is entered as the value of *arg3* (it is the entity about which the speaker informs the hearer).

In the full feature structure of a sentence the situation schema is inter-related with the syntax substructure in various ways. Fig. 3 shows the basic layout of full feature structures, while fig. 4 shows a simplified example.

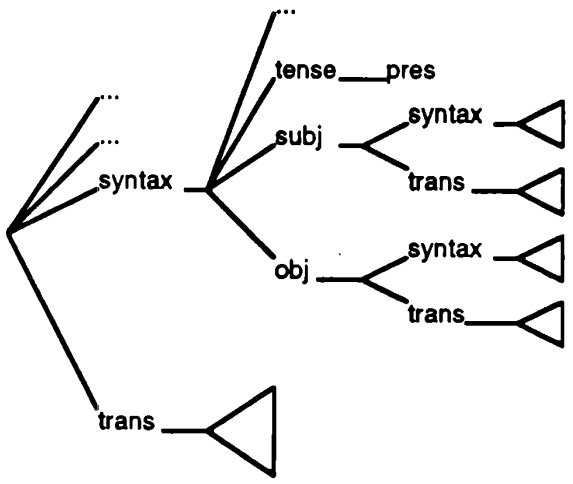


Fig. 3 Basic layout of PONS feature structures

The patterns of unifications between the syntax and semantics substructures express such things as linking relations between syntactic functions and semantic argument positions. In Mode 3 the situation schema is extracted as a kind of interlingua expression from the feature structure. In other words, linking relations and other syntactic information from the source string parse are forgotten about, since they are source language specific. Target strings are then generated from the situation schema, on the basis of the target grammar, in three stages (cf. fig. 1, right-hand column of boxes):

1. Target stems expressing the relation values in the schema are retrieved from the lexicon, and their feature structures are unified into the situation schema. The result is that target language specific syntactic features are added, giving a set of full feature structures, linking relations and all, as output.

2. Syntactic rules are predicted top-down, the predictions being constrained and governed by the information in the feature structures, which in a sense "know all about" the constructions about to be generated. The ouput is a set of trees, with lexical stems at the terminal nodes.

3. Word forms are entered at the terminal tree nodes, to the extent that their feature structures are compatible with what is already there. The output is a set of target strings – the set of strings, that is, to which the target grammar assigns situation schemata compatible with the source situation schema. To the extent that I have been able to capture the properties that define translational equivalence in the situation schemata, the strings are useful translations.

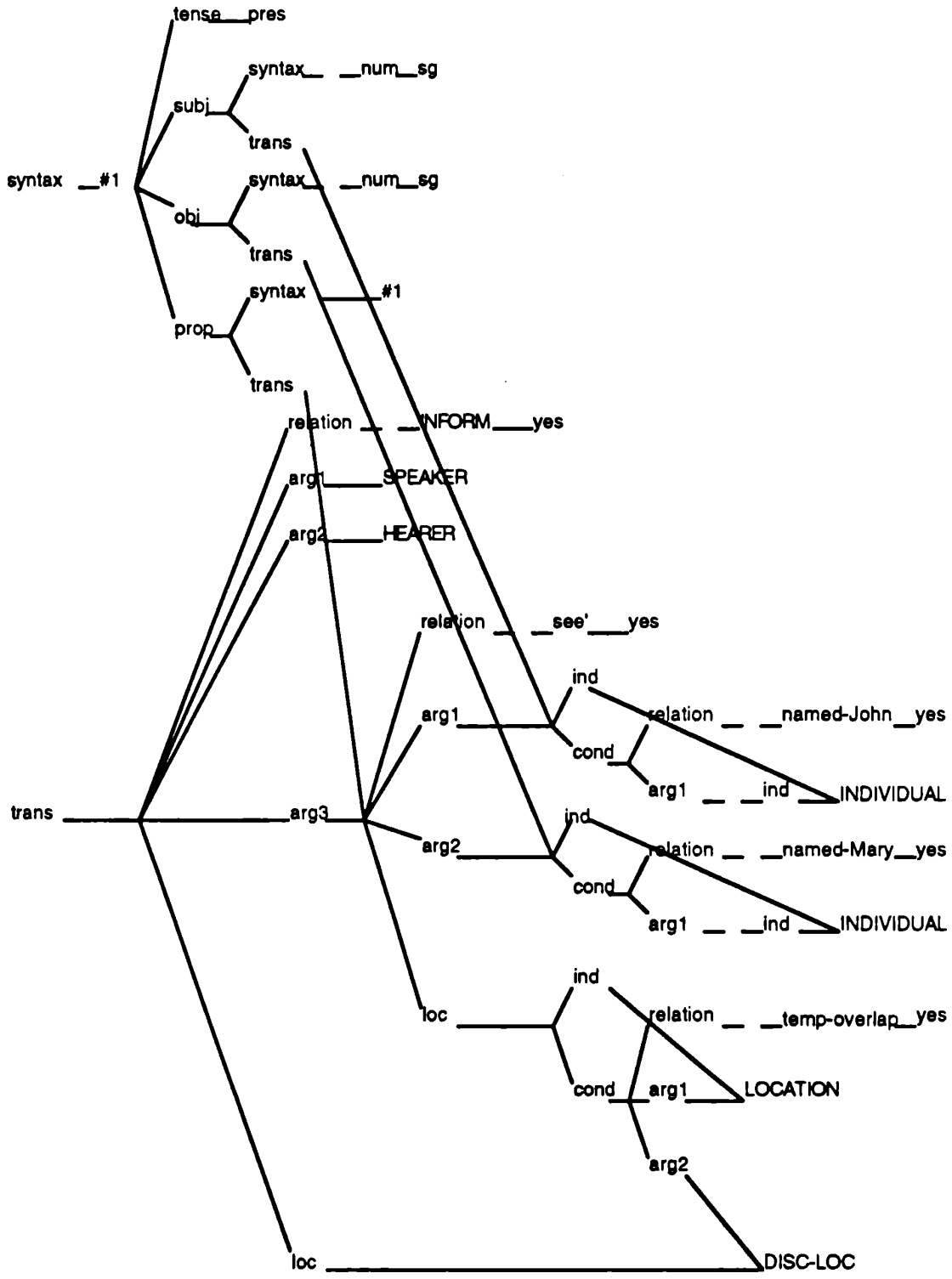


Fig. 4 Feature structure with interrelated syntax and trans substructures

6. *Modes 1 and 2: Shortcuts by exploiting similarity*

This, then, was the most elaborate mode. Now let us consider the modes that allow the system to exploit structural similarities between the languages – modes 1 and 2 in fig. 1. The idea is that the system itself should be able to decide that *this* chunk corresponds to a target construction word-by-word, and hence is to be tackled in Mode 1, whereas *this* chunk corresponds in a slightly more indirect fashion, with discrepancies in constituent order and grammatical formatives, and hence Mode 2 is in order, while, finally, *this* one does not correspond in any interesting way at all, so here we have to resort to Mode 3 and retrieve a situation schema. The basis for making such decisions will have to be language pair specific information. At the outset, however, the language descriptions are independent of any specific translation partner. The languages are described in their own terms – except for the obvious fact that the same grammatical apparatus has been employed, ensuring that the same type of grammatical phenomenon is described in the same terms in every language. But on the basis of such descriptions grammar comparisons can be performed automatically. An algorithm has been implemented which to some extent pre-compiles grammar pairs: it compares the entities in the two grammars and the two lexicons and adds information about the results of the comparison to them. Thus after the comparison the language descriptions will contain information about their partners. Let us briefly consider how this will be utilized in Mode 1 and Mode 2.

Mode 1 is the default mode from which the system departs only when triggered to do so. It is the mode of word-by-word correspondence. Even so, some syntactic analysis has to be performed, in order to achieve disambiguation of homonymous forms, and in order to detect cases where Mode 1 is insufficient. Therefore a syntactic tree is derived. The rules and lexical entries utilized in this derivation will contain information to the effect that an identical tree with the same compositional properties can be derived by the target language rules and lexical entries. This means that we do not need to do that: when the correspondence between source and target string is that close, we can forget about the target grammar; the source tree itself can be used. Then the source lexical entries at the terminal nodes are replaced by corresponding target entries, whereupon the resulting structure enters stage 3 of the synthesis procedure (cf. fig. 1) during which compatible target word forms are found. This procedure will appropriately handle gender clashes and various kinds of one-to-many correspondences between the two lexicons. Mode 2 is in order in cases where source and target string still correspond with respect to the sets of sense-carrying words they contain, but differ in constituent order, in the presence of sense-less grammatical formatives, or both. In this case the source tree cannot be used, but the source feature structure, including its grammatical information (its f-structure in LFG terms) is still useful. An example of such a correspondence would be Norwegian "den nye bilen min" vs. Swedish "min nya bil" 'my new car'. The grammar comparison has identified cases where a rule in one language corresponds to a rule in the other with respect to sense-carrying daughters and compositional properties, but not with respect to constituent order or sense-less daughters. In such cases pointers have been constructed between the rules. Hence the procedure is that the system recognizes the pointer, retrieves the target rule, overwrites its feature structure over the source feature structure, and starts top-down prediction of target rules constrained by the modified source feature structure. In other words, it enters stage 2 of the full synthesis procedure.

Finally, Mode 3 is triggered when the parse reveals either a rule or a lexical entry with no Mode 1 or Mode 2 correspondent in the target language.

This was a very brief sketch of a system which operates partly on interlingua principles – in Mode 3 – and partly on transfer principles – in Modes 1 and 2. I have suggested that a semantic representation ideally should be able to capture everything in terms of which we define translational equivalence (at least to the extent that such equivalence is reducible to pre-established correspon-

dences between elements in the language descriptions), and that such a semantic representation can be seen as a kind of theoretical interlingua expression. How does this work out in the case of the kind of equivalence relations that are captured in Modes 1 and 2? In Mode 1 we utilize common tree structures in the two languages, in Mode 2 we utilize common syntactic feature structures. In other words, in these modes we have equivalence with respect *to types of linguistic signs:* the same linguistic devices are employed. As we have seen, this is one kind of equivalence we want to capture, since it is involved in achieving connotative and formal equivalence. But if translational equivalence may involve such grammatical equivalence, and semantics is to characterize all aspects of translational equivalence, then semantics has to be able to refer to aspects of grammatical form. Is this reasonable?

7. Language as Part of the World

Several situation semanticists (Robin Cooper, Stanley Peters, Jean Mark Gawron) have been working towards the development of a situation theoretic account of grammar itself. The problems addressed by Cooper do not pertain to translation, but rather to the situation theoretic characterization of the information states attainable by human and artificial information processors. A grammar, according to his analysis, is a relation between possible linguistic utterances and possible types of facts in the world around us. This is no place to go into details; suffice it to say that this approach involves treating linguistic entities – words, sentence constructions, grammatical categories etc. – as "things in the world" along with other things that human beings can talk about and relate to. This is a desirable perspective; semantics ought to be able to account for language as something which occurs in the same world as the one language relates to, and not as something occurring in a domain entirely of its own. This is obviously necessary in order to account semantically for reflexive language use, that is, use of language to talk about language, in expressions like "the word 'horse'", "the expression you just used", "that was a bad sentence", etc. Another example is the usual situation semantic analysis of *names* (fig. 5).

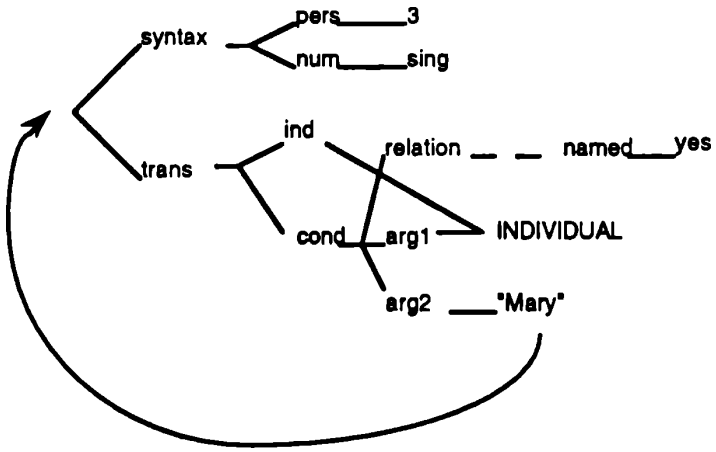


Fig. 5 Circular representation of a name

The analysis specifies that a precondition for using the name to refer to an individual is that the individual enters into the relation *named'* with the name itself. In other words, the name enters into the specification of its own semantics, as part of the type of described situation it relates to. There is an element of circularity in this. If we interpret the feature structure as a representation of the name itself, the result is that the feature structure will contain a circular path.

In this case the linguistic entity itself occurs in its described situation, which is slightly special. But the occurrence of linguistic entities in *discourse situations* is quite typical, of course, and discourse situations are also part of what enters into the specification of linguistic meaning in a situation theoretic framework. Hence the element of circularity could become quite widespread if we develop the situation theoretic account in some detail: linguistic signs become essentially circular structures, containing themselves as constituents of their "semantic" subparts. This is a consequence of understanding the meaning of an expression as the way it constrains the relation between discourse situations in which it can be used and objects it can describe. Intuitively, the circular structure expresses the fact that an essential property of any discourse situation in which sign S is used, is that S is a constituent in it.

A consequence of this is that if we assume that translational equivalence is to mean equivalent situation schemata, then a sign will only be translationally equivalent with itself. This captures in a fairly brutal way the unique, non-translatable character of linguistic signs: the insight of many translators that the perfect translational equivalence relation across languages is empty. What we do in practice is search for target expressions that are equivalent with respect to *subsets* of the information in the schemata. In fact, I believe that we do want a theory of translation that accounts for its basic impossibility, while at the same time allowing for various approximations to the unattainable ideal.

Considerations like these might be a starting point for developing a semantic theory of translation – a study of the semantics *of* translation rather than simply working with semantics *for* translation. As I have stressed, the theoretical constructs of such a theory are not necessarily to be taken as models for direct implementation, but they might still provide the work with some theoretical basis.

8. References

Cooper, R. 1989: "Information and Grammar", unpublished manuscript.

Dyvik, Helge 1990: *The PONS Project. Features of a Translation System.* Skriftserie, nr. 39, Department of Linguistics and Phonetics, University of Bergen.

Fenstad, J.E., P.-K. Halvorsen, T. Langholm and J. v Benthem 1987: *Situations, Language and Logic.* Dordrecht: Reidel.

Gawron, J.M., and S. Peters 1990: *Anaphora and Quantification in Situation Semantics.* CSLI Lecture Notes Number 19, Center for the Study of Language and Information, Stanford.

Koller, W. 1983: *Einführung in die Übersetzungswissenschaft,* Heidelberg: Quelle & Meyer.

Helge Dyvik
Department of Linguistics and Phonetics
University of Bergen
Sydnesplass 9
N-5007 Bergen, Norway

Tel +47 5 212261
E-mail dyvik@hf.uib.no

En homografseparator baserad på sannolikhet

Gunnar Eriksson
Stockholms universitet

0. Inledning

Homografi eller lexikal flertydighet blir ett stort problem så snart man börjar försöka analysera större mängder text. Analys av verkliga texter från skilda genrer kräver ett generellt lexikon med hög täckningsgrad och för att den lexikala analysen ska kunna ligga till grund för annat än triviala lingvistiska iakttagelser krävs att antalet möjliga analysdistinktioner (t ex ordklassser, subkategorier eller morfologiska egenskaper) är stort. Båda dessa förutsättningar försvårar den lexikala analysen: Antalet homografer ökar!

Några triviala exempel:

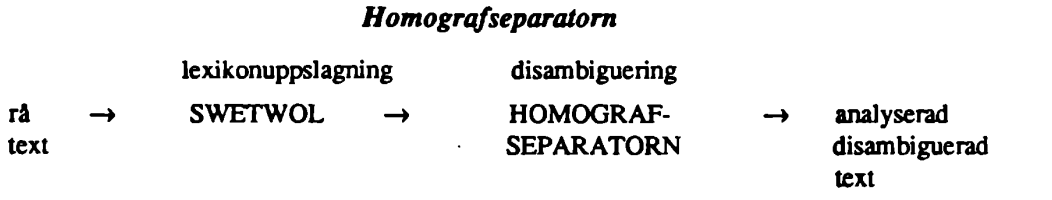
När ett lexikons täckningsgrad ökar, ökar också chansen att en mindre frekvent tolkning av en ordform finns som alternativ i detta lexikon. Om det använda kategorisystemet skiljer mellan finita och infinita verbformer uppstår i många svenska verbböjningsmönster homografi mellan en preteritum- och en participtolkning.

Redan vid ett ganska blygsamt antal analysdistinktioner kan graden av flertydighet bli besvärande. Som jag redovisar nedan ger ett lexikon med hög täckning och ett system av 11 kategorier (i huvudsak motsvarande de traditionella ordklasserna) till resultat att 36% av orden i de använda texterna är minst tvåfalt homografa. Om man utökar analyskategorierna till att omfatta även traditionella morfologiska drag kryper homografigraden för texterna upp över 50%.

I arbetet med att bygga upp Stockholm-Umeå Corpus, SUC, konfronterades vi snart med detta problem. Projektet (se Källgren 1990) har satt sig före att bygga upp en korpus motsvarande de engelska Brown- och LOB-korpusarna. Korpusen ska innehålla (minst) 1 milj. ord och dessa ska vara försedda med en (entydig) lexikal och morfologisk analys. I största möjliga utsträckning ska analysen utföras automatiskt - även om den första miljonen ord kommer att vara kontrollerad av mänskliga ögon - och för att nå fram till detta mål ska projektet även utvärdera olika metoder för automatisk lexikal disambiguering. Flera metoder och algoritmer ska testas, såväl lingvistiskt regelbaserade som sådana som baseras på frekvens hos tidigare analyserade korpusar. Här kommer att rapporteras om en pågående utvärdering av en metod av den senare typen.

1. Homografseparatorn

I fig. 1 nedan visas homografseparatorns plats i analysprocessen. En text tilldelas i det första steget alla möjliga analysalternativ av lexikon. I nästa steg filtreras de mindre sannolika alternativen bort av homografseparatorn.



Figur 1

Det lexikon, SWETWOL, som projektet har fått möjlighet att använda baseras på Koskenniemis tvånivåmodell och är uppbyggt vid enheten för datorlingvistik vid institutionen för allmän språkvetenskap i Helsingfors (se Koskenniemi 1983 resp. Karlsson 1992).

1.1 Utgångsmaterial för statistik

För att kunna utarbeta en metod som ska baseras på tidigare förekomst i text behövs naturligtvis tillgång till tidigare analyserade texter. Jag har kunnat använda mig av Nusvensk frekvensordbok, NFO (Allén 1970), som innehåller frekvensuppgifter om 1 miljon löpord i Press 65-materialet och av Skrivsyntax, en mindre korpus bestående av texter från fyra olika genrer (se Teleman 1974). Under det fortsatta arbetet med SUC-korpusen kan de redan analyserade texterna i korpusen tillsammans med ovanstående material utgöra underlag för ny (och förbättrad) statistik.

Från dessa två källor har två sorters statistik hämtats. Från NFO har hämtats alla ord som enligt min kategoriuppsättning är homografa, 2404 ord. Från Skrivsyntax har jag hämtat statistik om samförekomst mellan kategorier. Skrivsyntax' korpus innehåller, i den form jag har använt den, 105 529 "lexikala enheter" dvs ord, lexikaliserade fraser och skiljetecken. I exemplen (1) och (2) nedan ges exempel på dessa olika typer av frekvenser.

(1)	*toppar*	verb pres	4
		substantiv pl	8
		totalt	12

(2)	substantiv – verb	4354
	pronomen – verb	3260
	substantiv – substantiv	1112

I (1) kan man alltså notera att ordformen *toppar* förekommer totalt 12 gånger i Press 65, därav 4 gånger som substantiv i plural, de övriga 8 gångerna som verb i presens. I (2) visas att ett verb föregås av ett substantiv i 4354 fall i Skrivsyntaxkorpusen, pronomen föregår verb i 3260 fall och ett substantiv föregås av ett annat substantiv i 1112 fall.

1.2 Övergångssannolikhet och lexikal sannolikhet

Processen kan översiktligt beskrivas på följande sätt: Homografseparatorn väljer den bästa tolkningen av en viss ordform genom att av den relativa frekvensen dra slutsatser om den mest sannolika tolkningen. Nedan är alltså nämnaren i exempel (3) den totala förekomsten av ordformen *toppar* och i (4) det totala antalet substantiv. De två typerna av information kombineras i (5) genom att de multipliceras samman och separatorn väljer den tolkning som har den högsta sammanslagna sannolikheten.

(3) lexikal sannolikhet (P-lex):

toppar	
V	4/12 = 0.33
N	8/12 = 0.67

(4) övergångssannolikhet (P-överg):

<entydigt ord>	<flertydigt ord>	
N	V	4354/21562 = 0.20
N	N	1112/21562 = 0.05

(5) kombination av lexikal sannolikhet och övergångssannolikhet:

<entydigt ord>	*toppar*	
N	V	0.33 * 0.20 = 0.066
N	N	0.67 * 0.05 = 0.034

1.3 Disambiguering av räckor av flertydiga ord

Val av den troligaste tolkningen för vart och ett av orden i en längre räcka, ett "spann" (se (6) nedan), av flertydiga ord kan i princip fortgå på liknande sätt. Sannolikhetsvärdet för varje tolkning beräknas som ovan och värdet för varje möjlig sekvens av analyser beräknas genom att de ingående alternativens värden multipliceras. Algoritmen beskrivs punktvis nedan.

(6) Ett spann innehållande N antal ord och kategoriema A-I. ... föreslagna av lexikon.

ord1	ord2	ord3	...	ord(N-1)	ordN
	B	D	...	G	
A	E		...		I
	C	F	...	H	

1) Bilda alla möjliga sekvenser av kategoriförslag för spannet:

A	B	D	...	G	I
A	B	D	...	H	I
A	B	E	...	G	I
...			...		
A	C	F	...	H	I

2) Beräkna sannolikhetsvärdet för varje sekvens genom att multiplicera de enskilda alternativens. Nedan visas alternativsekvensen A-B-D-...-G-I.

ord1	A	
ord2	B	[P-lex(B,ord2) * P-överg(A,B)] *
ord3	D	[P-lex(D,ord3) * P-överg(B,D)] *
...	...	[...] *
ord(N-1)	G	[P-lex(G,ordN-1) * P-överg(G,...)] *
ordNI		[P-överg(G,I)] = sannolikheten för kombination 1.

3) Välj kombinationen med det högsta sannolikhetsvärdet.

En algoritm som denna är mycket resurskrävande. Homografseparatorn använder en annan algoritm presenterad i DeRose 1988. Hans algoritm, VOLSUNGA, bygger på "dynamisk programmering" och löser problemet mer effektivt. Den baseras på följande iakttagelse: Om ett visst analysalternativ ingår i den bästa sekvensen av alternativ för hela spannet måste även den bästa delsekvens som föregår analysalternativet ingå i den bästa sekvensen. Detta innebär att man för varje tolkningsalternativ bara behöver spara den bästa sekvensen av tolkningar fram till detta

alternativ. I spannet i (6) ovan innebär användningen av DeRoses algoritm att vid processningen av ord4 endast två olika kombinationer av tolkningar (nämligen den bästa vägen till D och den bästa vägen till F) behöver analyseras för vart och ett av detta ords alternativ. Detta ska jämföras med de sex sekvenser som den primitivare algoritmen måste hantera.

2. *Utvärdering*

En utvärdering av denna metod för lexikal disambiguering har påbörjats. Utvärderingen ska baseras på homografseparatorns analys av en större mängd texter. Dessa texter är helt skilda från de texter som utgör bas för den statistik som separatorn använder. I denna artikel utvärderas analysen av ett mycket litet antal ord, 3103 st, fördelade på två olika texter.

Den kategoriuppsättning som har använts är begränsad och innehåller 11 kategorier som ungefärligt motsvarar de traditionella ordklasserna och 3 kategorier för olika typer av skiljetecken. Detta innebär som tidigare nämnts att homografseparatorn ges en enklare uppgift jämfört med disambigueringen av alternativ från ett mer finindelande system. Det finns två skäl till detta. Det första är rent praktiskt: De använda kategorierna är skärningsmängden av de kategorier som används av SWETWOL, NFO och Skrivsyntax. Det andra skälet är att vi ville undersöka hur mycket ett så pass grovmaskigt nät kunde sålla bland alternativen i ett mer omfattande system. Kategorierna är alltså inte desamma som används för analys av SUC-korpusen. Fig 2 visar kategorisystemet och fig. 3 resultaten av denna första utvärdering.

Figur 2 Använda kategorier

substantiv	N	konjunktion	KONJ
verb	V	preposition	PREP
adjektiv	A	pronomen	PRON
adverb	ADV	räkneord	NUM
interjektion	INTERJ	infinitivmärke	INFMARK
egennamn	<Prop>		

och

"stora" skiljetecken (.?!:)	CB
komma (,)	IK
övriga skiljetecken	IG

Figur 3 Resultat (3103 ord)

instanser av ambiguitet / totalt antal ord	1109/3103	36%
korrekt analys / ambiguiteter	974/1109	87%
felaktig analys / totalt antal ord	135/3103	4%
kvarvarande ambiguitet		0%

Man kan notera att homografseparatorn vid analys av de ca 3100 orden väljer en riktig analys i 87% av de fall den utsätts för. Till detta kommer 1994 ord som får en entydig analys av lexikon och där homografseparatorn alltså inte appliceras. Av dessa får ett fåtal en felaktig lexikonanalys men totalt sett är ändå närmare 96% av orden i texten korrekt analyserade efter separatorns arbete.

2.1 Några exempel på analyser

Nedan redovisar jag några exempel från analysen. A-exemplen visar det flertydiga spannet före disambigueringen, b- och c-exemplen vilken analys homografseparatorn valt.

I (7a) visas ett spann som innehåller 7 flertydiga ordformer i följd och de sammanlagt 288 möjliga analyserna för spannet. (7b) visar homografseparatorns korrekta analys.

(7a)

iakttar	*vad*	*som*	*händer*	*på*	*andra*	*håll*	*i*	*världen*
V	ADV	KONJ	V	PREP	NUM	N	PREP	N
	N	PRON	N	ADV	PRON	V	ADV	
	PRON				V			

(7b)

iakttar	*vad*	*som*	*händer*	*på*	*andra*	*håll*	*i*	*världen*
V	PRON	PRON	V	PREP	PRON	N	PREP	N

I (8a) kan man notera att en ordform, *bedragnade*, inte ges något förslag till analys av SWETWOL-lexikonet. Homografseparatorn hanterar av lexikonet oanalyserade ord på enklast tänkbara sätt. En ordform som saknar analysförslag görs av homografseparatorn till ambiguöst mellan alla kategorier i det använda kategorisystemet, f.n. 14 stycken. Detta innebär alltså att disambigueringsprocessen i detta fall ska ta ställning till 112 möjliga analyser för spannet. (8b) visar den (nästan) korrekta analysen. Strikt bedömt skulle *bedragnade* ha analyserats som V enligt mitt kategorisystem men jag väljer att godkänna även adjektivanalysen eftersom att bestämma perfekt participformers ordklasstillhörighet inte är ett problem bara för homografseparatorn!

(8a)

ej	*bedragnade*	*band*	*av*	*dubbla*	*löften*
ADV		V	PREP	V	N
		N	ADV	A	

(8b)

ej	*bedragnade*	*band*	*av*	*dubblal*	*öften*
ADV	A	N	PREP	A	N

(9) är exempel på en typisk svaghet hos separatorn. Att avgöra om *som* ska analyseras som KONJ eller PRON klarar den sällan!

(9a)

världen	*som*	*blott*	*en*	*illusion*
N	KONJ	ADV	ADV	N
	PRON	KONJ	N	
		A	PRON	

(9b)

världen	*som*	*blott*	*en*	*illusion*
N	KONJ	ADV	PRON	N

Som nämnts hanteras oanalyserade ordformer mycket enkelt. För att förbättra homografseparatorns prestation är det möjligt att använda sig av ett flertal heuristiska regler av typen: identifiering av ändelser, användning i texten av versaler/gemena, etc. Ingenting av detta är implementerat i den nuvarande versionen av homografseparatorn. C-exemplen nedan visar att effektiviteten kan förbättras genom att bara begränsa de kategorier som kan tillskrivas ett oanalyserat ord. I dessa exempel har separatorn bara bedömt sannolikheten för att ett oanalyserat ord ska tillhöra någon av de öppna kategorierna: N, A, V, <Prop> eller ADV, medan b-exemplen väljer bland 14 kategorier inklusive skiljeteckenkategorier som CB (clause boundary).

(11a)					
.	*Nibelungen*	*Alberich*	*rövar*	*rhenguldet*	*och*
CB			V		KONJ
			N		
(11b)					
.	*Nibelungen*	*Alberich*	*rövar*	*rhenguldet*	*och*
CB	N	**PREP**	N	**CB**	KONJ
(11c)					
.	*Nibelungen*	*Alberich*	*rövar*	*rhenguldeto*	*ch*
CB	N	**V**	V	N	KONJ
(12a)					
.	*Nibelungarnas*	*fångenskap*			
CB		N			
(12b)					
.	*Nibelungarnas*	*fångenskap*			
CB	PRON	N			
(12c)					
.	*Nibelungarnas*	*fångenskap*			
CB	N	N			

3. Sammanfattning

Redan dessa preliminära försök visar att man kan nå överraskande goda resultat med denna enkla metod. Vi är medvetna om att när vi gradvis utökar antalet kategorier kommer de goda resultaten att försämras men samtidigt kommer korpusen ständigt att växa vilket ger oss ett växande material att basera och därmed förbättra statistiken på. Det allt större materialet ger oss också möjlighet att pröva mer sofistikerade metoder för statistiskt baserad disambiguering. Vi bedömer det vara fullt möjligt att uppnå 96-97% ordklasskorrekthet även med ett utökat kategorisystem, men hur mycket längre kan man komma?

Det har, oss veterligt, heller ännu inete någonstans, gjorts några försök att disambiguera morfologiska särdrag med samma metoder som här har använts för disambiguering av ordklasskategorier. Sådana försök står vi i begrepp att utföra inom SUC-projektets ramar.

Referenser

Allén, S., 1970. *Nusvensk frekvensordbok.* Almqvist & Wiksell. Stockholm.

DeRose, S.,1988. *Grammatical Category Disambiguation by Statistical Optimization.* Computational Linguistics, 14,1.

Karlsson, F., 1992. *SWETWOL: A Comprehensive Morphological Analyzer for Swedish.* Nordic Journal of Linguistics 1:1992.

Koskenniemi, K., 1983. *Two-level Morphology: A General Computational Model for Word-form Recognition and Production.* Publications of the Department of General Linguistics, University of Helsinki, No. 11.

Källgren, G. 1990. 'The first million is hardest to get'. COLING 1991.

Teleman, U., 1974. *Manual för grammatisk beskrivning av talad och skriven svenska.* Studentlitteratur. Lund.

Gunnar Eriksson
Institutionen för lingvistik
Stockholms universitet
S-106 91 Stockholm
E-mail: gunnar@ling.su.se

Syntax and prosody in a Danish Text-to-Speech System

Peter Molbæk Hansen & Ebbe Spang-Hanssen
University of Copenhagen

0. Introduction

This paper is structured as follows:

In section 1. we present the basic structure of a Danish text-to-speech system (henceforth TTS-system), and illustrate some of the special needs of TTS-systems as compared with other natural language processing systems (NLP-systems). In this connection we exemplify some cases of syntactically determined stress loss in Danish.

In section 2. we present and discuss some preliminary results of an empirical investigation of the relation between syntactic structure and pause distribution in read-aloud Danish prose.

In section 3. we describe a specialised grammar formalism for expressing syntactic as well as morphological and phonological structure, and we outline the parsing strategy used to transform input sentences to phonetically consistent and prosodically adequate transcriptions with special emphasis on how to achieve robustness.

1. Syntactic Aspects of the TTS-System

APPENDIX 1 is a schematic overview of the data flow through the serially connected components which converts strings of ASCII-symbols to sound waves in a stepwise fashion. The component labeled SPECIALIZED PARSER (SSPS FORMALISM INTERPRETER) takes care of the lexico-morphological and syntactic analysis needed in order for the system to extract sufficient linguistic information from the input text. This component, henceforth referred to as SSPS (Surface Structure Parsing System), is the one which we hope to be of most interest to computational linguists, and we will not comment on the other components at all, although most of them are of course most important for the text-to-speech conversion process.

The task of parsing text in TTS-systems is different from that of parsing text in systems aiming at assigning semantic structure to input text. APPENDIX 2 highlights some important differences between TTS-systems and other (typically semantically oriented) NLP-systems. The existence of such differences has, of course, rather direct consequences for the design of parsers, since the information to be extracted from the input text is partly different. The task of SSPS is basically to convert a text string representing an input sentence into a representation which contains all phonetically relevant properties of the sentence (prosodic as well as segmental) in string form, i.e. as some sort of systematic phonetic representation. In most older TTS-systems this task is solved

by a set of letter-to-phoneme rules, see e.g. Carlson & Granström (1975), which have no (systematic) access to higher-level linguistic information as such. Although several early TTS-systems can produce comprehensible output utterances practically without any syntactico-semantic knowledge, they do not sound "intelligent", and there is much to be gained in synthetic speech quality by introducing such information, since in many cases semantic, pragmatic, and syntactic structure is more or less directly reflected in the prosodic structure of the spoken sentence, determining e.g. the distribution and relative length of pauses, the distribution and relative prominence of stresses, and the intonation contours of whole utterances.

The ideal TTS-system would therefore have to include at least as much linguistic information as the ideal monolingual "text-to-conceptual structure" system. Since, however, a good comprehensibility will suffice for many practical applications of TTS-systems, it seems reasonable to set the level of ambition somewhat lower, and try to see how far syntactic (and of course lexical and morphological) knowledge will get us in the direction of producing intelligent sounding output, before attempting to include semantic and pragmatic information.

In Danish this makes all the more sense, since there are regular patterns of stress loss in nouns and verbs and of potential pause placement at syntactic boundaries which are almost completely predictable from the syntactic context. APPENDICES 3 and 4 illustrate the most important stress removal patterns of this kind.

Note that the conditioning factors as far as stress loss is concerned are signaled in syntactic surface structure: properties such as the definiteness of noun phrases and the transitivity of verbs are easily extractable from text, once morphosyntactic and lexical information is made available through lexico-morphological and syntactic parsing. Note also that the type of stress loss considered here is not a matter of phonetic degree, but a discrete syntactico-phonological matter, viz. absence vs presence of main word stress in the stressable syllable of the words in question. Most languages (including Danish) have regularly unstressed function words such as articles and prepositions, but stress loss in major class words (verbs and nouns) is a pecularity of the Scandinavian languages, the Danish version being probably the clearest example of this phenomenon. This fact in itself clearly necessitates a syntactic analysis.

Needless to say, the actual phonetic manifestations of stressed and unstressed syllables of such words also depend in part on the syntactic context, and this is also true of the rules for placing pauses, but here basic empirical research is needed to determine these rules themselves. This will be the subject of the next section.

2. Pause Placement

2.1 General Principles

It is not surprising that all linguists who have worked on pause location in speech synthesis agree that the good use of pauses highly increases the quality of the speech produced by the machine. This is also our experience. Until now, we have only manually made some pauses in a few sentence representations, but it is quite clear that even a very limited use of pauses makes synthetic speech much more natural.

This means that our work with pause location has a good practical justification, but it should be mentioned that it also is of considerable theoretical interest. The problem concerns the relations between the various components of language:

- pragmatics: discourse structure.
  Do notions like topic and focus play a role in the determination of pause sites?

- semantics: logic structure.
  Does predicate argument structure play a role?
  Are pauses more likely to occur between a
  predicate and its modifier than between a predicate and its arguments?

- syntax: constituent structure.
  Are pauses most likely to be found at important
  phrase boundaries?

- prosody: prosodic units determined by length and stress.
  Is rythm more important than syntax?

At first, it seems reasonable to think that pragmatics and semantics must be the decisive factors. It is a common view that pauses and units of meaning are connected.

This view is reflected, for instance, in the official Danish guide to punctuation. The traditional Danish punctation, whereby we understand here the use of commas, is a grammatical punctuation that Denmark has in common with Germany and the Eastern European countries, whereas the Western European countries, including Norway and Sweden, use a so-called pause punctuation or meaning punctuation.

In the discussion about the commas we clearly see the opposition between on one hand syntax, which is accused of having only poor links to pause location, and on the other hand the meaning, which is connected to the true pause location. However, when the authorities work out the practical rules for pause punctuation, these rules are coined as very syntactic looking rules: put the comma between sentences, in enumerations, around appositions and so forth. At least, syntax is used to indicate where it is possible to put the commas.

As computational linguists we must hope that syntax plays an important role, since syntax is much easier to handle in a machine than semantics, and, as a matter of fact, it turns out that recent research seems to show that purely morphological and syntactic means will be sufficient to make useful pause determination. This does not mean that traditional German-Danish grammatical comma rules are reliable in all respects relevant for pause placement, nor that it is possible to make the ideal pauses without an understanding of the text. What we claim is that it is possible to solve an interesting part of the problems in a purely syntax based system.

2.2 A Short Look at Recent Research

The point of view expressed by J. Bachenko and E. Fitzpatrick (1990) represents one major trend in American research in this field:

Our current analysis rests on two ideas. First it is possible to describe a level of prosodic phrasing that is independent of discourse semantics. Second, this discourse-neutral phrasing depends on a mix of syntactic and nonsyntatic factors; chiefly, syntactic constituency, left-to-right word order, and constituent length. There is no necessary fit between syntactic structure and phrasing, since prosodic phrasing may ignore major syntactic boundaries in order to satisfy the constraints on phrase length. Bachenko and Fitzpatrick, (1990, p. 155).

The authors thus continue the work done by E. Selkirk (1984) and Gee and Grosjean (1983). In their opinion, the most important component of language in this respect is prosody, but prosody, in its turn, is based on syntax, word order and length. This means that syntax is considered as an important, but indirect factor.

At present, we are inclined to think that the American prosodic rules are not fit for our purposes. In order to build the prosodic phrases, the American linguists work in a very buttom-up fashion, starting with very small phonological units. They end up building a prosodic phrase structure which is about as fine-grained as a normal constituent structure. However, if we consider only the observed

pauses, this whole structure seems very hypothetical. By far the majority of the prosodic boundaries they set up are not marked by any pause when people read aloud at a normal speed.

The hypothetical character of these rule systems can be illustrated by the basic rule in Gee and Grosjean (1983), the verb balancing rule, which has been taken over by Bachenko and Fitzpatrick. The verb balancing rule says that the first grouping of major constituents is made around the verb. If the constituent to the left of the verb is short it is grouped with the verb, and a major boundary is made after the verb. On the other hand, if the constituent preceding the verb is longer than the constituent that follows, then a break is made before the verb, and the verb is grouped together with the following unit:

A. (This little incident) // gives (a new zest) (to our investigation)
B. (Chickens) were eating // (the remaining green vegetables)

In this notation, the parentheses indicate the primary prosodic units that have first been identified. The double slash marks the boundary produced by the verb balancing rule.

This rule explains of course why there is often a break before the verb when the preceding phrase is long. But it is impossible to see that there ever is a boundary after the verb when the preceding phrase is short. As a matter of fact, there seldom is a break after the verb. But that is explained in the theory of Gee and Grosjean by the existence of a subsequent rule that builds this first big phrase into still bigger phrases, see APPENDIX 5.

The bigger the constituent, the more important the boundary. The algorithm builds the phrases mechanically in a left-associative way, and the last boundary will always be the most important. Thus, it seems to us that the main effect of the much praised verb balancing rule is to say that there is a break before the verb if the first constituent is long, and that the chances of a new break increase the further away we get from the first break. Bachenko and Fitzpatrick do not hide the fact that it is difficult to observe the verb balancing rule:

During subsequent processing, this balancing effect is usually lost since neither length nor adjacency to a verb play any further role in Gee and Grosjean's analysis (1990, p.162).

Nevertheless, Bachenko and Fitzpatrick stick to the verb balancing rule, to which they propose some amendments, in particular a verb adjacency rule that measures the distance from the verb. They attach so much importance to the prosodic phrasing that they think it rules out clausal constituency. Examples like the following show, they think, that clausal boundaries are unimportant:

Even my fiancee // believes it's only my imagination.

2.3 Our Own Investigations

Our preliminary investigations seem to show that parsing problems play a somewhat more important role than recognized by the American linguists we have mentioned. This difference might be due to the languages we have examined. The Danish rules may be different from the English rules. But our disagreement with these American linguists may also stem from the fact that they identify the syntactic factor with syntactic constituency, whereas we think that the syntactic factor is a matter of how people parse sentences.

Pauses do not serve to indicate the boundaries of syntactic constituents. Perhaps they mostly serve to indicate where complex constituents end, and thus facilitate parsing. Going down into an embedded clause does not cause any parsing problem, but it is often a problem to pop up again at the right place. In parsing terms: it is often more difficult to pop than to push.

We have investigated 3 political reports from the Danish radio, corresponding to 8 pages of written text, and 2 articles from magazines, corresponding to another 8 pages of text. The political reports

from the radio were registered with a tape recorder and later on written down. The two articles were read by a colleague who did not know the purpose of the investigation.

Our main findings are shown in APPENDIX 6. There are boundaries where speakers nearly always make a pause (obligatory sites), and there are others where it is just possible to do it (potential sites). In the latter case, speakers will use the opportunity to make a pause, if the boundary comes a long time after the preceding pause.

We have listed the various cases in an order corresponding to the relative frequency with which the type of boundary in question is used for locating a pause. Pauses occur very frequently around appositions, quotes and the like (37 to 8), somewhat less frequently after a coordinate term that is not an S, etc.

The categories enumerated from 3 to 8 can be subdivided so that some subcategories will get a very high priority. It seems to be the case that there is a pause before a clause if this clause in its turn starts with an embedded clause, as is the case in example 7a below (at når..).

As to PP's, it does not seem to matter that much whether the PP is a complement or a modifier. It is more important how far it is from the head of the construction, i.e. the verb or the head noun in an NP.

The examples listed in APPENDIX 7 illustrate the 8 main syntactic boundaries which we have examined and which are listed above. Very few pauses can not be classified as belonging to any of the 8 categories.

3. The SSPS Formalism and Parsing Strategy

3.1 General Features

Once the rules governing stress and pause placement are established, they must be expressed in a way which is compatible with the "normal" syntactic rules which treat linguistic expressions as strings of terminal symbols in the usual Chomskyan way, and the effects of such rules must be introduced in the parsed output. To this end we use a specialised formalism: SSPS.

An early version of the SSPS formalism is described in Molbæk Hansen (1989) and in Molbæk Hansen (1991). The current version is outlined in Molbæk Hansen et al. (1991). APPENDIX 8 illustrates some of the main features of the current version, which is basically a phrase structure grammar augmented with facilities for expressing restrictions on sister constituents and for expressing feature percolation from daughter nodes to mother nodes. Each rewrite rule has optionally associated with it one or more indented lines expressing such restrictions and/or percolations. The first restriction associated with the rule

NP ::= DET? AP* NOUN

viz. the indented line

NOUN > DEF C "DEFI"

expresses a **restriction**, namely that if the DEF and NOUN constituents are both present, then the values of the feature attribute DEF (definiteness) for DET (DET > DEF) must be compatible (C) with the absolute value DEFI (indefinite)

The last indented line associated with the same rule, viz. the indented line

NP > DEF : "DEFD"

expresses a **percolation**, namely the assignment to the mother node NP of the absolute value

DEFD (definite) of the feature attribute DEF. The reader is referred to Molbæk Hansen et al. (1991) for a fuller description of the notational conventions. In the current version attribute registers whose scope is global within a whole (sub)tree (cf. the hold registers known from e.g. Augmented Transition Networks) are not implemented, but this facility can be easily added, if the need arises.

In SSPS the same formalism is used to express both morphological and surface syntactic constituent structure, and Danish syntax and morphology actually form two sections of the same grammar delimited by a single line of the form

MORF

marking the beginning of the morphological section.

The chart-based parser analyses the input text according to this grammar and its associated morph lexicon, and it schedules its analysis as shown in APPENDIX 9. First each word of an input sentence is analysed in a top-down, first-rule-first fashion by the morphological section of the grammar, and in this morphological mode up to 4 interpretations are accepted for a word. When all words are analysed, the parser mode is switched to syntactic mode, in which the results of the morphological analyses are taken to be terminal edges from the outset. This strategy permits the use of optimization facilities such as precompiled left-corner tables and Kilbury-strategy compatible with bottom-up parsing, cf. Wiren (1987). Note, however, that the parser will only switch mode if the existence of a morphological grammar section is indicated in the grammar. Otherwise, the whole input string (including white space) will be parsed according to the grammar considered as a monolithic whole. Under normal circumstances, however, the grammar prescribed separation of syntax and morphology makes for faster parsing.

3.2 Demarcation of Prosodic Structure

Syntactically conditioned phonological phenomena (like stress loss, pauses, intonation markers etc.) are introduced in the parser output as linearized output symbols of lexical entries matching with zero-length input symbols, which means that they can be introduced everywhere in the string. The technical details of this facility are described in Molbæk Hansen (1991). APPENDIX 10 shows a parse of the input sentence "en sur tjener taber altid ansigt" with stress loss on the finite verb. The situation after morphological parsing is that each word has several interpretations, each with a unique combination of string representation and morphosyntactic feature values (the abbreviations enclosed in < >). In the example the third input word has two noun interpretations viz. the stressed and unstressed variants of the (deverbal) noun "tjener" ('waiter') and two verb interpretations, viz. the stressed and unstressed variants of the present tense verb form "tjener" ('earns'). The unstressed variants have string representations ending in , (comma) and the value STR0 of the feature attribute STR. The syntactic rule

```
S ::= SUBJ VERB ADVPHR* OBJ?
      OBJ > DEF C "DEFI"
      VERB > STR C STR0
      ?
      VERB > STR C STR1
```

states that a sentence may consist of a subject + a finite verb + optional adverbials + an optional object. The restrictions state that if the object is present and has the value DEFI, then the verb must have the value STR0, otherwise the verb must have the value STR1. Such a rule will ensure that the variant tab*0r, will be chosen in the example sentence, and the comma is interpreted by the phonological system as a symbol which prevents stress from being assigned to the preceding word. (For details of the cooperation of the parser component and the phonological transformer component of APPENDIX 1, see Molbæk Hansen (1991)).

3.3 Robustness

A TTS-system must be robust, both in the sense that unsuccesful parses should not be tolerated, and in the sense that the output should be sensible in cases of unidentifiable input. Many modern parsers take care of robustness in a preprogrammed way, typically by chosing the longest partial subcharts in cases where there are unparsed islands in the input, see e.g. Russi (1991). In SSPS **robustness is obtained in grammar prescribed ways.** The grammar writer may base his rules on e.g. his knowledge of frequent constructions. He may then write meaningful rules for such constructions and write **default rules** for configurations not compatible with these rules. Default rules are rules which are matched by any material at the relevant level if they are allowed to apply, that is, if the "structured" rules are not matched. This presupposes a first-rule first strategy.

APPENDIX 11 illustrates the principle, showing a morphological and a syntactic default rule. This principle enables the grammar writer to experiment with various arrangements (orderings) of the same rules and to evaluate the overall performance of each arrangement on a representative sample of test sentences. We are at present engaged in developing such a test sample.

References

J. Bachenko and E. Fitzpatrick 1990: A Computational Grammar of Discourse Neutral Prosodic Phrasing in English, *Computational Linguistics* Volume 16 Number 3 September 1990 p. 155-170.

R. Carlson and B. Granström 1975: A Text-to-Speech System based on a Phonetically Oriented Programming Language, *Speech Transmission Laboratory, KTH Stockholm, Quarterly Progress and Status Report* 1/1975 p. 17-26.

J.P. Gee and F. Grosjean 1983: Performance Structures: A Psycholinguistic and Linguistic Approach. *Cognitive Psychology* 15 1983 p. 411-448.

F. Grosjean and J.P. Gee 1987: Prosodic structure and spoken word recognition, *Cognition* 25 1987 p. 135-155.

F. Grosjean and M. Collins 1979: Breathing, Pausing and Reading, *Phonetica* 36 1979 p.98-114.

F. Grosjean, L. Grosjean and H. Lane 1979: The Patterns of Silence: Performance Structures in Sentence Production, *Cognitive Psycology* 11 1979 p. 58-81.

P. Molbæk Hansen 1989: Syntax, Morphology, and Phonology in Text-to-Speech Systems, *ARIPUC 23 = Annual Report of the Institute of Phonetics, University of Copenhagen* No 23 1989 p. 119-152.

P. Molbæk Hansen 1991: The Linguistic Components of the Danish Text-to-Speech System, *Copenhagen Working Papers in Linguistics* volume 1. 1990/91 p. 153-162.

P. Molbæk Hansen, N. Reinholt Petersen, J. Rischel, and C. Henriksen 1991: Higher Level Linguistic Information in a Text-to-Speech System for Danish, *EUROSPEECH 91 = Proceedings og The 2nd European Conference on Speech Communication and Technology, Genova, Italy, 24-26 September 1991*, Volume 3 p. 1243-1246.

T. Russi 1991: Robust and Efficient Parsing for Applications such as Text-to-Speech Conversion, *EURO-SPEECH 91 = Proceedings og The 2nd European Conference on Speech Communication and Technology, Genova, Italy, 24-26 September 1991* Volume 2 p. 775-778.

E.O. Selkirk 1984: *Phonology and Syntax: The Relation between Sound and Structure.* MIT Press, Cambridge, 1984.

M. Wirén 1987: A Comparison of Rule-Invocation Strategies in Context-Free Chart Parsing, *Proceedings of The Third Conference of the European Chapter of the Association for Computational Linguistics, Copenhagen, 1-3 April 1987* p. 226-233.

Peter Molbæk Hansen
Institut for Almen og Anvendt Sprogvidenskab
Københavns Universitet
Njalsgade 80
DK-2300 København S

E-mail: pm@cphling.dk

Ebbe Spang-Hanssen
Institut for Almen og Anvendt Sprogvidenskab
Københavns Universitet
Njalsgade 80
DK-2300 København S

E-mail: esh@cphling.dk

APPENDIX 1: THE DANISH TEXT-TO-SPEECH SYSTEM:

ASCII text: De spiste altid middag kl. 19.

TEXT NORMALIZER (A TRIVIAL PREPROCESSING PROGRAM)

Normalized: de spiste altid middag klokken nitten

SPECIALIZED PARSER (SSPS FORMALISM INTERPRETER)

Morphophonemic: di,spis*t0,all+# tidd mIdd-a klokk-0n,nItt-0n

PHONOLOGICAL TRANSFORMER (SPL FORMALISM INTERPRETER)

Phonetic: disbi:sd0'all,tiD?m'edaklCg^nn'ed^n

PHONETIC TRANSFORMER (SPL FORMALISM INTERPRETER)

parameterized:

P1 P1 P1 P1 P1 P1 P1 P1 P1 .... P1 P1
P2 P2 P2 P2 P2 P2 P2 P2 P2 .... P2 P2

Pn Pn Pn Pn Pn Pn Pn Pn Pn .... Pn Pn

SYNTHESIZER (SOFTWARE SIMULATION)

Wave image: Fr1 Fr2.....Fr 1649 Fr1650

D/A-CONVERTER

Electric oscillations:

LOUDSPEAKER

Sound waves:

APPENDIX 2: TASKS IN TEXT-TO SPEECH SYSTEMS (TTS) AS OPPOSED TO OTHER SYSTEMS (O)

1. TTS: surface structure 1 -> surface structure 2
vs
O: surface structure -> some other structure

Example:
word inventory and order must be identical in input and output in TTS. (TTS-systems involve both analysis (identification) and generation.

2. TTS: Semantic structure is often irrelevant
vs
O: Semantic structure is usually essential

Example: semantic structure essential in translation systems

3. TTS: Prosodic features often essential
vs
O: Prosodic features often redundant

Example 1:
"Jens drak mælk" is distinguished from "Jens drak mælken" both morphologically and prosodically, but for O-parsers the prosodic difference is irrelevant.

Example 2:
In Danish many but not all stød susceptible monosyllabic first parts of compounds exhibit stød-loss. Thus "halgulv" with stød retension and "balsal" with stød loss have a different morphophonological structure. This difference is completely irrelevant in O-systems.

APPENDIX 3: STRESS LOSS IN NOUN PHRASES

(comma => next word unstressed)
(single quote => next word stressed)

Noun phrase internal constructions of various semantic subtypes:

a) Determiner phrase ending in indefinite noun:
ˌet 'stort ˌbundt 'friske 'radiser
to ˌfed 'hvidløg

b) Constructions with indefinite noun + proper name
ˌprofessor Spang-'Hanssen
ˌslagter 'Olsen
ˌtante 'Agate
ˌfætter 'Jens

c) Constructions with indefinite noun + numeral
ˌnummer 'et
ˌindgang '2
ˌtrappe '6

d) Constructions with indefinite noun + "place name"
ˌrestaurant Den 'Gyldne 'Fortun
ˌKap Det 'Gode 'Håb

e) Hypotactic proper name phrases:
ˌJens 'Peter
ˌOle 'Svendsen
ˌJens ˌPeter 'Jensen

APPENDIX 4: STRESS LOSS IN VERB PHRASES

(comma => next word unstressed)
(single quote => next word stressed)

a) Verb with indefinite direct object, irrespective of intervening material:
Jens ‚drikker aldrig 'fadøl
Jens ‚leder efter 'penge

vs

Verb with definite direct object, irrespective of intervening material:
Jens 'drikker aldrig ‚sin 'mælk
Jens 'leder efter 'pengene

b) Verb + adverbial complement, possibly with intervening direct object:
Jens ‚satte sin cykel 'ind i skuret

vs

Verb phrases with peripheral locative adverbial:
Jens 'satte sin cykel 'inde i skuret'

c) A combination of a) and b):
Jens ‚støvede 'bordet 'af

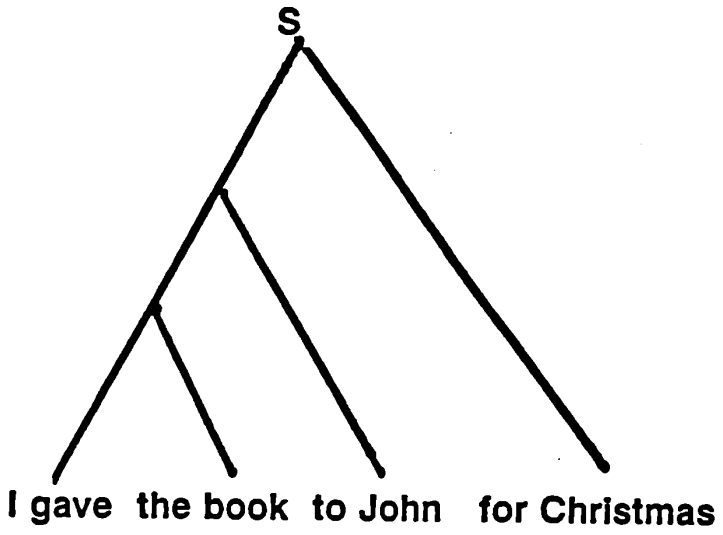
vs

Jens ‚støvede 'borde ‚af

APPENDIX 5: VERB BALANCING RULE

A. (This little incident) // gives (a new zest) (to our investigation)

B. (Chickens) were eating // (the remaining green vegetables)



APPENDIX 6: PAUSES IN WRITTEN DANISH READ ALOUD (16 PAGES)

	YES	NO
OBLIGATORY SITES		
1. After coordinated S	65	5
2. After embedded S	40	2
POTENTIAL SITES		
3. Around appositions, quotes	37	8
4. After coordinated non-S	25	17
5. After complex initial phrase	42	39
6. After complex inverted subject	10	6
7. Before clause	89	75
8. Before PP	41	252
TOTAL NUMBER OF PAUSES	343	
TOTAL NUMBER OF UNUSED SITES		404

APPENDIX 7: SAMPLE PAUSE LOCATIONS

1. Som barn elskede hun kirsebær, // og hun tog imod
posen med en selvfølge, der først bagefter har forundret
mig,// og gik rundt blandt de øvrige passagerer // og bød af sin rigdom. (D,2,1)

2. Men jeg oplevede tre ting den sidste dag, hun levede,// der stadig
fylder mig med undren. (D,1,8)

3. Ved Hitlers magtovertagelse den 30. januar 1933,//marcherede
brunskjorterne,// SA-korpsene //, i et firetimers fakkeltog
gennem porten. (R,3,5)

4. I mange landbrugsegne / er hele landsbyer på flugt// på
grund af mangel på mad og vand // og af frygt for epidemier. (R,6,27)

5. Men ved retten i München // kører en proces mod en af
Wolffs tidligere underordnede,// generalmajor Schutt. (R,5,1)

6. Så skulle tohundredeårsdagen for åbningen af Brandenburger Tor i
dag // være anderledes problemløs.(R,1,13)

7a. En nabo, der ønskede at hjælpe hende,// beklagede
sig over,// at når han kom forbi på vejen / og gerne ville
hilse på hende,// så vendte hun ryggen til. (D,3,11).

7b. Da han erfarede, at jeg endnu ikke havde sovet,//
insisterede han på at give mig en indsprøjtning at sove på. (D,3,28)
Derfor stod Markus Wolff naturligt nok højt på ønskesedlen hos de
vesttyske myndigheder // ved genforeningen. (R,4,13)

APPENDIX 8: SSPS GRAMMAR SECTION

```
NP ::=      DET? AP* NOUN
      NOUN > DEF C "DEFI"
      DET > DEF C AP > DEF
      DET > AN C AP > AN
      DET > AG C AP > AG
      AP > AN C NOUN > AN
      AP > AG C NOUN > AG
      DET > AN C NOUN > AN
      DET > AG C NOUN > AG
      NP > DEF : "DEFD"

DET ::=   QUANT? ART? ENUM?
      QUANT > DEF C ART > DEF
      QUANT > AN C ART > AN
      QUANT > AG C ART > AG
      ART > DEF C ENUM > DEF
      ART > AN C ENUM > AN
      ART > AG C ENUM > AG

QUANT    ::=   WRD
      WRD > W c "WKVA"

ART ::=   WRD? NP
      WRD > W C "WPRON"
      NP > W c "WGEN"
      ART > DEF : "DEFD"
      ART > A : "A"

ART ::=   WRD
      WRD > W C "WDEFR"

ENUM     ::=   WRD
      WRD > W c "WNUM"
```

APPENDIX 9: SSPS PARSER ALGORITHM

```
FOR EACH INPUT SENTENCE
BEGIN
 INPUT A SENTENCE.

 SWITCH TO MORPHOLOGICAL MODE = TOP-DOWN, DEPTH-FIRST,
 FIRST-RULE-FIRST, EACH CHARACTER SURROUNDED BY
 TWO CONSECUTIVE VERTICES.

 FOR EACH WORD
 BEGIN
 PARSE THE WORD ACCORDING TO MORPHOLOGICAL
 PART OF GRAMMAR, ACCEPTING AT LEAST 1 AND
 AT MOST 3 INTERPRETATIONS.
 END
 ARRRANGE THE RESULTS AS TERMINAL EDGES
 IN A CHART FOR THE SYNTACTIC MODE.

 SWITCH TO SYNTACTIC MODE = BOTTOM-UP, LEFT-CORNER,
 EACH WORD-RESULT SURROUNDED BY
 TWO CONSECUTIVE VERTICES.

 PARSE THE SENTENCE ACCORDING TO SYNTACTIC PART OF
 GRAMMAR, ACCEPTING EXACTLY 1 INTERPRETATION.

 OUTPUT INTERPRETATION IN
 LINEARIZED MORPHOPHONEMIC FORM.
END
```

APPENDIX 10: EXAMPLE OF ANALYSIS

Input: En sur tjener taber altid ansigt.

MORPHOLOGICAL ANALYSIS:

0-1 en <WNUM ANUMS AGENC DEFI>
0-1 en, <WDET ANUMS AGENC DEFI>
1-2 surr, <WVIMP STR0>
1-2 surr <WVIMP STR1>
1-2 sur! <WADJ ANUMS AGENC DEFI>
2-3 tjEn=0r, <WSUB ANUMS AGENC DEFI STR0>
2-3 tjEn=0r <WSUB ANUMS AGENC DEFI STR1>
2-3 tjEn+0r, <WVFIN STR0>
2-3 tjEn+0r <WVFIN STR1>
3-4 tab=0r <WSUB ANUMS AGENC DEFI>
3-4 tab*0r, <WVFIN STR0>
3-4 tab*0r <WVFIN STR1>
4-5 all+# tidd <WADV>
4-5 altid, <WSUB ANUMS AGEN DEFI STR0>
4-5 altid <WSUB ANUMS AGEN DEFI STR1>
4-5 altid <WADJ ANUMS AGENC DEFI>
5-6 anh%sIgt, <WVIMP STR0>
5-6 anh%sIgt <WVIMP STR1>
5-6 anh=sIgt, <WSUB ANUMS AGENN DEFI STR0>
5-6 anh=sIgt <WSUB ANUMS AGENN DEFI STR1>

SYNTACTIC ANALYSIS:
en,sur tjEn=0r tab*0r, all+# tidd anh=sIgt

OUTPUT FROM PHONOLOGICAL TRANSFORMATION:
ens'u:?rtj'E:nCtæ:bC'al?,tiD?'an,segd

APPENDIX 11: ROBUSTNESS IN THE SSPS-SYSTEM

UNSUCCESFUL PARSES ARE AVOIDED BY
1) FIRST RULE FIRST STRATEGY
2) DEFAULT RULES:

MORPHOLOGICAL DEFAULT RULE:

```
STEM      ::=   letr+
      STEM > M : "MNSAC MCC0 MNPX"
```

I.E. ANY LEXICALLY UNIDENTIFIABLE
PART OF AN INPUT WORD WILL
DEFAULT TO
1) A NOUN STEM,
2) OF THE COMMON GENDER,
3) WITH PLURAL IN -ER,
4) WITHOUT -E- OR -S- AS 1ST PART OF COMPOUND
5) WITH SPELLING = MORPHOPHONEMIC REPRESENTATION.

E.G. AN INPUT WORD LIKE "SMULPERNE"
WILL BE OUTPUT AS:

"smulp+0rn0" <WSUB ANUMP DEFD>

SYNTACTIC DEFAULT RULE:

```
S     ::=   WRD+
```

I.E. "A SENTENCE IS A SEQUENCE OF WORDS"

THIS PRINCIPLE ALLOWS ANY DEGREE OF MORPHOLOGICAL
AND SYNTACTIC EXPLICITNESS TO BE COMBINED WITH
A GIVEN PRACTICAL IMPLEMENTATION.

A new Dictionary of Swedish Pronunciation

Per Hedelin and Dieter Huber
Chalmers University of Technology

Abstract

This paper describes some aspects of a pronunciation dictionary for Swedish, "Svenskt Uttalslexikon" (SUL), which is presently developed at our department. This dictionary provides, among other items, three kinds of information about Swedish pronunciation that are not included in standard dictionaries: information on **variants**, on **inflected forms** and **compounds**, and on **proper names**. SUL is organized as a machine-readable lexical database which in its present form contains approximately 100.000 headwords. The run-time system comprises four separate processing modules: an inflection engine, a compound engine, the transcription engine, and the dictionary search algorithm. In addition to phonetic and phonological information, SUL also aims to supply various kinds of paradigmatic, syntagmatic and statistical information, needed for the linguistic processing stages in text-to-speech synthesis and automatic speech recognition.

1. Introduction

Large-scale, high-quality, machine-readable pronunciation dictionaries are a fundamental prerequisite for progress in various areas of speech and language technology. For instance, modern speech synthesis systems require access to a dictionary in order to look up the pronunciation of individual words and compounds which are to be converted from orthographic text to speech. The other way round, automatic speech recognition systems need pronunciation information from large-vocabulary lexical databases in order to match the incoming speech signal against possible word candidates.

However, while spelling-to-sound conversion (and its reverse) admittedly constitutes one important reason for the use of specialized pronunciation dictionaries in computer speech synthesis and recognition, it is certainly not the only one. Linguistic processing modules that perform syntactic and semantic parsing are becoming more and more integrated into both synthesis and recognition systems, with the purpose to supply relevant information for e.g. prosodic processing, pause marking, disambiguation, anaphoric resolution, and the detection and correction of erroneous input. In order to be able to support these modules in an efficient way, i.e. integrating both the speech processing and natural language processing stages of the overall system, various kinds of morphological, syntactic, semantic and even context-oriented information need to be included in the dictionary.

Ordinary dictionaries do not provide these different types of information within one single, comprehensive data structure that is immediately accessible to spoken language processing (SLP) in computer speech synthesis and recognition. For instance, conventional dictionaries, in as far as

they comprise any phonological information at all, usually include only one, i.e. the most common and least stigmatized (not necessarily the same!) lexical pronunciation for each headword in some kind of broad phonemic transcription. They thus ignore the inherent variability of natural human speech, dictated by influences of sociolect, dialect, and various features of speaker idiosyncrasy. Even more importantly, they do not provide any systematic information about alternative pronunciations that typically occur in continuous, connected speech, i.e. including different kinds (and degrees!) of assimiliation, reduction, neutralization, and elision of individual speech sounds, caused by influences of coarticulation, speech rate, and stress.

The intended user of a conventional dictionary is a human being. With this in mind, lexicographers have over the centuries aspired to optimize the selection, interpretation and presentation of the material to be collected in a dictionary. Machine-readable dictionaries, on the other hand, are not written primarily for intelligent human users, but have to deal with the natural language processing and speech signal processing modules of computer systems. This apparently trivial observation has a number of far-reaching consequences concerning both the selection and the organization of the lexical items. Thus, conventional dictionaries *implicitly* assume that their users have access to other sources of linguistic and factual knowledge, and are able to extract and combine information from various sources in a sophisticated and experienced manner. On the contrary, when computers access data in a machine-readable dictionary, all information must be stated *explicitly*, i.e. either by way of exhaustive listing, or retrievable by applying a set of well specified, formalized rules. As a consequence, dictionaries for applications in the fields of speech and language technology compared with traditional, human-oriented lexica, have to meet considerably higher de-mands in *explicitness*, *exhaustiveness* and *consistency*.

This paper discusses some aspects of a pronunciation dictionary for Swedish, "Svenskt Uttalslexikon" (SUL), which is presently developed at our department for use in computer speech synthesis and recognition. Several early versions of the work have been documented, i.e. (Hedelin 1986), (Hedelin 1987) and (Hedelin 1989). Certain additional aspects of this dictionary are also described in Huber (1990). This present article focuses mainly on the formalization and representation of *phonological* (chapter 3) and *morphological* (chapter 4) information within the SUL lexical database. The treatment of syntagmatic information in future extensions of SUL is discussed in a separate paper published within this volume (cf. Huber 1992).

2. *Lexicon Structure*

In standard computer terminology, SUL constitutes a lexical database which is organized in records and fields. Each record comprises a collection of various types of information pertaining to one lexical item, the lexeme. The records are again subdivided into fields storing, for instance:

- the *headword* in normal orthography;
- a *phonemic transcription;*
- information on the *word class (part-of-speech);*
- one (or several) *paradigm codes*.

In addition to these obligatory entries, SUL also contains fields for the storage of information on alternative spellings and pronunciations, on inflectional irregularities (both in the written and in the spoken domain), on semantics and etymology, homonym indices, collocations, and different kinds of statistical data. On the whole, each record in the SUL database comprises a total of 17 data fields. Two of the more important of these fields are briefly identified below:

Headword

The headword is listed in normal orthographic spelling. Acronyms, abbrevations and blends are treated as separate records. Homonyms (i.e. both homophones and homographs) are handled by using separate headwords marked with a homonym index and a number.

Word Class (Part-of-Speech)

Each headword is classified according to its normal usage into one of the categories *noun, pronoun, verb, adverb, adjective, numeral, preposition, conjunction, interjection,* and *article.* Words having multiple usage, i.e. pertaining to two of more word class cate-gories, are duplicated as headwords, i.e. one entry for each category. Subcategories (e.g. auxiliary verb, subordinating conjunction, personal pronouns) are not specified. Proper nouns, acronyms and abbreviations are marked accordingly.

The representation of information about pronunciation and morphology is described in detail in the following sections of this article. A brief authentic excerpt from the SUL database (displaying only the first 5 of the total of 17 fields) is shown for illustration purposes in table 1 below.

anammelse	/a'nammelsə/	s5	Avldng anamma+-else	
anamnes	/anam'ne:s/	s3		*medicin: sjukdomshistoria, sjukbikt*
anamnestisk	/anam'nɛstisk/	a3	Avldng anamnes+-tisk	*anamnes-*
anamnetisk	/anam'ne:tisk/	a3	Avldng anamnes+-isk	*anamnes-*
anamorfos	/anamor'fo:s/	s3		*ombildning*
ananas	/'annanas/	s9		*botanik: släkte av familjen ananas-växter; en frukt*

Table 1. Excerpt from the SUL database

3. Pronunciation

3.1 Background

Conventional dictionaries of Swedish, in as far as they provide any phonological information at all, generally seem to treat pronunciation in a rather ad hoc and untheoretical way. That is, older dictionaries, dating back to the beginning of this century or even earlier (e.g. Dalin 1850, SAOB 1898, Lyttkens & Wulff 1911) often use notation systems that do not conform with our present standards of either broad "phonemic" or narrow "phonetic" (i.e. allophone-oriented) transcription. More recent, general-purpose dictionaries (e.g. Molde 1977, Östergren 1981, Allén et al 1986, Malmström et al 1988, Collinder 1989) usually provide pronunciation information for a subset of headwords only, and often use notation systems based on graphemic analogies rather than on established phonetic transcriptions (e.g. IPA 1949). Only few dictionaries of Swedish include systematic information on pronunciation usage for all lexical entries, all of them, however, covering only a very limited and highly specialized subset of present-day Swedish vocabulary such as for instance geographical names (e.g. SprN 1977,1979, 1981). The only more recent, general-purpose Swedish dictionary which provides transcriptions systematically for all headwords is Lexin (1984), a 12.000-headword lexicon intended primarily for use in teaching Swedish as a second language. SFL-oriented information on the pronunciation of Swedish can also be found in a number of multi-lingual, foreign-learner dictionaries published abroad.

Until recently, the Swedish Language Council (Svenska Språknämnden) in Stockholm was preparing the publication of a major pronunciation dictionary of modern Swedish. This project has, however, been abandoned.

Clearly, there is an urgent need for the compilation of a large-scale, high-quality pronunciation dictionary of modern Swedish, which meets the standards of present-day phonetics and lexicology, and which needs to be provided both as a machine-readable electronic database for use in computer speech applications, and as a conventional hard-copy lexicon for interested human users. Similar dictionaries are already available or under development for a number of European (e.g. Wells 1990, Molbæk Hansen 1990) and non-European (e.g. EDR 1990) languages.

3.2 Notation

Principally, the transcription of pronunciation can be either narrow (allophone-oriented) or broad (phoneme oriented). For some applications (e.g. speech synthesis or speaker verification) a narrow, allophonic transcription may be desired, whereas for other applications (e.g. automatic speech recognition) a broad transcription may be more useful (cf. Huber 1990 for a more detailed discussion). Based on these requirements, SUL follows an essentially dual approach, viz:

- all transcriptions listed explicitly in the SUL database are basically *phoneme-oriented*,
- *allophone-oriented* transcriptions are generated by rule whenever required.

Phonetic (i.e. allophone-based) transcriptions of inflected word forms and compounds are thus performed first in the phoneme domain, before handling phoneme-to-allophone conversion by rule in the automatic transcription system. The resulting three-step procedure for generating both phonemic and phonetic transcriptions of inflections and compounds is illustrated below in table 2 for the inflected Swedish noun "gatorna" (Eng: the streets):

database storage	phoneme string	/ˈga:ta/
word inflexion	phoneme string	/ˈga:ta/ + /-ɷrna/ = /ˈga:tɷrna/
conversion rules	allophone string	[ˈgɑ:tɷɳa]

Table 2. Phoneme-to-allophone conversion in SUL.

3.2.1 Broad Transcription

The notation system used in SUL for transcriptions at the phoneme level is based on the following phoneme concepts:

Vowels:	/a/ /e/ /i/ /ɷ/ /ʉ/ /y/ /o/ /ɛ/ /ø/ /ə/
Consonants:	/b/ /d/ /g/ /m/ /n/ /ŋ/ /j/ /r/ /l/ /v/ /p/ /t/ /k/ /f/ /s/ /ʃ/ /ʂ/ /ç/ /h/
Prosody:	/ˈ/ /ˌ/ /²/ /ˈ/ /ˈ/ /ˑ/ /~/ /:/

Note that in order to be able to perform the phoneme-to-allophone conversion automatically and in both directions in a consistent and unequivocal manner, we had to make some con-cessions to the traditional way of defining the phoneme system of Swedish in standard textbooks of phonology. These concessions refer (i) to the extensive use of prosodic markers, and (ii) to the unorthodox classification of /ə/ and /ʂ/ as phonemes during the phonemic processing stages. On the whole, our

SUL phoneme transcriptions are more detailed than what is customary for a phoneme-based transcription system in other applications (cf. for instance Elert 1970 and Garlén 1988). As stated earlier, however, all phoneme labels in SUL are written with the inherent problems of automatic allophone conversion in mind. The overriding consideration in devising our system has been to provide for simple and safe phoneme-to-allophone mapping in both directions.

The phoneme strings for each headword are stored in the transcription field of the SUL database, using normal, lowercase ASCII characters: one ASCII character per symbol.

Note also that our SUL broad transcriptions mark stress and accent with the symbols placed in the *initial* position of a syllable, as illustrated in the following examples of Swedish words: "springer" (i.e. /'sprɪŋ ə r/), "springor" (i.e. / 'sprɪŋɔr/) and "springbrunn" (i.e. / 'sprɪŋ~²brʉn/). IPA (1949) follows the same convention, whereas Swedish phoneticians often tend to place the corresponding symbols at (or after) the stressed vowel.

The symbol /~/ as a prosodic marker is used for the dual purpose: (i) to classify a word as a prosodic compound, thereby indicating that the word is pronounced with a special version of the grave accent (accent II), and (ii) to mark the prosodic boundary. SUL thus writes for instance:

/'a:v¹ga:s~²rø:r/ and ['a:v¹ga:s~²ræ:r]

for the Swedish noun compound "avgasrör" (Eng: exhaust pipe) in the phoneme and in the allophone domain respectively. Likewise the dictionary lists:

/be'ta:lnɪŋ/,	[bɐ'to:lnlŋ]
/be`ta:lnɪŋs~ba²lans/,	[bɐ`to:lnlŋs~ba²lans]
/be`ta:lnɪŋsba'lans~prø²ble:m/	[bɐ`to:lnlŋsba'lans~prø²ble:m]

for compounds pronounced with the Swedish word accent II. On the other hand, whenever a morphological compound does correspond to a prosodic compound, as for instance in the Swedish words "trädgård" (i.e. /'trɛggord/), "Bergman" (i.e. /'bɛrjman/, note the prosodic distinction between the proper noun indicating the name "Bergman" and the ordinary noun "bergman" indicating an occupational group corresponding to the English "miner"), "Alingsås" (i.e. /alɪŋs'o:s/) and "Dalby" (i.e. /'da:lby/), the prosodic marker /~/ is *not* used. It must be appreciated in this context that our definition of a prosodic compound is not necessarily equivalent to a compound in the morphological sense.

Finally, the distinction between long and short *syllables* is maintained by the use of prosodic stress markers, whereas the (binary) distinction between long and short *vowels* is marked in the SUL phoneme system by placing the duration sign /:/ after the vowel symbol. The distinction between long and short *consonants*, on the other hand, is not maintained in the SUL system at all, i.e. the duration sign /:/ is used for vowels only. However, for any long syllable containing a short vowel and only one consonant, the consonant symbol is repeated in all phoneme transcriptions, such as for instance in /'kall/, /'takk/ and /'takka/, but not in /'kalt/ or /'takla/.

3.2.2 Narrow Transcription

SUL uses 46 allophone labels to describe the inherent variability of the Swedish phoneme system. These labels are listed and exemplified below in the tables 3 (vowels) and 4 (consonants). The authors apologize that for technical reasons it has not been possible to correctly reproduce all the diacritic symbols in this report (cf. appendix C).

[ɑ:]	*mat*	[a]	*att*		
[e:]	*vet*	[ɛ]	*vett*	[ə]	*året*
[i:]	*vit*	[I]	*vitt*		
[u:]	*bo*	[ɔ]	*bott*		
[ɯ:]	*hus*	[ɵ]	*hund*		
[y:]	*byt*	[Y]	*bytt*		
[o:]	*gå*	[ɔ]	*gått*		
[ɛ:]	*säl*	[ɛ]	*vätt*		
[æ:]	*här*	[æ]	*kärr*		
[ø:]	*hö*	[ø]	*höst*		
[œ:]	*hör*	[œ]	*förr* ·		

Table 3. SUL notation for the vowel allophones

[b]	*bo*	[d]	*dag*
[d]	*bord*	[f]	*fisk*
[g]	*gata*	[h]	*hel*
[j]	*ljuv*	[k]	*kunna*
[l]	*leta*	[l]	*kärl*
[m]	*meta*	[n]	*natt*
[ɳ]	*barn*	[ŋ]	*ting*
[p]	*peta*	[r]	*reta*
[s]	*sol*	[ʂ]	*fors*
[ʃ]	*sjö*		
[t]	*tak*	[ʈ]	*fart*
[ç]	*tjuv*	[v]	*veta*

Table 4. SUL notation for the consonant allophones

In addition to these symbols, SUL also provides a special notation to mark the pronunciation of diphthongs, which, although they do not form part of the Swedish vowel system, frequently occur in loan words of foreign origin, such as for instance *Audi*. These occurrences are transcribed on the allophone level by using (i) the two vowel allophones that most accurately describe the standard Swedish pronunciation of these sounds, combined with (ii) a subscript ligature, resulting in for instance [a͡ɷ] in the previous example.

Note also that the transcriptions in our dictionary frequently use the allophone [ɔ] in foreign loan words where the orthographic spelling uses "o" (e.g. in "abolition"). Some transcribers might prefer to use the allophone [ɷ] in these contexts. Another fine distinction in Swedish concerns the difference between the short allophones [ɛ̟] and [ɛ̠], which are often confused due to the dialect of Stockholm and its preference for [ɛ̟] in both /e/ and /ɛ/ contexts. In both cases, i.e. for the /ɷ/ - /o/ as well as for the /e/ -/ɛ/ distinction, SUL attempts to describe Standard Swedish (Rikssvenska) and follows the conventions established in Lyttkens & Wulff (1911).

4. Morphology in SUL

The inclusion and adequate representation of morphological information in a pronunciation dictionary, both for man and machine, serves several important purposes. Word inflection, both in the orthographic and in the phonetic domain is the main motivation. The generation of derived forms and compounds, again both in a written and a spoken version, is almost equally important. Computer systems without access to morphological data would need to resort to an exhaustive listing of all the potentially possible inflected forms for each lexeme in order to correctly associate incoming word candidates with their lexical heads as they appear in the dictionary. This kind of explicit listing is not only highly uneconomical (and unintelligent!) in practical applications, it would also be outright impossible with regard to the practically unlimited number of compounds and derivations contained in languages like Swedish.

Swedish is both an inflectional language and a language with a strong potential to form new words by compounding and by derivation. All three of these morphological processes may (and often do!) change the phonological realization of a word in a variety of ways, including for instance stress shifts, reductions, assimilations, elisions, syncope and epenthesis. These changes are governed by a large set of rules and, unfortunately, an even larger list of exceptions.

In analogy with the approach presented by Hellberg (1978), SUL attempts to capture the morphophonemic variability of Swedish in a representational format which integrates the changes of the words induced by compounding and derivation with their inflection. It must be appreciated in this context that the inflectional and derivational behaviour of words in Swedish, like in many other languages, is considerably more complex in the phonological than in the orthographical domain. For instance, stress shifts and changes in word accent patterns are manifested in spoken language only. The same applies to phenomena like vowel shortening or the realization of the mute final "e" in words like "garage". Morphological systems for use in spoken language processing therefore require a significantly larger number of paradigms than systems which operate solely in the domain of written language such as that of Hellberg.

SUL needs to operate in both domains, i.e. it needs to handle inflection, derivation and compounding both in written *and* in spoken language, in order to be useful in applications such as text-to-speech synthesis and speech-to-text conversion (i.e. automatic speech recognition). The morphological system implemented in our dictionary therefore comprises in its present form a total of 500 paradigms, as compared for instance with Hellberg's text-based system of 235 paradigms (Hellberg 1978). In order to facilitate the rule-based selection of the correct paradigm during morphological analysis and generation, these 500 paradigms are grouped into 24 *paradigmfamilies* (the paradigm family codes used in SUL are given in the brackets):

- 13 for the nouns (**s0, s1...s11, s20**);
- 6 for the adjectives (**a1...a6**);
- 5 for the verbs (**v1...v5**).

A summary of these 23 paradigm families, each exemplified by a word following the main paradigm within the family, is listed below in table 5.

A complete list of all 500 paradigms grouped into the 24 paradigm families is attached to this paper as appendix A.

Observant readers will notice that the division into paradigm families implemented in SUL follows, to a certain degree, the traditional concepts of *declension* (for nouns and adjectives) and *conjugation* (for verbs), established in standard textbooks of Swedish grammar. Thus, the paradigms included in the paradigm family **s1**, for instance, emulate the inflection patterns of the or-declension, **s2** the ar-declension, **v1** the first conjugation, etc. However, irregular inflection

paradigms, which are often simply listed as exceptions in traditional descriptions of Swedish morphology, are here collected and classified into separate paradigm families, in order to permit their strictly rule-based generation and analysis within the SUL inflection and compounding system.

s0 *vädjan*	**v1** *bada*	**a1** *stark*
s1 *lampa*	**v2** *väga*	**a2** *röd*
s2 *burk*	**v3** *blänka*	**a3** *defekt*
s3 *bild*	**v4** *tro*	**a4** *död*
s4 *fett*	**v5** *bedja*	**a5** *framstående*
s5 *sko*		**a6** *bra*
s6 *dike*		
s7 *träd*		
s8 *plock*		
s9 *tro*		
s10 *jägare*		
s11 *bok*		
s20 *pengar*		

Table 5. Paradigm families used in SUL

Table 6 below shows a short, authentic excerpt which illustrates the main paradigm of paradigm family **s1**, generating all possible inflections of the Swedish noun "flicka" (Eng: girl) in the graphemic, phonemic and phonetic domains:

s1	**a**	**/*/**			-	
			flicka	/flɪkka/	-	[ˈflɪk:a]
			flickan	/flɪkkan/	-	[ˈflɪk:an]
			flickor	/flɪkkɷr/	-	[ˈflɪk:ɷr]
			flickorna	/flɪkkɷɳa/	-	[ˈflɪk:ɷɳa]
genitiv:			*flickas*	/flɪkkas/	-	[ˈflɪk:a]
			flickans	/flɪkkans/	-	[ˈflɪk:an]
			flickors	/flɪkkɷʂ/	-	[ˈflɪk:ɷʂ]
			flickornas	/flɪkkɷɳas/	-	[ˈflɪk:ɷɳas]

Table 6. Inflexion paradigm example.

Right hand column giving an allophone translation is included for clarity only.

A complete listing of the four paradigms belonging to the paradigm family **s1** is shown in appendix B.

During operation in the run-time system, the paradigms are automatically converted into executable rules for stem generation and for suffix merge. A dual set of inflections is thus generated, one for the orthographic and one for the phonetic domain.

5. Status and Goals

The Swedish pronunciation dictionary (SUL) described in this paper comprises in its present form a total of approximately 100.000 headwords, each with a full grapheme representation (including relevant alternative spellings), a homograph index (if required), a word class specification, the paradigm family code, and a broad "phonemic" transcription (including accentuation and word level stress) representing standard pronunciation usage. Proper names are further classified as names of persons, places, products, etc. Some etymological information is included, mainly for words of foreign origin and for "outdated" Swedish words still in usa (e.g. in literary language).

The dictionary incorporates an inflection engine, a compound engine and an transcription engine, which together are capable of generating all possible inflected forms (i.e. 800.000) and a theoretically infinite number of compounds, in the orthographical as well as in the transcription domain.

Evaluation of the dictionary in practical text-to-speech synthesis showed that, depending on the material, on the average between 95 and 99 % of the orthographic words commonly found in Swedish newspaper texts are correctly handled. Word inputs missed by SUL comprise mainly proper nouns and foreign names and expressions.

These figures are based on the evaluation with text input. We have not yet evaluated the performance of SUL in the reverse task, viz. for spoken language input in applications such as automatic speech recognition.

Our goals for the future include (i) to extend the number of headwords from the present 100.000 entries to approximately 150.000 entries, (ii) to increase the number of variant transcriptions representing alternative pronunciations, and (iii) to incorporate various kinds of syntagmatic information (cf. Huber 1992).

Finally, we would like to mention that earlier versions of SUL have already been acquired by other groups of researchers working in the field of spoken language processing, both inside and outside academia. We are presently working with compiling a Macintosh based version of SUL which we hope to be able to make available to other researchers within the nearest future. We have also been approached with the suggestion to provide a hard-copy edition of SUL for the general public, based on a selection from the larger file of lexical data compiled at our department (cf. appendix C).

Acknowledgements

Our work with compiling SUL has been considerably facilitated by our cooperation with the Department of Computational Linguistics (Språkdata) at Gothenburg University. We want to thank Professor Sture Allén and Christian Sjögreen for their continuous support and for making available to us several of their lexical databases. We are also greatly indepted to Margareta Westman and Claes Garlén of the Swedish Language Council (Svenska Språknämnden) who helped us with matters of phonological representation and transcription.

References

[Allén 1986] S. Allén et al, "Svensk Ordbok", Stockholm: Esselte Studium, 1986.

[Collinder 1989] B. Collinder, "Stora Ordboken" (nyutgåva), Lund: Liber Förlag/Läsförlaget, 1989.

[Dalin 1850] A.F. Dalin, "Ordbok öfver svenska språket", 1850-1852.

[EDR 1990] EDR Technical Report, "An Overview of the EDR Electronic Dictionaries", TR-024, Japan Electronic Dictionary Research Institute, Tokyo, 1990

[Elert 1979] C.-Ch. Elert, "Ljud och Ord i Svenskan", Almqvist & Wiksell, 1970.

[Garlén 1979] C. Garlén, "Svenskans Fonologi", Lund: Studentlitteratur, 1988.

[LEXIN 1984] "Svenska Ord med Uttal och Förklaringar", LEXIN-projektet, Stockholm: Esselte Studium, 1984.

[Hellberg 1978] S. Hellberg, "The Morphology of Present-day Swedish", Data linguistica, Stockholm: Almqvist & Wiksell, 1978.

[Hedelin 1986] P. Hedelin & A. Jonsson, "Svenskt uttalslexikon; 1:st edition", Technical Report, Chalmers University of Technology, 1986.

[Hedelin 1987] P. Hedelin & A. Jonsson, "Svenskt uttalslexikon; 2:nd edition", Technical Report, Chalmers University of Technology, 1987.

[Hedelin 1989] P. Hedelin, A. Jonsson & P. Lindblad, "Svenskt uttalslexikon; 3:rd edition", Technical Report, Chalmers University of Technology, April 1989.

[Huber 1990] D. Huber, "An Electronic Dictionary for Computer Speech Applications", Proceedings of the International Workshop on Electronic Dictionaries, pp.130-140, Oiso, Japan, 1990.

[Huber 1992] D. Huber, "Integrating Syntagmatic Information in a Dictionary for Computer Speech Applications", 1992 (this volume).

[IPA 1949] IPA, "The Principles of the International Phonetic Association, London, 1949

[Lyttkens and Wulf, 1911] Lyttkens & Wulf, "Svensk Ordlista", Gleerup, 1911.

[Malmström, 1988] Malmström, Györki och Sjögren, "Bonniers Svenska Ordbok" (nyutgåva), Stockholm: Bonniers, 1988.

[Molbæk Hansen 1990] P. Molbæk Hansen, "Udtaleordbog", Copenhagen: Gyldendal, 1990.

[Molde, 1977] B. Molde, "Illustrerad Svensk Ordbok", Natur och Kultur, 1977.

[SAOB, 1898] "Svenska akademiens ordbok (SAOB)", Gleerup, 1898ff.

[SprN, 1979] J. Sahlgren & G. Bergman, "Svenska Ortnamn med Uttalsuppgifter", Stockholm: Esselte Studium, 1979.

[SprN, 1977] G. Bergman, "Språknämndens Uttalsordlista", Esselte Studium, (1965) 3:e uppl. 1977.

[SprN, 1991] C. Garlén & A-C Mattisson, "Svenska Ortnamn: Uttal och Stavning", Norstedts, 1991.

[Wells 1990] J.C. Wells, "Longman Pronunciation Dictionary", Longman, 1990

[Östergren, 1981] O. Östergren, "Nusvensk Ordbok" (nyutgåva), Stockholm: Esselte Studium, 1981.

Per Hedelin and Dieter Huber
Department of Information Theory
Chalmers University of Technology
S-412 96 Gothenburg, Sweden

Appendix A

Excerpt of the paradigm data-base

s 1	**a , /*/**			-
		flicka	/flikka/	-
		flickan	/flikkan/	-
		flickor	/flikkɷr/	-
		flickorna	/flikkɷrna/	-
genitiv:		*flickas*	/flikkas/	-
		flickans	/flikkans/	-
		flickors	/flikkɷş/	-
		flickornas	/flikkɷrnas/	-
s 1	**el , /əl/**			
		toffel	/toffəl/	-
		toffeln	/toffəln/	-
		tofflor	/tofflɷr/	-
		tofflorna	/tofflɷrna/	-
genitiv:		*toffels*	/toffəls/	-
		toffelns	/toffəlns/	-
		tofflors	/tofflɷş/	-
		tofflornas	/tofflɷrnas/	-
s 1	**s , /s/**			-
		ros	/rɷ:s/	-
		rosen	/rɷ:sən/	-
		rosor	/rɷ:sɷr/	-
		rosorna	/rɷ:sɷrna/	-
genitiv:		*rosens*	/rɷ:sənş/	-
		rosors	/rɷ:sɷş/	-
		rosornas	/rɷ:sɷrnas/	-
s 1	*** , /*/**			-
		våg	/vo:g/	-
		vågen	/vo:gən/	-
		vågor	/vo:gɷr/	-
		vågorna	/vo:gɷrna/	-
genitiv:		*vågs*	/vo:gs/ -	
		vågens	/vo:gəns/	-
		vågors	/vo:gɷş/	-
		vågornas	/vo:gɷrnas/	-

Appendix B

The list below illustrates all paradigms by examples of words. Each example represents one paradigm.

s0 *Alex, Alice, vädjan*
s1 *flicka, toffel, ros, våg*
s2 *choke, båge, hummer, lämmel, nagel, öken, dager, serve, sky, gran, mun, sjukdom, gam, kam, sommar, caterpillar, lapp*
s3 *aveny, dekor, professor, doktor, modul, konsul, byrå, balsam, muskel, kansler, fiber, kollega, ven, vän, dam, parallellogram, geranium, akademi, kö, analys, bild*
s4 *bageri, gage, fängelse, sekel, studium, drama, fett*
s5 *ensamble, creme, apache, damejanne, pavane, cape, grace, manege, champagne, offside, pose, bagette, sko*
s6 *altare, huvud, tabu, knä, spö, bi, dike*
s7 *garage, bekymmer, fönster, faner, kummel, kapitel, album, system, gram, ton, bäcken, vatten, träd*
s8 *natrium, minimum, berör, flimmer, bladder, tummel, babbel, vatten, kaffe*
s9 *agio, lappri, sperma, bagage, college, avantgarde, plock, insyn, service, grape, massage, ampere, chamotte, chartreuse, tro, judendom, rom, gödsel, fräken, liter, purpur, radar, peppar, dollar, terror, fosfor, humor, standard*
s10 *kärande, bagare, kritiker, rododendron*
s11 *farao, cesar, bok, tång, fot, öra, bekant, fröken, botten, braxen, öga, hinder, finger, man, middag, söndag, måndag, tisdag, onsdag, torsdag, fredag, lördag, vardag, namnsdag, middag, land, and, baryton, morgon, son, farbror, bror, mor, mormor, farfar, far, broder, moder, fader, afton, officer, toker, hammare, dotter, spektrum, center, tema, dekanus, eforus, amaryllis, emeritus, deputerad, lus, gås, get, gnet, nöt, ledamot, rot, natt, bokstav, historia, stad, verkstad, verkstad, bonde, jävul, jävel, predikan, duo, albino, lexikon, examen, spann, snö, orden, sägen, sjö, pilgrim, altare, genre, himmel, papper, lager, pansar*
s20 *pengar*
v1 *bada, andas*
v2 *tända, föda, skilja, följa, skämmas, stämma, kännas, känna, böja, sörja, föra, mala, tåla, kräva, trivas, blygas, väga*
v3 *mäta, lyfta, minna, begynna, synas, blåsa, blänka*
v4 *bräs, tro*
v5 *begrava, drita, be, bedja, binda, bita, bjuda, bli, bliva, bringa, brinna, brista, bryta, byta, bära, böra, dimpa, dra, draga, dricka, driva, drypa, duga, dväljas, dyka, dö, dölja, falla, fara, förfaras, finna, fisa, flyga, flyta, fnysa, frysa, få, förgäta, förnimma, gala, ge, gifta, giva, gjuta, glida, gnida, glädja, gripa, gråta, gå, gås, göra, ha, hava, heta, hinna, hugga, hålla, idas, kliva, klyva, knipa, knyta, koka, komma, krypa, kunna, kvida, kväda, le, lida, ligga, ljuda, ljuga, ljuta, lyda, lysa, lägga, låta, löpa, minna, minnas, mysa, må, måste, niga, njuta, nypa, nysa, pipa, pysa, rinna, rysa, rådas, se, simma, sitta, sjuda, sjunga, sjunka, skina, skita, skjuta, skola, skrida, skrika, sprida, spörja, strida, vrida, rida, skriva, riva, trivas, skryta, tryta, ryta, skälva, skära, slinka, slinta, slippa, slita, sluta, slå, slåss, smita, smyga, smälla, smälta, smörja, snika, snyta, sova, spinna, spricka, springa, spritta, sticka, stiga, stinka, stjäla, stryka, ryka, strypa, stå, stöda, stödja, suga, supa, svida, svika, svälja, välja, svälta, svära, svärja, säga, säja, sälja, sätta, taga, tiga, tjuta, töras, tvinga, två, tälja, tämja, tör, töras, vara, varda, veta, vika, vilja, vina, vinna, vänja, växa, äta, ta*
a1 *vacker, ädel, ljummen, ilsken, kavat, tunn, frigid, brydd, blond, sakta, tam, dum, ledsam, allmän, vig, avig, trygg, svensk, dov, stark*
a2 *fri, blöt, rädd, god*
a3 *beige, disträ, olovlig, alternativ, komisk, ansedd, hybrid, aktad, anlagd, brunnen, kommen, svulten, avlång, fjäderlätt, jättevacker, femdubbel, halvgammal, perifer*
a4 *arbetsfri, benvit, död, skyhög*
a5 *framstående*
a6 *bra, dålig, gammal, grov, hög, liten, låg, lång, stor, trång, tung, ung*

Appendix C

a subst, s6 ['ɑ:]
|| aet['ɑ:ət], pl an['ɑ:n] ana['ɑ:na]
• *alfabetets första bokstav.*
• Böjningar skrivs ofta a:et, a:n osv.

à prep ['a]
• *till, för.*
(< av Fra: *à*, till, av Latin: *ad*)

a capella adverb [aka'pɕl:a]
alt: [aka'pɕl:a].
• *utan ackompanjemang; med endast sångstämmor.*
(< av Ital: *a capella*, som i kapellet, av Latin: *capella*, liten get, diminutiv av *capra*, get)

a conto adverb [a'kɔntɔ]
alt: [a'kɔntɔ]; el [-tɔ]. • Eller *a konto*.
• *i räkning.*
Jmfr: MSEOB [a'kɔntɔ]. (< av Ital: *a conto*)

a dato adverb [a'dɑ:tɔ]
alt: [-tɔ].
• *från i dag; från utgivningsdagen.*
Jmfr: SAOB [-'dɑ:tɔ]. MSEOB [a'dɑ:tɔ]. (< av Latin: *dato*, av *datum*)

à jour adjekt, a ob [a'ʃu:r]
• *uppdaterad; underrättad om; från denna dag.*
(< Fra: *à jour*)

à la adverb ['ala]
• Nästan alltid obetonat[ˌala].
• *enligt.*
• I förbindelser som *biff à la Lindström*, *ris à la Malta*. (< Fra: *à la*)

à la carte subst, s ob [ala'kaɽ:]
• *rätt på meny som tillagas på beställning.*
(< Fra: *à la carte*)

a posteriori adverb [aˌpɔstɛri'ɔ:ri]
alt: [-ˌpɔstær-].
• *i efterhand.*
(< Latin *a posteriori*, ej från det efterföljande)

a priori adverb [apri'ɔ:ri]
• *på förhand.*
Jmfr: SpN /aprárˈi/. (< Latin *a priori*, ej från det föregående)

Aage subst, s0 ['o:gə]
• *egennamn: mansnamn.*

Aalborg subst, s0 ['o:lˌbɔrj]
• *egennamn: ort på Jylland.*

Aarhus subst, s0 ['o:rˌhu:s]
• *egennamn: ort på Jylland.*

Aasa subst, s0 ['ɔ:sa]
• *egennamn: kvinnonamn.*

Aasarum subst, s0 ['ɑ:sa~ˌrɛm:]
• Eller *Asarum*.
• *egennamn: ort i Blekinge län.*

Aase subst, s0 ['ɔ:sə]
• *egennamn: kvinnonamn.*

Aatos subst, s0 ['ɑ:tɔs]
• *egennamn: mansnamn.*

ab förkortn ['a~ˌbe:]
• Skrivs ofta med versaler *AB*.
• *aktiebolag.*

ab- förled [ˌab-]
• *latinskt prefix med betydelse ifrån, isär.*

ABAB förkortn ['a:bab]
• *allmänna bevaknings AB.*

abakus subst, s3 ['ab:akɛs]
alt: ['a:bakɛs]. • Skrivs även *abocus*. || abakusen['ab:akɛsən], pl abakuser['ab:akɛsər] abakuserna['ab:akɛsɛŋa]
• *1. en kulrum använd som räknehjälpmedel. 2. arkit: platta överst på bl a kolonnkapitäl.*
Jmfr: SAOB ['a:bakɛs]. SvOB ['a:b-]. SAOL[11] ['ab:-]. (< av Latin *abacus*, av Grek: *abax*, något hoptrogat)

abandon subst, s9 [abaŋ'dɔŋ:]
alt: [aban-]. || abandonen[abaŋ'dɔŋən]
• *ledighet, otvungenhet i uppträdandet; uppsluppenhet.*
Jmfr: L&W (1911) [ˌabaŋ'dɔŋ]. SAOB [ˌabaŋ'dɔŋ]. SvOB [abaŋ'dɔŋ]. SAOL[11] [-aŋ'dɔŋ] el [-an'dɔŋ]. (< av Fra: *abandon*, med samma betydelse, av *être à bandon*)

abasi subst, s9 [aba'si:]
|| abasin[aba'si:n]
• *medicin: oförmåga att gå.*
(< av Grek: *a-*, icke- + *basis*, gång)

abbé subst, s3 [a'be:]
|| abbén[a'be:n], pl abbéer[a'be:ər] abbéerna[a'be:əŋa]
• *titel för abbott (också använt för katolsk präst i allmänhet).*
Jmfr: L&W (1911) [a'be:]. SAOB [a'be:]. DMTOB [a'be:] (< av Fra: *abbé*, abbot, av Latin: *abbas*)

-abbedissa subst, s1; avid [abɛ'dɪsa]
|| abbedissan[abɛ'dɪsan], pl abbedissor[abɛ'dɪsɔr] abbedissorna[abɛ'dɪsɔŋa]
• *föreståndarinna för nunnekloster.*
Jmfr: L&W (1911) [ˌabɛ'dɪsa]. SAOB [ˌabɛ'dɪsa]. SvIOB /-ɛs'-/ el /-ɛs'-/. (< Fsv: *abbedissa*, av senmedeltida latin *abbatissa*)

Abbekås subst, s0 [abe'kɔs]
alt: ['ab:ə~ˌkɔs].
• *egennamn: ort i Malmöhus län.*

abborrart subst, s3; ssg ['ab:ɔr~ˌa:ɽt]
-abborrartad adjekt, a3; avid['ab:ɔr~ˌa:ɽad]

-abborre subst, s2,c1 ['a~ˌbɔr:ə]
• *zool: en sötvattensfisk (Perca fluviatilis).*
Jmfr: L&W (1911) ['abˇ~ˌbɔrə]. SAOB ['abˇ~ˌbɔrə] el ['abɔrə] el ['a~ˌbɔrə]. (< Fsv: *aghborre*, av *agh*, spetsigt föremål) **-abborrfisk** subst, s2; ssg ['ab:ɔr~ˌfɪsk] • *zool: ett släkte av fiskar.* **-abborrgrund** subst, s7; ssg ['ab:ɔr~ˌgrənd] **-abborrpinne** subst, s2; ssg ['ab:ɔr~ˌpɪn:ə] • *vard: liten abborre.*

Abborrträsk subst, s0 ['ab:ɔ~ˌtrɛsk]
• *egennamn: ort i Västerbottens län; i Norrbottens län.*

abbot subst, s2 ['ab:ɔt]
alt: ['ab:ɔt]. || abboten['ab:ɔtən], pl abbotar['ab:ɔtar] abbotarna['ab:ɔtaŋa]
• *föreståndare för kloster; munk.*
Jmfr: L&W (1911) ['ab:ɔt]. SAOB ['ab:ɔt] el ['a'bu:t]. även ['ab:ɔt] el ['a'bɔ:t], sällan['ab:ˇ~ˌbɔ:t]. SvOB ['ab:-]. SpN /ab'ɪt/. DMTOB ['ab:ɔt]. (< Fsv: *abote*, över latin, av arameiska *abba*, fader) **-abbotdöme** subst, s6; avid ['ab:ɔt~ˌdø:mə] • Eller *abbotsdöme*. • Stundom med kort ö-vokal, [-ˌdœmːə]. **-abbotsdöme** subst, s6; avid ['ab:ɔts~ˌdø:mə] • Eller *abbotdöme*. • Stundom med kort ö-vokal, [-ˌdœmːə]. **-abbotskap** subst, s7; avid ['ab:ɔt~ˌska:p] **-abbotsstift** subst, s7; ssg ['ab:ɔt~ˌstɪft] alt: tydligt ['ab:ɔts-ˌstɪft]. • Eller *abbotstift*. **-abbotstift** subst, s7; ssg ['ab:ɔt~ˌstɪft]• Eller *abbotsstift*.

abbreviation subst, s3; avid [ˌabrɐvɪa'ʃu:n]
alt: (över)tydligt [ˌabrɐˌvɪ'a-], red [aˌbrɐvja-]. || abbreviationen [ˌabrɐvɪa'ʃu:nən], pl abbreviationer [ˌabrɐvɪa'ʃu:nar] abbreviationerna[ˌabrɐvɪa'ʃu:nəŋa]
• *förkortat skrivsätt.*
Jmfr: L&W (1911) [ˌabrɐˌvɪ'a'ʃu:n]. SAOB [ˌabrɐˌvɪ'a'ʃu:n] el [a-ˌbrɐ-]. (< av Fra: *abbréviation*, förkortning, av Latin: *abbreviare*, förkorta, av *brevis*, kort)

-abbreviatur subst, s3; avid [ˌabrɐvɪa'tu:r]
alt: (över)tydligt [ˌabrɐˌvɪ'a-], red [aˌbrɐvja-]. || abbreviaturen [ˌabrɐvɪa'tu:rən], pl abbreviaturer [ˌabrɐvɪa'tu:rər] abbreviaturerna[ˌabrɐvɪa'tu:rəŋa]
• *förkortad beteckning: symbol som ersätter flera skrivtecken.*
Jmfr: L&W (1911) [ˌabrɐˌvɪ'a'tu:r]. SAOB [ˌabrɐˌvɪ'a'tu:r] el [a-ˌbrɐ-]. (< av Tyska: *abbreviatur*, av Latin: *abbreviare*, förkorta)

-abbreviera verb, v1 [ˌabrɐvɪ'e:ra]

On the Coverage of a Morphological Analyser based on "Svensk Ordbok" [A Dictionary of Swedish]

Anna Sågvall Hein
Uppsala University

Introduction

In the project a *Lexicon-oriented Parser for Swedish* a stem dictionary (Sågvall Hein & Sjögreen 1991; Sjögreen, forthcom.) covering the 58,536 entry lemmas of *Svensk Ordbok* (1986) along with a complete inflectional grammar of Swedish (Sågvall Hein, forthcom.) was generated. This language description together with the *Uppsala Chart Processor*, UCP (Sågvall Hein 1987) constitute a morphological analyzer of Swedish, henceforth referred to as *SMU*, short for Swedish Morphology in the UCP framework.

So far, there are no word formation rules in the SMU grammar, and words outside the scope of Svensk Ordbok don't get an analysis[1]. Eventhough closed in its present version, the coverage of SMU is well-defined; prior to any processing we may consult Svensk Ordbok to find out for any word form whether it will get an analysis or not; the dictionary provides an intuitive, familiar format through which we may explore the (present) competence of the SMU analyser without any prior knowledge of its formalisms or operation. SMU is also well-defined in the sense, that for any of its lemmas, Svensk Ordbok provides links to the corresponding lexemes (basic senses), and for each lexeme a definition.

In our ongoing work on a machine-tractable dictionary for Swedish, we are approaching problems concerning the distinction between general and domain specific vocabulary, and the present coverage of SMU is our starting-point for delimiting a general Swedish vocabulary. For an evaluation of the generality of the dictionary, the analyser has been applied to different sets of Swedish text. For one of them, consisting of the 10,224 most frequent types of the 7,3 million word newspaper corpus of The Language Bank (Gellerstam 1989) the words outside he scope of the analyser have been examined at some detail. Here we will present the results achieved so far, and also discuss their impact on our continued work on the dictionary.

First, however, we will briefly characterize the SMU analyser with regard to morphological descriptions, and dictionary representation of inflection.

The morphological descriptions of the SMU analyser

The morphological descriptions generated by the analyser are expressed as attribute-value structures (Sågvall Hein & Ahrenberg 1985; cf. directed acyclic graphs, dags, for short, Shieber 1986). For a first illustration, we present the description of the noun *festernas* [*of the parties*] (see fig. 1).

It comprises four general attributes, i.e. LEM for lemma, WORD.CAT for word category (part of speech), DIC.STEM for dictionary stem, and INFL for inflection type), and, four attributes specific to the nouns i.e. GENDER, NUMBer, FORM (species), and CASE. The general attributes are present in the descriptions of all the words, regardless of part of speech (noun, adjective, pronoun, verb, adverb, numeral, preposition, conjunction, interjection, article, and infinitive marker).

```
FESTERNAS   :
(* = (       LEM=FEST.NN
              WORD.CAT=NOUN
              INFL=PATTERN.FILM
              DIC.STEM=FEST
              GENDER=UTR
              NUMB=PLUR
              FORM=DEF
              CASE=GEN)
```

Figure 1. An analysis of the noun festernas [of the parties]

The value of the lemma attribute is identical to the basic form of the lemma with a (two letter) word class marker for the distinction between homograph lemmas, i.e. *springa1.nn* [chink; slot] and *springa2.vb* [run]. In addition, the homographs are numbered as they are in Svensk Ordbok (cf. **1spring/a** subst. [noun] and **2spring/a** verb), whereby immediate reference to this background material is facilitated. If there are two homograph lemmas of the same word category, the homograph number alone will keep them apart, e.g. *bok1.nn* [book] (plur. *böcker*) and *bok2.nn* [beech (tree)] (plur. *bokar*). The well-defined lemma marker supports the distinction between external and internal homography, a basis for subsequent lemmatization. Further, it provides a basis for the selection of domain tuned lemma dictionaries from a general dictionary; the lemmas specific to the domain are recognized by the morphological analysis of texts typical of that domain. The dictionary stem attribute may serve the same function in building a domain tuned stem dictionary from a general stem dictionary.

The value of the inflection attribute is a *pattern word* which is also the name of an *inflectional rule* defined in the grammar. The inclusion of this information in the morphological descriptions provides a basis for frequency studies of inflectional types in current text.

In addition to the general attributes, each part of speech (except for the prepositions and the infinitive marker) is characterised by its own set of attribute value pairs. In fig. 2 we present the morphological descriptions resulting from the analyses of an adjectival paradigm, the adjective *festlig [festive; grand]*. The descriptions illustrate, among other things, the representation of *internal homography*. The form *festliga* gets two descriptions, one corresponding to an analysis of it as a definite singular positive form of the adjective *festlig.nn*, the other one as a plural positive form of the same adjective. In other words, this is a case of internal homography. The forms differ with regard to number (singular versus plural) and form (definite in the singular case, unstated in plural). The description is *underspecified* in the sense that the form value is left out in plural with the effect that it will unify with definite as well as with indefinite contexts in a unification-based syntactic analysis; it allows for a definite or an indefinite reading (cf. Karlsson forthcom. representing underspecification by means of composite values, e.g. DEF/INDEF). A fully specified representation would have implied the establishement of two descriptions, one plural indefinite and one plural definite. There are many such cases in the Swedish inflectional system, consequently, underspecification has a substantial impact on representational economy. In Table 1 we summarize the attributes and values specific to the different parts of speech.

```
FESTLIG     :
(* = (       LEM=FESTLIG.AV
            WORD.CAT=ADJ
            INFL=PATTERN.BLEK
            DIC.STEM=FESTLIG
            GENDER=UTR
            NUMB=SING
            FORM=INDEF
            COMP=POS)

FESTLIGT    :
(* = (       LEM=FESTLIG.AV
            WORD.CAT=ADJ
            INFL=PATTERN.BLEK
            DIC.STEM=FESTLIG
            GENDER=NEUTR
            NUMB=SING
            FORM=INDEF
            COMP=POS)

FESTLIGA    :
(* = (       LEM=FESTLIG.AV
            WORD.CAT=ADJ
            INFL=PATTERN.BLEK
            DIC.STEM=FESTLIG
            COMP=POS
            NUMB=SING
            FORM=DEF)

(* = (       LEM=FESTLIG.AV
            WORD.CAT=ADJ
            INFL=PATTERN.BLEK
            DIC.STEM=FESTLIG
            COMP=POS
            NUMB=PLUR)

FESTLIGE    :
(* =         (LEM=FESTLIG.AV
            WORD.CAT=ADJ
            INFL=PATTERN.BLEK
            DIC.STEM=FESTLIG
            GENDER=UTR
            NUMB=SING
            FORM=DEF
            SEX=MASC
            COMP=POS)

FESTLIGARE  :
(* = (       LEM=FESTLIG.AV
            WORD.CAT=ADJ
            INFL=PATTERN.BLEK
            DIC.STEM=FESTLIG
            COMP=COMP

FESTLIGAST  :
(* = (       LEM=FESTLIG.AV
            WORD.CAT=ADJ
            INFL=PATTERN.BLEK
            DIC.STEM=FESTLIG
            COMP=SUP
            FORM=INDEF))

FESTLIGASTE :
(* = (       LEM=FESTLIG.AV
            WORD.CAT=ADJ
            INFL=PATTERN.BLEK
            DIC.STEM=FESTLIG
            COMP=SUP
            FORM=DEF))
```

Figure 2. Analyses of the adjective festlig [festive; grand]

Part of speech	Attribute	Values		
NN	GENDER	Neutr	Utr	
	NUMB	Sing	Plur	
	FORM	Indef	Def	
	CASE	Basic	Gen	
	PROPR	+		
	ABBREV	+		
AV	GENDER	Neutr	Utr	
	NUMB	Sing	Plur	
	FORM	Indef	Def	
	COMP	Pos	Comp	Sup
	FUNC	Attr	Pred	
	SEX	Masc		
	(+ CASE for the description of nominalized adjectives)			
VB	TENSE	Pres	Sup	Pret
	INFF	Inf	Pp	Ap
	VOICE	Pass	Act	Depon

	IMP	+		
	CONJ	+		
	(+ NUMB, GENDER, FORM, FUNC, SEX, CASE for the description of the participles)			
AL	GENDER	Neutr U	tr	
	NUMB	Sing	Plur	
	FORM	Indef	Def	
PN	PRON.TYPE	Pers	Poss	Rel
	ATTR.TYPE	Select	Quant	Comp
	DET.TYPE	Tot	Det	
	GENDER	Neutr	Utr	
	NUMB	Sing	Plur	
	FORM	Indef	Def	
	CASE	Basic	Gen	Obj
	SEX	Masc		
	PARTITIV	+		
	DUAL	+		
	NOUN.INDEF	+		
NL	ATTR.TYPE	Select	Quant	
	GENDER	Neutr Utr		
	NUMB	Sing	Plur	
	FORM	Indef	Def	
	CASE	Basic	Gen	
	SEX	Masc		
AB	COMP	Pos	Comp	Sup
KN	SUBJU	+		

Prepositions, interjections, and the infinitive marker have no attributes.

Table 1. An Overview of the attributes assigned by the SMU analyser

Inflection in Svensk Ordbok and in the SMU dictionary

In fig. 3 we present the inflectional variety of the *-ar-declination* for an illustration of the relation between the inflectional format of SOB and that of the SMU stem dictionary. In the simplest case, there is only one stem, in SOB and in SMU, and, further, the stem is identical to the basic (lemma) form of the word (see **1** and **2**). In such cases, the entry form of SOB represents at the same time the lemma and the stem, whereas in the SMU formal dictionary the two concepts have to be individually represented. The pattern rules of the SMU dictionary cover the inflectional information given in SOB in terms of word class and significant endings, and, also, the grammatical background information to which these morphological keys refer, i.e. morphotactic rules determining the inflectional behaviour of the nouns. Sometimes, SOB recognizes a stem identical to an initial substring of the lemma form and delimits it from the succeeding ending by a slash, as in **3 – 8**, and **11**. The stem concept thus adopted is the *technical stem* (Hellberg 1978). The SMU analyser treats these cases in a different way, i.e. by means of a general rewriting rule (**4 – 8**) or stem handling operations in the pattern rules (as in **3** and **11**). In the first case, the non-vowel stem alternant is regarded as the canonical form of the stem, and its vowel counterpart is reduced to this form by a secondary vowel deletion rule[2]. Thus there is only one stem in dictionary, and, what is more important, we don't have to accept as endings strings such as *eln, lar, nen, nart* etc. In the second case, SMU analyses the stem in two steps, i.e. as a *dictionary stem* followed by a stem building element, e.g. *pojk-e, dröm-m, läm-mel*. A word form such as *pojken* is analysed *pojk-e+n, lämmeln* as *läm-mel+n*, etc. where the corresponding pattern rules (pattern.gosse and pattern.lämmel) are

responsible for the recognition of the stem building segments and their distribution in the paradigm (see further Sågvall Hein, forthcom.).

In all, there are 132 (stem) pattern rules for the nouns, 40 for the adjectives, 89 for the verbs, 1 for the articles, 30 for the pronouns, 5 for the numerals, 9 for the adverbs, 2 for the conjunctions, and one each for the prepositions, the interjections, and the infinitive marker. In most cases the analysis is based on one stem[3], and in these cases, the stem with its pattern rule is a sufficient characterisation of the inflectional behaviour of the lemma as such. If, on the other hand, the lemma is represented by more than one stem in the dictionary (cf. **17** and **18** in fig. 3), the *set of stems* involved along with their *pattern words* determine the inflection of the lemma, the lemma inflection as opposed to the stem inflection. Frequency data on the inflectional types of the SMU words have been presented elsewhere (Sågvall Hein & Sjögreen 1991).

SOB	stem	SMU lemma	pattern
1 gård subst. *~en ~ar*	gård	gård.nn	.stol
2 sjö subst. *~n ~ar*	sjö	sjö.nn	.fru
3 pojk/e subst. *~en ~ar*	pojk-e	pojke.nn	.gosse
4 speg/el subst. *~eln ~lar*	speg(e)l	spegel.nn	.nyckel
5 sock/en subst. *~nen ~nar*	sock(e)n	socken.nn	.öken
6 bott/en subst. *~nen* el. =, ~nar	bott(e)n	botten.nn	.botten
7 myrt/en subst. *~en ~nar*	myrt(e)n	myrten.nn	.fröken
8 fing/er subst. *~ret* el. ~em *~rar*	fing(e)r	finger.nn	.finger
9 dröm subst. *~men ~mar*	dröm dröm-m	dröm.nn	.kam
10 läm/mel subst. *~meln ~lar*	läm-mel läm-l	lämmel.nn	.lämmel
11 sum/mer subst. *~mern ~rar*	sum-mer sum-r	summer.nn	.hummer
12 sommar subst. *~(e)n somrar*	som som-mar	sommar.nn	.sommar
13 hammare subst. *hammar(e)n*, = el. *hamrar*, best. plur.	ham-r ham-mar ham-mare	hammare.nn	.kammare
14 himmel subst. *himmel(e)n* el. *himlen, himlar*	him-mel him-l	himmel.nn	.himmel
15 afton subst. *aftonen aftnar*	aft-on aft-n	afton.nn	.morgon
16 djävul subst. *~en djävlar*	djäv-ul djäv-l	djävul.nn	.djävul
17 moder äv. [1]**mor** subst. *modern mödrar*	moder mödrar mor	moder.nn moder.nn moder.nn	.moder .mödrar .far
18 tok subst. *~en ~ar* el. **tok/er** *~ern ~ar*	tok toker	tok.nn tok.nn	.stol .toker
19 stadgar subst. plur.	stadg	stadgar.nn	.vägnar

Figure 3. Inflection in SOB and in SMU. An example: the -ar-declination.

The Scope of Svensk Ordbok and the SMU analyser

For the recognition, and subsequent examination, of missing entries in the dictionary, we apply the SMU analyser to different sets of Swedish text. So far, we have analysed four text materials of substantial size, i.e. the 10,224 most frequent types of the 7,3 million word newspaper corpus of the Swedish Language Bank (Gellerstam 1989), referred to as *PressFreq*, the pharmacological text of the Swedish drug catalogue *FASS* (1985) (660,000 current words), the *Professional Prose* corpus of the *Skrivsyntax* project (Teleman 1974), referred to as *ProfProse* (78,0366 current words), and, finally, the corpus of the definitions of Svensk Ordbok, referred to as *DefVoc* (360,144 current words). ProfProse consists of four types of text of equal size (textbooks, newspapers, debate books, and brochures).

corpus	tokens	types	ty/to	covered	uncovered
Press-Freq	5,091,965	10,224	0,02	8,424 (82%)	1,790 (18%)
Prof-Prose	78,036	13,766	0,18	10,083 (73%)	3,683 (27%)
DefVoc	360,144	43,934	0,12	31,350 (71%)	12,584 (29%)
FASS	664,314	39,884	0,06	9,767 (25%)	30,117 (75%)

Table 2. Results of the application of the SOB-based SMU analyser to four sets of Swedish text

As might be expected, the analyser covers best in relation to the high frequent words of the newspaper text and worst with respect to the highly domain-specific pharmacological text. In between comes the more general text of ProfProse and that of the definition corpus, covered, basically, to the same extent. In tables 3 to 6 we present more detailed data on the results of the processing of each of the text materials. They include information on homography of three kinds, i.e. (lemma) internal, (lemma) external, and mixed (internal and external). (For each text material, we also have detailed data on the different subtypes of homographies that were found, e.g. int. AV, ext. AB/AB, ext. AB/KN, ext. AB/NN, and their frequencies.)

Number of parses	lexical coverage types	%	textual coverage tokens	%
0	1,790	17,5%	299,235	5,9%
1	6,142	60,0%	2,270,961	44,6%
2 (int.)	593	5,8%	152,334	3,0%
2 (ext.)	959	9,4%	1,391,623	27,3%
3 (int.)	158	1,5%	24,934	0,5%
3 (ext.)	191	1,9%	435,066	8,5%
3 (mix.)	206	2,0%	54,176	1,1%
4 (ext.)	33	0,3%	43,814	0,9%
4 (mix.)	69	0,7%	24,544	0,5%
5 (ext.)	14	0,1%	258,713	5,1%
5 (mix.)	54	0,5%	20,382	0,4%
6 (ext.)	6	0,1%	83,669	1,6%
6 (mix.)	7	0,1%	32,118	0,6%
7 (mix.)	1	0,0%	96	0,0%
8 (mix.)	1	0,0%	280	0,0%
Total:	10,224	100,0%	5,091,965	100,0%

Table 3. Results of the application of the SOB-based SMU analyser to PressFreq

The lemma can be unambiguously determined for 6,893 types (67,4%) and 2,447,956 tokens (48,1%). In **PressFreq words outside the scope of SOB** below, we give an account of the kinds of words that got no parses.

Number of parses	lexical coverage [4] types	%	textual coverage tokens	%
0	3,683	26,8%	6,806	8,7%
1	7,828	56,9%	38,015	48,7%
2 (int.)	800	5,8%	2,693	3,5%
2 (ext.)	829	6,0%	17,638	22,6%
3 (int.)	161	1,2%	282	0,4%
3 (ext.)	149	1,1%	6,068	7,8%
3 (mix.)	168	1,2%	758	1,0%
4 (ext.)	24	0,2%	455	0,6%
4 (mix.)	63	0,5%	384	0,5%
5 (ext.)	8	0,1%	3,282	4,2%
5 (mix.)	44	0,3%	266	0,3%
6 (ext.)	3	0,0%	1,201	1,5%
6 (mix.)	4	0,0%	186	0,2%
7 (mix.)	1	0,0%	1	0,0%
8 (mix.)	1	0,0%	1	0,0%
Total:	13,766	100,0%	78,036	100%

Table 4. Results of the application of the SOB-based SMU analyser to ProfProse

The lemma can be unambiguously determined for 8,789 types (63,8%) and 40,990 tokens (52,5%). Roughly 13% (479) of the words that got no analysis are numerical expressions.

Number of parses	lexical coverage[4] types	%
0	12,584	28,6%
1	26,348	60,0%
2 (int.)	1,654	3,8%
2 (ext.)	2,205	5,0%
3 (int.)	251	0,6%
3 (ext.)	258	0,6%
3 (mix.)	372	0,8%
4 (int.)	1	0,0%
4 (ext.)	40	0,1%
4 (mix.)	137	0,3%
5 (ext.)	9	0,0%
5 (mix.)	63	0,1%
6 (ext.)	3	0,0%
6 (mix.)	7	0,0%
7 (mix.)	1	0,0%
8 (mix.)	1	0,0%
Total:	43,934	100,0%

Table 5. Results of the application of the SOB-based SMU analyser to DefVoc

The lemmas of 28,234 of the types (64%) can be determined unambiguously. Only 430 (~3,5%) of the 0-parses are numerical expressions.[5]

Number of parses	lexical coverage[4] types	%
0	30,117	75,5%
1	7,598	19,1%
2 (int.)	874	0,2%
2 (ext.)	766	0,2%
3 (int.)	147	0,0%
3 (ext.)	116	0,0%
3 (mix.)	142	0,0%
4 (ext.)	17	0,0%
4 (mix.)	52	0,0%
5 (ext.)	6	0,0%
5 (mix.)	42	0,0%
6 (ext.)	3	0,0%
6 (mix.)	3	0,0%
8 (mix.)	1	0,0%
Total:	39,884	100,0%

Table 6. Results of the application of the SOB-based SMU analyser to FASS

The lemmas of 8,619 of the types (22%) can be determined unambiguously. 11,148 (27%) of the 0-parses are numerical expressions or hybrids of numbers, special signs, and single letters, such as *75-80, 85%, 4$8, F100, E218, C++, 0,5-0,7, 20:e* etc. (A pharmacological stem dictionary covering the non-numerical words outside the scope of SOB has been built (see Sågvall Hein et al., forthcom.)) In fig. 4 we present a drug description from FASS to illustrate the special character of this text.

Abboticin®
Abbott

Dosgranulat 200 mg ℞

Antibiotikum Grupp 7B 3005

Deklaration. *1 dosgranulat innehåller:* Erythromycin. aethylsuccin. respond. erythromycin. 200 mg, mannitol. 1,5 g, constit. et aroma q. s.
Egenskaper. Dosgranulaten innehåller erytromycinetylsuccinat motsvarande 200 mg erytromycin. Granulatet löses i litet vatten (2–3 dessertskedar=20–30 ml). Beredningsformen är speciellt avsedd för barn och är även lämplig som jourförpackning. Beredningen är sockerfri och har körsbärssmak.
Erytromycinetylsuccinat är en ester av erytromycin och efter absorption sker hydrolys till fritt aktivt erytromycin. Se f ö ABBOTICIN tabletter.
Indikationer. Se ABBOTICIN tabletter.
Kontraindikationer. Se ABBOTICIN tabletter.
Försiktighet. Se ABBOTICIN tabletter.
Graviditet och amning. Se ABBOTICIN tabletter.
Biverkningar. Se ABBOTICIN tabletter.

Dosering. *Dosen för barn beräknas efter 30–50 mg per kg kroppsvikt och dygn fördelat på 2–4 doseringstillfällen.* 1 dosgranulat=200 mg erytromycin. För barn upp till 4 kg beräknas dosen i det enskilda fallet. Vid kroppsvikt överstigande 4 kg kan om dygnsdosen fördelas på två doseringstillfällen följande schema vanligen tillämpas:

Vikt kg	Dygns-dosering dosgranulat	mg/kg/dygn	Lämplig förpackning för 10 dagars behandling
4– 7	½×2	29–50	1×30
8–14	1×2	29–50	1×30
15–24	2×2	33–53	2×30
25–34	3×2	35–48	2×30

Inträffar gastrointestinala problem rekommenderas uppdelning av dygnsdosen på 3 eller 4 administreringstillfällen. För vuxna och barn över 35 kg ges 3 dosgranulat 3 gånger per dygn. Optimal absorption erhålles om dosen intages omedelbart före måltid.
Interaktion. Se ABBOTICIN tabletter.
Förpackningar och priser. *Dosgranulat 200 mg*
30 st 55:30

Figure 4. An example of a drug description from FASS 1985

PressFreq words outside the scope of SOB

Proper nouns	1,433	(80,1%)
Abbreviations	137	(7,7%)
Compounds	127	(7,1%)
Numerical expressions	45	(2,5%)
Derivatives	20	(1,1%)
Foreign words	17	(0,9%)
Syntagmatic words	5	(0,3%)
Partial phrases	4	(0,2%)
Inflectional forms	2	(0,1%)
Total	1,790	(100%)

Table 7. Kinds of uncovered types in Pressfreq

Proper nouns. The dominating category is that of the proper nouns (including a small number of proper noun abbreviations, e.g. *ABF, AIK, DN, DDR, KFUM*). No normalization of spelling variation has been carried out, so far, so each appearance has been counted as an independent unit, e.g. *Anna, anna,* and *ANNA*; *Erik* and *Eric; Bernard* and *Bernhard, Bengtsson, Bengtson,* and *B-son, Lidingö* and *Lid-ö* etc. Roughly, half the number of the proper names refer to people (approximately, 440 first names and 280 second names.)

The high number of proper nouns, in specific, personal proper nouns, seems to be a characteristic feature of newspaper text. Most of the proper nouns that we found will be entered into the dictionary with a marking of their origin (PressFreq) as a clue to future work on domain.

Abbreviations. The abbreviation category comprises abbreviations (excl. those of the proper nouns) in various orthographic shapes, e.g. *bl.a* and *bl a, FEB, feb.* and *febr, kr* and *kr.* etc. They will all be included in the core of our dictionary (see Östling, this volume), and some of them be treated as functional core phrases (Sågvall Hein et al. 1990) and represented in the dictionary as such, e.g. *bl a* and *bl. a., d. v. s.* and *d v s* etc. The figures presented in table 9 are, however, based on individual text words, for instance *bl, bl., a, a.* etc.

Compounds. As is well-known, the Swedish compounds make up an open category, and in table 10 we present an overview of the different kinds of PressFreq compounds that were found to be outside the present scope of the SMU dictionary.

Type of compound	No of members	Examples
NN-NN → NN	97	arbetsuppgifter, hälso-
NL-NN → NN	10	andraplats, 30-talet
NN-AV → AV	6	medelstora, nordöstra
AV-NN → NN	4	nypris
NL-AV → AV	3	50-årig
NL-NN-NN → NN	2	50-årsåldern
AV-VB → VB	2	nybyggda
NN-NN → AB	1	förhoppningsvis
P-NN → AB	1	härområet
P-NN → NN	1	överåklagare
P-VB → VB	1	efteranmäld
NN-NN-VB → VB	1	text-TV-Textat
Total	129	

Table 8. PressFreq: Uncovered compounds

Following Blåberg (1988) we refer the participles to the verb category. Further, prepositions, and adverbs, are treated as members of one common category (in the compound context), denoted P.

Many of the productive compound types are indicative of domain (economy, politics, social security, sport, culture, weather, TV and broadcasting). The effect of compounding on domain is an important issue in our future work on a domain-sensitive extension of the dictionary. It is one of the criteria that should be taken into account when considering a rule-based (as opposed to a lexicalized) treatment of compounds.

Numerical expressions. The members of the numerical category are 39 expressions consisting of Arabian numerals, and hybrids of numberals, special signs, and single letters, such as *031/119840, 14.30, 25:E, and D0502.* (Some Roman numerals were also found, e.g. *VII*, but referred to the proper noun category as candidates for (parts of phrasal) proper nouns.)

Derivatives. Most of the derived words outside the scope of SMU (14 of 20) can be found in Svensk Ordboks as *morphological examples* (see table 11). This means, that their existence is confirmed, and that their meanings (definitions) should be derivable from the definitions of the words (lexemes) that they illustrate. When a lemma has more than one lexeme (definition), the morphological example tells us, on what definition the meaning of the derived word should be based, at least primarily (see *osäker*, in table 11). Five derivatives are not presented as morphological examples, but derived from one-lexeme lemmas, and so, the definition on which to base their derived meaning is uniquely determined. The remaining case, however, will cause overgeneration. *spelmässigt* is derivd from a noun with 9 definitions and an adjectival suffix with 2 definitions; the derivational power of the lexeme is in no way constrained, and we will have to consider them all equally well fit as bases of the derived words. (In all, there are 39,831 morphological examples in SOB.)

Type	Entry in SOB	Morph. ex.
avveckling avvecklingen	**avveckla** verb	+
mobbning	**mobba** verb	+
utvisning utvisningar	**utvisa** verb	+
sänkning	**sänka** verb	+ (1st lexeme)
pensionering pensioneringen	**pensionera** verb	- (1 lexeme)
enighet	**enig** adj.	+
osäkerhet	**osäker** adj.	+ (3rd lexeme)
skicklighet	**skicklig** adj.	+
trovärdighet	**trovärdig** adj.	- (1 lexeme)
öppenhet	**öppen** adj.	+ (4th lexemes)
författarinnan	**författare** subst.	+
socialdemokratisk socialdemokratiska	**socialdemokrat** subst.	+
mittfältare mittfältaren	**mittfält** subst.	- [6](1 lexeme)
spelmässigt	**spel** subst.	- (9 lexemes)
	-mässig	- (2 lexemes)

Table 9. Uncovered derivatives in PressFreq

Foreign words. To the foreign word catagory we have referred foreign words, not immediatly recognized as proper nouns. It includes function words (*der, des, die, the, til, to, with, you*) as well as content words (*chiffres, glasnost, labour, lettres, new, outs, science, télévisé, week*). We assume, that the function words, and most of the content words, are parts of phrasal proper nouns or other

phraseological expressions, and so they won't be entered as individual entries into the dictionary. *glasnost*, alone, will be entered into the SMU dictionary, and marked with respect to origin (PressFreq).

Syntagmatic words. Four missing types are examples of varying writing conventions, i.e. *iväg* (cf. *i väg*), *godnatt* (cf. *god natt*), *långtifrån* (cf. *långt ifrån*), *framförallt* (cf. *framför allt*). To the same category we refer the colloquial *gomiddag* (cf. *god middag*.) The one word variants will all be included in the SMU dictionary (the last one marked as colloquial).

Partial phrases. Four missing types are old inflectional forms appearing in phraseological expressions only, i.e. *godo* (till godo [to someone's credit etc.]; i godo [amicably etc.]), *sjöss* (till sjöss [at see]), *vintras* (i vintras [last winter]), and *somras* (i somras [last summer]).(*till godo* and *till sjöss* can be found among the examples of [1]**god** and [1]**sjö**.) The four expressions will be entered into the SMU dictionary as phrases.

Inflectional forms. Two underived inflectional forms were found to be unaccounted for, i.e. *måst* (supine of the verb *måste* [must]) and *törs* (present tense of the verb *tör/as* or *tord/as* [dare]), eventhough *törs* is part of an example of the verb. Frequent as they are found to be in newspaper text, both forms will be included in the SMU dictionary.

Conclusions

The SMU analyser, operating on Swedish text, works well as a tool for distinguishing between general vocabulary, as defined by the lemma entries of Svensk Ordbok (i.e. its explicitly defined vocabulary), and words outside that scope. As a result of the morphological analysis, members of the general vocabulary are identified and described in terms of lemma, part of speech, and form, and homographies are recognized in accordance with the lemma distinctions made in SOB.

The processing of four different Swedish materials has shown, that the SOB lexical coverage (in terms of types) ranges from 82% to 25%. The highest figures are valid for highfrequency words of newspaper text, PressFreq, and the lowest ones for highly specialized pharmacological text. In between we find some good 70% relating to general LSP (Language for Special Purpose) text.

The words outside the scope of the analyser indicate domain and type of the analysed text. The big amount of numerical expressions (and hybrids of numerals, special signs and single letters), i.e. 27% of the unanalyzed words, stand out in the pharmacological text as does that of proper nouns (close to 80% of the unanalyzed words) in PressFreq.

The PressFreq zero-parses have been examined in some detail, and categorized into: proper nouns, abbreviations, compounds, numerical expressions, derivatives, foreign words, syntagmatic words, partial phrases, and simple inflectional forms. Abbreviations, syntagmatic words, partial phrases, and inflectional forms form a, basically, closed set of a general character (in all, less than 150 items). They will all be included in the general part of the SMU dictionary (as one-word units or as phrases).

Most of the foreign words (except for *glasnost*) seem to be part of phraseological expressions (proper nouns), and so far, they will be disregarded, but *glasnost* be entered in the dictionary, marked by origin (PressFreq), as a first clue to domain. The proper nouns, forming a big, but, basically, closed and domain-related category, will be handled in the same manner. The numerical expressions, forming an open category, will be handled by means of rules, defined in the SMU grammar (see Sågvall Hein 1987).

Among the zero-parse derivatives, six types were found, i.e. verb-to-noun by means of the suffix *-ing* (the process; 8 cases), adj-to-noun by means of the suffix *-het* (presence of the property; 5 cases), noun-to-adj by means of *-isk* (the property; 2 cases), noun-to-noun by means of the suffix *-inna* (feminine; 1 case), noun-to-noun by means of *-are* (agens; 1 case), and, noun-to-adj by means

of *-mässig* (according to the noun etc.; 1 case). The first four types will be handled by means of word formation rules in the grammar, whereas the remaining two cases will be entered into the dictionary, marked by origin. This treatment is supported by SOB, presenting the first four types as morphological examples.

The most difficult category to handle is that of the compounds, being an open, productive category, with a complex semantics, dominating the zero-parses of general LSP text (see Sågvall Hein 1990). Further, compounding has a bearing on domain. In our continued work on the dictionary we will approach the problems of the compounds from the point of view of the effect of compounding on domain. The material presented by the application of the SMU analyser to text of different type and domain is a valuable source for such studies.

Notes

1 CF. SWETWOL by Karlsson (forthcom.) performing rule-based structural analysis of compounds and derivatives.

2 The secondary vowel deletion rule in part of the inflectional grammar and invoked by the dictionary search process.

3 in the sense of dictionary stem

4 Textual frequency data were not at hand when the analysis was carried out, so only lexical coverage can be accounted for here.

5 In a pilot study of a fragment of (2,500) types) of DefVoc the (572) types outside the scope of SOB were examined (see Sågvall Hein 1990).

6 Eventhough *mittfältare* doesn't appear as a morphological example, the relative *mittfältsspelare* does.

References

Blåberg, O. 1988. A study of Swedish compounds. Umeå University. Department of General Linguistics. Report No 29.

FASS. Farmaceutiska specialiteter i Sverige. 1985. [Pharmaceutical Specialties in Sweden.] LINFO.

Gellerstam, M. 1989. The Language Bank. The Department of

Computational Linguistics. University of Gothenburg.

Hellberg, S. 1978. *The morphology of present-day Swedish.* Stockholm. Karlsson, F. SWETWOL: A comprehensive morphological analyzer for Swedish. Forthcoming.

Östling, A. A Swedish Core Vocabulary for Machine Translation. This volume.

Shieber, S. 1986. An introduction to unification-based approaches to grammar. CSLI. Lecture Notes Number 4. Sjögreen, C. 1988. Creating a dictionary from a lexical database. In: *Studies in computer-aided lexicology.* Stockholm. Pp. 299-338.

Sågvall Hein, A. 1987. Parsing by means of Uppsala Chart Processor, (UCP). In: L. Bolc (ed.) *Natural language parsing systems.* Berlin & Heidelberg. Pp. 203-266.

Sågvall Hein, A. 1988. Towards a comprehensive Swedish parsing dictionary. In: *Studies in computer-aided lexicology.* Stockholm. Pp. 268-298.

Sågvall Hein, A. 1990. Lemmatising the definitions of Svensk Ordbok by morphological and syntactic analysis. A pilot study. In: J. Pind & Rögnvaldsson, E. (eds.) *Papers from the seventh Scandinavian conference of computational linguistics.* Reykjavík. Pp. 342-357.

Sågvall Hein, A. The SMU inflectional grammar. Uppsala University. Department of Linguistics. Forthcoming.

Sågvall Hein, A. & Ahrenberg, L. 1985. A parser for Swedish. Status Report for Sve.Ucp. June 1985. Uppsala University. Center for Computational Linguistics. UCDL-R-85-2.

Sågvall Hein, A. & Sjögreen, C. 1991. Ett svenskt stamlexikon för datamaskinell morfologisk analys. En översikt. [A Swedish stem dictionary for computational morphological analysis. An overview.] In: M. Thelander et al. (eds.) *Svenskans beskrivning 18.* Lund. Pp. 348-360.

Sågvall Hein, A., Östling, A. & Wikholm, E. 1990. Phrases in the Core Vocabulary. Uppsala University. Center for Computational Linguistics.

Sågvall Hein, A., Starbäck, P. & Wikholm, E. A pharmacological stem dictionary based on FASS, Pharmacological Specialties in Sweden 1985. Uppsala University. Department of Linguistics. Forthcoming.

Svensk Ordbok. 1986. *[A Dictionary of Swedish.]* Stockholm.

Teleman, U. 1974. *Manual för beskrivning av talad och skriven svenska.* Lund.

Anna Sågvall Hein
Uppsala University
Department of Linguistics
Computational Linguistics
Box 513
S-751 20 Uppsala
E-mail: uduas@seudac21.bitnet

Integrating Syntagmatic Information in a Dictionary for Computer Speech Applications

Dieter Huber
Chalmers University of Technology

Abstract

Conventional dictionaries, albeit they often comprise an impressive amount of *paradigmatic* information on various aspects of linguistic description, usually pay only little attention to the representation of *syntagmatic* information. Admittedly, apart from spelling conventions and rules of inflectional agreement, the co-occurrence of individual lexical items will not normally change the orthographic shape of a word when it appears in written text. In spoken language, however, the phonetic realization of words is heavily influenced by context and may change dramatically in a variety of ways, including segmental as well as prosodic features. These changes need to be taken into account in both computer speech synthesis and automatic speech recognition. In this paper, therefore, we argue for the inclusion of syntagmatic information in dictionaries which are developed for the special purpose of spoken language processing in computer speech applications. Two kinds of syntagmatic information will be considered in more detail: *Case Frames* and *Collocations*.

1. Introduction

Spoken language differs from written language in several important respects. For one, natural human speech does not normally present itself in the acoustical medium as a simple linear string of discrete, well demarcated and easily identifiable symbols (i.e. representing the letters and signs of some specified alphabet) and blocks of symbols (i.e. representing individual words separated by blanks), but constitutes a continuously varying signal which incorporates virtually unlimited allophonic variations, assimilations, reductions, elisions, repairs, overlapping segmental representations, grammatical deficiencies, and potential ambiguities at all levels of linguistic description. There are no "blanks" and "punctuation marks" to define words or indicate sentential boundaries in the acoustical domain. Important components of the total message are typically encoded and transmitted by non-verbal and even nonvocal means of communication. Syntactic structures, at least in spontaneous speech, are often fragmentary or highly irregular, and cannot be described in terms of established grammatical theory.

Given these differences, clearly, the computational models, tools and techniques developed for natural language processing (NLP) of written material are not immediately and automatically applicable to spoken language processing (SLP) of human speech. In particular, SLP for practical computer speech applications such as for instance text-to-speech synthesis (TTS) and automatic

speech recognition (ASR) requires lexical information that is not normally contained in conventional (including machine-readable and machine-tractable) dictionaries.

In an earlier paper (Huber 1989) a speech parsing algorithm has been presented which exploits the prosodically cued chunking present in the acoustical speech signal and uses it to perform speaker-independent segmentation and broad classification of continuous speech into functionally defined information units. A fundamental objective associated with this approach is to integrate speech signal processing and natural language processing techniques (both linguistic and stochastic) in order to fully exploit the combination of partial information obtained at various stages of the analysis. The contents and structure of the lexicon component to be used with this SLP parsing system have been described in (Huber 1990). In that paper, the overall format of the lexicon has been defined as a medium-sized Swedish monolingual pronunciation dictionary incorporating four kinds of lexical information:

- *phonological information*, i.e. narrow phonetic transcriptions reflecting both standard pronunciation usage and principle variants;
- *paradigmatic information* on various aspects of linguistic analysis specially relevant for SLP purposes;
- *syntagmatic information* reflecting the use of words in context;
- *statistical information*, i.e. data on the frequency of occurrence in a large corpus of language material.

The representation of phonological and paradigmatic information is described in more detail in a separate paper published in this volume (Hedelin & Huber 1992). In this paper I shall focus on the use of syntagmatic information in spoken language processing for practical computer speech applications such as speech synthesis and automatic speech recognition.

2. Syntagmatic Information

Words are normally applied in context, i.e. they co-occur with other words preceding and following them, with which they combine into larger structures like phrases, clauses, and sentences to express the intricate meanings of language. These co-occurrence relations are linguistically predictable to a greater or lesser extent, and can be specified in either grammatical or lexical terms as tactic (context-sensitive) rules and restrictions. For instance, adverbs co-occur with verbs (but not necessarily the other way round), transitive verbs require a direct object as complement (at least), and *kill* collocates with *animal, poison, victim*, etc, but not with, say, *sky, paper* and *blue*.

Conventional word dictionaries, albeit they often comprise an impressive amount of *paradigmatic* information on various aspects of linguistic analysis, pay only little attention to these kinds of *syntagmatic* relationships that exist between individual lexical items. In as far as co-occurrence data are included in dictionaries at all, they are usually found among the "examples of usage" listed under one or another of the component lexemes, i.e. stated implicitly, randomly, and without any claims of consistency and comprehensiveness. This imbalance between paradigmatic and syntagmatic information in most of today's dictionaries reflects not only the "division of labour" commonly assumed to exist between lexicographers on one side and grammarians on the other. Even more so, it reflects the traditional preoccupation of both lexicographers and grammarians with written language. Admittedly, apart from spelling conventions (e.g. sentence initial capitalization, use of a hyphen in certain compounds) and rules of inflectional agreement (e.g. he *works* versus they *work*) valid in some of the worlds languages, co-occurrence relationships will not normally change the orthographic shape of a word as it appears in written text. In spoken language, however, the phonetic realization of individual words is heavily influenced by context (both co-textual and

situational) and may change dramatically in a variety of ways, including segmental (e.g. assimilation across word boundaries, reduction, elision, sandhi) as well as prosodic (e.g. speech rate, duration, pausing, intonation, accentuation, stress) features. For example, the actual pronunciation of the Swedish word:

NATURLIGTVIS
(naturally)

which in isolated (lexical) usage may be transcribed phonemically as:

/nɑ'tɯrligt'vis/

and in a narrow phonetic transcription as:

[na'tɯ:[ɪt'vi:s]

may be realized phonetically in informal conversation as:

[n'ates]

or as:

[n'aes]

or even as:

[hanɦɑ:ɳatesˈɳɔŋ:a]

when it appears in the context:

HAN HAR NATURLIGTVIS MÅNGA...
(He has naturally many...)

thus involving not only various kinds of assimilation across word boundaries, but also a shift of accentuation with concomitant reductions and elisions of the unstressed syllables, and an increase of duration in both the stressed and in the phrase-final syllables.

Clearly, this kind of information about the phonetic variability of words in different kinds of co-textual and situational context is of paramount importance for all aspects or speech signal processing (analysis, synthesis, transmission, coding, compression, enhancement, etc) and computer speech technology (text-to-speech, speech recognition, speaker identification and verification, etc). For instance, text-to-speech systems using standard syntactic parsers designed to find "major syntactic boundaries" at which cross-word coarticaluation needs to be interrupted and the intonation contour has to be broken into separate units that help the listener to decode the message, invariably come up with the same two kinds of problems: (1) they tend to produce not one (the most probable, semantically most plausible) but several alternative parses, and (2) they produce too many boundaries at falsely detected or inappropriate sentence locations (e.g. Klatt 1987). Perceptual evaluation of these synthesized contours reveals that listeners get distracted and often even plainly confused by too many, prosodically and/or segmentally marked boundaries, while too few breaks just sound as if the speaker is simply talking too fast. This shows not only that the amount of segmentation and the correspondence between syntactic and prosodic units are dependent on the rate of speech, but also that listeners apparently neither expect, nor need, nor even want prosodically cued information about all the potential richness in syntactic structure described by modern syntactic theories, in order to decode the intended meaning of an utterance.

In order to be able to handle these kinds of phenomena in practical computer speech applications, we propose to include co-occurrence information into our Swedish pronunciation dictionary, which

has been designed to provide continuous lexical support to the language processing components in text-to-speech synthesis and automatic speech recognition. In the following two sections, two kinds of co-occurrence data will be discussed in more detail: *case frames* and *collocations*.

3. Case Frames

The grammar formalism adopted for the SLP parsing algorithm presented in Huber (1989) is based on Fillmore's case grammar (Fillmore 1968). The following reduced set of cases for verb entries has been adopted from Stockwell, Schachter and Partee (1973):

AGENT	- animate instigator of the action
DATIVE	- animate recipient of the action
INSTRUMENTAL	- inanimate object used to perform the action
LOCATIVE	- location or orientation of the action
NEUTRAL	- the thing being acted upon

According to this approach, a caseframe is thus defined as an ordered array composed of the entire set of cases:

```
caseframe = array[agent...neutral]
```

in which each case can be either required (`req`) or optional (`opt`) or disallowed (`dis`), and must be marked accordingly.

Since different verbs often share the same particular kind of caseframe, we propose to store the entire set of 3^5 logically possible caseframes as an indexed list, using the indices as pointers (identifiers) with the respective verb entries in the lexicon. Thus, instead of listing the complete case-frame specification together with the lexical entry, as in the following example for the Swedish verb "`hacka`" (to chop):

```
hacka 3    type: verb
           infl: v1
           freq: 4
           tran: [2hak:a]
           case: agent - req
                 dative - dis
                 instrumental - opt
                 locative - opt neutral - opt
```

we propose to use the indexed representation format, which results in the more space-economic and search-effective structure:

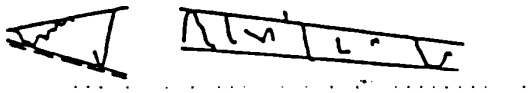
```
hacka 3    type: verb
           infl: v1
           freq: 4
           tran: [2hak:a]
           case: 97
```

Further work in sematic case frame representation is presently directed towards the extension of individual case states marked as with "`req`" or "`opt`" lexical hypotheses derived from KWIC-studies of coherent speech.

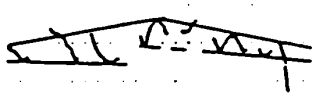
4. Collocations

Collocations constitute the recurrent combinations of words as they appear in context. These combinations include both phraseologically relevant collocations such as *när det gäller att* (Eng: as far as ... is concerned) and phraseologically irrelevant collocations, like *och han* (Eng: and he). While only the former, i.e. phraseologically relevant class of collocations has so far attracted some interest in the linguistic research community, both groups are interesting and important from an information theoretical point of view.

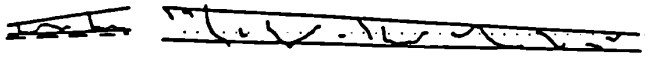
In a sense, collocations can be seen as occupying the border area between grammatical and lexical descriptions of language. In most of today's NLP applications they are handled by the syntactic parsing component with little or no support from the lexicon. We argue, on the other hand, that collocations are primarily a lexical phenomenon that can be handled more effectively by the lexicon component and thus should be included in the dictionary. After all, it makes little sense and would indeed be extremely uneconomical for real-time computer speech applications, to generate these recurrent constructions over and over again from their component words. Also, there is convincing evidence from psycholinguistic research indicating that collocations are indeed part of the mental lexicon that a speaker has at his or her disposal. In natural human speech, collocations are normally processed as coherent lexical blocks equivalent to prosodic phrases (cf. Selkirk 1984), i.e. signaled by a high degree of coarticulation between their component words and an integral melodic pattern. This latter tendency has been found to be particularly prevalent in sentence initial collocations, sentence final collocations, and idioms (i.e. collocations whose collective meaning can not be derived on the basis of the meanings of its component words). Figure 1 below shows examples from our Swedish speech material (cf. Hedelin & Huber 1990) which illustrate the effects of prosodic marking for sentence initial collocations.



Text 1; Sentence 30; Speaker BMH (female)
Dagen därpå undertecknades i klostret ett fördrag...



Text 1; Sentence 30; Speaker LR (female)
Dagen därpå undertecknades i klostret ett fördrag...



Text 2; Sentence 24; Speaker LO (male)
Men å andra sidan har 40 procent av städernas och köpingarnas...

Figure 1. Intonation contours for the sentence initial collocations

Based on these findings, we propose to incorporate the most frequently occurring collocations in the dictionary, including both their orthographic shape and narrow phonetic transcription(s). Special attention is directed towards a complete and comprehensive representation of sentence initial, e.g.

TROTS DETTA

[trɔtsˋɖɛtːa]

(despite of that)

sentence final, e.g.

ÅR SEDAN

[ˈoːsẹn]

(years ago)

and idiomatic constructions, e.g.

I OCH FÖR SIG

[iːɔfœˈsẹj]

or

[iːfœˈsẹj]

(as a matter of fact)

both because of their recurrent status as prosodically cued blocks, and their high potential for the kind of expectancy-driven island parsing approach advocated in (Huber 1989). Consequently, collocations are listed under each component head word, which enables the parsing system to effectively guide the word segmentation and identification process by generating expectations resulting from partial linguistic analyses.

Complete listings of the Swedish collocations in descending order of frequency, based on statistical analysis of a one-million-words corpus of news texts (cf. Allén 1975), have been obtained from the Department of Computational Linguistics, University of Gothenburg. These data are stored in machine-readable format which allows direct transfer and incorporation of the orthographic representations into the pronunciation dictionary.

In addition to these listings, we also propose to include two statistical measures: (1) the observed frequency (F_{col}) for each collocation, i.e. indicating the number of occurrences in the accumulated corpus material, and (2) the *constructional tendency* (F_b), i.e. a measure which indicates for each component word its tendency to be bound in collocational constructions. F_b thus represents the ratio between the instance frequency of the word in collocational constructions and the total frequency of the same word. Both measures are included primarily in order to generate expectations for nearest-word neighbours during the SLP parsing and recognition process.

5. Present Status

The syntagmatic information described in this paper is presently incorporated into our Swedish pronunciation dictionary (cf. Huber 1990, Hedelin & Huber 1992). Regarding the inclusion of statistical data concerning the frequency of occurrence of collocations and their constructional tendency, we are of course aware of the three-fold discrepancy:

- that the frequency data provided by Språkdata reflect solely the distribution over the accumulated one-million-words corpus, thus ignoring the dispersion between different text types and genres;
- that these frequencies are based on the evaluation of written news texts, and do not ideally represent the statistical distribution of words in comparable news speech;
- that word frequencies in general require continuous updating both in a macroperspective (i.e. to reflect language changes and the concomitant expansion and reorientation of vocabularies) and in a microperspective (i.e. to capture the prevalence of thematically dictated terminology in application specific domains).

References

Allén, S. (1975) "Frequency Dictionary of Present-Day Swedish based on Newspaper Material, Part 3: Collocations", Almqvist & Wiksell, Stockholm

Fillmore, Ch.J. (1968) "The Case for Case", in: E Bach and R T Harms, Universals in Linguistic Theory, Holt, Rinehart & Winston, Chicago

Hedelin, P. & D. Huber (1990) "The CTH Speech Database: An Integrated Multi-Level Approach", *Speech Communication* Vol. 9, No. 4, pp.365-374

Hedelin, P. & D.Huber (1992) "A Pronunciation Dictionary for Swedish", this Volume

Huber, D. (1989) "Parsing speech for structure and prominence", Proceedings of the First Inter-national Workshop on Parsing Technologies, CMU, Pittsburgh, Penn.

Huber, D. (1990) "An Electronic Dictionary for Computer Speech Applications", Proceedings of the International Workshop on Electronic Dictionaries, Oiso, Japan

Klatt, D.H. (1987) "Review of Text-to-Spech Conversion for English", *Journal of the Acoustical Society of America* Vol. 82, No. 3, pp.737-793

Selkirk, E.O. (1984) Phonology and Syntax: The Relation between Sound and Structure, The MIT Press, Cambridge, Massachusetts

Stockwell, R.P., P. Schachter & B.H. Partee (1973) "The Major Syntactic Structures of English", Holt, Rinehart & Winston, New York

Dieter Huber
Department of Information Theory,
Chalmers University of Technology
S-412 96 Gothenburg,
Sweden

PC-phonetics: A help or a strain for the philologist?

Anna K. Lysne
University of Bergen

The title is an EITHER – OR question. The paper will probably reveal that the answer must be BOTH – AND.

In this paper a report will be given of the work with the project 'Norwegian learners' problems with English prosody. Norwegian-English intonation'.[1] The focus will be on hindrances met with, when changing from an auditive to an acoustic analysis of intonation, at the same time as there will be an attempt to show how modern research can be a help to overcome difficulties. Finally, results arrived at from a detailed study of the use of pitch range by Western Norwegians will be presented and compared with other studies.

An auditive phonetic error analysis of the intonation of Norwegian students of English was started some years ago. The material used was a selection of readings, consisting of a 3-minutes' diagnostic test for 1st term students, collected at Engelsk institutt, Universitetet i Bergen, from 1977 to 1989.

Results from the auditive analysis (of 160 samples, representing readers from all parts of the country) have been published earlier (cf. Lysne 1985, 1988). As it is generally acknowledged that to rely only on auditory analysis of speech is not satisfactory, it was tempting to test the auditive results instrumentally. However, in the literature, for instance in the works of the 'Dutch School' (cf. Willems 1982; de Pijper 1983; Collier 1989; and t'Hart et al. 1990) it is emphasized that instrumental analysis of intonation is far from satisfactory, too.[2] In brief, the 'Dutch' method of analysing intonation amounts to starting with a detailed acoustic analysis of the fundamental tone (Fo), the physical correlate of PITCH, which is what we hear as speech melody or INTONATION. By the aid of aural control and stylized resynthesized curves Collier and Terken (1987) arrive at a 'close copy' of the Fo curve by eliminating excursions in the original curve which can be deleted without affecting the auditory impression of 'sameness'/naturalness. Their 'close copy' variant 'mainly eliminates the Micro-intonational modulations from the Fo curve (cf. below) and respects whatever variation there is in the overall shape of the individual pitch movements ...' (ibid. p. 166).

This kind of research reveals that there is a great amount of irrelevant information in an acoustic representation of intonation, something also underlined by Crystal (1969) and Gibbon (1976) among others. Thus faced with instrumental analysis, the inexperienced performer can easily be confused. However, consolation is found in e.g. Collier (1989). Collier advocates the 'perceptual detour', underlining strongly that 'perceptual verificaton' is of major importance in the analysis of intonation. It is apparent that some of the information in the sound curve that may be labelled irrelevant may lie above the frequency threshold of hearing, but our perception apparatus seems to work like a filter, ignoring irrelevant signals. The computer, on the other hand, does not ignore anything, and this is where the problems lie for an only auditively trained philologist, who is used to drawing simple lines to illustrate intonation patterns (cf. fig. 1). Nevertheless, what is perceived by a trained ear, may prove to be just what is needed in a pedagogical situation. Actually, very little

experience in studying sound curves on the PC is needed in order to understand warnings or recommendations like 'You must rely on your ear and not on what you see on the screen' (Collier, personal communication), and 'Trust your ear' (Dickson, personal communication).

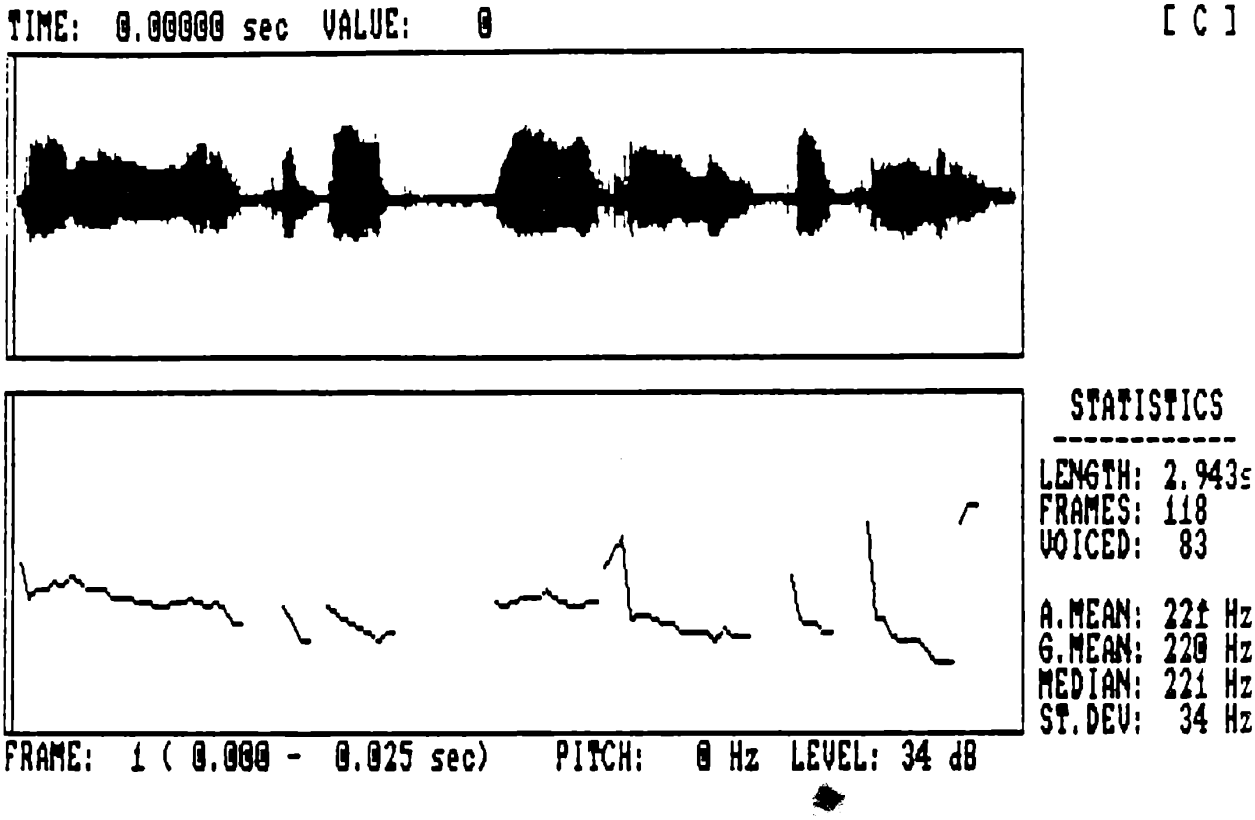


Figure 1: MSLPITCH curve representing 'William went to bed very early that evening' (not cleane or repaired).

Extralinguistic and practical problems

Before looking at what is perceived of the speech signal (or how we perceive), it might be useful to look at extralinguistic and practical hindrances encountered in digitizing speech (and in acoustic analysis):

At the worst, signals digitized and appearing on the screen might represent noise from the equipment used. Further, unsatisfactory quality of the original tape recording might be a problem in case the material was not meant to be used for acoustic analysis.[3] Wrong clock speed of the PC in relation to speech tempo might be another problem.[4] In this last case cautious adjustment is essential in order to secure correct and identical processing of all the material. It is the digitizing-process which is sensitive to correct clock speed, the process of analysis of data is not affected.

The Fo limits of the normal human voice are well known, as are also the perceptual limits of tonal frequency (cf. Ladefoged 1972). However, in relation to the perception of running speech Clark and Yallop (1990) strongly underline that perception is relative. This also applies to 'the perception of sound, studied under the heading of PSYCHOACOUSTICS' (ibid. p.208). Concerning this problem t'Hart et al. (1990:10) maintain that '...the psychoacoustically defined thresholds do not

account for the selective sensitivity of the listener when extracting pitch information from the speech signal. ... the listener applies selective attention in the perception of pitch in speech'.

From what has been said above one might conclude that the computer with its acoustic analysis brings forth and shows details in an Fo curve which are linguistically irrelevant and which we are liable to ignore as listeners.

What we perceive of tone in speech depends on several factors. Quantity/timing is a decisive factor.[5] According to Fay (1966:16) 'Time... is extensively involved in the many phases of the total auditory process.' Gibbon (1976:52) says 'Prosodic features ... are generally defined with reference to the time dimension, ... which leads to the vexed question of the linguistic relevance of static trajectories or points in trajectories compared with dynamic trajectories'. Gibbon (ibid. p. 55) also says 'pitch perception requires at least 20 – 25 msecs even for static trajectories ..., and thus cannot be abstracted from the time dimension'. Gibbon's 'vexed question' refers to the fact that if a moving tone ('dynamic trajectory') occurs immediately before or after a level tone ('static trajectory') it is the relative duration of the trajectories which determines whether we hear a moving tone or a level tone (cf. Brandt 1976). A sudden peak of short duration is not heard in a smooth sequence.

In the present study it has been necessary to take into account the time dimension especially when cleaning Fo curves. This is necessary in cases where the shape of the curve does not correspond with what we hear of TONE when listening to the signal. From what has been stated so far it should be clear that it might be necessary fairly often to repair or clean the acoustic curve, which can contain such a lot of faulty information, in order to make it appear more like what it sounds. Obvious noise is easily detected and deleted, and so is strong aspiration and breath wrongly computed as voiced (cf. figs. 1 and 2). Further, we find the so-called OCTAVE JUMPS: If Fo lacks intensity one of the HARMONICS might show up instead[6] (cf. end of fig.1). The pitch perceived will still be that of the fundamental, which means that we hear a lower tone than the one appearing in the Fo curve, and reparation is necessary (cf. fig. 2).[7] This is often the case in **weakly pronounced endings**: A low tone is heard and the signal is registered to have a higher Hz-value, (sometimes showing a sharp rise which is not there – and 'which we cannot make or hear' (Henning Reetz, personal communication).[8]

WEAK ENDINGS posed a great unforeseen problem in the pitch-range analysis. Here one has to find the highest and the lowest Hz-value in the unit, and the lowest value will occur at the end of a long unit, especially so in a statement conveying new information. Sørensen and Cooper (1980:407) affirm that 'The general direction of Fo in a sentence is downward. ...This general fall has been termed Fo declination.' Clark and Yallop (1990:285, 286) say that there 'appears to be an almost universal tendency in language, namely a moderate progressive fall in pitch from the beginning to the end of any sequence of speech of appreciable length... . There has been considerable debate about the status and causes of declination. It has also been suggested that declination effects are observed mainly in reading aloud,... and that they are much less noticeable in the patterns of informal speech'.[9] An important point in this discussion would be that the linguistic units in informal speech will be shorter than sentences (cf. Brown and Yule 1983). Thus it is reasonable that we will get longer units in reading, which is based on written text containing complete sentences. This would make the downdrift more noticeable in reading. According to Lehiste (personal communication) 'The longer the unit the higher the onset has to be.' Sørensen and Cooper (1980) verify that there is a tendency for Fo to start higher at the beginning of a longer constituent. Cf. also Brown et al. (1980:132): '... it does seem that when the speaker has more ground to cover she begins higher'.

MICRO INTONATION (cf. p.1 above) has to be tackled as a special phenomenon. The term refers to segmental influence on the Fo curve of which there are two types: the intrinsic pitch of vowels and the influence of a consonant on the pitch of a following vowel (cf. Lehiste 1970, Ladd and Silverman 1984, and Reinholt Petersen 1986). Close vowels will have higher pitches than open vowels. 'Inherent Fo' is another term for this (Scheffers, personal communication). According to

Lehiste (ibid.) intrinsic pitch appears to be a physiologically conditioned universal. The pitch of a vowel will in addition be increased by a preceding voiceless consonant and lowered by a voiced consonant in the same position. This type is referred to as 'contextual Fo' (Scheffers, ibid.), and 'coarticulatory Fo fluctuations' (Reinholt Petersen, ibid.). Clark and Yallop (1990:282) state that 'The reasons for this conditioning of pitch are not fully understood.' And t'Hart et al. (1990:14-15) contend that micro intonation 'causes perturbations, not intended by the speaker. Evidently, such perturbations ... cannot be considered as constituents of the pitch contour as a linguistic entity; but they make the interpretation of Fo curves in terms of underlying intonation patterns all the more difficult.'

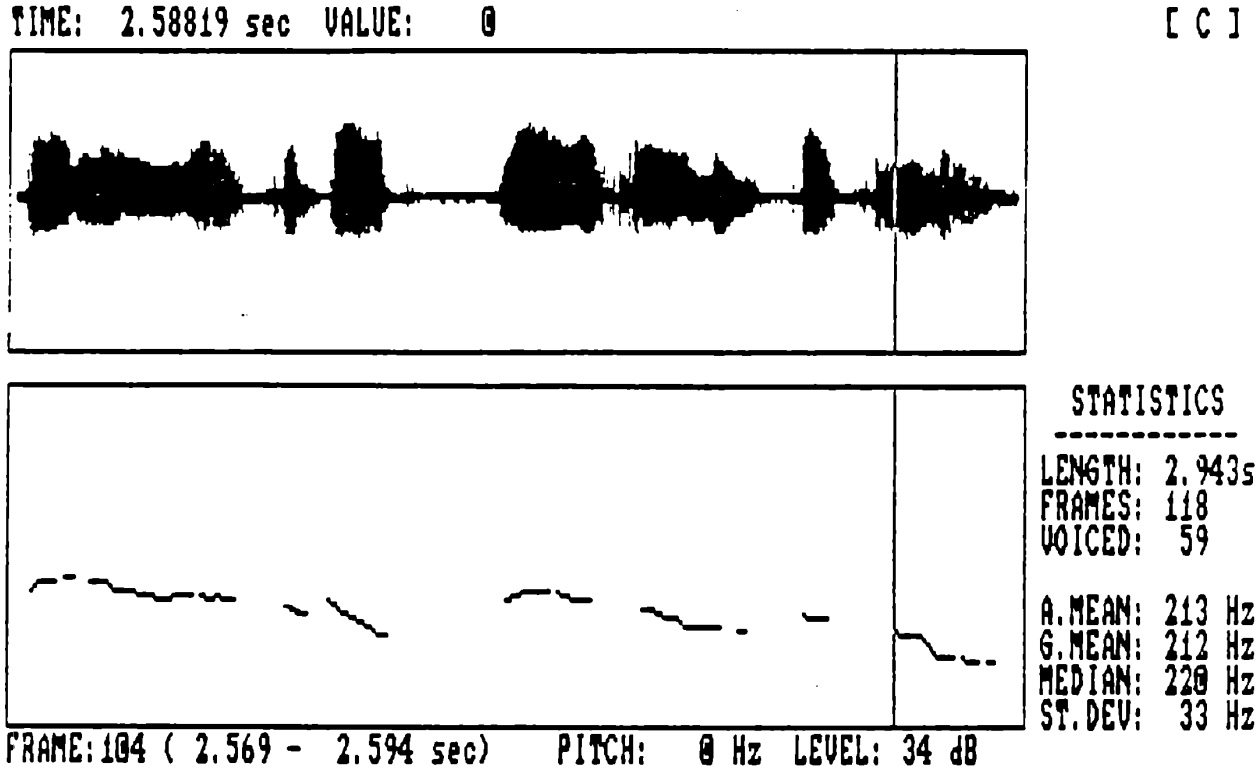


Figure 2: MSLPITCH curve representing 'William went to bed very early that evening' after being cleaned and repaired.

Inexperienced in the field of acoustics I welcomed the following statements on micro intonation by M. Scheffers (ibid.), 'These effects are relatively small in size (some percent) and difficult to distinguish in the Fo contour of a phrase because they are of the same order of magnitude as the inaccuracy of the Fo analysis method. ... no analysis method is perfect' and 'you have to learn to detect errors in Fo measurements. ... In teaching intonation in a foreign language, however, you can safely ignore these minor Fo excursions since it is generally agreed that they are involuntary and physiologically driven'. Obviously, in a learning situation anything physiologically driven will come naturally. In my analysis I have followed Collier and Terken (1987) deleting signals which are impossible to detect auditively (cf. fig. 1). (However, in one of my sentences (nr 7) it seems as if intrinsic pitch and co-articulatory Fo might have an impact to be counted with (cf. below).

We shall now turn to a more detailed presentation of a survey of the use of pitch range (PR) of students of English from Western Norway (Hordaland more precisely; 47 male and 74 female

informants). As stated above, findings from the earlier auditive analysis were to be controlled by looking at the acoustic features of problems observed and listed.[10] The reasons for looking into PR first were mainly practical. Intonation studies demand units covering a certain stretch of speech. And thus the first sentence of the story read was digitized, representing an 'intonation unit' in most of the readings.[11] Intonational 'errors' listed (cf. note 10) were expected to appear within the unit. (To start with digitizing this sentence was a part of learning to use the MSL programme. In addition it was a way of becoming familiar with the unknown activities of digitizing speech and analysing acoustic sound curves).

Even at a very early stage I had a notion that analysing only the first sentence would be unsatisfactory as a basis for PR-studies. Therefore, for some informants two similar sentences (4 and 6) were digitized as well.[12] These have not been systematically analysed yet. Professor Lehiste (personal communication) suggested that looking at two more sentences (7 and 12) might be interesting. The reasons given were that sentence 7 represents a yes-no question, and sentence 12 would give the lowest possible Fo value, marking finality.[13] Actually, as for sentence 7 an extra high onset might be expected, caused by a maximum effect of micro intonation: a voiceless fricative /ʃ/ preceding the close vowel /i:/. But, seen in relation to sentence 1, the PR-value of sentence 7 varies from informant to informant as 'The effect of vowel intrinsic pitch is negligible in unaccented syllables' (Ladd and Silverman 1984:31). And some readers have stressed *she* whereas others have not.[14]

It is worth noting that the readings used for the auditive analysis (five from each area) were not all suitable for instrumental analysis. In spite of the fact that these had been carefully selected as 'good readings' and fit for auditive analysis, very few could be used for acoustic analysis. Therefore, the final results reported here cover quite a few other informants, and many more informants from each area. According to Professor Victor Zue, MIT, (lecture, Edinburgh, 1988) one needs a great amount of data in order to draw conclusions from acoustic measurements.[15] Thus the plan of using readings by native RP speakers among our students for comparison had to be dropped for lack of enough data. The UCL study (Barry et al. 1989; cf. fig. 3) was welcomed for comparison.[16]

Results of the study reported here compared with those of the UCL-study are illustrated in fig.4. The histogrammes illustrate a marked difference in the use of pitch range of the Norwegian students from Hordaland and RP speakers.

Table 2a. 90% range in octaves from 1st-order Dx analysis for 2 male and 2 female native speakers in 6 languages

	Speakers			
	F.1	*F.2*	*M.1*	*M.2*
Danish	0.67	0.59	0.51	0.47
Dutch*	0.71	0.67	0.47	
English	1.30	1.15	1.03	1.03
French	0.47	0.99	0.55	0.79
German	0.71	0.55	0.59	0.51
Italian	0.83	0.67	0.71	

Figure 3. Table borrowed from Barry et al. 1989.

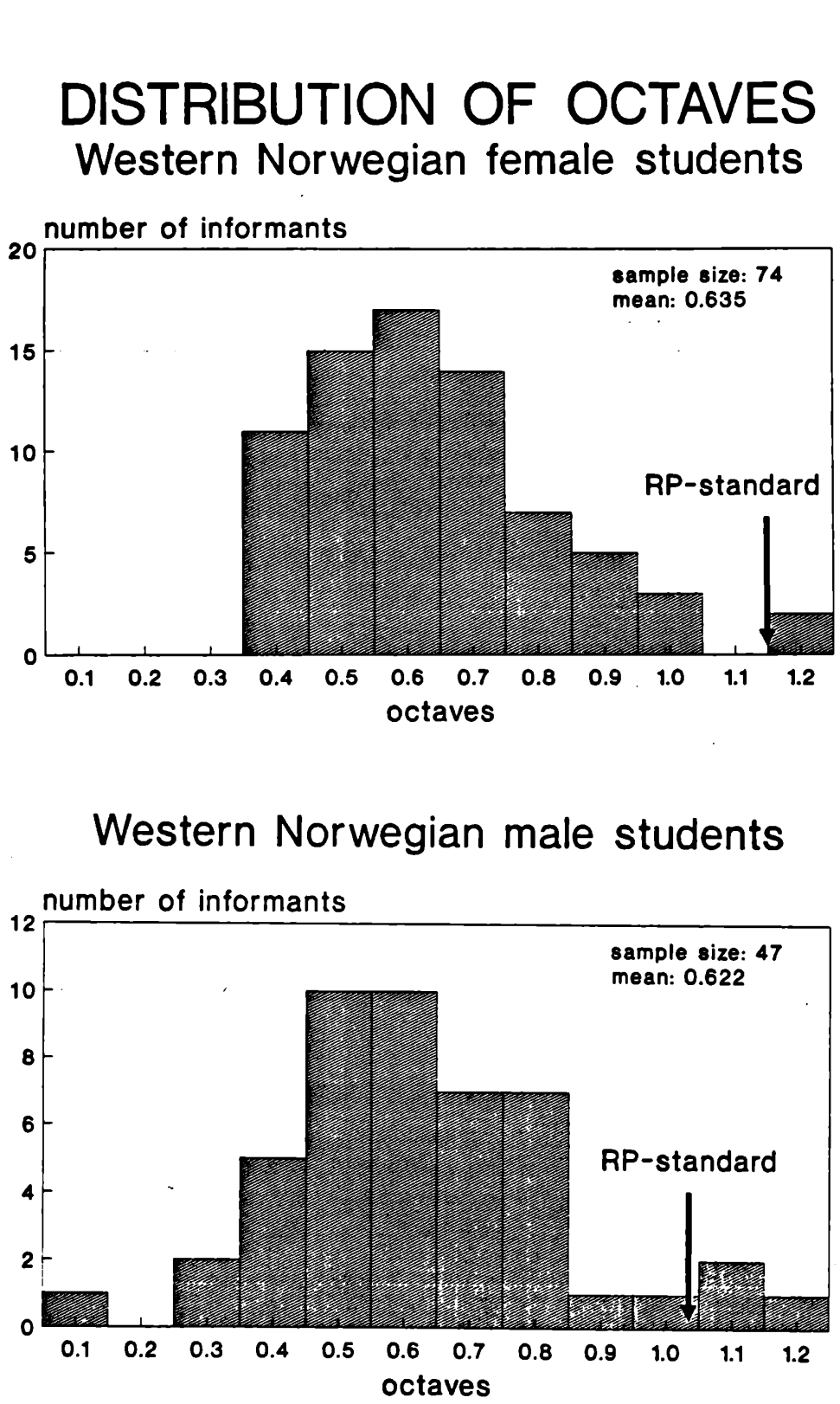


Figure 4: The lower values given for English (RP) (fig. 3) are indicated here. Both histogrammes show that more than 90% of the Norwegian informants use a pitch rang lower than that indicated for English. (A few falsetto voices cause the mean value in the male-student histogramme to appear too high).

It is also interesting to see that the octave values for Danish in fig. 3 come very close to the mean PR values of the present study of Norwegian speakers.

The validity and value of the work referred to can be discussed. The problem of the use of only the 1st sentence (so far) is one point in question (cf. notes 11 and 13).

Pitch range will vary in the use of one and the same speaker at the same time as it is obviously language specific. Hall (1972) states that the use of a wide PR is one of the most characteristic features of British English. Further de Pijper (1983:91) points to the difference between British English and Dutch in the total PR covered, indicating a ratio of 12:6 semitones, i.e. 1:1/2 octave.

PR is recognized as a paralinguistic feature; nevertheless, according to Esling and Wong (1983) it can influence intelligibility. A too narrow range gives an impression of weariness (cf. Abe 1980). Thus lack of interest, which might amount to impoliteness, could be demonstrated unintentionally by a Norwegian speaking English. According to Crystal (1969) PR is a regular and important means of emotional expression. In spite of Professor Lehiste's contention that PR is is of no interest linguistically (personal communication), it must be said to be of interst communicatively. In a symposium on 'Intonation: models and parameters' at the XII International Congress of the Phonetic Sciences, Aix-en-Provence, August, 1991, Professor Nooteboom claimed that phoneticians still know too little about the so-called paralinguistic features. The point is that synthetic speech reveals what features are vital for 'naturalness'. Thus knowledge about PR differences must also be of interest pedagogically.

Finally, a few concluding remarks are in order.

The PC has made phonetic research more accessible, as such studies can be pursued without assistance and access to a laboratory. How demanding this is will depend on the background of the analyst. It is difficult to be a multi-specialist. I would like to contend that thorough linguistic and phonetic knowledge of the language to be studied is a must, as this takes a long time to acquire. The acoustics of speech must be said to be a more limited area. It seems a less time-consuming task to acquire the knowledge of acoustics needed to be able to use the programmes available today for this kind of research. Nevertheless, a visit to a phonetic laboratory would no doubt be useful. To be acquainted with studies in speech perception is a necessity. Here help is to be found in recent literature on the subject.

Notes

1. *Prosody* comprises suprasegmental features of speech such as *stress*, *tone*, and *quantity*. The term *intonation* refers to the tonal movement of an utterance, i.e. variations in *pitch*, of which the fundamental tone (Fo) is the physical correlate. The Fo curve pictures the frequency of the vibrations of the vocal cords. For this project sound curves have been produced on the PC screen by the aid of the digitizing programme MSL (cf. Dickson 1985; Lysne 1989). The sound signal was transferred from a JVC cassette recorder/player. Detailed Fo analysis has been performed by the additional programme MSLPITCH (cf. Dickson 1987).

   Sentences digitized for analysis:
   1. William went to bed very early that evening.
   4. He hadn't heard from Jane since he sent her the wire.
   6. He had said something about a missing coat.
   7. Could she have found out so soon?
   12. After all it didn't matter.
2. Cf. t'Hart et al. 1990:24: 'The satisfaction which phoneticians may feel with the enormous progress made during the last decade in the automatic determination of Fo is overshadowed by

the fact that the finer the performance of the techniques applied, the more it becomes apparent how irregularly the vocal folds vibrate'.

3. The recordings used as material here (a by-product of a diagnostic test) were actually recorded in a speech-lab (class room) with 20 booths/readers). The equipment was the cassette recorder of Tandberg IS8, type: TCR 5500, 2 tracks, with attached headphone.
4. In the MSL-programme this is corrected by adjusting the "configuration file" as one listens in while changing/rectifying the appropriate values. In our case this was necessary when changing to a computer with greater capacity than the one first applied.
5. Cf. comment heard in a Norwegian radio (NRK) programme on music: 'Tid er musikkens verktøy' (= time is the tool of music). This might also apply to speech and hearing.
6. Very often this will be the second harmonic (with a frequency like Fo multiplied by 2, which will be the same as the OCTAVE VALUE of the fundamental (cf. Clark and Yallop 1990:210: 'Complex sounds pose a curious problem. ...It does not matter what the amplitude of that fundamental is in relation to the other harmonic components of the sound. ... even if the fundamental is removed by some form of electronic processing..., a pitch corresponding to the fundamental, known as the 'phantom fundamental,' will still be perceived ... .
7. What is originally digitized by the MSL programme will still be in the file. Cleaning/repairing an Fo curve in MSLPITCH affects only the visual curve, and the statistics conforms to this. (The original signal can always be reloaded and analyse again e.g. with changed parameters).
8. At an early stage in my work I blamed this 'end-problem' on poor recordings. However, the fact seems to be that this is a general problem in accoustic research of speech, as endings, especially in longer units, are very often weak because of lack of intensity.
9. According to Collier (1989) this downdrift was named DECLINATION by Cohen and t'Hart (1965); cf. also Thorsen (1980); Cohen, Collier, and t'Hart (1982); de Pijper (1983); t'Hart, Collier, and Cohen (1990).
10. The most frequent problems detected and listed as interesting were:
    1. The un-English finishing off of an intonation unit. What can be wrong with the fall/rise of a Norwegian speaking English?
    2. The starting-point/'onset' is not high enough (as verified by teaching experience).
    3. Is there a lack of use of 'pitch range' in 'Norwegian English'? (a phenomenon closely related to point 2; also verified by teaching experience).
    4. The overall tone pattern in a unit is not particularly English; do we move about/jump up on unstressed syllables where native speakers keep a steady course?
    5. What are the reasons for un-English rythm in 'Norwegian English'? (Is this a question of timing/lack of deaccenting?)
11. The choice of the first sentence for PR-studies has been questioned by experts, as the first sentence will normally be pronounced on a higher pitch than what follows (Noven, Slethei, and Lehiste) (personal communication). But in the recording-situation the informants had to read a sentence in Norwegian introducing themselves as a start. (Ex.: "Dette er student nr B 18 V 79." (10-12 sylls, statement, new information; as for the code see below). How far this introduction in Norwegian influenced the immediately following performance in English will be left here as an open question).

    However, considering the fact of the higher pitch of the first unit, this cannot be said to be negative, or make the work presented here particularly unreliable, as it does not support, but rather refute the underlying hypothesis and findings of this study: that Norwegians, westerners in particular, do not use enough PR when speaking English as their onset is not high enough.

    The original code of the student/reader is used as filename in the PC directory: B refers to lab-number, *18* to the booth, *V79* refers to term (Spring 1979). The number of the sentence is

added to this: B18V791 will be the filename of the sentence 'William went to bed very early that evening' of a reading that can be traced back to the tape.

12. Similarity here refers to kind of information, length, and grammatical structure.
13. It would be enormously time consuming and demanding as for PC-capacity to go into all these sentences in detail. Therefor, only a few spot checks have been made, revealing PR-values slightly different from those of sentence 1 as expected; but so far the hypothesis behind this work has not been disproved. Octave values of other units (sentences 4, 6, 7, and 12) analysed so far (about 20 in number) have been below those reported for RP (cf. fig. 4.)
14. 'As for vowel-intrinsic pitch, the difference in Fo between close and open vowels may amount to three semitones (0.25 octave). Thus, it is plausible that such a difference can be perceived, even in the dynamic situation of speech melody, and even though it has been reported ... that listeners tend to at least partially compensate for the effect' (t'Hart et al. 1990).
15. Professor Fry (UCL, 1976, personal communication) suggested three informants per area for aural analysis of speech. Here one need not distinguish between male and female speakers. In an acoustic analysis, on the other hand, these have to be kept apart, because of physiological differences. In this study I aimed at five male and five female voices from each area. For some areas the number is lower because of the lack of acceptable recordings.
16. This study was approved of by Professor Lehiste as valid for comparison (personal communication). Very little is to be found about PR in the literature. The reason for this is easy to see, as this kind work would be difficult to do without the computer. Experts only would be able to appreciate the amount of work behind such a study. It is particularly interesting that at the European Conference on Speech Technology, Edinburgh, September 1987, my inquiries about PR studies in English yielded no results. Two years later at the European Conference on Speech Communication and Technology, Paris, September 1989, the UCL study was reported.

Bibliography

Abe, I. 1980. 'How vocal pitch works'. In Waugh and van Schooneveld (eds.), 1980:1-24.

Abercrombie, D., Fry, D.B., MacCarthy, P.A.D., Scott, N.C., and Trim, J.L.M. (eds.), 1964. In *Honour of Daniel Jones*. London: Longman.

Ainsworth, W.A. and Holmes, J.N. (eds.), 1988. Speech'88. Proceedings, 7th FASE Symposium, Edinburgh.

Bannert, R. 1979. *Ordprosodi i invandrarundervisningen*. Praktisk Lingvistik 3-1973, Lund: Lunds Universitet.

Barry, W. Grice, M. Hazan, A. Fourcin, A. 1989. Excitation distributions for synthesised speech. In Tubach and Mariani (eds.), 1989.

Black, J.W. 1975. Introduction: sidelights on measurements. In Singh (ed.), 1975:1-15.

Bolinger, D. 1958. 'A theory of pitch accent in English'. Word, 14, 2-3.

Bolinger, D. 1970. 'Relative height'. In Bolinger (ed.), 1972: 137-153.

Bolinger, D. (ed). 1972b. *Intonation*. Harmondsworth: Penguin.

Bolinger, D. 1975. *Aspects of language* (2nd edn). New York: Harcourt Brace Jovanovich.

Bolinger, D. 1986. *Intonation and its parts*. London: Edward Arnold.

Brandt, J.F. 'Perceptual psychophysics: speech and hearing'. In Lass (ed.), 1976: 459-483.

Brown, A. (ed.), 1991. *Teaching English pronunciation*. London and New York: Routledge.

Brown, G. 1977. *Listening to spoken English*. London: Longman.

Brown, G., Currie, L. and Enworthy, J. 1980. *Questions of intonation*. London: Croom Helm.

Brown, G. and Yule, G. 1983. *Teaching the spoken language.* Cambridge: Cambridge University Press.

Chafe, W.L. 1970. *Meaning and the structure of language.* Chicago: The University of Chicago Press.

Clark, J. and Yallop, C. 1990. *An introduction to phonetics and phonology.* Oxford: Basil Blackwell.

Cohen, A. and 'T Hart, J. 1967. 'On the anatomy of intonation'. *Lingua*, 19:177-192.

Cole, R.A. (ed.), 1980. *Perception and production of fluent speech.* Hillsdale, New Jersey: Laurence Erlbaum Associates, Publishers.

Collier, R. 1989. 'Intonation analysis: The perception of speech melody in relation to acoustics and production.' In Tubach and Mariani (eds.), 1989, vol. 1: 38-44.

Collier, R. and Terken, J. 1987. 'Intonation by rule in text-to-speech applications. In Laver and Jack (eds.), 1987: 165-168.

Couper-Kuhlen, E. 1986. *An introduction to English prosody.* London: Edward Arnold.

Cruttenden, A. 1986. *Intonation.* Cambridge: Cambridge University Press.

Crystal, D. 1969a. *Prosodic systems and intonation in English.* Cambridge: Cambridge University Press.

Crystal, D. 1969b. 'The intonation system of English'. In Bolinger (ed.), 1972: 110-136.

Crystal, D. 1980. 'The analysis of nuclear tones'. In Waugh & van Schooneveld (eds.), 1980: 55-70.

Dauer, R.M. 1983. 'Stress-timing and syllable-timing reanalyzed'. *Journal of Phonetics*, 11, 1983: 51-62.

Dickson, C. 1985. *User's manual for Micro Speechlab.* Victoria, B.C.: Software Research Corporation.

Dickson, C. 1987. *User's manual for the fundamental frequency analysis program: MSLPITCH.* Victoria,B.C.: CSTR Society, University of Victoria.

Esling, J.H. and Wong, R.F. 1991 'Voice quality settings and the teaching of pronunciation.' In Brown (ed.), 1991: 288-295.

Fay, W.H. (1966). *Temporal sequence in the perception of speech*, The Hague.

Faure, G., Hirst, D.J., and Chafcouloff, M. 1980. 'Rhythm in English: isochronism, pitch, and perceived stress'. In Waugh and van Schooneveld (eds.), 1980: 71-79.

Fintoft, K. 1970. *Acoustical analysis and perception of tonemes in some Norwegian dialects.* Oslo: Universitetsforlaget.

Fry, D.B. 1958. 'Experiments in the perception of stress'. *Language and Speech*, vol. 1: 126-152.

Fry, D.B. 1968. 'Prosodic phenomena'. In Malmberg (ed.), 1968: 365-410.

Gibbon, D. and Richter, H. (eds.), 1984. *Intonation, accent and rhythm.* Berlin: Walte de Gruyter.

Gibbon, D. 1976. *Perspectives of intonation analysis.* Frankfurt/M. und München: Herbert Lang Bern, Peter Lang.

Gimson, A.C. 1980. *An introduction to the pronunciation of English* (3rd edn), London: Edward Arnold.

Gustafson, K. 1987. *Elementær akustikk for lingvister.* Bergen: UiB, Institutt for fonetikk og lingvistikk.

Gårding, E. 1981. 'Contrastive Prosody: a model and its application'. In AILA 81, Proceedings II, *Studia Lingvistica*, vol. 35, 1-2. Lund: Gleerup.

Hadding-Koch, K. and Studdert-Kennedy, M. 1964. 'An experimental study of some intonation contours'. *Phonetica*, 11: 175-185.

Hall, Jr. R.A. 1953. 'Elgar and the intonation of British English'. In Bolinger (ed.), 1972: 282-285.

Hammerly, H. 1982. 'Contrastive phonology and error analysis'. *IRAL*, vol. XX/1, 1982: 17-32.

Haugen, E. and Joos, M. 1952. 'Tone and intonation in East Norwegian'. In Bolinger (ed.), 1972:414-436.

Haugen, E. 1955. 'Tonelagsanalyse'. *Maal & Minne*, 70-80.

Heffner, R-M.S. 1950. *General phonetics*. Madison: University of Wisconsin Press.

Jakobson, R., Fant, C.G.M., & Halle, M. 1963. *Preliminaries to speech analysis*. Cambridge, Mass.: M.I.T. Press.

Jenner, B.R.A. 1987. 'Articulation and phonation in non-native English: the example of Dutch-English'. *Journal of the International Phonetic Association* 17:2, 125-138.

Ladefoged, P. 1962. *Elements of acoustic phonetics*. Chicago & London: The University of Chicago Press.

Ladefoged, P. 1982. *A course in phonetics* (2nd edn). New York: Harcourt Brace Jovanowich.

Ladd, D.R. and K.E.A. Silverman. 1984. 'Vowel intrinsic pitch in connected speech'. *Phonetica* 41: 31-40.

Lado, R. 1957. *Linguistics across cultures*. University of Michigan Press.

Lass, N.J. (ed.), 1976. *Contemporary issues in experimental phonetics*. New York: Academic Press.

Laver, J. 1970. 'The production of speech'. In Lyons (ed.), 1970.

Laver, J. 1980. *The phonetic description of voice quality*. Cambridge University Press.

Laver, J. and Jack, M.A. (eds.), 1987. *Speech technology. Procedings*. European Conference, Edinburgh 1987. Edinburgh: CEP Consultants.

Lehiste, I. 1970. *Suprasegmentals*. Cambridge: MIT Press.

Lehiste, I. 1977. 'Isochrony reconsidered'. *Journal of Phonetics*, vol. 5: 253-263.

Lehiste, I. 1979. 'Temporal relations within speech units'. *Proceedings*, vol. 3. 9th International Congress of Phonetic Sciences. *Copenhagen*.

Lieberman, P. and Michaels, S.B. 1962. 'Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech'. In Bolinger (ed.), 1972: 235-249, Harmondsworth: Penguin Books.

Lieberman, P. 1967. *Intonation, perception, and language*. Massachusetts: M.I.T. Press.

Lindblom, B. 1978. 'Final lengthening in speech and music'. Paper at the 1st Symposium, The Prosody of Nordic languages, Lund, 1978.

Lyons, J. (ed.), 1970. *New horizons in linguistics*. Harmondsworth: Penguin Books.

Lysne, A. 1985. 'Prosody. The bugbear of foreign language teaching and learning, with reference to English'. *Språk og språkundervisning*, 1985, nr.1: 34-39.

Lysne, A. 1989. 'Analyse av intonasjon ved hjelp av PC-programmet MICRO SPEECH LAB'. In *Humanistiske data*, 1/2 1989: 185-192.

Monsen, R.B. and Engebretson, A.M. 1977. 'Study of variations in the male and female glottal wave'. *Journal of the Acoustical Society of America*, 62: 981-93.

Moray, N. 1969. *Listening and attention*. Middlesex: Penguin Books Ltd, Harmondsworth.

Peters, R. 1975. 'The measurement of temporal factors in auditory perception'. In Singh (ed.), 1975: 157-175.

Pierrehumbert, J. 1979. 'The perception of fundamental frequency declination'. *Journal of the Acoustic Society of America*, 62: 981-93.

de Pijper, J. R. 1983. *Modelling British English intonation*. Dordrecht: Foris Publications.

Pike, K.L. 1972. 'General characteristics of intonation'. In Bolinger (ed.), 1972:53-82.

Reinholt Petersen, N. 1986. 'Perceptual compensation for segmentally conditioned fundamental frequency perturbation'. *Phonetica* 43: 31-42.

Roach, P.J. 1982. 'On the distinction between "stress-timed" and "syllable-timed" languages'. In Crystal (ed.), 1982: 73-79.

Scheffers, M.T.M. 1988. 'Automatic stylization of Fo contours'. In Ainsworth, W.A. and Holmes, J.N. (eds.), 1988:981-987.

Shoup, J.E. and Pfeifer, L.L. 'Acoustic characteristics of speech sounds'. In Lass (ed.), 1976: 171-224.

Singh, S. (ed). 1975. *Measurement procedures in speech, hearing, and language.* Baltimore: University Park Press.

Sondhi, M.M. 1975. 'Measurement of the glottal waveform'. *Journal of the Acoustical Society of America*, 57: 228-232.

Sørensen, J. and Cooper, W. 1980. 'Syntactic coding of fundamental frequency in speech production'. In Cole (ed.), 1980: 399-440.

Strandskogen, Å.B. 1979. *Norsk fonetikk for utlendinger.* Oslo: Gyldendal Norsk Forlag.

Takefuta, Y. 1975. 'Method of acoustic analysis of intonation'. In Singh (ed.), 1975: 363-378.

Terken, J.M.B. and Collier, R. 1989, 'Automatic synthesis of natural-sounding intonation for text-to-speech conversion in Dutch'. In Tubach and Mariani (eds.), 1989. Vol. 2: 357-359.

T'Hart, J. Collier, R. and Cohen, A. 1990. *A perceptual study of intonation.* Cambridge: Cambridge University Press.

Trim, J.L.M. 1959. 'Major and minor tone groups in English'. *Le Maître Phonétique*, vol. 112: 26-29.

Tubach, J.P. and Mariani, J.J. (eds.), 1989. *Eurospeech 89.* Procedings, European Conference on Speech Communication and Technology. Paris, 1989.

Vanvik, A. 1975. *English phonetics for norwegian students.* Oslo: Universitetsforlaget.

Waugh, L.R., and van Schooneveld, C.H. (eds.), 1980. *The melody of language.* Baltimore: University Park Press.

Willems, N. 1982. *English intonation from a Dutch point of view.* Dordrecht: Foris Publications.

Anna K. Lysne
Engelsk institutt
HF-bygget
Sydnesplass 9
5007 Bergen

Determinisme og syntaktisk flertydighet

Torbjørn Nordgård
Universitetet i Bergen

1. Innledning

Man forstår gjerne en *deterministisk* innretning som et redskap som finner raskeste vei til en, og bare en, løsning av en oppgave, eksempelvis analysen av en setning. *Syntaktisk flertydighet*, på den annen side, behandles oftest ved at grammatikken prediserer et *sett* av syntaktiske representasjoner til en ordsekvens. I det som følger skal jeg vise hvordan determinismebegrepet kan modifiseres på en måte som bevarer determinismens fordelaktige sider, samtidig som syntaktiske flertydigheter kan analyseres korrekt.

2. Syntaktisk flertydighet

Den lingvistiske litteraturen inneholder utallige eksempler på syntaktiske flertydigheter. Typisk er analysen av den syntaktiske rekkevidden til preposisjonsfraser, som i (1), men også varianter som i (2) og (3):

(1) Jens vil se mannen med kikkerten (2 lesninger)

a. Jens vil [VP se [NP mannen [PP med kikkerten]]]
"Jens vil se mannen, og mannen har kikkerten"

b. Jens vil [VP se [NP mannen] [PP med kikkerten]]
"Jens vil se mannen, og Jens vil bruke kikkerten"

(2) Hvem elsker Marit? (2 lesninger)

a. Hvem [VP elsker [NP Marit]]
"For hvilke individer x forholder det seg slik at x er elsker av Marit"

b. [CP Hvem $_i$ [C' elsker $_j$ [S Marit [VP ej e_i]]]]^1
"For hvilke individer x forholder det seg slik at Marit er elsker av x"

(3) Jens så mannen med kikkerten (4 lesninger)

a. [S Jens [VP så [NP mannen] [PP med kikkerten]]]
"Mannen ble sett av Jens, og Jens brukte kikkerten"

b. [S Jens [VP så [NP mannen [PP med kikkerten]]]]
"Mannen som hadde kikkerten ble sett av Jens"

c. [CP Jens$_i$ [C'så$_j$ [S [NP mannen [PP med kikkerten]] [VP ej e_i]]]]
"Jens ble sett av mannen som hadde kikkerten"

d. [CP Jens$_i$ [C'såj [S [NP mannen] [VP e$_j$ e$_i$ [PP med kikkerten]]]]]
"Jens ble sett av mannen, og mannen brukte kikkerten"

3. Syntaktisk flertydighet og automatisk setningsanalyse

En maskin som foretar setningsanalyse, dvs. en *parser*, bør finne frem til alle syntaktiske lesningene av et gitt input. Merk den noe beskjedne modaliteten; den kommer av at det kan gis argumenter som motiverer parsere som bare finner *en* syntaksrepresentasjon for hvert input. Jeg tenker her på parsere for formelle språk, f.eks. programmeringsspråk, der man aldri tillater noen form for flertydigheter (en datamaskin må ha entydige instruksjoner). Det kan også argumenteres for at parsere for naturlige språk ikke skal returnere mer enn en strukturell analyse. Dette gjelder parsere som bare skal finne den mest plausible lesningen, gitt at den også har tilgang til intonasjonsmessig relevant informasjon, diskursinformasjon, osv. De sistnevnte forutsetningene er problematiske når man stiller seg oppgaven å implementere en parser idag. Intonasjonsinformasjon forutsetter enten at talegjenkjenningsproblemet i vid forstand er løst, hvilket det ikke er, eller at intonasjonsinformasjon på en eller annen måte kodes sammen med parserens input. Dertil trenger vi en troverdig og generell teori om diskursrepresentasjon, og til tross for lovende ansatser finnes ikke en slik teori idag. Et siste argument for parsere som bare kommer ut med ett resultat, kan være antagelser om at der finnes et rent syntaktisk definert hierarki mellom syntaktisk flertydige lesninger som gir korrekte preferanser. Abney (1987) er et relevant arbeid i denne sammenheng, men jeg må innrømme at jeg synes det er vanskelig å se at man kan skille mellom lesningene av setning (2) på et rent syntaktisk grunnlag.

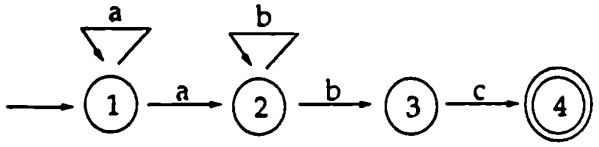
For å summere opp; dersom man ønsker å lage en parser for naturlige språk kommer man ikke utenom at den av og til må kunne returnere mer enn ett resultat. Dermed distanserer vi oss fra en "naiv" deterministisk parser som bare finner en analyse av et input.

4. Determinisme og setningsgjenkjenning

La oss først klargjøre hva man normalt forstår med termen "determinisme". Vi går da til automatteori og skillet mellom deterministiske og ikke-deterministiske automater (en automat er en slags "abstrakt" maskin som kan utføre nærmere definerte oppgaver). La oss for enkelhets skyld se nærmere på en deterministisk og en ikke-deterministisk automat som gjenkjenner språket $a^n b^m c$ for n,m > 0; altså et (kunstig) språk der setningene alltid består av et vilkårlig antall *a*'er etterfulgt av et vilkårlig antall *b*'er samt en *c* til slutt. Automat A er ikke-deterministisk og automat B er deterministisk:

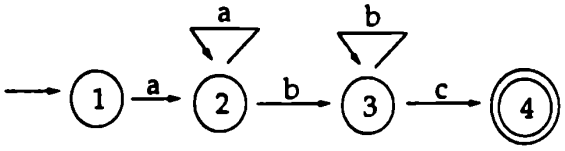
(4) **Automat A:**

Initiale tilstander:	{1}
Finale tilstander:	{4}
Instruksjoner:	Fra 1 til 1 ved symbol a
	Fra 1 til 2 ved symbol a
	Fra 2 til 2 ved symbol b
	Fra 2 til 3 ved symbol b
	Fra 3 til 4 ved symbol c



Automaten fremstilt som transisjonsdiagram:

Automat B:	Initiale tilstander:	{1}
	Finale tilstander:	{4}
	Instruksjoner:	Fra 1 til 2 ved symbol **a**
		Fra 2 til 2 ved symbol **a**
		Fra 2 til 3 ved symbol **b**
		Fra 3 til 3 ved symbol **b**
		Fra 3 til 4 ved symbol **c**



Automaten fremstilt som diagram:

Den deterministiske automaten trenger en enkel algoritme til å fortolke og benytte den i en analyse:

1. Finn en instruksjon der første symbol i inputstrengen er lik det symbol som er nevnt i instruksjonen, og der instruksjonens "fra"-tilstand er en initial tilstand (en tilstand som er element i mengden av initiale tilstander).
2. Når en slik instruksjon *i* er funnet, les inputstrengens neste symbol. Let etter en instruksjon *ii* som er slik at dette symbolet er lik symbolet som er nevnt i instruksjon *ii*, og *is* "til"-tilstand er den samme som *iis* "fra"-tilstand. Gjenta denne prosessen inntil hele inputstrengen er lest eller det ikke finnes matchende instruksjoner.
3. Når step 1 og 2 er ferdig, avgjør om følgende kriterier er sanne: (i) instruksjon *iis* "til"-tilstanden er element i mengden av finale tilstander, og (ii) alle symbolene i inputstrengen er lest. Hvis begge kriteriene er oppfylte, er setningen akseptert. Hvis ikke er den ikke en setning i språket slik det er definert av automaten.

5. Ikke-determinisme og setningsgjenkjenning

Den ikke-deterministiske automaten (Automat A), derimot, krever en noe mer sofistikert fortolkningsalgoritme fordi den må utstyres med en *hukommelse* som gjør det mulig å gå tilbake og prøve alternative muligheter fra en gitt analysesituasjon. Det finnes flere velkjente teknikker for å oppnå dette; den kanskje mest kjente og konseptuelt enkleste å forstå er *backtracking*. Den deterministiske algoritmen modifiseres på følgende måte når den gjøres til en (noe forenklet) backtracker:

1. Finn alle instruksjoner der første symbol i inputstrengen er lik det symbol som er nevnt i instruksjonen, og der instruksjonens "fra"-tilstand er en initial tilstand (en tilstand som er element i mengden av initiale tilstander).
2. Når de er funnet, **velg** en av dem. Kall den *i*. De andre tas vare på sammen med informasjon om hvilket symbol som var i ferd med å bli lest. Les inputstrengens neste symbol. Finn alle instruksjoner *ii* som er slik at dette symbolet er lik symbolet som er nevnt i instruksjon *ii*, og *is* "til"-tilstand er den samme som *iis* "fra"-tilstand. Gjenta denne prosessen inntil inputstrengen er tom eller det ikke finnes matchende instruksjoner.
3. a. Når step 1 og 2 er ferdig, avgjør om følgende kriterier er sanne: (i) det finnes minst en instruksjon *ii* hvis "til"-tilstand er element i mengden av finale tilstander, og (ii) alle symbolene i inputstrengen er lest. Hvis (i) og (ii) er oppfylte, er setningen akseptert.
   b. Undersøk om det finnes noen alternativer som ikke er forsøkt ennå. Hvis alternativer finnes, velg det som sist ble tatt vare på, og fortsett analysen fra punkt 2.

c. Hvis kriteriene (i) og (ii) under 3.a. aldri er blitt tilfredsstilt under analysen, samt at det ikke finnes noen uprøvede alternativer, er strengen ikke en setning i språket slik det er definert av automaten.

Merk at denne algoritmen finner frem til *alle* mulige veier frem til korrekte analyser av en streng. Det er presis en slik egenskap vi er ute etter for å kunne gripe syntaktisk flertydighet i *parsing* av setninger i naturlige språk. Forskjellen mellom en gjenkjenner og en parser er at gjenkjenneren kun svarer benektende eller bekreftende på om en foreslått streng er grammatisk eller ikke, mens parseren produserer strukturelle representasjoner av velformede setninger. Dersom mengden av representasjoner etter parsingen er tom, er setningen ugrammatisk. Dersom den har ett element, er setningen entydig, to elementer viser tvetydighet, osv.

Den "deterministiske" algoritmen bør ikke appliseres på ikke-deterministiske automater fordi den ikke er istand til å ta vare på alternative instruksjoner i en gitt situasjon. Derfor klarer den ikke å fortolke Automat A på en måte som gjør at språket $a^n b^m c$ gjenkjennes. Den fortolker Automat B som vi ønsker nettopp fordi automaten er deterministisk og det aldri er mer enn en mulighet videre fra et punkt i analysen.

Backtrackeren oppviser altså den egenskap å kunne håndtere ikke-deterministiske automater. Den kan også benyttes i fortolkningen av mer velkjente PS-grammatikker for naturlige språk. Alternativer til backtrackere finnes; i datalingvistikken er særlig datastrukturer som representerer et *chart* etterhvert blitt populære, særlig fordi chartet organiserer informasjon på en mer effektiv måte enn en standard backtracker gjør, samt at det gir bedre muligheter til å studere selve analyseprosessen.

Likevel, ikke-deterministiske parsere har enkelte negative egenskaper. En har å gjøre med effektivitet; når mengden av PS-regler blir stor og setningene som skal analyseres er lange, kan slike parsere bli trege, selv om eksponensialitet unngås, jf. det kompleksitetsteoretiske uttrykket $|G|^2 \cdot n^3$ der $|G|$ er grammatikkstørrelsen (mengden av symboler i grammatikken) og n er setningslengden. Det bør innskytes at å *verifisere* om en foreslått analyse er korrekt oftest er fort gjort, mens det å *finne* den/de korrekte analysene er et adskillig hardere problem. Dersom man går utover det rene PS-regelformatet og introduserer metaregler, ID-regler, LP-restriksjoner, trekkperkolering, unifikasjon, Kasusmarkering osv., kan man få parsere med ukjente kompleksitetsegenskaper. De kan bli noe raskere, men de kan også bli markant tregere. En slik kompleksitetsdiskusjon er ikke emne for denne artikkelen og jeg lar den derfor ligge, men den grunnleggende *årsaken* til kompleksiteten må påpekes, og den har å gjøre med *mengden av hypoteser som må konsulteres, eller rettere, kombineres, under analysen*; derfor den treffende termen "blind combinatorial search". Dersom en formalisme eller et analyseredskap (en parser) ikke setter noen grense for mengden av slike hypoteser, er den *potensielt* ineffektiv, selv om den ikke trenger å være det i praktiske applikasjoner.

6. Determinisme-egenskaper og deterministiske parsere

Med dette utgangspunktet synes det rimelig å hevde at dersom muligheten til kombinatorisk søking begrenses, vil årsakene til komputasjonell ineffektivitet langt på vei elimineres, kanskje elimineres helt. Determinisme er en nærliggende mulighet, men som vi husker av diskusjonen ovenfor, har determinismen den uheldige egenskap at den gjør det umulig å frembringe mer enn en analyse av et input. Man kan da spørre om det er mulig å operere med en alternativ definisjon av determinisme som både ivaretar ønsket lingvistisk dekningsgrad og som unngår potensielle kompleksitetsproblemer. La oss forsøke å komme frem til et slikt determinismebegrep.

En deterministisk maskin har følgende essensielle egenskaper:

A: Den kan ikke "glemme" eller "ødelegge" hypoteser på veien mot en løsning av en oppgave eller et problem.

B: Bare én regel (eller instruksjon) kan slå til når parseren (eller automaten) er i en gitt tilstand.

Alle valg den foretar i løpet av analysen er ugjenkallelige. I automatteori er denne egenskapen ivaretatt ved transisjonsfunksjonen, dvs. beskrivelsen av hvordan instruksjonene ser ut. For en finitt automat (fa) går den fra mengden av tilstander og inputalfabetet til mengden av tilstander (fra S x A til S der *S* er en finitt mengde tilstander og *A* er alfabetet). Siden man her snakker om en *funksjon* vil det for ethvert par av tilstander og symbol fra alfabetet finnes en unik tilstand. Enhver deterministisk fa må tilfredsstille dette kravet. Determinismekravet til en pushdown-automat (pda) må også ta pushdown-stacken i betraktning. En pda er deterministisk dersom det for enhver kombinasjon av tilstand, inputsymbol og stackelement kun finnes ett par av tilstander og stack-symboler, dvs. en mapping fra S x A x Γ til S x Γ der *S* er en finitt mengde tilstander, *A* et finitt sett symboler som utgjør alfabetet og Γ et finitt sett av stack-symboler. (For enkelhets skyld ser vi bort fra ε-transisjoner).

De automatene vi kjenner fra automatteori har i tillegg et par andre underforståtte og viktige egenskaper som skiller dem fra deterministiske Turingmaskiner (dvs. vanlige "serielle" datamaskiner). For det første kan de ikke benytte en vilkårlig mengde "kladdetaper" til å beregne hjelpehypoteser underveis. Bare pda'er, lineært avgrensede automater og stack-automater tillates å benytte slike ekstrataper, men de bruker kun én. Videre kan automatene ikke skrive nye symboler på eller vilkårlig fjerne symboler fra inputtapen. For det tredje kan de ikke bevege skrivehodet vilkårlig frem og tilbake på tapen; de kan bare se det symbol som til enhver tid står lengst til venstre på tapen. Et fjerde poeng er at de ikke tillates å scanne gjennom hele strengen først, ta vare på det de ser, og deretter starte analysen fra begynnelsen.

Dersom vi beveger oss over til parsingslitteraturen finner vi at de fleste av disse egenskapene etterleves, men ofte med visse modifikasjoner. Såkalte LR(k) parsere kan "se" *k* symboler videre inn i inputstrengen. En LR(k,t) parser kan også se *k* symboler inn i inputstrengen, men den kan i tillegg utsette tilordningen av *t* ferdig analyserte delkonstituenter (f.eks. nominalfraser og adjektivfraser). Parseren til Marcus (1980) er påstått å være LR(2,2)[2]. LR(k) og LR(k,t) parsere gjenkjenner deterministisk kontekstfrie språk. I tillegg kommer LR* parserne som kan se ubegrenset langt fremover i inputstrengen.

Alle disse parserne er deterministiske, men ingen av dem er istand til å håndtere syntaktisk flertydighet av den art som ble beskrevet innledningsvis.

7. Determinisme og syntaktiske flertydigheter

7.1. To typer determinisme

La oss anta to typer deterministiske innretninger; "løse" ("sloppy") og "strenge" deterministiske maskiner. En streng deterministisk maskin forstår vi heretter som en "klassisk" deterministisk maskin som aldri returnerer mer enn ett analyseresultat. En løs deterministisk maskin kan derimot returnere mer enn et resultat, men den er underlagt restriksjoner som vi skal presisere i det som følger.

Vi trenger et term for å omtale mengden av korrekte analyser for et gitt input. Derfor skal vi snakke om mengden av gyldige analyser for streng s i henhold til grammatikk *G*, forkortet som MGA(s,G). Et eksempel; MGA(s,G) for s lik Jens så mannen med kikkerten og *G* lik grammatikk (5) skal være mengden av frasemarkører i (6):

(5) S → NP VP
NP → N | N PP
VP → V | V NP | V NP PP
PP → P NP

(6) {[S [NP [N Jens] [VP [V så [NP [N mannen]] [PP [P [NP [N kikkerten]]]]]]]
[S [NP [N Jens] [VP [V så [NP [N mannen] [PP [P [NP [N kikkerten]]]]]]]]}

La oss sette termen MGA i relasjon til standard automatteori før den appliseres på parsing. Vi er interessert i en beskrivelse av hvordan automaten *fortolkes* slik at den kommer frem til MGA. For enkelhets skyld holder vi oss i første omgang til finitte tilstandsautomater. Vi beskriver egenskapene til en backtracker-fortolker av en fa: Før analysen starter, skrives et sett av sett av ordnede par av initial tilstand og ukonsumert input, f.eks. slik:

{{(1, aaabbc)}}

Kall dette settet for *mengden av analysestier*, forkortet til Σ. Analysestiene består av en mengde ordnede par av tilstander og ukonsumert input. Denne mengden er i sin tur ordnet etter lengden på ukonsumert input.

Backtrackeren prøver stiene i tur og orden. Etter hver utført instruksjon legger den til et nytt ordnet par bestående av ny tilstand og ukonsumert input minus symbolet lengst til venstre i inputstrengen. Dersom instruksjonen gir opphav til mer enn en ny tilstand, lages det kopier av stien frem til "nåværende situasjon" slik at to eller flere distinkte stier oppstår. Et eksempel:

{{(1, aaabbc) (1, aabbc)}
{{1, aaabbc) (2, aabbc)}}

I dette tilfellet er det understrekede paret i den siste stien en kopi fra den første stien.

Etterhvert som Σ øker, konsulteres og analyseres hver sti på samme måte. Enhver sti har en endestasjon, dvs. et par som kan eller ikke kan ekspanderes videre. Dersom endestasjonen i en sti er et par bestående av en final tilstand og Ø, kaller vi den en terminert endestasjon. Enhver sti med en terminert endestasjon er et element i MGA, og MGA inneholder kun slike terminerte endestasjoner. Dersom Σ ikke inneholder stier med terminerte endestasjoner, er ikke inputstrengen velformet. Inneholder den mer enn èn sti med terminert endestasjon, er den flertydig.

Vi har nå beskrevet egenskapene til en backtracker, og vi illustrerer med følgende eksempel (vellykkede stier er markert med fete typer):

La Automat 4.B gjenkjenne strengen aaabbc. Automaten er deterministisk og produserer bare en sti, altså kun ett medlem av Σ:

Σ= { {(1,aaabbc)(2,aabbc)(2,abbc)(2,bbc)(3,bc)(3,c)(4,Ø)} }

Denne stien er også element i MGA(aaabbc, Automat B). La Automat 4.A gjenkjenne samme streng. Resultatet blir

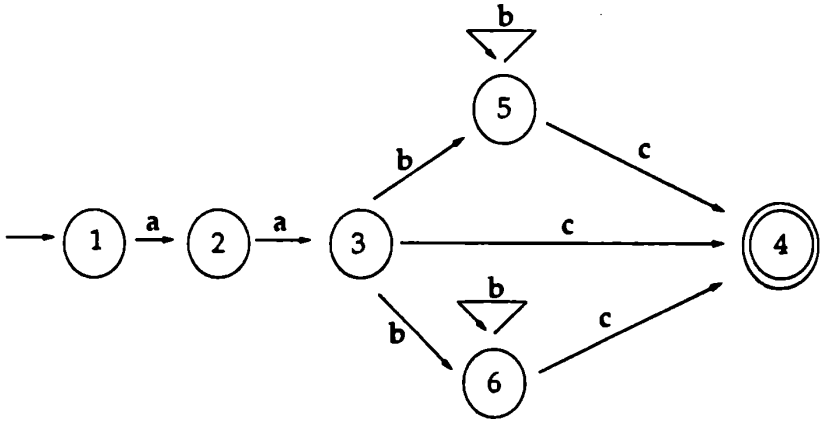
Σ= {	(1,aaabbc) (1,aabbc) (1,abbc) (1, bbc)}	Feil
	{(1,aaabbc) (1,aabbc) (1,abbc) (2,bbc) (2,bc) (2,c)}	Feil
	{(1,aaabbc) (1,aabbc) (1,abbc) (2,bbc) (2,bc) (3,c) (4,Ø)}	**Suksess**
	{(1,aaabbc) (1,aabbc) (1,abbc) (2,bbc) (3,bc)}	Feil
	{(1,aaabbc) (1,aabbc) (2,abbc)}	Feil
	{(1,aaabbc) (2,aabbc)}	Feil
	}	

I dette tilfellet er bare stien {(1,aaabbc) (1,aabbc) (1,abbc) (2,bbc) (2,bc) (3,c) (4,Ø)} element i MGA(aaabbc, Automat A). Alle de øvrige stiene er blindveier. Dette resultatet er typisk for ikke-deterministiske automater (og parsere), dvs. at MGA bare er et lite subsett av Σ.

Så langt har vi ikke sagt noe nytt, men la oss nå se på en løs deterministisk automat, som vi kaller Automat C:

(7) Automat C: Initiale tilstander: {1}
Finale tilstander: {4}
Instruksjoner:

1. Fra 1 til 2 ved a.
2. Fra 2 til 3 ved a.
3. Fra 3 til 5 ved b.
4. Fra 3 til 4 ved c.
5. Fra 3 til 6 ved b.
6. Fra 5 til 5 ved b.
7. Fra 5 til 4 ved c.
8. Fra 6 til 6 ved b.
9. Fra 6 til 4 ved c.



Automaten fremstilt som transisjonsdiagram:

Denne automaten gjenkjenner språket aab^nc for $b \geq 0$, og den gir to analyser av enhver streng i aab^nc for $b > 0$, som er et subsett av språket aab^nc for $b \geq 0$. Strengen aac, derimot, får kun en analyse. Vi lar automaten analysere *aabbc*. Resultatet blir

Σ= { {(1,aabbc)(2,abbc)(3,bbc)(5,bc)(5,c)(4,Ø)}
{(1,aabbc)(2,abbc)(3,bbc)(6,bc)(6,c)(4,Ø)} }

I dette tilfellet inneholder Σ to stier, men merk at *begge* er elementer i MGA(aaabbc, Automat C). Altså, Σ = MGA(aaabbc, Automat C).

Vi er nå istand til å definere et determinisme-hierarki: For en vanlig deterministisk automat gjelder at Σ aldri har kardinalitet > 1, mens en løs deterministisk automat tillater at Σ har kardinalitet > 1. For begge determinismetyper gjelder at automaten aksepterer input bare når hvert element av Σ også er element i MGA. For vanlige ikke-deterministiske automater er disse restriksjonene irrelevante.

Dette betyr at en løs deterministisk maskin alltid må være sikker på at enhver instruksjon som anvendes produserer et tilstand/input par som er medlem av en en sti som er element i MGA. Det er presis denne intuisjonen som er uttrykt i Nordgård (1991) der det kreves at parseren må være sikker på at inputstrengen er strukturelt flertydig før tilstandskopiering finner sted.

Et par kommentarer er på sin plass. Illustrasjonene ovenfor viser at løs determinisme ikke er det samme som blind kombinatorisk søking siden alle elementer i Σ er korrekte analyser. Blind kombinatorisk søking vil typisk føre til haugevis av mislykkede stier i Σ, og desto flere mislykkede stier, desto tregere analyse. Et annet poeng er at løs determinisme slik den er fremstilt her benytter samme prosesseringsmaskineri som blind kombinatorisk søking, selv om maskineriets muligheter bare utnyttes i begrenset omfang. Ideelt sett ønsker vi et maskineri som håndterer løs determinisme, men som ikke kan brukes til blind kombinatorisk søking.

7.2. Parseren beskrevet i Nordgård (1991)

I Nordgård (1991) diskuteres en parser som har de essensielle determinisme-egenskapene nevnt ovenfor, dvs. at den ikke kan fjerne eller glemme strukturer den har bygget underveis, og bare en regel slår til i en gitt tilstand. Men ulikt LR-parserne kan den også analysere strukturelle flertydigheter. Den har likheter med andre deterministiske analyseredskaper:

(a) Den kan ikke gå frem og tilbake i inputstrengen, flytte eller legge til symboler i inputstrengen.
(b) Den minner om en LR* parser siden den til enhver tid kan se hele inputstrengen som et regulært uttrykk, altså en større uttrykkskraft enn LR(k), LR(k,t) og Marcus-parserne.
(c) I likhet med en stack-automat kan den se hele "stack'en" den har bygd (dvs. de konstituentene i trestrukturen som dominerer den konstituenten som er i ferd med å bygges), altså mer enn en pda og Marcus-parseren.

Til forskjell fra en stack-automat kan parseren modifisere innholdet på stack'en (her forstått som node i trestrukturen som dominerer den noden som er i ferd med å bygges), f.eks. ved å legge til bindingsrelevant informasjon ("node x binder node y").

I tillegg har den et par andre egenskaper som gjør den spesiell. For det første kan den bygge et "templat" som den kan etterfylle tomme plasser i. Den kan således bevege seg inne i stack'en og modifisere dens innhold uten hele tiden å måtte forholde seg til "øverste" stackelement, jf. Marcus-parseren og pda'er. Heller ikke dette affekterer determinismen, men det kan angå generativ kraft (dvs. hvilke formelle språk den er istand til å gjenkjenne), noe som kan være verdt å diskutere ytterligere. Den andre særegenheten er at den kan søke oppover og nedover i treet vha. deterministiske fa-teknikker. Heller ikke dette angår determinismen, men det fortjener å settes i relasjon til effektivitet. Ingen av disse egenskapene rører ved den essensielle determinisme-egenskapen, men de er relevante for parserens eventuelle psykologisk plausibilitet og effektivitet.[3]

7.3 En løs deterministisk analyse av syntaktiske flertydigheter

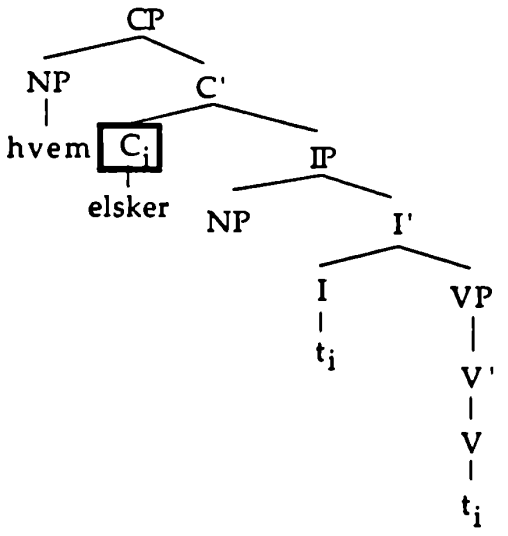
La se hvordan parseren beskrevet i Nordgård (1991) håndterer syntaktiske flertydigheter vha. løs determinisme.

Vi ser nærmere på setning (2) igjen:

(2) Hvem elsker Marit? (2 lesninger)

Parseren vil, når den er i følgende tilstand, konsultere en heuristisk regel, kall den regel #1:[4]

(8)



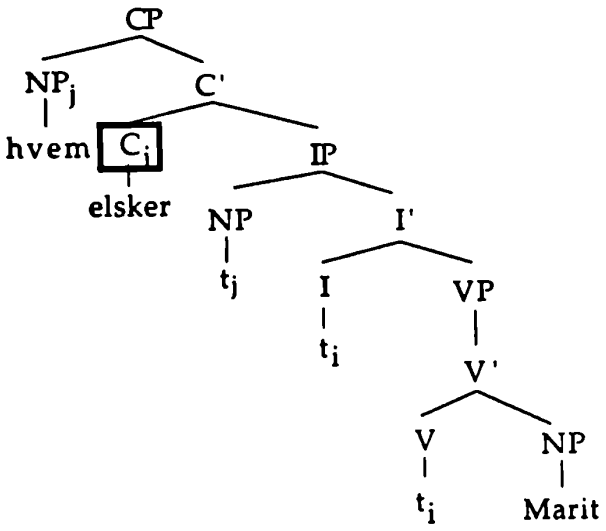
Input: **# Marit #**

(9) Heuristisk regel #1:

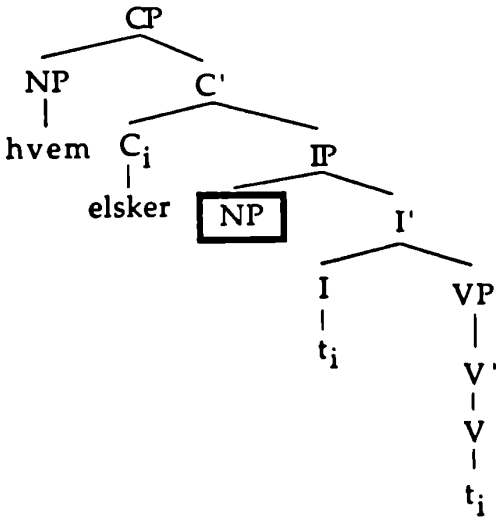
Tree Condition:	↑ : [Head CP], (Transitive) = yes. ∧ [Spec (category C)], (category) = N, (barlevel) = 2, (binder-of) = { }.
String Condition:	From !1 to !1 by Adv From !1 to 0! by N From !1 to 0! by Det From !1 to 0! by A
Action:	CopyState (PopQWP) EC (∧ [Spec (category C)] (∨ [Spec, (category I)]

Regelen krever at parserens oppmerksomhet er rettet mot posisjonen [Head, CP] og at den inneholder et transitivt verb ("oppmerksomhet" er symbolisert ved "↑"). Dessuten skal [Spec,CP] inneholde en NP som ikke binder noe.[5] Videre skal inputstrengen inneholde et nomen eller en determinativ eller et adjektiv eller et adverb etterfulgt av N, Det eller A.[6] Dersom disse kravene holder, skal tilstanden kopieres sammen med en instruksjon om at oppmerksomhen skal flyttes til neste ventende posisjon ([Spec,IP]). I den originale strukturen skal det plasseres et spor i [Spec,IP] bundet fra [Spec,CP]. Resultatet av den første parsingen blir som i (10), mens kopien som starter opp fra tilstand (11) returnerer (12):

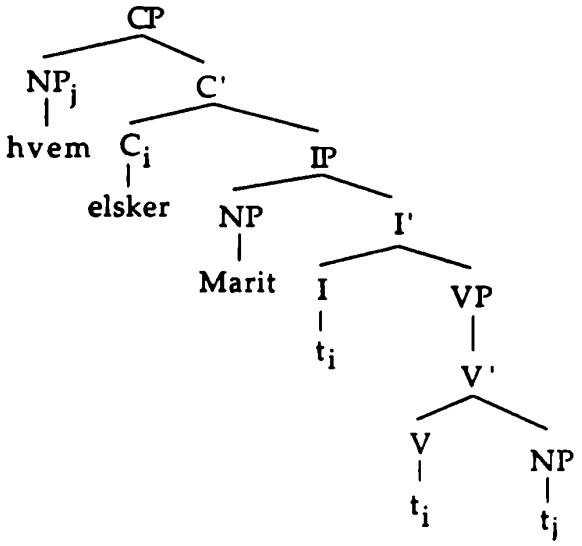
(10)


CP
NP$_j$
C'
hvem
C$_j$
IP
elsker
NP
I'
t$_j$
I
VP
t$_i$
V'
V
NP
t$_i$
Marit

(11)


CP
NP
C'
hvem
C$_i$
IP
elsker
NP
I'
I
VP
t$_i$
V'
V
t$_i$

(12)



Detaljene i parsingen går vi ikke inn på her; leseren henvises til Nordgård (1991, 1992). Det sentrale poeng er imidlertid at parseren ikke har laget mer struktur under parsingen enn nettopp de to ønskede analysene. Derfor er determinismekravet tilfredsstilt, og parseren har bygget to representasjoner vha. løs determinisme.

8. Løs determinisme og ikke-deterministiske språk

I formell språkteori relateres gjerne distinksjonen determinisme/ikke-determinisme til distinksjonen deterministiske/ikke-deterministiske språk. Et ikke-deterministisk språk synes å kreve ikke-deterministiske analysealgoritmer. Språket karakterisert som mengden av speilbildestrenger over {a, b} kan ikke gjenkjennes av en deterministisk pda, mens en ikke-deterministisk pda kan gjøre den jobben. Men det er et velkjent problem at det er vanskelig å avgjøre om et språk generert av en ikke-deterministisk pda kan genereres av en deterministisk pda. Situasjonen er den samme for deterministiske og ikke-deterministiske lineært avgrensede automater.

Et naturlig spørsmål for vår diskusjon er om alle ikke-deterministiske kontekstfrie språk kan kan genereres vha. løs determinisme. Svaret er negativt siden det er vanskelig å se hvordan speilbildespråket over {a,b}* kan gjenkjennes vha. en løs deterministisk pda.[7]

Løs determinisme synes ikke å være spesielt relevant for diskusjoner av generativ kraft. Begrepet er derimot egnet til å karakterisere et subsett av ikke-deterministiske maskiner eller grammatikker som som *ikke* utnytter ikke-determinismens muligheter til å forfølge feilaktige hypoteser. Denne klassen av grammatikker er interessante både i et komputasjonelt og psykolingvistisk perspektiv. Den komputasjonelle fordelen ligger i at det ikke blir utført mer arbeid enn nødvendig under prosesseringen. Hva psykolingvistikken angår burde man vha. løs determinisme kunne lage mer troverdige modeller der også syntaktiske flertydigheter gis adekvate analyser.

9. Oppsummering

Jeg har i det foregående diskutert ulike sider ved determinisme, og jeg har introdusert begrepet løs determinisme. Det er vist hvordan løs determinisme forstås i relasjon til finitte automater og hvordan begrepet kan benyttes i parsing. Et gjenværende spørsmål er hvorvidt potensiell ineffektivitet lurer i bakgrunnen i prosesseringen av løst deterministiske grammatikker. Svaret er både ja og nei. Negativt fordi parseren aldri kan plukke opp alle kompatible utveier fra en gitt tilstand, altså at flere regler matcher en parsingstilstand. Positivt fordi prosesseringen kan føre til mislykkede resultat, enten ved kopien eller den "originale" strukturen. Men uansett er ett viktig aspekt ved blind kombinatorisk søking eliminert. Det andre kan bare ivaretas ved at grammatikkreglene er etterlever determinismekravet, altså et spørsmål om grammatikkskriverens dyktighet.

Noter

1 Dette er en GB-analyse der hv-frasen **hvem** har flyttet fra objektsposisjonen e_i og verbet har flyttet fra hodeposisjonen e_j i verbfrasen.

2 Se Berwick (1985).

3 For en diskusjon av generativ kraft og effektivitet til parseren beskrevet i Nordgård (1991), se Nordgård (1992b).

4 Plassen tillater ikke en fremstilling av hvordan struktur (8) er produsert. Leseren henvises til Nordgård (1991) eller Nordgård (1992).

5 Operatoren "^" betyr at det skal søkes oppover i treet fra "oppmerksomhetsposisjonen". "v" fortolkes som søking nedover i treet.

6 En forklaring av notasjonskonvensjonene til strengautomaten er påkrevet: !n betyr initial tilstand og n! betyr final tilstand. !n! er både initial og final tilstand.

7 Et annet spørsmål er om speilbildespråket over {a,b} kan gjenkjennes av en løs deterministisk pda med ubegrenset lookahead. Se Nordgård (1992b) for en diskusjon av dette og enkelte andre emner relatert til løs determinisme.

Referanser

Abney, S. (1987) "Licensing and Parsing", i *Proceedings of NELS 17, Volume 1*. Dept. of Linguistics, University of Massachusetts, Amherst.

Berwick, R. (1985) *The Acquisition of Syntactic Knowledge*. MIT Press.

Nordgård, T. (1991) A GB-Related Parser for Norwegian. Upublisert doktoravhandling, Institutt for fonetikk og lingvistikk, Universitetet i Bergen.

Nordgård, T. (1992) *A GB-Related Parser for Norwegian*. Det historisk-filosofiske fakultets publikasjonsserie, Universitetet i Bergen.

Nordgård, T. (1992b) "On the efficiency of E-Parser", upublisert manuskript, Institutt for fonetikk og lingvistikk, Universitetet i Bergen.

Torbjørn Nordgård
Institutt for fonetikk og lingvistikk
Universitetet i Bergen
Sydnesplass 9
N-5007 Bergen
E-mail: nordgaard@hf.uib.no

Anaphora and Intensionality[1] *in Classical Logic*

Jørgen Villadsen
Technical University of Danmark

Abstract

The dynamic perspective on natural language suggests that texts change the value of contextual parameters in much the way computer programs change the value of program variables. One phenomenon where dynamic aspects emerge is that of anaphora (introduction of referents for later anaphoric links), but intensionality and tense are other candidate phenomena. However, most logics used in computer science for the study of the semantics of program are often rather complicated and different from classical logic. We describe a representation directly in classical logic using a categorial grammar. In the logic a theory is defined (via axioms) to describe how in a text a pronoun manages to pick up a referent that was introduced by a determiner. Besides (nominal) anaphora we discuss how to handle intensionality present in (belief and knowledge) attitude reports.

Introduction

We present a theoretical study of central semantical problems of anaphora and intensionality in natural language texts (English). A basic knowledge of anaphora, intensionality and logic is assumed.

As texts we take sequences of declarative and disambiguous sentences; hence each text has a unique meaning and the objective is to define a semantical theory specifying, in some way and to some extent, the meaning of each text.

We think that in order to use the texts to communicate our information about facts (or assumptions) concerning "the universe around us" the theory must handle anaphora and intensionality as amply present in natural language. For instance, consider the following *Love Story* text:

A man kisses a woman.

She believes that he loves her.

The pronouns (he, she,...) imply anaphoric links across sentences and attitude reports (believe that,...) imply intensional properties for embedded sentences.[2]

We require that the theory must be a so-called correspondence theory of meaning – that a relation between the language and certain "things" independent of the language must be specified. We also

require that the theory is computational – that a "calculus" for manipulations of meanings of texts must be specified.[3] Both requirements can be directly satisfied using the model theory and the proof theory, respectively, of a formal logical system. For instance, returning to the first sentence of the *Love Story* we might tentatively represent it as the following formula in classical predicate logic:

$$\exists x\,(man\,x \wedge \exists\, y\,(woman\,y \wedge kiss\,x\,y))$$

The model theory says that man and woman denote sets of objects, that kiss is a relation between two objects and that the variables x and y can be instantiated to certain objects satisfying the given constraints. The proof theory, on the other hand, lists axioms and inference rules for deriving or rewriting the formula as such.

The representation of the second sentence is less straightforward. Even ignoring the proper representation of the embedded sentence construction (symbolised by $\boxed{\ldots}$) it seems necessary to redo the already completed representation of the first sentence to accomodate the second sentence:

$$\exists x\,(man\,x \wedge \exists y\,(woman\,y \wedge kiss\,x\,y \wedge belief\,y\ [\,love\,x\,y]\,))$$

This seems to render impossible a systematical and compositional analysis where the meaning of an expression is systematically composed from the meanings of its parts.

Also, a seemingly innocent change (conjoining the two previous sentences to a single conditional sentence) implies a drastic change for the representation (where the previous representations cannot be reused in any straight-forward manner):

If a man kisses a woman then she believes that he loves her.

$$\forall x\,(man\,x \rightarrow \forall y\,(woman\,y \rightarrow \ldots))$$

Of course, the representation in the logical system above is rather ad hoc. Discourse representation theory (see [2,3,4] for references) deals with translations of texts in a systematical way, but, as the name suggests, uses a separate representation level and does not yield a direct compositional analysis in a logical system.

Our claim is that a compositional analysis in a logical system is possible, but that texts are complicated and are best understood if one moves from static semantics (as above) to dynamic semantics (to be fully explained later).

It must be emphasized that we can do dynamic semantics for natural language in a logical system where the logical language itself has a static semantics! This can be done by choosing suitable representations and axioms in the logical system. We briefly illustrate this point by considering a recent non-classical logical system, namely dynamic predicate logic [3], which is based on the same language as classical predicate logic, but where the following entailment holds (ϕ and ψ any formulas):

$$(\exists x\phi[x]) \rightarrow \psi\,[x] \models \forall x(\phi[x] \rightarrow \psi\,[x])$$

Notice that the quantifier binds the variable outside the scope of the quantifier. The quantifiers and operators have different behavior: existential quantification is externally dynamic, universal quantification is externally static, and implication is internally dynamic (between operands) but externally static.

The entailment above allows for a compact and elegant treatment of many anaphora problems using dynamic predicate logic, but compared to classical predicate logic it is less flexible and transparent.

For instance the following simple entailment does not hold:

$$\phi \models \phi$$

However, a restricted version holds (when no overlap between active quantifier variables and free variables in the formula). See [3] for details about dynamic predicate logic as well as references to the extended logical system called dynamic Montague grammar.

Overview

First the general idea of logical semantics is described, subsuming both static and dynamic semantics. Then static semantics is defined (based on the notion of truth values) and hereafter dynamic semantics (based on the notion of information state changes).

Information states can be structured in many ways – we propose to let an information state be a set of alternatives.[4] Even without saying more about what an alternative is we obtain interesting definitions of vacuous, absurd and definite information states. Moreover, reductions are possible if the information state changes are distributive and eliminative.

We argue that by a proper construction of the set of alternatives anaphora can be handled in an elegant way. We propose a representation in logical type theory using a categorial grammar. A tiny example is provided. Finally we discuss how to incorporate intensionality in this framework.

A few words on notation. In schemas we take $\Leftrightarrow$ to mean: if and only if (for all appropriate instantiations with expressions of the shown schema variables). We have two equality symbols: $\equiv$ means equality by definition and = means equality by calculation (hence $\equiv$ implies =).

Logical Semantics for Natural Language

The main purpose of logical semantics for natural language is, we think, to provide a precise description of the intuitive notion of a valid argument.[5]

Arguments can be built in several ways. We take an argument to consist of a sequence of texts $t_1, \ldots, t_n$ (the premises) and a single text t_* (the conclusion). We arrange the argument as in the following schema with the symbol $\therefore$ in front of the conclusion:

$$\begin{array}{l} t_1 \\ \vdots \\ t_n \\ \therefore t_* \end{array}$$

We translate each text t in natural language to a formula $t^\bullet$ in a suitable logical language – a logic.[6] By a logic we mean a formal language endowed with a notion of entailment $\models$, which we require must match the notion of a valid argument in the following sense:

$$\begin{array}{l} t_1 \\ \vdots \\ t_n \\ \therefore t_* \end{array} \quad \Longleftrightarrow \quad t_1^\bullet, \ldots, t_n^\bullet \models t_*^\bullet$$

A formal language is a language with a precise syntax and semantics.[7] The syntax defines the set of formulas (for instance via formation rules). The semantics defines for each formula its meaning (for instance a mathematical object). In general we write the meaning of the formula $t^\bullet$ as $[t]$. This hides the translation, but no confusion can arise. We shall also refer to $[t]$ as the meaning of the text t.

We require that although entailment is introduced as relation $\models$ between formulas (syntactical objects) it must correspond to a relation $\mathcal{R}$ between the meanings of the formulas (semantical objects):

$$t_1^\bullet, \ldots, t_n^\bullet \models t_*^\bullet \iff \mathcal{R}([t_1], \ldots, [t_n], [t_*])$$

Since we aim at using mathematical objects as the meanings, our goal is to find objects with enough structure to make the definition of such a relation $\mathcal{R}$ possible. In the static semantics the objects are rather coarse-grained, but in the dynamic semantics to be described later, the objects are more fine-grained.

Static Semantics

In static semantics the meaning of a text is a truth value (0/1 for false and true) with respect to a so-called model. It would actually be more appropriate to say that the meaning then is a function from models to truth values, but since it turns out that the model parameter is common to all meaning compositions (based on the formation rules) it makes sense to use the "with respect to" phrase.

We write $[t]_M$ for the meaning of the text t with respect to the model $M = \langle D, d \rangle$, where D is the domain (a set of objects) and d is the denotation (a function from constants to mathematical structures of objects). There can be many kinds of constants: individual constants (denoting objects), first order predicate constants (denoting relations overobjects), second order predicate constants (denoting relations over objects as well as over relations over objects), etc. The denotation of a predicate constant is often called its extension. The so-called order of a logic is not determined by the orders of the constants present, but by the orders of the variables of quantification.

We assume that the formula $t^\bullet$ is closed (no occurrences of free variables).[8] To account for free variables in sub-formulas the meaning must be given with respect to an assignment a (a function from variables to mathematical structures of objects) besides the model M, but this complication can be ignored here. As for constants there can be many kinds of variables. In first order logic, quantification is allowed over individual variables only, whereas in higher order logic quantification is allowed over predicate variables too.

Entailment is defined as:

$$\begin{array}{l} t_1 \\ \vdots \\ t_n \\ \therefore t_* \end{array} \iff \text{For all } M\text{: if } [t_1]_M = \ldots = [t_n]_M = 1 \text{ then also } [t_*]_M = 1$$

Dynamic Semantics

Dynamics imply the (continuous or discrete) change of something – in case of natural language semantics we suggest to consider the change from one information state to another information state of an agent (human or program) processing the texts.[9] We first present the general idea leaving the details of information states unspecified. We then consider different specific structures as information states.

We intuitively have the following distinguished information states:

Vacuous information state
The agent has no information at all.

Absurd information state
The agent has contradictory information.

Definite information state
The agent has maximal information.

Often there is only one vacuous and one absurd information state, written I_0 and I_∞, whereas there are numerous definite information states.[10] We often have a partial ordering of information states corresponding to getting "better" informed.

In dynamic semantics we let $[t]$ be a function from information states to information states. Observe that there is no need to see the meaning relative to a model as in the static semantics (although it is possible to do so). If the initial information state is I_0 we have the following information state changes:

$$I_0 \xrightarrow{[t_1]} I_1 \xrightarrow{[t_2]} I_2 \xrightarrow{[t_3]} \cdots \xrightarrow{[t_n]} I_n \xrightarrow{[t_*]} I_*$$

We define the relevant resulting information states as follows:[11]

$$I_n \equiv [t_n](\ldots[t_1](I_0)) \qquad I_* \equiv [t_*](I_n)$$

There are several different definitions of entailment possible. We here follow the slogan: entailment makes explicit implicit information. More precisely: The conclusion t_* is implicit in the premisses $t_1, \ldots, t_n$ if no (important) change in information state occurs when t_* is processed after $t_1, \ldots, t_n$ have been processed:

$$\begin{array}{l} t_1 \\ \vdots \\ t_n \\ \therefore t_* \end{array} \quad \Longleftrightarrow \quad I_n \approx I_*$$

Here $\approx$ is a suitable equivalence relation that compares information states for (important) changes; if all changes are important then we can use $=$ for $\approx$.

Information States as Sets of Alternatives

One way to specify information states is to let each information state be a set of alternatives that has to be taken into account. Growth of information then comes down to elimination of alternatives, up to the level of definite information where all but one of the alternatives have been eliminated.

Let $A \equiv \{a_1, a_2, a_3, \ldots\}$ be a set of not further specified alternatives and let an information state I be a set of alternatives ($I \subseteq A$). The definitions of the vacuous and absurd information states are:

$I_0 \equiv A \quad I_\infty \equiv \emptyset$

The definite information states are $I = \{a\}$, $a \in A$.
The information states have a partial ordering via the subset relation $\subseteq$. For example:

$$\{a_1\} \subseteq \{a_1, a_2\} \subseteq \{a_1, a_2, a_3\}$$

$$\{a_1, a_2\} \not\subseteq \{a_2, a_3\}$$

For a particular information state change $[t]$ there are some interesting properties it may have:

Distributive information state changes $[t](I) = \bigcup_{a \in I} [t](\{a\})$ for all I

Eliminative information state changes $[t](I) \subseteq I$ for all I

Distributive information state changes mean that the agent works "point-wise", i.e. based on the definite information states. This property does not hold for agents dealing with defaults etc. An illustrative example:

$t \equiv$ **John might lie.**

$$[t](I) = \begin{cases} I & \text{if there is an } a \in I \text{ such that "John lies".} \\ I_\infty & \text{otherwise.} \end{cases}$$

Note that if one has the information that "John does not lie" it is absurd to process t. It is easily seen from the above definition that the information state change is not distributive, since the test $a \subseteq I$ does not work point-wise. However, the information state change is eliminative, since $I_\infty \subseteq I$ for all I.

Eliminative information state changes means that the agent gets "better informed" towards the definite information states (with the exception that the only change from a definite information state is to the absurd information state). This property does not hold for agents dealing with revision etc.

When these two properties are "universal" – that is fulfilled for all texts t considered – optimisations are possible:

- Distributive information state changes:
  Instead of meaning [t] as a function from information states to information states, that is a function from sets of alternatives to sets of alternatives, we just specify a relation $\rightarrow t$ between alternatives:

$$a \rightarrow_t a' \iff a' \in [t](\{a\})$$

The meaning $[t]$ is then obtained as:

$$[t](I) = \bigcup_{a \in I} \{a' \mid a \rightarrow_t a'\}$$

The condition $a \rightarrow_t a'$ corresponds to the following idea in computer science:[12]

Non-deterministic execution of program t when in program state a yields program state a'.

- Distributive and eliminative information state changes:
  The relation $\rightarrow_t$ now only holds between identical alternatives; hence we just specify a set S_t of alternatives.

We then simply have:

$$a \rightarrow_t a' \iff a = a' \wedge a \in S_t$$

Observe that if the information state changes are eliminative, but not distributive, optimisations of the abovementioned kinds are not possible (although these information state changes are conceptually simplified).

Some final observations: Classical logic, which has a static semantics, can be viewed as having an optimised dynamic semantics, where sets of models play the role as information states and all information state changes are both distributive and eliminative. Also, and this is the line we shall follow here, if the structure of the alternatives is not too complicated, it ought to be possible to make denotations of classical logic function as information states. Since information states are sets of alternatives we need a logical system that makes sets (for instance as characteristic functions) available as first class citizens. Logical type theory is such a logic.

Logical Type Theory

A type theory asserts certain terms to be of certain types. A type theory can make other assertions too; for instance about the equality of terms or types. Terms and types belong to formal languages. These language will be distinct here (but are often the same in advanced type theories).

In the simple type theory a type τ is either basic (interpreted as a set of objects) or of the form $\tau \rightarrow \tau''$ *(interpreted as the set of functions from type* τ' to type τ'').[13] The terms are so-called λ-terms from the λ-calculus.[14]

In (simple) logical type theory the types are simple types with t as the basic type of propositional formulas. The set of object of type t are the set of (classical) truth values $\{0,1\}$. Axioms and inference rules must be given to ensure that the λ-calculus can be seen as a higher-order logic based on functions and predicates. The latter are characteristic functions (functions returning truth values). For our purposes it suffices to regard the terms as formulas in classical first-order predicate logic augmented with λ-terms. We write $\phi : \tau$ when the formula ϕ has type τ. See [5] for further details about logical type theory.

Anaphora

The strategy for anaphora handling is to let each alternative be a pair: world and environment, resembling the model and the assignment in case of static semantics. Hence a world is a function from constants to mathematical structures of objects and an environment is a function from stores to objects.[15]

We have a separate store for each (anaphoric) discourse markers indicated by a natural number index:

A_1 man kisses a_2 woman.

She_2 loves him_1.

This gives distributive information state changes, since the different environments can be seen point-wise. If we apply the abovementioned optimisation for distributive information state changes then we get non-eliminative information state changes, since the definite information states must also be used and not only the information states resulting from processing the texts.

Note: The chain of information states $I_0, I_1, \ldots I_n$ has eliminative information state changes with respect to $t_1, \ldots t_n$; hence for $1 \leq i \leq n$ we have:

$I_i \equiv [t_i](I_{i-1}) \subseteq I_{i-1}$

Representation in Logical Type Theory

Besides the basic type t of truth values we introduce the following basic types:

- e Entities
- i Indices

The variables x,y,z and v,w are of types e and i respectively.

The worlds are represented via models for logical type theory and the type of entities.[16] The environments are represented via the type of indices using a special constants for each store, that is for each (anaphoric) discourse marker:[17]

$m_1, \ldots, m_k : i \to e$

The following definition (for $1 \leq i \leq k$) expresses the equality of two environments on the values of all stores except store j:

$$\sim_i \equiv \lambda v w \Big(\bigwedge_{\substack{1 \leq j \leq k \\ j \neq i}} m_j v = m_j w \Big)$$

The following axioms (for $1 \leq i \leq k$) expresses that we can change each value of each store in all environments independently:

$\forall v \forall x \, \exists w \, (v \sim_i w \wedge m_i w = x)$

Example using a Categorial Grammar

We use the following categorial grammar (here written as a phrase structure grammar, but see [1] for details and a more substantial fragment):

S	$\to$ NP VP	Sentence
VP	$\to$ walk | whistle | ...	Verb Phrase
NP	$\to$ DET CN | he_1 | ... | he_k	Noun Phrase
DET	$\to$ a_1 | ... | a_k	Determiner
CN	$\to$ man | ...	Common Noun

Consider the following tiny example (morphology ignored):

$t \equiv$ **A_1 man walks.**

Translation of lexical entries (variables X,Y of type $e \to i \to i \to t$):

$$\begin{aligned}(\mathsf{a}_i)^\bullet &\equiv \lambda X Y \lambda v w \exists v' v''(v \sim_i v' \wedge X\,(m_i\,v')\,v'\,v'' \wedge Y\,(m_i\,v')\,v''\,w)\\ (\mathsf{he}_i)^\bullet &\equiv \lambda X \lambda v w (X\,(m_i\,v)\,v\,w)\\ (\rho)^\bullet &\equiv \lambda x \lambda v w (v = w \wedge \rho^\dagger\,x)\end{aligned}$$

where $\rho^\dagger \equiv$ walk | whistle | man | ... are constants of type $e \to t$ corresponding to $\rho =$ walk | whistle | man | ...

Based on the grammar, where string concatenation corresponds to function application (shown as juxtaposition as usual), we obtain the following meaning after λ-conversions:

$\to_t = (\mathsf{a}_i)^\bullet\,(\mathsf{man})^\bullet\,(\mathsf{walk})^\bullet = \lambda v w (v \sim_1 w \wedge \mathrm{man}\,(m_1\,w) \wedge \mathrm{walk}\,(m_1\,w))$

Analogously:

$t' \equiv$ **He_1 whistles.** $\to_{t'} = \lambda v w (v{=}w \wedge \mathrm{whistle}(m_1,w))$

Sentence conjunction uses the operator $\oplus \equiv \lambda p q \lambda v w \exists v' (pvv' \wedge qv'w)$ (variables p,q of type $i \to i \to t$) and using the above-mentioned axioms we have the desired result:

$\to_{tt'} = \oplus\,t^\bullet\,t'^\bullet = \lambda v w (v \sim_1 w \wedge \mathrm{man}\,(m_1\,w) \wedge \mathrm{walk}\,(m_1\,w) \wedge \mathrm{whistle}\,(m_1\,w))$

All there remains is to spell out the details of a suitable relation $=$ and we have our entailment $\models$ using a systematical and compositional analysis in a classical logic.

Intensionality

A common approach to intensionality rests on the notion of "possible worlds" (see 1), usually manipulated by new operators like $\Box$ and $\Diamond$ Diamond. Such a "possible world" corresponds to the world component of our alternatives in an information state. However, in our representation in logical type theory we have kept a fixed world via models for the logical type theory and the type of entities (the type of indices is used for the environments together with the discourse markers), but this setup will not work any longer. One way out is to introduce a new basic type w of worlds and let $\rho^\dagger$ from above be constants of type $w \to e \to t$ instead of type $e \to t$ just. A special marker m_{world} of type $i \to w$ would then select the current world. We intend to spell out the details in a future paper.

Notes

1) Notice that "intention" means purpose, aim, etc. whereas "intension" here (as well as in logic and the philosophy of language) is the direct opposite of "extension" (cf. the distinction between sense and reference)!
2) Neither the anaphora resolution nor the intensionality characterisation are too important here. For the latter the following example might help: Even though the properties of being an unmarried man and being a bachelor are extensional equivalent (fulfilled by the same indivi-

duals) they are not (always) intensional equivalent; believing that a particular individual is an unmarried man without believing that the same individual is a bachelor is entirely plausible.

3) However, if meanings are specified as mathematical objects and structures then in principle it is always possible to use, say, an object-language axiomatisation of the number and set theory of the meta-language.
4) It is most convenient to think of an alternative as a so-called (possible) world – that is, a complete description of a "state of affairs" – but other choices can be made as to be shown.
5) We do not require that such a description must provide an explanation of how humans perform the judgement.
6) Remember that we assume the text to be disambiguous. If the text was ambiguous the proper translation would depend on, but not necessarily be determined by, the context.
7) A formal calculus has just a precise syntax – possibly with some syntactical operations defined. Occasionally the syntactical operations are regarded a having semantical content themselves.
8) Anaphora in texts is not to be handled by free variables in formulas.
9) An information state is also called a knowledge state or a belief state.
10) One sure way to obtain contradictory information is to gather all possible and impossible pieces of information – this explains the ∞ sign.
11) Better notation if argument before function, namely $I_n \equiv I_o[t_1]...[t_n]$ etc.
12) Program states and information states are different notions. Also in computer science we have a fixed interpretation in mind. See [5] for the application of dynamic logics to computer science.
13) There are other interpretations for the simple type theory than the element/set one given here, for example proof/proposition and program/specification (due to the similarity in the explanation of these pairs of notions).
14) The λ-calculus comes in many guises – also without types.
15) More precisely from store names to objects (likewise with constant and variable it symbols).
16) The type of entities plays much the same role in dynamic semantics as the (single) domain for static semantics. This changes when intensionality is added.
17) The environment/store matrix is transposed – instead of environments taking a store we have stores taking an environment (or rather an index).

References

Jørgen Villadsen: Combinatory Categorial Grammar for Intensional Fragment of Natural Language. In Scandinavian Conference on AI '91, 328-339, IOS Press, 1991.

Reinhard Muskens: Anaphora and the Logic of Change. In Logics in AI '90, 412-427, Springer Lecture Notes in Computer Science 478, 1991.

Jeroen Groenendijk and Martin Stokhof: Dynamic Predicate Logic. Linguistics and Philosophy, 14, 39-100, 1991.

Johan van Benthem: Semantic Parallels in Natural Language and Computation. In Logic Colloquium '87, 331-375, North-Holland, 1989.

Robert Goldblatt: Logics of Time and Computation. Lecture Notes 7, Center for the Study of Language and Information, Stanford University, 1987.

Peter Andrews: An Introduction to Mathematical Logic and Type Theory: To Truth through Proof, Academic Press, 1986.

Jørgen Villadsen
Department of Computer Science
Technical University of Denmark
ID/DTH, Building 344
DK-2800 Lyngby, Denmark

E-mail: jv@id.dth.dk

Übersetzungstheorie und maschinelle Übersetzung

Eva Wikholm
Uppsala University

1. Einleitung

Übersetzen ist eine Tätigkeit, die zum Verständnis zwischen Sprachgebieten beiträgt und jahrtausendelang beigetragen hat. Solange Sprachbarrieren haben überbrückt werden müssen, sind Übersetzungen gemacht worden. Obwohl die Kunst des Übersetzens eine lange und reiche Tradition aufweist, hat die systematisch wissenschaftliche Beschäftigung mit der Theorie erst in den fünfziger Jahren dieses Jahrhunderts begonnen. Unter den Pionieren ist hier E.A. Nida mit seiner Bibelübersetzungsarbeit zu erwähnen.

Aus den ersten theoretischen Ansätzen auf Basis von Erfahrungen praktischer Übersetzungsarbeit ist allmählich ein eigenes Forschungsfeld entwickelt worden, und zwar die *Übersetzungswissenschaft*. In der Natur der Übersetzungswissenschaft liegt es, daß sie Elemente anderer Wissen-schaften beinhaltet, z.B. der Sprachwissenschaft, Literaturwissenschaft, Philosophie und Kommunikationsforschung. Dementsprechend weist sie auch ein vielfältiges Spektrum von Teilgebieten auf, unter denen *Übersetzungstheorie* und *linguistisch-sprachenpaargebundene* bzw. *textbezogene Übersetzungswissenschaft* zu finden sind (Koller 1983; Ingo 1991).

Eine der wichtigsten Aufgaben der Übersetzungstheorie ist es, den Übersetzungsprozeß zu beschreiben. Außerdem beschäftigt sie sich damit, Übersetzungsprobleme zu systematisieren und Methoden für deren Lösung zu erarbeiten. Im Rahmen eines linguistisch-sprachenpaargebundenen Ansatzes werden kontrastive Sprachbeschreibungen aufgestellt und die theoretischen Grundlagen dafür gegeben. Die textbezogene Übersetzungswissenschaft betrachtet das Übersetzen aus der Sicht der Textlinguistik und berücksichtigt dabei auch die Textgattung sowie die Stellung des Textes innerhalb einer Texttypologie. Andere Teilgebiete der Übersetzungswissenschaft sind wissenschaftliche Übersetzungskritik, angewandte Übersetzungswissenschaft, theorie-, übersetzungs- und rezeptionsgeschichtliche Übersetzungswissenschaft und Didaktik des Übersetzens (ebd.).

In der folgenden Darstellung werden in erster Linie übersetzungstheoretische und linguistische Aspekte der Übersetzungswissenschaft im Mittelpunkt stehen, indem folgende Fragen erörtert werden: wie sich der Übersetzungsprozeß einteilen läßt, auf welcher sprachlichen Ebene sich die Übersetzung vollzieht und was eine Übersetzungseinheit kennzeichnet. Die Ergebnisse der traditionellen Übersetzungswissenschaft auf diesen Gebieten werden in eine Umgebung maschineller Übersetzung eingegliedert. Dagegen kann hier nicht auf spezifische Methoden für die Auseinandersetzung mit Übersetzungsproblemen infolge von Sprachverschiedenheiten eingegangen werden. Solche Probleme und Methoden werden an anderer Stelle behandelt (Wikholm, in Vorbereitung II).

2. Der Übersetzungsprozeß

Das Wort *Übersetzen* weist viele Verwendungsweisen auf, weswegen es für übersetzungswissenschaft-liche Zwecke eine eindeutige Abgrenzung nötig hat. Für diese Zwecke möchten wir weiterhin mit dem Begriff Übersetzen den *Vorgang* bezeichnen, bei welchem ein Text der Ausgangssprache (AS) in schriftlicher Form in einen äquivalenten Text der Zielsprache (ZS) von einem Menschen umgesetzt wird (vgl. Koller 1983, S.106). Natürlich bezieht sich der Begriff dabei auch auf sprachliche und textbezogene Aspekte des Übersetzens.

Der Übersetzungsvorgang mit seinen Bestandteilen wird in der übersetzungswissenschaftlichen Literatur beschrieben und modelliert (u.a. Ingo 1991; Koller 1983; Nida 1969; Wilss 1977). Die Beschreibungen unterscheiden sich darin, wie sie den Vorgang gliedern und welche Faktoren sie dabei hervorheben.

W. Wilss (1977) gliedert den Vorgang aus der Sicht des Übersetzers in zwei Hauptphasen, eine *Verstehensphase* und eine *Rekonstruktionsphase*, indem er das Übersetzen als einen Textverarbeitungs- und Textreverbalisierungsprozeß auffaßt. In der Verstehensphase analysiert der Übersetzer den AS-Text auf Inhalt und Stil hin. In der Rekonstruktionsphase reproduziert der Übersetzer den AS-Text in der Zielsprache "unter optimaler Berücksichtigung kommunikativer Äquivalenzgesichtspunkte" (ebd. S.72; Koller 1983, S.111). O. Kade (1968) und G. Jäger (1975), die auch den Vorgang in zwei Teile gliedern, versuchen klarzulegen, wie die beiden Phasen miteinander verbunden sind. Die Rekonstruktionsphase kann demgemäß auf zwei verschiedene Weisen vollzogen werden, nämlich durch *Neukodierung* oder *Umkodierung*. Bei Neukodierung sind der gemeinte Sachverhalt und der Bewußtseinsinhalt Ausgangspunkt für die Rekonstruktion. Bei Umkodierung liegt primär eine sprachliche Analyse der Zuordnung von AS- und ZS-Ausdrücken zugrunde (vgl. Koller 1983, S.113). Bei einer zweiphasigen Gliederung umfaßt die Rekonstruktions-phase sowohl die Zuordnung von Äquivalenten als auch die Herstellung und Bearbeitung des ZS-Textes, wobei die beiden Momente so eng miteinander verbunden sind, daß sie nicht getrennt werden können.

Eine andere Betrachtungsweise des Übersetzungsprozesses, die ihren Ursprung in der Tradition der generativen Transformationsgrammatik hat, gliedert dagegen den Prozeß in drei Phasen. In einer solchen dreigliedrigen Beschreibung werden (1) Analyse des AS-Textes, (2) Umsetzung eines AS-Textes in einen äquivalenten ZS-Text und (3) Bearbeitung des ZS-Textes als drei selbständige Schritte betrachtet.

Die drei Phasen haben in verschiedenen Darstellungen unterschiedliche Benennungen. Zu Vertretern eines dreigliedrigen Modells wollen wir hier die Übersetzungswissenschafter E.A. Nida und R. Ingo machen. Nida (1969) nennt die drei Phasen *analysis*, *transfer* und *restructuring*, wofür Ingo (1991) die Benennungen *analys* (Analyse), *överföring* (Umsetzung) und *bearbetning* (Bearbeitung) verwendet.

Zu dieser Gliederung macht Nida (1969, S.484) folgende Bemerkung: "However, a careful analysis of exactly what goes on in the process of translating, especially in the case of source and receptor languages having quite different grammatical and semantic structures, has shown that, instead of going directly from one set of surface structures to another, the competent translator actually goes through a seemingly roundabout process of analysis, transfer, and restructuring. That is to say, the translator first analyses the message of the source language into its simplest and structurally clearest forms, transfers it at this level, and then restructures it to the level in the receptor language which is most appropriate for the audience which he intends to reach."

Laut Nida (ebd.) und Ingo (1991) ist die *Analyse* als ein komplexes Vorgehen zu verstehen. Dabei müssen verschiedene Aspekte berücksichtigt werden, wie z.B. grammatische, semantische, stilistische und pragmatische (vgl. Ingo 1991, S.39 ff). In der *Analyse* wird der AS-Text auf einfachere

oder generellere Strukturen in der AS hin analysiert, die möglichst wenige einzelsprachspezifische Züge aufweisen und eine Art Tiefenstruktur im generativen Sinne ausmachen.

Auf die Analyse folgt die *Umsetzungsphase*. In dieser Phase wird der analysierte AS-Text zu einer äquivalenten Struktur in der ZS überführt. Die Vorteile der Zuordnung von elementaren Strukturen bestehen darin, daß die Verhältnisse zwischen den Konstituenten auf einer tieferen Ebene einfacherer und klarer sind, und daß sich die Sprachen auf dieser Ebene ähnlicher sehen als in der Oberflächenstruktur (vgl. Nida 1969, S.487). Das Ergebnis der Umsetzung soll eine semantisch exakte Wiedergabe der entsprechenden AS-Struktur sein (Ingo 1991, S.165). Die Umsetzung läßt sich allerdings nicht vollständig beschreiben, weil der Umsetzungsprozeß nicht genau beobachtbar ist, da er sich im Gehirn des Übersetzers vollzieht (ebd.). Ingo hebt doch die Unentbehrlichkeit der Umsetzungsphase hervor (freie Übersetzung nach Ingo 1991, S.92, 165): "Die große Bedeutung der Umsetzungsphase bei allen Arten von Übersetzungen spricht jedoch für ihre Selbständigkeit als Phase im Übersetzungsprozeß, ..., obwohl sie teilweise von der Bearbeitungsphase des ZS-Textes schwer zu trennen ist."

Nach der Umsetzung fehlt noch die endgültige Form des ZS-Textes, die ihre Vollendung erst in der *Bearbeitungsphase* erhält. In dieser dritten Phase des Übersetzungsprozesses werden die elementaren ZS-Strukturen stilistisch zu einem ZS-Text bearbeitet, der den Normen der ZS idiomatisch, stilistisch und empfängerbezogen angemessen ist (vgl. Ingo 1991, S.216; Koller 1983, S.119).

Die oben erörterten Gliederungsvorschläge für einen zweiphasigen bzw. dreiphasigen Übersetzungsvorgang stimmen in ihren Grundzügen mit den Modellen der *maschinellen Übersetzungsansätze* überein. Die ganze prozeßbezogene Betrachtungsweise des Übersetzens ruft eigentlich den Gedanken einer Automatisierung dieses Prozesses hervor. Eine zweiphasige Einteilung kann mit einem *Interlinguamodell* (vgl. z.B. Slocum 1985) verglichen werden, während eine dreiphasige Gliederung grundsätzlich einem *Transfermodell* (ebd.) gleichkommt.

In einem Interlinguamodell sind die zwei Phasen beim Übersetzungsvorgang *Analyse* des AS-Textes und *Synthese* des ZS-Textes. Synthese wird oft bei Interlingua *Generierung* genannt. Der AS-Text wird zuerst zu einer weitmöglichst interlingual konstanten Repräsentation analysiert. Die sprachneutrale Repräsentation soll den Inhalt invariant behalten. Von dieser Repräsentation ausgehend wird dann der ZS-Text generiert. Hier liegt der Vergleich mit der Gliederung von Wilss (1977) nahe, da zwischen Analyse und *Verstehensphase* ebenso wie zwischen Generierung und *Rekonstruktionsphase* gewisse Ähnlichkeiten bestehen (siehe oben). Bei Kade (1968) und Jäger (1975) kommt der Interlinguaansatz am ehesten dem Gedanken der *Neukodierung* gleich[1].

In einem transferbasierten Übersetzungssystem können dagegen drei Phasen unterschieden werden, und zwar *Analyse*, *Transfer* und *Synthese*. Dieser Ansatz entspricht weitgehend den oben angeführten Einteilungen von Nida (1969) und Ingo (1991). Analyse würde *analysis* bei Nida und *analys* (Analyse) bei Ingo entsprechen, wie Transfer *transfer* bzw. *överföring* (Umsetzung) und Synthese *restructuring* oder *bearbetning* (Bearbeitung).

Die Ähnlichkeiten, die zwischen den beiden Forschungsfeldern Übersetzungstheorie und maschinelle Übersetzung vorliegen, deuten darauf hin, daß interessante Beiträge für die Gestaltung interlingua- bzw. transferbasierter maschineller Übersetzungssysteme in der Übersetzungstheorie zu finden sind.

3. Auf welcher Ebene vollzieht sich eigentlich die Übersetzung?

Der Punkt, in welchem im Übersetzungsprozeß der Wechsel von der AS zu der ZS stattfindet und die Art, wie dieser Wechsel vollzogen wird, unterscheiden, wie wir schon gesehen haben, die beiden Ansätze Transfer und Interlingua grundsätzlich voneinander.

Wie sich allerdings Transfer und Interlingua in dieser Hinsicht zueinander verhalten, beschreibt Tucker (1987) wie folgend: "A inherent feature of a transfer-based MT system as opposed to an interlingua system is that the structures to be transferred comprise elements typical of the source language. The deeper the analysis, the fewer the language specific elements, and vice versa, to the point where there are no more language specific features, and the structures representing the meaning of the source language text are interlingual, and, no transfer step (or, as it were, no translation) has to be carried out. What remains is a generation phase."

Die Abgrenzbarkeit der Umsetzungsphase ist also nur bei Transfermodellen vorhanden, da bei einem vollständig durchgeführten Interlinguaansatz keine eigentliche Umsetzung vorkommt, wie auch Tucker betont. Bei transferbasierter Übersetzung stellt sich dann unvermeidlich die Frage, *auf welcher sprachlichen Ebene* sich diese Umsetzung vollzieht. Wir haben schon mit Nida (1969) festgestellt, daß bei einem dreiphasigen Ansatz in der traditionellen Übersetzungstheorie die Zuordnung von äquivalenten AS- und ZS-Ausdrücken nicht direkt in der Oberflächenstruktur der Sprachen stattfindet, sondern auf einer wenig sprachspezifischen Ebene. Die darauf folgenden Fragen sind, wie sprachneutral diese Struktur sein müsste, d.h. wie tief sich diese Ebene befinden würde, und wie eine solche Struktur tatsächlich aussehen könnte.

Ingo (1991, S.165) erkennt hier eine besondere Art von Struktur als Ausgangspunkt für die Umsetzung, die er *zwischenliegende Struktur (mellanstruktur)* nennt. Diese Zwischenstruktur entspricht etwa der Tiefenstruktur in der generativen Grammatik, die von der Basiskomponente einer solchen Grammatik generiert wird. Ingos zwischenliegende Struktur liegt in der Form von solchen primitiven Sätzen, sogenannten *Kernsätzen*, vor. Der Kernsatz generalisiert also die strukturelle Variation und zeigt das vorhandene Satzmuster. Der Sinn des obenflächenstrukturellen Satzes wird aber durch ihn bewahrt, indem er eine Art Paraphrase von dessen Inhalt ausmacht[2]. Die zwischenliegende Struktur entsteht, indem die AS-Sätze auf Kernsätze hin in der AS analysiert (oder rücktransformiert) werden. Die Kernsätze werden definiert, nachdem die Konstituenten des Satzes identifiziert sind, und sie werden dann angemessen gruppiert, damit ihre Verbindungen zueinander explizit werden. Ingo meint, daß die Kernsätze dem Übergang zwischen Sprachen entsprechen, wo es möglichst wenig sprachspezifische Züge gibt, oder, wie er es auf Seite 166 ausdrückt, "wo man die Grenze zwischen den Sprachen an der engsten Stelle überschreitet" (eigene Übersetzung)[3].

Nida und Taber (1974) unterscheiden sieben Haupttypen von Kernsätzen im Englischen (Ingo 1991, S.97):

1. John ran quickly. 2. John hit Bill. 3. John gave Bill a ball. 4. John is in the house. 5. John is sick.
6. John is a boy. 7. John is my father.

Die Anzahl der Kernsätze ist von Sprache zu Sprache verschieden, im allgemeinen sind sie aber auf nicht mehr als zehn Typen (ebd.). In der Umsetzungsphase werden die analysierten AS-Strukturen äquivalenten Zwischenstrukturen der ZS auf Kernsatzebene zugeordnet[4]. Aus einem Kernsatz soll es möglich sein, die verschiedenen Satztypen mit Hilfe von Transformationen zu bilden.

Ein Vergleich mit transferbasierten maschinellen Übersetzungssystemen liegt hier nahe, da sich dieser Ansatz auf die Umsetzung von abstrakten Strukturen gründet. Es gibt in den Transfersystemen ein Modul, das die aktuellen Strukturen umsetzt. Dieses Transfermodul macht eine sprachenpaarbezogene Komponente aus. Es gibt aber keine universale Transferstruktur, sondern diese sieht in verschiedenen Systemen unterschiedlich aus. Die Transferstrukturen liegen gewöhnlicherweise in der Form von syntaktisch-semantischen Repräsentationen vor. Die Vermittlung zwischen Transferstrukturen erfolgt mit Hilfe von Transferregeln. Beim Übergang von der einen Sprache zur anderen auf einer generelleren Ebene kann die Anzahl der Transferregeln geringer sein, als wenn der Übergang auf einer sprachspezifischeren Ebene stattfindet.

4. Was für Einheiten werden übersetzt?

Wir haben festgestellt, daß im Übersetzungsprozeß eine Zuordnung von AS- und ZS-Ausdrücken auf eine Art genereller Ebene stattfindet und diese Zuordnung nur in einem transferbasierten Ansatz relevant ist. Aber was für Ausdrücke werden eigentlich übersetzt?

Die historisch gesehen erste Methode für maschinelle Übersetzung, die direkte Methode, gründet sich auf eine Wort-für-Wort-Übersetzung zwischen zwei Sprachen. Unten wird ein Beispiel angeführt, das die Problematik eines solchen Ansatzes beleuchtet:

Schw. Löneförhöjning får nu arbetarna avstå från Dt. Auf Lohnerhöhung müssen nun die Arbeiter verzichten

Nach einer morphologischen Analyse der einzelnen Wörter des schwedischen Satzes sind ohne Berücksichtigung des Kontextes u.a. folgende potentielle Äquivalente vorhanden:

löneförhöjning:	Lohnerhöhung, Gehaltserhöhung
får:	[nn][5] Schaf
	[vb] dürfen, müssen, mögen, bringen, bekommen, erhalten
nu:	[ab] nun, jetzt;
	[nn] Augenblick, Nu
arbetarna:	die Arbeiter
avstå:	verzichten
från:	[pp] von, aus, von...her, von...aus, von...an/ab, ab, seit;
	[ab] aus

Wie aus diesem Beispiel hervorgeht, bringt die Zuordnung auf Wortebene viele Wahlmöglichkeiten mit sich. Die einzelnen Wörter sind erst in ihren Textzusammenhängen, Kontexten, disambiguierbar, und eine wortbasierte Übersetzungsmethode stellt sich natürlich als unzulänglich heraus.

Mit Hilfe einer syntaktischen Analyse können die Wörter in gewissem Maße disambiguiert werden. Dies ist der Fall mit *får*, *nu* und *från* in unserem Beispielsatz. Die syntaktische Analyse erkennt *får* als finites Modalverb, das zusammen mit dem Infinitiv *avstå* das Prädikat des Satzes ausmacht. Damit kann die Funktion von *får* als Nomen sowie die nicht modalen Bedeutungen des Verbs, die die deutschen Äquivalente *bringen*, *bekommen* oder *erhalten* tragen, nicht in Frage kommen. *Nu* hat die Funktion als Adverb, und seine unbetonte Stellung deutet darauf hin, daß *nun* lieber als *jetzt* gewählt werden soll. Die Wahl des Äquivalents für *från* unter den Präpositionen erfolgt durch Valenzinformation.

Um die richtige Wahl eines deutschen Äquivalents für *löneförhöjning* und *får*, unter den drei Modalverben *dürfen*, *müssen* oder *mögen*, zu treffen, sind dagegen zusätzliche semantische Informationen notwendig. Das Verb *får* muß im Zusammenhang mit *avstå*, was negativ zu verstehen ist, gesehen werden. Die Wahl des Äquivalents für *lön* in *löneförhöjning* ist davon abhängig, ob hier ein Arbeiter oder ein Beamter gemeint ist. Syntaktische und semantische Analyse des Ausgangssatzes als Ganzes, die ein unumgänglicher Schritt für die Umsetzung ist, ist im Transferansatz zentral.

Wie die lexikalische Äquivalenz in der Transferkomponente beschrieben werden soll, läßt sich nicht generell beantworten. Sowohl Ein-Wort-Lexeme als auch Mehr-Wort-Lexeme *(z.B. avstå från)* müssen ausgedrückt werden können. Die Frage ist, welche sprachliche Größe die angemessene Einheit für die Äquivalenzbeziehungen zwischen lexikalischen Einheiten ausmacht, wenn nun das einzelne Wort sich generell als nicht ausreichend erwiesen hat. Die traditionelle Übersetzungswissenschaft hat sich auf Grund praktischer Übersetzungsarbeit mit der Problematik der Ambiguität und Äquivalenz auseinandergesetzt. Durch solche kontrastiven Studien ist dort eine besondere Art von sprachenpaarbezogener, primär semantischer, Einheit erkannt worden, die grundlegend für die Umsetzung scheint. Diese Einheit wird *Übersetzungseinheit* (ÜE) genannt (z.B. Koller 1983,

S.116). Besonders nützlich hat sie sich bei der Herstellung von Äquivalenzbeziehungen zwischen lexikalischen Einheiten infolge von Sprachverschiedenheiten erwiesen. Laut Koller (ebd.) hat man sich aber "theoretisch wenig mit dem Problem der Übersetzungseinheit beschäftigt, obwohl sie bei sprachenpaarbezogenen Beschreibungen eine wichtige Rolle spielt".

A. Malblanc (1968) und J.-P. Vinay/J. Darbelnet (1971) verwenden für die ÜE den Terminus *unité de pensée*. Damit meinen sie "le plus petit segment de l'énoncé dont la cohésion des signes est telle qu'ils ne doivent pas être traduits séparément" (ebd.). Gemäß dieser Definition behauptet man, daß die ÜE als Sinneinheit in der AS unabhängig von der ZS festgelegt werden kann (Koller 1983, S.116). Es geht jedoch hier um eine Einheit, eine Folge von Zeichen, die dadurch gekennzeichnet ist, daß sie bei Übersetzung *als ein Ganzes* aufgefaßt werden soll, und ist demzufolge nicht sprachenpaarunabhängig. Es scheint unmöglich, diese Einheit festzulegen, ohne eine spezifische oder hypothetische ZS vor Augen zu haben.

O. Kade (1968, S.90) definiert die ÜE gerade unter Berücksichtigung sowohl der AS als auch der ZS: "Die Übersetzungseinheit ist das jeweils kleinste Segment des AS-Textes, für das dank der potentiellen Äquivalenzbeziehungen ein Segment im ZS-Text gesetzt werden kann, das die Bedingungen der Invarianz auf der Inhaltsebene erfüllt."

Da die Einteilung der Sinneinheiten von Sprache zu Sprache unterschiedlich ist (Koller 1983, S.116), müssen wir uns darüber im klaren sein, daß ÜE sprachpaarbezogen sind, und daß es für jedes Sprachenpaar einen Bestand an ÜE gibt[6].

Eine ÜE kann aus einem Wort, partiellen oder vollständigen Syntagma, Satz oder Textabschnitt bestehen (vgl. Koller 1983, S.116), wie folgende schwedisch-deutsche Beispiele zeigen:

Wort:

arbetare	Arbeiter
hydraulmatare	hydraulische Vorschubeinrichtung

Partielles Syntagma:

avstå från	verzichten auf

Syntagma:

slangarnas sträckning	Schlauchstreckung
i enlighet med	laut
be en bön	ein Gebet sprechen
ta under övervägande	in Erwägung ziehen

Satz:

Förbjudet att luta sig ut.	Nicht hinauslehnen.
Nu gäller det!	Es geht um die Wurst.
Tro det eller inte.	Sage und schreibe. / Ob du's glaubst oder nicht.
Allt tar sin tid.	Rom ist auch nicht an einem Tag erbaut worden.

Textabschnitt:

Über allen Gipfeln	Över alla toppar ro.	Alla höjder vilar
Ist Ruh	Knappt ett sus.	i ljus -
In allen Wipfeln	Alla vatten spegla klart	i skogen silar
Spürest du	stillnat ljus.	knappt ett sus
Kaum einen Hauch;		bland träden fram.
Die Vöglein schweigen	Skogens fåglar i sitt bo tystna nu.	Var fågel är tyst när
im Walde.		det svalkas.
Warte nur, balde	Vänta, snart	Vänta - snart nalkas
Ruhest du auch.	vilar också du.	den ro du förnam.
(Goethe: Wanderers	(von Heidenstam)	(Nils Johansson)
Nacht lied)		

Die Notwendigkeit des Begriffes ÜE beschränkt sich nicht auf traditionelle nichtautomatisierte Übersetzung, sondern die ÜE ist eine Erscheinung, die auch bei computergestützter und maschineller Übersetzung von höchster Relevanz zu sein scheint. Eine eindeutige Definition ist gerade für diesen Bereich sogar noch wichtiger, weil die Wahl und Abgrenzung von ÜE Rückwirkungen auf die kontrastive Sprachbeschreibung hat. Die Äquivalenzbeziehungen müssen im Übersetzungssystem so genau beschrieben werden, daß sie mit formalen Transferregeln ausgedrückt werden können. Im folgenden Abschnitt wird ein Beispiel aus aktueller Forschung gegeben, die sich mit solchen Fragestellungen auseinandersetzt.

5. Ein Beispiel

Die Arbeit an dem Projekt *Multilingual Support for Translation and Writing* an der Universität in Uppsala kann als Beispiel dienen, um die Formalisierung der lexikalischen Wahl in einem System für maschinelle Übersetzung zu beleuchten. Dieses Projekt wird von der Professorin in Computerlinguistik Anna Sågvall Hein am Institut für Linguistik geleitet. Das Projekt hat zum Ziel, ein Werkzeug für computergestützte Übersetzung zu entwickeln. Es soll die Möglichkeit schaffen, eine automatische Übersetzung der vom menschlichen Übersetzer angegebenen Abschnitte eines Textes zu erhalten. Die Ausgangssprache ist Schwedisch. Als ZS sind in erster Linie Deutsch, Englisch, Französisch und Russisch vorgesehen. Das System, welches das maschinelle Übersetzen besorgt, ist transferbasiert. Die schwedische Transferstruktur liegt als eine Analyse des syntaktischen und morphologischen Parsers, *Uppsala Chart Processor* (Sågvall Hein 1987), mit einem Lexikon und einer Grammatik für Schwedisch (Sågvall Hein & Sjögreen 1991), vor. Durch Transferregeln wird eine äquivalente Struktur der ZS ermittelt. Transfer erfolgt durch Unifierung von Transferstrukturen (Beskow, im Druck).

Die Arbeit, ÜE zu identifizieren, ist schon eingeleitet worden. Bisher sind nicht flektierbare funktionale Phrasen (wie *in Übereinstimmung mit, in letzter Minute, zum Beispiel*) (Sågvall Hein, Östling & Wikholm 1990) und Zusammensetzungen (Wikholm, in Vorbereitung I) in kontrastiver Hinsicht studiert worden. Diese Arbeit wird weitergeführt werden. Nachdem die ÜE definiert sind, werden sie und ihre Äquivalenzbeziehungen zu den ZS-Entsprechungen mit Hilfe von Transferregeln ausgedrückt. Der Transferformalismus macht keinen Unterschied zwischen lexikalischen und strukturellen Regeln. Die Beschreibung der ÜE ist demzufolge in diesem Formalismus strukturell, während ihre Definition aber lexikal/semantisch basiert ist. Unten werden einige Beispiele für Regeln gegeben. Die Regeln sind von Anna Sågvall Hein definiert worden.

Beispiel 1 bis 3 unten zeigen die Notation von drei lexikalischen Transferregeln. LABEL nennt den Namen der Regel. Im AS-Teil (SOURCE) wird die lexikalische Einheit, das *Lexem*, angegeben, sowie Wortklassenangabe (Sågvall Hein 1987). Unter TARGET wird das entsprechende ZS-Lexem genannt, mit notwendigen ZS-bezogenen Auskünften versehen. Für eine ausführlichere Beschreibung des Formats der Transferregeln siehe (Beskow, im Druck).

Die Regeln beschreiben die schwedischen ÜE *hydraulmatare* und *i enlighet med* und die deutschen Äquivalente *hydraulische Vorschubeinrichtungen* bzw. *laut*.

Hydraulmatare ist eine zweiteilige Zusammensetzung und entspricht im Deutschen der nominalen Wortgruppe *hydraulische Vorschubeinrichtungen*, wo dem Nomen ein attributives Adjektiv vorangeht. Die plurale Bedeutung ist hier beabsichtigt. Wir können auch feststellen, daß das Adjektiv starke, oder pronominelle (Freund & Sundqvist 1988), Deklination aufweist.

Beispiel 1. Lexikalische Transferregel: Zusammensetzung

```
LABEL
HYDRAULMATARE
SOURCE
  <* LEX> = HYDRAULMATARE.NN.X
  <* WORD.CAT > = NOUN
TARGET
  <* PHR.CAT> = NG
  <* PHRASE > = +
  <* ATTR LEX> = HYDRAULISCH.AV.X
  <* ATTR WORD.CAT > = ADJ
  <* HEAD LEX > = VORSCHUBEINRICHTUNG.NN.X
  <* HEAD FEAT GENDER> = FEM
  <* HEAD WORD.CAT > = NOUN
TRANSFER
```

I enlighet med ist eine nicht flektierbare funktionale Phrase mit präpositionaler Funktion. Ein potentielles Äquivalent im Deutschen ist die Präposition *laut*, die den Dativ regiert.

Beispiel 2. Lexikalische Transferregel: funktionale Phrase

```
LABEL
  I+ENLIGHET+MED
SOURCE
  <* LEX > = I+ENLIGHET+MED.PP.1
TARGET
  <* LEX > = LAUT.PP.X
  <* WORD.CAT > = PREP
  <* CASE > = DAT
TRANSFER
```

Avstå från, deutsch *verzichten auf*, ist als partielles Syntagma zu betrachten. In dem vollständigen Syntagma sollte auch das Präpositionsobjekt (Freund & Sundqvist 1988) explizit ausgedrückt werden, z.B. *avstå från något*. Da das Objekt austauschbar ist, kann es nicht in eine generelle Transferregel mit einbezogen werden.

Beispiel 3. Lexikalische Transferregel: partielles Syntagma

```
LABEL
  AVSTÅ+FRÅN
SOURCE
  <* LEX > = AVSTÅ+FRÅN.VB.1
  <* WORD.CAT> = VERB
TARGET
  <* LEX > = VERZICHTEN+AUF.VB.X
  <* WORD.CAT> = VERB
  <* CASE > = ACC
TRANSFER
```

6. Zusammenfassung der Ergebnisse

Vergleiche von Ergebnissen der beiden Forschungsfelder traditioneller Übersetzungswissenschaft, die von Menschen gemachte Übersetzungen studiert, und maschineller Übersetzung haben

Ähnlichkeiten bei der Beschreibung des Übersetzungsprozesses als ein zweiteiliger bzw. dreiteiliger Vorgang aufgezeigt.

Ausgehend von einem dreigliedrigen Modell hat man sich innerhalb der beiden Felder mit der Frage von Ebene und Form der Transferrepräsentation auseinandergesetzt. Die Notwendigkeit einer genauen Beschreibung einer solchen Repräsentation in maschinellen Übersetzungssystemen spiegelt sich auch in der größeren Anzahl und Ausführlichkeit solcher Vorschläge in dieser Forschungstradition wider.

Als eine besonders große Schwierigkeit bei maschineller Übersetzung hat sich die Formalisierung der lexikalischen Wahl herausgestellt. In der traditionellen Übersetzungswissenschaft ist eine semantisch definierbare Einheit, Übersetzungseinheit (ÜE) genannt, identifiziert worden, die grundlegend für die Herstellung von Äquivalenzbeziehungen scheint. Dieser Begriff dürfte auch von Interesse in einer kontrastiven Sprachbeschreibung für maschinelle Übersetzungszwecke sein.

Im Rahmen des Projektes MULTRA in Uppsala wird daran gearbeitet, den Formalismus für maschinelle Übersetzung zu gestalten und Transferregeln in diesem Formalismus auszuprobieren anhand konkreter Übersetzungsarbeit. Die Handhabung von ÜE wird in dieser Umgebung studiert. Bisher sind Zusammensetzungen und nicht flektierbare funktionale Phrasen als ÜE untersucht worden, und der Begriff hat sich als nützlich erwiesen. Weitere Arbeit mit solchen Fragestellungen zielt darauf hin, den Begriff genauer zu definieren und, von praktischen Übersetzungsfällen ausgehend, verschiedene Arten von ÜE abzugrenzen und so sorgfältig zu beschreiben, daß es möglich wird, die Äquivalenzbeziehungen mit Transferregeln auszudrücken.

Anmerkungen

1) Umkodierung könnte mit der sogenannten direkten Methode für maschinelle Übersetzung verglichen werden
2) Ingo (1991) verzichtet darauf, etwas darüber zu sagen, wie z.B. stilistische und thematische Faktoren in der Zwischenstruktur gehandhabt werden.
3) Eine tiefgehendere Analyse bis zu einer sogenannten sprachunabhängigen Repräsentation, wie bei Interlingua, scheint aber nicht notwendig. Eine solche Analyse kann stattdessen laut Nida (1969, S.487) zu abweichenden Ergebnissen kommen: "In reality the transfer at the kernel level can generally be made with far less danger of skewing than if one follows the highly involved processes of going to the level of semantic universals and returning again to the kernel level."
4) Da die Strukturen der Kernsätze der verschiedenen Sprachen ähnlich sind, sollte demzufolge der Inhalt bei Umsetzung auf Kernsatzebene möglichst invariant bleiben (Nida 1969, S.489). Natürlich sind aber inhaltliche Veränderungen, die mit einem Mangel an semantischer Äquivalenz zusammenhängen, nicht auszuschließen (vgl. ebd.). Hier könnte man auch behaupten, daß lexikalische Ambiguität ein Grund zur inhaltlichen Diskrepanz bei der Umsetzung sein könnte.
5) Die Notation der Wortklassenbezeichnungen folgt (Allén et al. 1970): [ab] = Adverb, [nn] = Nomen, [pp] = Präposition, [vb] = Verb.
6) Hier kann ein Vergleich mit dem Umgang dieser Problematik innerhalb des Interlinguaansatzes gemacht werden. Dort versucht man, eine Einteilung in sprachunabhängige Sinneinheiten vorzunehmen. Eine universale Einteilung in Sinneinheiten ohne Berücksichtigung existierender Sprachen ist jedoch schwer vorstellbar.

Literaturverzeichnis

Allén, S. et al. *Nusvensk frekvensordbok 1*. Stockholm. 1970.

Beskow, B. Machine Translation in a Unification Based Framework. Presentation vid de Nordiska Datalingvistikdagarna 1991. Im Druck.

Catford, J.C. 1965. *A Linguistic Theory of Translation*. An Essay in Applied Linguistics. London (= Language and Language Learning).

Freund, F. & Sundqvist, B. 1988. *Tysk grammatik*. Natur och Kultur, Arlöv.

Hutchins, W.J. 1986. *Machine Translation*. Past, Present, Future. Chichester.

Ingo, R. 1991. *Från källspråk till målspråk*. Introduktion i översättningsvetenskap. Lund.

Jäger, G. 1975. *Translation und Translationslinguistik*. Halle (Saale) (= Linguistische Studien).

Kade, O. 1968. *Zufall und Gesetzmäßigkeit in der Übersetzung*. Leipzig (= Beihefte zur Zeitschrift Fremdsprachen, I).

Koller, W. 1983. *Einführung in die Übersetzungswissenschaft*. Heidelberg.

Malblanc, A. 1968. *Stylistique comparée du français et de l'allemand*. Essai de représentation linguistique comparée et Etude de traduction. Paris (4. Aufl.) (= Bibliothèque de stylistique comparée, II).

Nida, E.A. 1969. Science of Translation. In: *Language*. Baltimore. S. 483-498.

Nida, E.A. & Taber, Ch.R. 1969. *The Theory and Practice of Translation*. Leiden.

Slocum, J. 1985. A Survey of Machine Translation: Its History, Current Status and Future Prospects. In: *Computational Linguistics*. Volume 11, Number 1, January-March 1985.

Sågvall Hein, A. 1987. Parsing by means of Uppsala Chart Processor (UCP). In: Bolc, L. (Hrsg). *Natural Language Parsing Systems*. Springer Verlag.

Sågvall Hein, A., Östling, A. & Wikholm, E. 1990. Phrases in the Core Vocabulary. Center for Computational Linguistics. Uppsala University.

Sågvall Hein, A. & Sjögreen C. 1991. Ett svenskt stamlexikon för datamaskinell morfologisk analys. En översikt. I: Mats Thelander et al. (utg.) *Förhandlingar vid Artonde sammankomsten för svenskans beskrivning*. Uppsala den 25-26 oktober 1990. Lund. S.348-360.

Tucker, A. 1987. Current Strategies in Machine Translation Research and Development. In: Nirenburg, S. (ed). *Machine Translation*. Theoretical and Methodological Issues. Cambridge

Vinay, J.-P./Darbelnet, J. 1971. *Stilistique comparée du français et de l'anglais*. Méthode de traduction (nouvelle édition revue et corrigée, 1. Aufl. 1958). Paris (=Bibliothèque de stylistique comparée, I).

Wikholm, E. Kontrastivität auf lexikaler Ebene. Die Zusammensetzung als Übersetzungseinheit. Uppsala Universität. Institut für Linguistik. In Vorbereitung I.

Wikholm, E. Sprachwechsel - Perspektivewechsel? Methoden bei der Auseinandersetzung von Übersetzungsproblemen. Uppsala Universität. Institut für Linguistik. In Vorbereitung II.

Wilss, W. 1977. *Übersetzungswissenschaft*. Probleme und Methoden. Stuttgart.

Eva Wikholm
Uppsala University
Department of Linguistics
Computational Linguistics
Box 513
S-751 20 Uppsala
E-mail: uduew@seudac21.bitnet

A Swedish Core Vocabulary for Machine Translation

Annette Östling
Uppsala University

Abstract

The reasons for establishing a new Swedish core vocabulary are presented, and the steps taken in its establishment are described. Some conclusions are drawn as for the usefulness of frequency lists in this respect. The impact on the frequency lists of the nature of the corpus is illustrated. The necessity of introducing phrases in the core vocabulary is pointed out. The information to be associated with the entries to meet the requirements of the translation process is looked upon in the light of the definitions in a monolingual dictionary.

Introduction

The project *Multilingual Support for Translation and Writing* carried out at the Department of Linguistics/Computational Linguistics at Uppsala University aims at a computer tool for the processes of translation and writing, with professional translators as the main user group in view. The translation process is to be monodirectional with Swedish as the source language and English, German and French as the targets. Apart from providing possibilities for mechanical translation of parts of a document specified by the user, the purpose of the support is also to provide functions for selective dictionary look-up, allowing the user to specify the kind of information wanted about a requested word or phrase. The dictionary is to be organized as a lexical database, made up of monolingual databases for Swedish, English, German and French. The links between the Swedish database and the target language ones will be the equivalence relations between the translation units in a language pair. A translation unit is defined as the smallest possible unit in the source language that can be substituted by an equivalent in the target language, on the semantic level as well as on the stylistic level. It can be a word, a phrase or a larger segment of the text. It is a well-known fact that one-to-one correspondences between the lexical units, words or phrases of two languages are rare. Contrastive lexical studies, investigating the equivalents and their relations are central in translation theory, and hence are important also for machine translation. In Wikholm (1991) an account is given of the types of lexical differences that exist between languages and different types of equivalence relations are discussed, along with the implications for machine translation.

Within the monolingual databases of the LDB, there will be a distinction between a general, permanently stored vocabulary and domain-specific parts. The general part, the core dictionary, will cover high-frequency, common language words and phrases, whereas the users will incorporate the domain- or text-specific vocabularies in the LDB in accordance with their needs.

It is important to have in mind that the LDB is to be consulted by a human translator/writer as well as by the machine translation component of the support and must fulfill the demands of both kinds of users.

Since the primary units of the database are the translation units, the Swedish core dictionary must be structured in such a way that the establishment of translation relations to the target languages is possible. High-frequency words are often manyfold ambiguous, morphologically and semantically. For example, the English word 'in' can be either a preposition, an adverb or a noun, and there are obviously no one-to-one correspondences between 'in', 'do' and 'make' and their Swedish equivalents. In phrases the ambiguity can be resolved: 'in' is disambiguated as a preposition when used in the phrase 'in accordance with'. This fact will be taken advantage of by introducing phrases as an important part of the core vocabulary, and is one step in the development of a core vocabulary of translation units.

In this project, Swedish is the source language. Therefore, in a first phase, the Swedish core vocabulary will be established. Then translation links to the equivalents in English, German and French will be determined.

Is there a need for a new Swedish core vocabulary?

A first Swedish basic vocabulary was identified within the project *Nusvenskfrekvensordbok (NFO), 'Frequency Dictionary of Present-Day Swedish'* (Allén et al. 1970). The NFO basic vocabulary was derived from a corpus of Swedish newspaper text from 1965, consisting of one million running words. The basic vocabulary comprises 10,000 lemmatized graphic words[1]. Thus no compound expressions such as phrases or phrasal verbs are recognized as units in the NFO basic vocabulary, the basic reason for the decision to determine a new core vocabulary within this project. The size is another important issue: is 10,000 graphic words a reasonable size for a core vocabulary of the kind needed for a tool for translation and writing? Since the basic vocabulary of NFO is quite big, it comprises a high proportion of content words. It is in the nature of content words to reflect the topics treated in a text, and some of the current issues of 1965 newspaper articles are no longer of the same importance (the Vietnam war, for instance), and many new topics and phenomena have arisen since then, such as the concern for the environment, the media explosion with video and satellite television, etc. As shown below, even words in the frequency top can be quite domain-dependent. Thus part of the NFO basic vocabulary may be, in this respect, somewhat outdated.

The NFO basic vocabulary was determined as a general frequency top list, without any special use in mind. The core vocabulary aimed at in this project is to be a tool for all the users in view, regardless of their various fields of application. Hence one of the main criteria for the choice of which content words to be incorporated must be their neutrality with respect to domain, or, put in another way, the likelihood for them being used in different domains. It follows that it is natural to look not only at frequency lists of newspaper vocabulary, but also to compare with frequency lists of other types of texts. Lehmann (1991) describes the comparison of five word lists from very different domains, and shows that the common vocabulary is very small indeed. He also points out that the definition of a representative corpus, and hence a representative vocabulary, must be made with the research goals pursued in mind. It is quite clear that a core vocabulary for the purpose of being used as a translation tool as the one described here is not the same thing as a basic vocbulary for language learning, for instance. A dictionary for this latter purpose must cover the vocabulary and language functions for every-day life, since it aims at guiding the language student to what is the "threshold level" of another language, something which is not the same thing as domain-neutrality.

How can a core vocabulary be determined?

The need for core vocabularies for the purpose of natural language processing is pointed out in Lehmann (1991), who also describes the establishment of a German core dictionary.

The new Swedish core dictionary for the translation and writing tool being worked out within this project is based on the total of the newspaper material available via Språkbanken ('The Language Bank') at Språkdata, University of Gothenburg (Gellerstam 1989). In all, this material comprises approximately seven million current words from 1965, 1976 and 1987, the words from 1987 constituting the largest part: five million words. Comparisons have been made with the novel corpus from Språkbanken. This corpus consists of two parts, totalling 9.6 million words[2]: novels published in 1976 and 1977, comprising 5.6 million running words, and novels published in 1980 and 1981, four million running words.

The first step in establishing the core vocabulary was to make a morphological analysis of all the word forms from the Språkbanken newspaper corpus, henceforth NWP, with a frequency of 160 or more (the 2,926 most common word forms). The morphological analysis was performed with the help of Uppsala Chart Processor (UCP) (Sågvall Hein 1987), an inflectional grammar of Swedish in the UCP formalism (Sågvall Hein, forthcoming) and a stem dictionary of about 60,000 entries (Sågvall Hein & Sjögreen 1991). The stem dictionary is generated from *Svensk Ordbok*, *'A Dictionary of Swedish'* (1986) and thus covers the same vocabulary. It follows that the morphological analysis results in the same lemma distinctions as the ones made in *Svensk Ordbok*.

Homography is a very common phenomenon in Swedish, not only among high-frequency function words, and therefore the number of possible lemmas assigned to a word form by the morphological analyzer is often > 1. The frequency 160 was chosen ad hoc as a starting point, but proved to be a good guess (see below). The result of this parsing was a list of all the possible lemmas for each word form according to *Svensk Ordbok*, as shown by the following example. Here, the graphic word *perfekt* is given one noun analysis (the grammatical term 'perfect') and one adjective analysis ('perfect'):

perfekt: (freq: 311)
2 parses, 9 vertices.

```
PERFEKT :
(* =   (LEM = PERFEKT2.NN          (* =   (LEM = PERFEKT1.AV
       INFL = PATTERN.GRYN                INFL = PATTERN.BLEK
       DIC.STEM = PERFEKT                 DIC.STEM = PERFEKT
       WORD.CAT = NOUN                    WORD.CAT = ADJ
       GENDER = NEUTR                     COMP = POS
       FORM = INDEF                       NUMB = SING
       CASE = BASIC))                     FORM = INDEF))
```

Since this is a pursuit of a *core* vocabulary, all lemmas marked for a certain level of style or a specific domain were excluded from the list. Furthermore, the lemmas that only occur in compounds were given a special code.

The next step was to compare this list of graphic words tagged with their possible lemmas with a lemmatized frequency list in order to exclude unlikely lemma attributions. Thus a comparison, much like a manual probabilistic tagging, was made with *Nusvenskfrekvensordbok*, and the lemmas having a NFO frequency of 10 or less (i.e. the lemmas in question are not among the 7,150 most common ones) were excluded from the list. For instance, the noun analysis of *perfekt* was among the ones sorted out. Setting the limit as low as $f \leq 10$ may seem an unnecessary precaution, but in this early sorting out precaution is necessary in order not to rule out lemmas that are rare only in

NFO. Of course lemmas that should not have been excluded may still have been sorted out by mistake. Testing against the texts of the potential users will be necessary in order to find inconsistencies and to refine the vocabulary so far defined.

In order to study the frequency of the different parts of speech more in detail, the list of graphic words tagged with their possible lemmas was divided into 14 segments of 200 word forms each. Segment 1 consists of the 200 most common word forms, segment 2 comprises the next 200 and so forth. The last 126 graphic words were thereby left out, something which does not seem to be of much importance (see below). The following table shows the frequency span for the graphic words included in each segment (the corpus size is seven million current words):

	Graphic word frequency		
Segment 1	201670-2510	Segment 8	352-306
Segment 2	2497-1194	Segment 9	306-269
Segment 3	1192-820	Segment 10	269-243
Segment 4	818-617	Segment 11	243-215
Segment 5	616-497	Segment 12	215-196
Segment 6	497-412	Segment 13	196-182
Segment 7	412-352	Segment 14	182-167

The table below gives an account of the number of possible lemmas of every word class found in each segment. The lemmas with a NFO frequency of 10 or less have been excluded from the table. S1 etc. refers to segment number[3].

	Noun	Verb	Adj	Adv	Pron	Prep	Con	Subj
S1	26	55	30	67	39	29	19	1
S2	61	48	48	51	15	9	3	1
S3	90	50	35	32	9	3	1	1
S4	106	62	41	11	2	4	1	3
S5	89	60	54	20	1	2	0	1
S6	100	58	46	11	2	3	3	0
S7	99	52	45	10	1	0	0	1
S8	109	66	33	12	2	2	0	0
S9	108	66	39	11	1	2	0	0
S10	112	54	43	10	0	1	1	2
S11	91	76	38	6	1	2	0	0
S12	117	64	35	4	1	1	0	0
S13	108	55	36	10	1	1	0	0
S14	103	71	32	8	3	1	0	0
Σ	1406	837	555	263	78	60	28	10

NB that the number of possible lemmas in each segment is over 200. This is due to the fact that many graphic words have more than one lemma attribution. For instance, many graphic words in segment 1 have been analyzed as being both possible prepositions and adverbs (i.e. in this case verb particles). *får*, on the other hand, an extremely common verb form meaning, among other things, 'may', 'should' or 'has', is homograph with the noun meaning 'sheep'. Since the frequency of the noun lemma *får* in the NFO frequency list is less than 10, this lemma attribution has been discarded.

The segmentation procedure proved fruitful for many decisions, especially concerning the function words, since it made it possible to see in which segments they start disappearing. It also permitted a close study of where on the frequency scale the domain-specific content words begin outnumbering the more general ones, although this study still could not be decisive for which

content words to incorporate in the core vocabulary. The segmentation made it evident, though, that the distribution of the content word lemmas is fairly even throughout the segments (apart from the first two for the nouns). It is thereby clear that the content words give their color even to the frequency top, something to have in mind when discussing the representativity of a corpus for the goals in view. Below follows a brief discussion of each word class. For a list of all the entries proposed for the core vocabulary, see Östling (forthcoming).

Function words

Prepositions

All prepositions found in this 2,800 word form top list seem to qualify for inclusion in the core vocabulary. It is evident that after f 2,000, i.e. segment 10, the number of "new" prepositions is very low, something which permits the preliminary conclusion that the prepositions found here really are the core ones in Swedish – it seems unlikely that some important prepositions appear below this frequency. Testing against user texts and other corpora would of course corroborate or falsify this hypothesis.

Conjunctions

It is interesting to see that the pattern for the prepositions is repeated, and is even clearer, for the conjunctions. Below f 2,000 no conjunctions appear, which supports the hypothesis that neither for this part of speech will it be necessary to go further down the frequency list. The conjunctions among the top 2,000 can thus tentatively be claimed to be the conjunctions of the core vocabulary.

Subjunctions

The subjunctions are very few, but the pattern remains the same as for the two parts of speech above: there are no subjunctions below f 2,000, and the list of subjunctions established from the frequency top here is preliminarily established as the list of the core vocabulary.

Pronouns

Below segment 3, i.e. below f 600, only a handful of pronouns are to be found. A closer study, however, shows that some personal and possessive pronouns are missing. The use of these pronouns is highly dependent on the communicative function of the text (Lehmann 1991), and has nothing to do with the topics treated. It is evident that in a core vocabulary of the kind needed here the pronouns of all six grammatical persons should be included. In order to find them all in this frequency list, it would be necessary to go beyond the top 2,800, but this would be a superfluous task, since it is easy to check this consistency manually in the existing list of pronouns. (Of course it is not uninteresting to see where on the frequency list the missing ones appear – this could be of importance for core vocabularies proposed for other types of use, as for instance language learning).

Adverbs

Many Swedish adverbs are formed on an adjective stem, the ending -t signalling the adverbial function. This ending, however, is also the neutral gender ending of Swedish adjectives, and this latter use is the only one recognized by the UCP parser. Thereby no adverbs of this open-class kind are included among the core vocabulary adverbs so far, something which of course has to be remedied in due course of time.

The majority of the adverbs among the top 200 are commonly used as verb particles. This is true, for instance, for *in* (*komma in* – 'come in') and *över* (*ta över* – 'take over'). The number of "new" adverbs is diminishing throughout the segments, but the segmentation gives no evidence for the conclusion that basic adverbs will not be found below f 2,800. Thus the frequency list has to be

explored beyond this limit, and the procedure described for the content words (see below) will have to be used.

Content words

A quick glance at the frequency list reveals the fact that surprisingly many content words with a high frequency are somehow specialized and domain-specific. Some noun examples:

	Freq	Segment
regeringen ('the government')	2,813	1
matchen ('the match')	1,371	2
kommunen ('the local authorities')	1,284	2
VM ('world championship')	930	3

A newspaper corpus consists of articles of varied topics, but it is clear that some topics are more frequent than others, that the current topics vary with time, and that some topics hardly ever are treated in newspapers. In spite of this, newspaper corpora are often judged to give a reliable cross-section of contemporary language. As a somewhat provocative contribution to this debate, the nouns among the top 2,800 in the NWP that are not to be found in the same segment of the novel corpus are most illustrative. Here follows a list of some of them, randomly chosen:

aids aids	*bolag* company	*final* final	*motståndare* adversary
aktie share	*budget* budget	*försvar* defence	*personal* staff
argument argument	*daghem* day-nursery	*kongress* congress	*resurs* resource
befolkning population	*debatt* debate	*kostnad* expense	*satsning* venture
bidrag allowance	*expert* expert	*läsare* reader	*syfte* purpose

It is, in our opinion, quite clear that these words are typical of newspaper texts. The presence of *aids* in the list also shows that this corpus must be a fairly recent one. Intuitively it is easy to conclude that the words are not likely to be used to the same extent in fiction, and cannot be labelled domain-neutral. If, on the other hand, we take a closer look at some of the novel nouns that are not to be found among the top 2,800 in the NWP, we find another type of nouns, intuitively felt to belong to the world of fiction and almost forming a poem just by themselves:

blick look	*frukost* breakfast	*idiot* diot	*mörker* darkness
disk dishes	*gryning* dawn	*klänning* dress	*skratt* laughter
dröm dream	*gråt* tears	*köksbord* kitchen table	*smärta* pain
famn arms	*hemlighet* secret	*längtan* longing	*suck* sigh
fest party	*hud* skin	*måne* moon	*säng* bed

This latter list could have been made even more "fiction biased", comprising only parts of the body and emotions, for instance.

This clearly shows the importance of studying corpora of different nature, if some degree of domain-neutrality is to be achieved. "Domain-neutrality" may be a notion which does not exist in reality – no word is neutral when it is used, but on the contrary gets some of its significance from the context. Domain-neutrality is thus something which can only be aimed at and probably never achieved. In order to approach this goal, the common procedure for the nouns, adjectives and verbs was to determine the intersection set between the two corpora of the top 2,800 of the parts of speech of the content words. Thereby the words chosen for the core vocabulary are used frequently in newspaper texts as well as in fiction, a clear indication that they are commonly used in domains of different types.

Nouns

It resulted from the study of the NWP that the raw frequency figures were of no great help when it came to deciding where to draw the limit for the nouns to be included in the core vocabulary: domain-specific nouns as well as more general ones occur in each segment. One important conclusion from this study was that it is not possible just to draw a limit in a frequency list, but that frequency studies can merely be one part of the material needed for conclusions concerning the content words. Semantic classification and "common sense" also have to play important roles.

The intersection with the novel corpus resulted in a list with 472 noun lemmas (many lemmas were represented by more than one graphic word). This list was then checked against the lemmatized NFO list, and some unlikely lemma attributions could be deleted. The nouns were then grouped semantically: family, professions, time & season, nature & weather, communications, body, institutions & community, food & drink, house & home, directions, daily things, attitudes, materials, measurements, arts. These groups captured a good deal of the nouns, and permitted the spotting of inconsistencies in the frequency lists. It was found, for instance, that all the weekdays except *lördag* and *söndag* ('Saturday' and 'Sunday') were missing. This grouping was most useful in order to find missing hyperonyms and missing and superfluous hyponyms, and was thereby of great help in the creation of more homogeneous groups. Quite a few hyperonyms were among the nouns added to the list on its way towards a higher degree of generality. In spite of the usefulness of the intersection with the novel corpus and the semantic grouping, it has to be pointed out that common sense still must play an important role for many decisions.

The "final" list consists of 400 noun lemmas, but it is very likely that it has to be further restrained when tested against user texts and other corpora.

Adjectives

The same procedure as for the nouns was carried out for the adjectives. The intersection with the possible adjective lemmas in the novel corpus was established, giving a list of 243 adjective lemmas. The list was checked against the lemmatized NFO material to further exclude rare and unlikely adjective lemmas. A rough semantic grouping was then made to check that no elementary adjectives or basic colors were missing. In this way some inconsistencies were found, and after these procedures, plus the important use of common sense, the ameliorated list today contains 172 lemmas.

Verbs

The verbs have been less thoroughly studied than the other parts of speech. This is due to the fact that no study yet has been made taking into account the very common phrasal verbs (*komma in, ta över*). The procedure followed so far is equivalent with the one used for the adjectives and the nouns. The intersection group with the novel corpus comprises 271 verb lemmas, but these have not yet been checked against the NFO material, neither has a semantic classification been performed, and the list is also in these respects less worked out compared to the others. The very preliminary verb list established so far comprises 265 simple verb lemmas.

Phrases in the core vocabulary

As already pointed out, ambiguity is a very common phenomenon among high-frequency function words. A step towards the establishment of translation units between Swedish and the target languages is the introduction of phrases in the core vocabulary. There are lexicalized phrases of many types: invariable, continuous phrases, an example of which is *i enlighet med*, 'in accordance with', inflectible phrases including phrasal verbs such as *stänga av*, 'turn off' and discontinuous ones, as for example *antingen ... eller*, 'either ... or'. The phrases focussed on so far in the project are the invariable continuous ones, the *functional core phrases*. Criteria for the determination of this type of phrases have been worked out (Sågvall Hein, Östling & Wikholm 1990) and include the following points:

- a functional core phrase should form a syntactically motivated unit. A phrase can not include an element which is part of a previous or following constituent. It follows that the phrase must be continuous.
- it must have a specific grammatical function: prepositional, adverbial, adnominal or conjunctional.
- it should be semantically neutral. If the phrase contains a nominal element, this has to be non-referring.
- it should disambiguate one or more elements of the phrase with regard to meaning or part of speech (see the example mentioned above concerning 'in').

A pilot study showed that the structures which are potential functional core phrase candidates are limited in number, and for Swedish the following four structures are the only possible ones:

preposition + noun + preposition
i enlighet med, 'in accordance with'
inom ramen för, 'within the framework of'

preposition + adjective/pronoun + noun
på goda grunder, 'for excellent reasons'
i första hand, 'in the first place'

preposition + noun
med flit, 'on purpose'
i år, 'this year'

adverb + adverb
snarast möjligt, 'as soon as possible'
inte minst, 'not least'

An important source for the establishment of the Swedish functional core phrases is the *Frequency Dictionary of Present-Day Swedish, Volume 3, Collocations* (Allén et al. 1975). It proved that less than half the number of the collocations of the structures above could be included in the core vocabulary. The non-fulfillment of the criteria was often due either to specific reference of the noun or the adjective, or to the preposition being dependent on a preceding verb, thus invalidating the criterion of a syntactically motivated unit ('learn from | this development', for example, in which the preposition is dependent on 'learn'. In this particular example, the noun also has specific reference). The list of functional core phrases is being worked out, and will include some 1,000 entries (Östling & Wikholm, forthcoming).

In a later phase of the project, inflecting phrases, such as phrasal verbs and temporal expressions will be included in the core vocabulary, along with some common abbreviations.

What information is to be associated with the entries in the Swedish database?

The information associated with each entry is of crucial importance for the functionality of the database, for the human user as well as for the machine translation component. The ideas presented below are part of our working concept, and have not yet been implemented.

In general terms, it is a quite simple task to determine which function words that are to be included in the core vocabulary, but it is most difficult to determine what information is to be associated with them for a nicely functioning translation and writing support. For the content words, the opposite is true: it is most difficult to decide on which ones fulfill the criterion of being sufficiently domain-neutral to be included in the core vocabulary, but it is an easier task, although far from trivial, to decide on what information is to be associated with them.

Morphological information is to be an essential part of each entry, and should be consistent with the output of the UCP parser (Sågvall Hein 1987), as exemplified above for *perfekt*. The morphological analysis provides the lemma, LEM, in accordance with the classification in *Svensk Ordbok*. The part of speech of the lemma is given after the punctuation mark, in the example nn, 'noun', and av, 'adjective'. INFL gives information about which inflectional pattern the lemma belongs to. DIC.STEM specifies the stem, since there is sometimes a difference in Swedish between the lemma and its stem. The analysis also provides morphological information specific to each part of speech, in the case of nouns gender, number, form and case.

Each lemma is associated with a list of lexemes, classified according to *Svensk Ordbok* (Sågvall Hein 1988). Thus the lemma glas.nn ('glass') is linked to its lexemes in the following way:

```
(GLAS.NN        (LEX GLAS.NN.1)
                (LEX2 GLAS.NN.2)
                (LEX3 GLAS.NN.3))
```

One type of semantic information in the database is thereby the definitions of the lexemes.

The entry for glas is a good starting point for a discussion of the information to be associated with the noun entries:

glas subst. *~et* = **1** ett hårt, glänsande, genomskinligt ämne som anv. särsk. i dryckeskärl, fönsterrutor etc. (jfr porslin): *glasburk; glasdörr; glasgjutning; glaskupa; glasprisma; glasruta; glasskiva; glasskål; glasveranda; buteljglas; kristallglas; rubinglas; sodaglas; spegelglas; trådglas; blåsa ~; grönt ~; en lampa i slipat ~; en futuristisk lägenhet i stål och ~* □ äv. om (vissa) föremål av detta ämne (jfr **glas 2**): *förstoringsglas; koppglas; lampglas; returglas; timglas; tomglas; spräcka ett av ~en i glasögonen; sätta bilden inom ~ och ram; några ansikten rörde sig bakom ~et; odla under ~* □ äv. utvidgat om glasliknande ämne el. föremål: *vattenglas* **2** dryckeskärl av glas utan handtag, men ibl. med fot; vanl. anv. för kalla drycker (jfr **kopp**, **mugg 1**, **bägare**): *glasservis; dricksglas; kristallglas; nubbeglas; spetsglas; vattenglas; vinglas; fylla ~et till brädden; ta fram en tillbringare och fyra ~* □ äv. med tonvikt på innehållet: *bjuda på ett ~ likör; dricka ur ~et i ett svep* □ spec. i fråga om alkohol: *ta sig ett ~* □ äv. om likn. föremål: *plastglas* **3** halvtimma mätt med timglas och markerad med slag på klocka; på fartyg (hist.): *åtta ~ utgör en skeppsvakt*

The first two lexemes of *glas* are to be included in the core dictionary, whereas the third one has the comment <hist.>, 'historical', and hence must be discarded, since its usage is limited to a certain domain. The distinction between the two first lexemes is crucial from a translational point of view: glas.nn1 is the substance, whereas glas.nn2 is the vessel used for beverages. This information can

be extracted from the definitions in the following way: glas.nn1 is defined as an *ämne*, 'substance', and with a semantic parsing, an ISA-link can be established (Vossen et al. 1989, Hagman 1991) to the hyperonym, which is thereby made evident. For glas.nn2 the ISA-link is established to *dryckeskärl*, 'vessel used for beverages', the hyperonym in question. Especially useful for the writing-process is information concerning hyponyms. These are enumerated after the meaning descriptions in the definitions in *Svensk Ordbok*: for glas.nn2 "*dricksglas; kristallglas;*" ('drinking glass', 'crystal glass') etc. Synonyms can also be of use, for the translation as well as the writing process, and this relation can be extracted from the meaning description in the following way: if the ISA-attribute has a numerus value matching the lemma and has no modifiers, it is considered a synonym of the lemma (Hagman 1991). Sometimes synonyms are also explicitely listed in *Svensk Ordbok* after the meaning description, and are thus easily retrieved.

Syntagmatic information, such as the use of the lexeme in expressions, idioms and phrases, is most important in order to achieve idiomatic translations. *Svensk Ordbok* gives quite a few examples of the lexemes in use, and some information of this kind can thereby be accessed from the definitions. It is crucial for the functioning of the LDB that the syntagmatic information be looked upon and determined in the light of what the translation units between the languages are. Since the translation units are not always the same between the language pairs, the syntagmatic information has to be specified with regard to the different target languages.

As for the contrastive information, the reference from the Swedish lexeme to its equivalent(s) in the target languages is the only link between them. glas.nn1 has one equivalent in English, 'glass (mass noun)', and the equivalent of glas.nn2 is 'glass (countable)'. The lexeme distinction reflects the different translation equivalents in this case. Often the use of a word in a phrase affects the translation, and equivalents have to established between these larger unities in the definitions, thereby overriding the translation equivalent of the simple lexeme.

The entry for the preposition *av* illustrates the situation for the function words:

'**av** prep. **1** med utgångspunkt eller ursprung i: *därav; ärva ~ sin far; resultatet ~ undersökningen; framgå ~ en artikel* □ spec. för att ut-tryc-ka orsak e. d.: *sjuk ~ sorg; lida ~ huvudvärk* □ äv. med angivande av material e. d.: *en ring ~ guld* □ äv. som markering för del av ngn helhet: *nio filmstjärnor ~ tio* ✱ *(rädd) ~ sig* (rädd) till sin läggning; *nog ~* äv. ett ord hur som helst; *rent ~* äv. ett ord till och med **2** särsk. i pass. konstruktioner genom direkt inverkan från ngn el. ngt agerande: *domen överklagades ~ en släkting; boken är skriven ~ en engelsman; statyn har skänkts ~ en armenisk oljemagnat; hon stoppades ~ en bastant bakdörr; träffas ~ blixten* □ äv. för att uttrycka upphovsman e. d.: *en dikt ~ Fröding* **3** i riktning bort från så att resultatet blir ett avskiljande; konkret el. abstrakt (jfr ²av 3): *dra huden ~ oxen; båten löpte ~ stapeln* **4** vanl. efter verbalsubst. med inriktning på ngt föremål för den angivna verksamheten etc. (jfr ²av 2): *den fortsatta utbyggnaden ~ kärnkraften; de allierades erövring ~ Tunis; vården ~ missbrukare* **5** i ngt som kan anges med anv. för specificering: *en kostnad ~ två miljarder; ett spann ~ 1600 meter; en sås bestående ~ gräddfil och gräslök; publiken utgjordes ~ kvinnor i 35-årsåldern* ✱ *~ vikt* viktig; *~ värde* värdefull **6** i hänseende till angivande relativt lösa samband, anv. som ett slags genitiv: *vid slutet ~ året; de trafiksvaga delarna ~ nätet*

The lemma av1.pp has six lexemes. It is not always evident that the division in lexemes coincides with the different translation equivalents. The lexeme av1.pp1 has at least two equivalents in English, often dependent on the preceding noun or verb: *ärva av* – 'inherit from', *resultatet av* – 'the result of', *rent av* – 'even'. These few examples clearly show the need for the establishment of translation equivalents between phrases. av1.pp2 refers to a grammatical construction, the passive, and hence this lexeme should be more uniform in its behaviour in the target languages. av1.pp5 exemplifies the difficulties in establishing the translation equivalents with a Romance language like French, whose constructions are often quite different from those of the two Germanic target languages: *ett spann av 1600 meter* ('a span of 1,600 metres') – 'une arche longue de 1600 mètres'. In cases like this, the translation unit would be the whole phrase, impossible to translate without knowledge of the nature of the head noun – is the proper adjective high, tall, long? To conclude, the lexeme distinction of the function words is not always of help when it comes to

establishing translation equivalents. The syntagmatic information is often decisive for the establishment of the translation equivalents to the target languages, and has to be determined contrastively. The introduction of phrases as translation equivalents is necessary for the database to be of help for the human translator and for the functioning of the machine translation component.

Conclusion

It proved that the mere use of a frequency list based on newspaper material was not sufficient for deciding which content words to include in the Swedish core vocabulary. Since a more domain-neutral set of content words was desirable for the purpose in view, a comparison with novel corpus frequency lists was undertaken, and semantic classifications also proved necessary. For the function words, on the other hand, the segmentation of the frequency top of the newspaper material seems to be a sufficient procedure to establish which ones to be incorporated in the core vocabulary. The ambiguity, especially prevalent among the high-frequency function words, calls for the introduction of phrases into the core vocabulary, thus making possible the establishment of equivalence links between the translation units in Swedish and the target languages. As for the information to be associated with the entries in the database, much can be extracted from the definitions in *Svensk Ordbok* through a semantic parsing, especially as regards the content words.

Notes

1. A graphic word is defined as a segment delimited by blanks, punctuation marks or a line feed.
2. The novel is thus bigger than the newspaper corpus, something which makes a direct comparison between the figures impossible. Since the novel corpus is used only to balance, or neutralize, the material, mathematical accuracy is not of primary importance, and no exact calculations have been made.
3. Numerals, interjections, proper nouns and the infinitive marker have been excluded from the frequency study of the following reasons: there is only one infinitive marker, and it is obvious that is should be included in the core vocabulary. Proper nouns are domain-specific, and are thus generally disqualified for inclusion, possibly with the exception of the country names and nationality adjectives referring to the countries in which the database languages are the mother tongues. As for the interjections, *ja*, *nej* and *tack* ('yes', 'no' and 'thank you') are evident candidates for incorporations in the core vocabulary. The basic numerals must of course also be included.

References

Allén, S. et al. 1970. *Nusvensk frekvensordbok.* 1. Graford. Homografkomponenter. [Frequency Dictionary of Present-Day Swedish. 1. Graphic Words. Homograph Components]. Stockholm.

Allén, S. et al. 1975. *Nusvensk frekvensordbok.* 3. Ordförbindelser. [Frequency Dictionary of Present-Day Swedish. 3. Collocations]. Stockholm.

Gellerstam, M. 1989. The Language Bank. Department of Computational Linguistics. University of Gothenburg.

Hagman, J. 1991. Common and Odd Relations in Italian Dictionaries and Their Treatment in Taxonomy Building. ACQUILEX Working Paper, Istituto di Linguistica Computazionale del CNR, Università di Pisa.

Lehmann, H. 1991. Towards a Core Vocabulary for a Natural Language System. Proceedings of the 5th ACL Conference. Berlin.

Östling, A. A Proposal for a Swedish Core Vocabulary. Simple Units. Department of Linguistics, Uppsala University. Forthcoming.

Östling, A. & Wikholm, E. A Dictionary of Functional Core Phrases. Swedish, English, German and French. Department of Linguistics, Uppsala University. Forthcoming.

Sågvall Hein, A. 1987. Parsing by Means of Uppsala Chart Processor. (UCP). In: Bolc, L. (ed). *Natural Language Parsing Systems*. Springer Verlag.

Sågvall Hein, A. 1988. Towards a Comprehensive Swedish Parsing Dictionary. In: *Studies in Computer-Aided Lexicology*. Almqvist & Wiksell International. Stockholm.

Sågvall Hein, A. The LPS Inflectional Grammar. A Listing of the Rules. Department of Linguistics, Uppsala University. Forthcoming.

Sågvall Hein, A., Östling A. & Wikholm, E. 1990. Phrases in the Core Vocabulary. Center for Computational Linguistics. Uppsala University.

Sågvall Hein, A. & Sjögreen, C. 1991. Ett svenskt stamlexikon för datamaskinell morfologisk analys. En översikt. [A Swedish stem lexicon for computational morphological analysis. An overview]. In: Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning 18. Uppsala University.

Svensk Ordbok. 1986. [*A Dictionary of Swedish.*] Stockholm.

Wikholm, E. 1991. Übersetzungstheorie und maschinelle Übersetzung. Uppsala Universität. Linguistisches Institut.

Vossen, P., Meijs, W. & den Broeder, M. 1989. Meaning and structure in dictionary definitions. In: Boguraev, B. & Briscoe, T. (eds), Computational Lexicography for Natural Language Processing. Longman.

Annette Östling
Dept of Linguistics/Computational Linguistics
Uppsala University
Sweden

NORDISKE DATALINGVISTIKKDAGER

ved Universitetet i Bergen 28–30.11.1991

PROGRAM

Torsdag 28. november

Fra kl. 10.00 ***Registrering*** i Studentsenteret, Nygårdshøyden
11.45 – 12.45 *Lunsj* i "Grillen", Studentsenteret

13.00 – 13.15 *Konferansen åpnes* (i Auditorium A, Sydneshaugen skole)
13.15 – 13.55 Bente Maegaard: *Sprogteknologi*

Auditorium A	Auditorium D
14.00 – 14.40 Lars Ahrenberg [og Stefan Svedberg]: *Konceptuell textrepresentation för flersprdkig generering och översättning*	**14.00 – 14.40** Benny Brodda: *Generaliserad boolsk sökning i dokumentsökningesystem*
Kaffepause	
15.15 – 15.55 Bjørn Beskow: *Unifieringsbaserad transfer*	**15.15 – 15.55** [Yvonne Cederholm, Martin Gellerstam, Rudolf Rydstedt,] Christian Sjögreen: *Språkbankens lexikaliska databas*
16.00 – 16.40 Gudrun Magnusdottir: *Problems in the Perception of what a Machine Translation System is*	**16.00 – 16.40** Claus Huitfeldt: *Merking, presentasjon og analyse av komplekse tekstlige primærkilder*

18.15–20.00 Joseph Pentheroudakis: *What are the Limitations for Development of Machine Translation Systems?* (Auditorium A)

Fredag 29. november

09.00 – 09.40 Helge J Dyvik: *Linguistics and Machine Translation (i Auditorium A)*

09.45 – 10.25
Benny Brodda og Gunnar Eriksson: *Olika vägar til förbättrad träffmängd vid dokumentsökning*

10.30 – 11.10
Gregers Koch: *Discourse Representation Theory and Data Flow*

11.30 – 12.45 *Lunsj* i "Grillen", Studentsenteret

Auditorium A	Auditorium D
13.00 – 13.40 Jan Tore Lønning: *Computational Semantics – Lambda Terms or Feature Structures*	**13.00 – 13.40** Dieter Huber og Per Hedelin: *Ett svenskt uttalslexikon*
13.45 – 14.25 Margrethe H. Møller og Ellen Christoffersen: *Oversættelse af NP' er fra tysk til dansk*	**13.45 – 14.25** Gunnar Eriksson: *En homografseparator baserad på sannolikhet*
Kaffepause	
15.00 – 15.40 Jørgen Villadsen: *Anaphora and Intentionality in Classical Logic*	**15.00 – 15.40** Anna Sågvall Hein: *The Coverage of a Morphological Analyzer based on "Svensk ordbok"*
15.45 – 16.25 Annette Östling: *A Swedish Core Vocabulary for Machine Translation*	**15.45 – 16.25** Peter Molbæk Hansen og Ebbe Spang-Hanssen: *Syntaks og prosodi i et dansk tekst-til-talesystem*
Demonstrasjoner	

19.45 *Bankett i Schøtstuene – se eget program*

Lørdag 30. november

09.00 – 09.40 Magnar Brekke og Roald Skarsten: *Operational Machine Translation: Where Do We Meet the Wall?*	**09.00–09.40** Dieter Huber: *Integrating Syntagmatic Information in a Dictionary for Computer Speech Applications*
09.45 – 10.25 Eva Wikholm: *Översättningsteori och maskinöversättning*	**09.45 – 10.25** Anna Kalve Lysne: *Fonetikk på PC; hjelp eller belastning for filologen?*
Kaffepause	
11.00 – 11.40 Torbjørn Nordgård: *On Determinism and Ambiguity*	**11.00 – 11.40** Christian-Emil Ore: *En felles leksikografisk database for norsk*
11.45 – 12.25 Adams Bodomo: *A Unification Grammar of Serial Verb Constructions*	**11.45 – 12.25** Øystein Reigem: *RUTH – et konkordansbasert program for tekstkoding*

12.30 – 13.00 *Avslutning* (Auditorium A)