Machine Translation Strategies: A Comparison of F-Structure Transfer and Semantically Based Interlingua

Martha Thunes
Bergen

Abstract

Two machine translation (MT) systems which respectively utilize the transfer and interlingua strategies will be presented and compared, emphasizing design principles. Feature structures and unification-based grammar are common denominators for the two MT systems; in particular, both make use of Lexical-Functional Grammar (LFG). In the transfer system, Machine Translation Toolkit, developed by Executive Communication Systems, of Provo, Utah, transfer is based on LFG f-structure representations. In the interlingua system, PONS, constructed by Helge Dyvik, Department of Linguistics and Phonetics, University of Bergen, situation schemata representing the semantics of the source language text are employed as interlingua descriptions.

Introduction

The background for this paper is a study of these two MT systems where they are tested on English-to-Norwegian translation of technical text. The aim of the project is to find out to what extent the two different strategies, which have been employed in the systems, are able to maintain translational equivalence when put to the task of translating the same set of sentences. Since both applications are development environments for machine translation, and not ready made systems, the investigation will focus on potential for improvement and extendability, given the principles on which system design is based.

The notion of 'translational equivalence' denotes the relation that holds between source and target language expressions which are accepted as valid translations of each other. Translational equivalence is not an equivalence relation in formal terms: it is often the case that when translating between two given languages, translating a particular target expression back into the source language does not yield the original source expression as the optimal result.

The main difference between the strategies of transfer and interlingua can be described as follows: In transfer-based MT systems the translation process typically consists of three steps: analysis, transfer and generation. Analysis produces a source language dependent representation of input text. During transfer this is transformed into a target language dependent

Proceedings of NODALIDA 1993, pages 291-300

representation which is the basis for target text generation. In principle, language pair specific information is employed only during transfer. In an interlingua system source sentence analysis yields a representation of the input string which is, ideally, language neutral, or at least neutral between source and target language. Because it is language neutral it is referred to as an 'interlingua' representation. Target text generation can be based directly on the interlingua representation.

The MT systems presented here both draw on the framework of **Lexical Functional Grammar**, cf. Bresnan (1982). This is a generative, non-transformational, unification-based grammar formalism. Linguistic expressions are assigned two levels of syntactic representation (see fig. 1): constituent structure, or c-structure, describes hierarchical and linear ordering of syntactic constituents. C-structures are derived by phrase structure rules. In addition to c-structure, there is a functional structure, or f-structure, where grammatical functions are represented. Nodes in a c-structure are annotated with functional equations. Functional equations, together with functional information associated with lexical entries, relate c- and f-structure to each other. The relation between c- and f-structure is one of co-description rather than derivation: Partial descriptions of an f-structure become associated with c-structure nodes. The f-structure is not derived by performing operations on the c-structure.

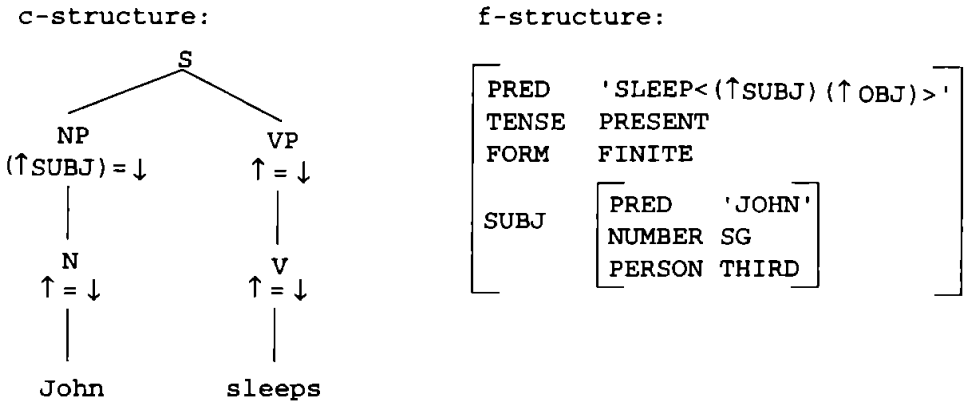

FIG 1 : A basic LFG representation of the sentence *John sleeps.*

A transfer system

Machine Translation Toolkit is a transfer-based MT system. Its grammars are designed in accordance with the LFG formalism. Lexical entries and grammar rules are coded as feature structures, or directed acyclic graphs (dags). A feature structure is a set of pairs of attributes and values. The f-structure representation of *John sleeps* in fig. 1 is an example of a feature structure. A linguistic representation language, LECS, has been developed

for the purpose of coding Toolkit language descriptions as feature structures. The structures that are built during the translation process are also represented as dags, coded in LECS. Information contained in the linguistic data base of the Toolkit system is mainly declarative, but there are also procedural elements in the linguistic descriptions. Firstly, monolingual lexical entries contain calls to structure-building operations that are employed during analysis and generation. Secondly, the bilingual transfer component consists of transfer entries, which contain translations as well as transfer rules. Transfer rules specify procedures, or dag-modifying functions, for transforming source sentence representations into corresponding target sentence representations. (1) is a sample transfer entry, written in LECS. In (1) the transfer rule named STD-TEN-P calls a function that substitutes the source language value of the attribute PFORM ('preposition, word form') with the value specified for PFORM in the corresponding target lexical entry.

(1) bilingual transfer entry mapping English *from* onto Norwegian *fra*:

```
en_from :: [ WORT  {  [ TECH  # GENERAL #
                FORM  "fra" ]  }
        \ STD-TEN-P ]
```

In the Toolkit system the analysis stage of the translation process outputs an f-structure representation of the source sentence, as illustrated in FIG. 2.

```
CAT     SBAR
FS      PRED        # SUBJ #
        FORM        burn
        VOICE       PASSIVE
        TENSE       PRES
        SUBJ        FORM        gas
                    PERSON      THIRD
                    NUMBER      SG
                    DEFINITE    PLUS
                    SPFORM      the
        ADJUNCTS    C 1         FORM        flare
                                PCASE       # SORC #
                                PFORM       from
                                QUANTIFIER  C 1   FORM      a
                                                  NUMBER    SG
                                                  DEFINITE  MINUS
```

FIG 2 : Toolkit, simplified source f-structure: *The gas is burned from a flare.*

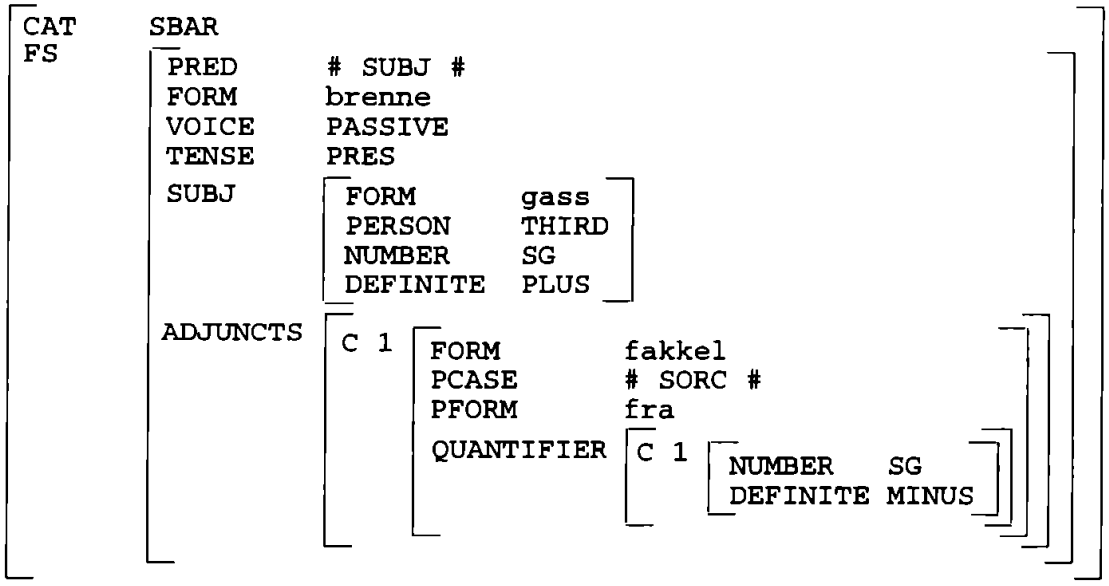

FIG 3 : Toolkit, simplified transfer dag:
The gas is burned from a flare. –> Gassen brennes fra en fakkel.

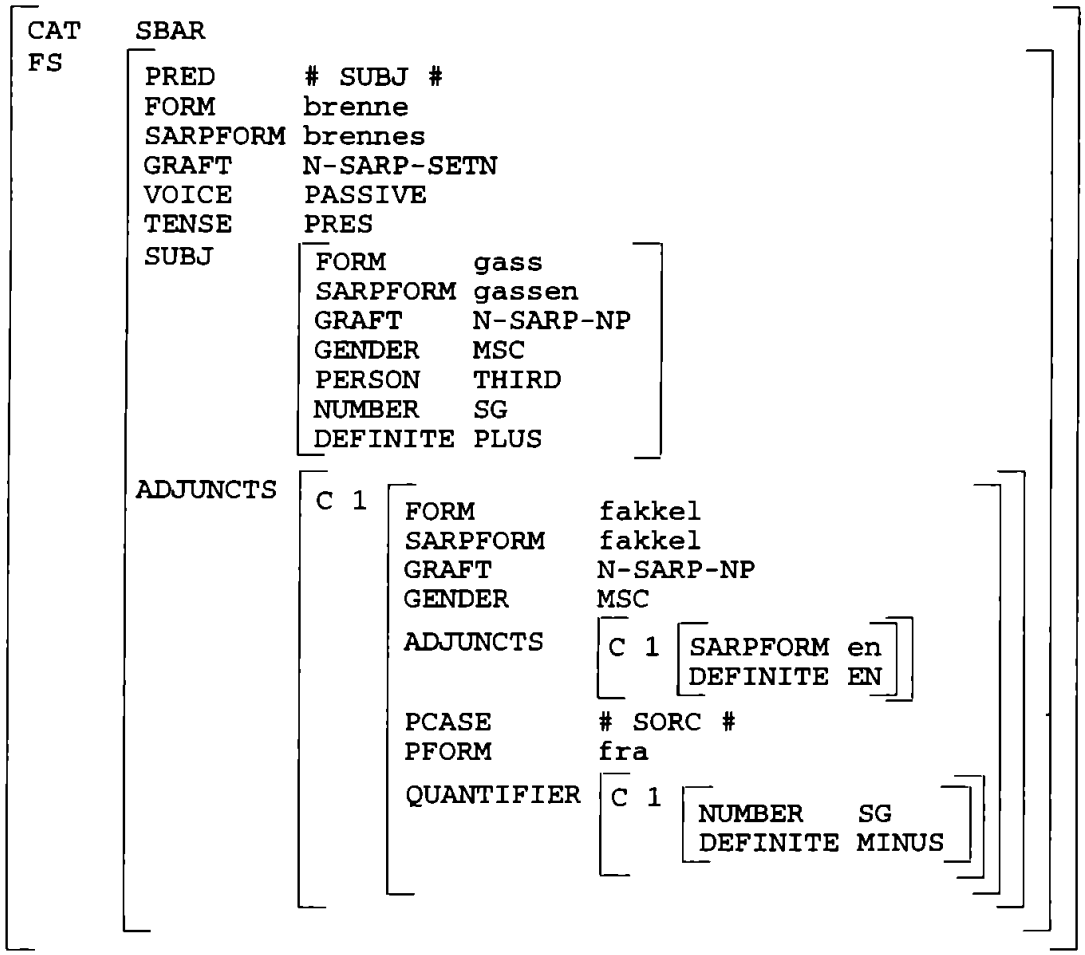

FIG 4 : Toolkit, simplified f-structure representation of target sentence:
Gassen brennes fra en fakkel.

In this particular source dag base forms of the words in the input sentence are given as values of the attributes FORM, SPFORM and PFORM. These values are pointers to a set of transfer entries (en_burn, en_flare, en_gas, en_from, en_a, en_the) which are processed during transfer. As a result transfer rules are executed, modifying the source dag into a transfer dag (fig. 3). In the transfer dag transfer rules have substituted English word forms with corresponding Norwegian forms. Also, transfer rules have deleted certain attribute-value pairs containing source language information which should not be carried over to generation. The target word forms in the transfer dag point to target lexical entries (nW_brenne, nW_gass, nW_fakkel, nW_fra). The information contained in these entries is added to the transfer dag, creating a target f-structure (fig. 4). The target dag contains inflected word forms which have been computed by applying morphological rules referred to in the target lexical entries. Lexical entries also point to syntactic rules, which build constituents. Syntactic constituent order is determined by functional ordering rules, which project grammatical functions onto syntactic constituents. Such rules are introduced either by monolingual lexical entries or transfer entries, and they apply only during generation. In the target dag they are referred to by the values of the attribute GRAFT.

An interlingua system

The PONS system is an experimental interlingua system for automatic translation of unrestricted text. 'PONS' is in Norwegian an acronym for "Partiell Oversettelse mellom Nærstående Språk" (Partial Translation between Closely Related Languages).

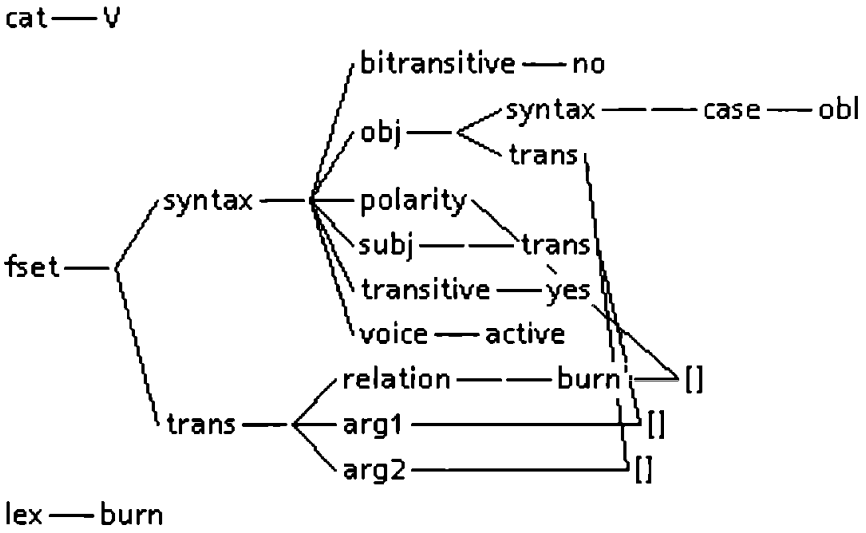

FIG 5 : PONS, simplified feature structure representing the word stem *burn*.

Translation is based on semantic analysis; however, a central principle is to exploit structural similarities between languages in cases where information about the syntactic structure of the source sentence can be used directly in target sentence generation. As a consequence of this, the PONS system has three different modes of operation: they vary with respect to the level of analysis at which translation is done. Interlingual translation is carried out only in the mode where translation is based on semantic representation. Linguistic descriptions in PONS are implemented in an extended version of D-PATR (Karttunen 1986). All grammatical and semantic information is coded as feature structures or directed graphs. The feature structure in fig. 5 is a graph representation of a sample lexical entry. All linguistic information in PONS is declarative; there are no procedures contained in the data base.

Before starting the translation process, different kinds of pointers are established between rules and word stems in source and target grammars. This is done automatically by a routine built into the system. The pointers describe a set of correspondences between representations of linguistic units in the two languages. These correspondences are exploited in cases where structural similarities between source and target language allow translation to be based on syntactic representation. The input sentence must be parsed before mode of translation can be chosen. Parsing yields one or more constituent trees. Attached to the topmost node in the tree is a feature structure representing the whole sentence; an example is given in fig. 6. Substructures of this structure are associated with individual nodes in the parse tree. A feature structure in PONS has essentially two components: *syntax* contains syntactic information, whereas *trans* is a semantic representation. Links between syntactic functions and semantic roles are expressed by giving shared values to specific attributes of the two substructures. E.g., *trans* of the syntactic subject is unified with *arg2* of the semantic relation *burn'*.

The parse tree also contains pointers to corresponding rules and word stems in the target grammar. The complexity of translation is automatically determined by the kinds of pointers that are contained in the parse tree. **Mode 1** performs word-for-word translation. It is necessary that the source and target stems express the same semantic relations and that the target pointers at each node show that source and target sentences are identical in syntactic structure. During translation terminal nodes in the parse tree are substituted with corresponding target word stems (fig. 7a). Inflected word forms must be found which are compatible with the feature structures associated with terminal nodes. However, if there are any word order differences between source and target expression, mode 1 will be insufficient, and **mode 2** may be employed. Mode 2 exploits correspondences between syntactic rules in source and target grammar. Differences in constituent order are allowed, but it is required that there is direct corres-

pondence between sense-carrying words (such as noun, verb, adjective) in source and target string. Fig. 7b) illustrates mode 2: During translation that subpart of the parse tree which represents the rule NP –> POSS N' is substituted with a subtree representing the target rule NP –> N' POSS.

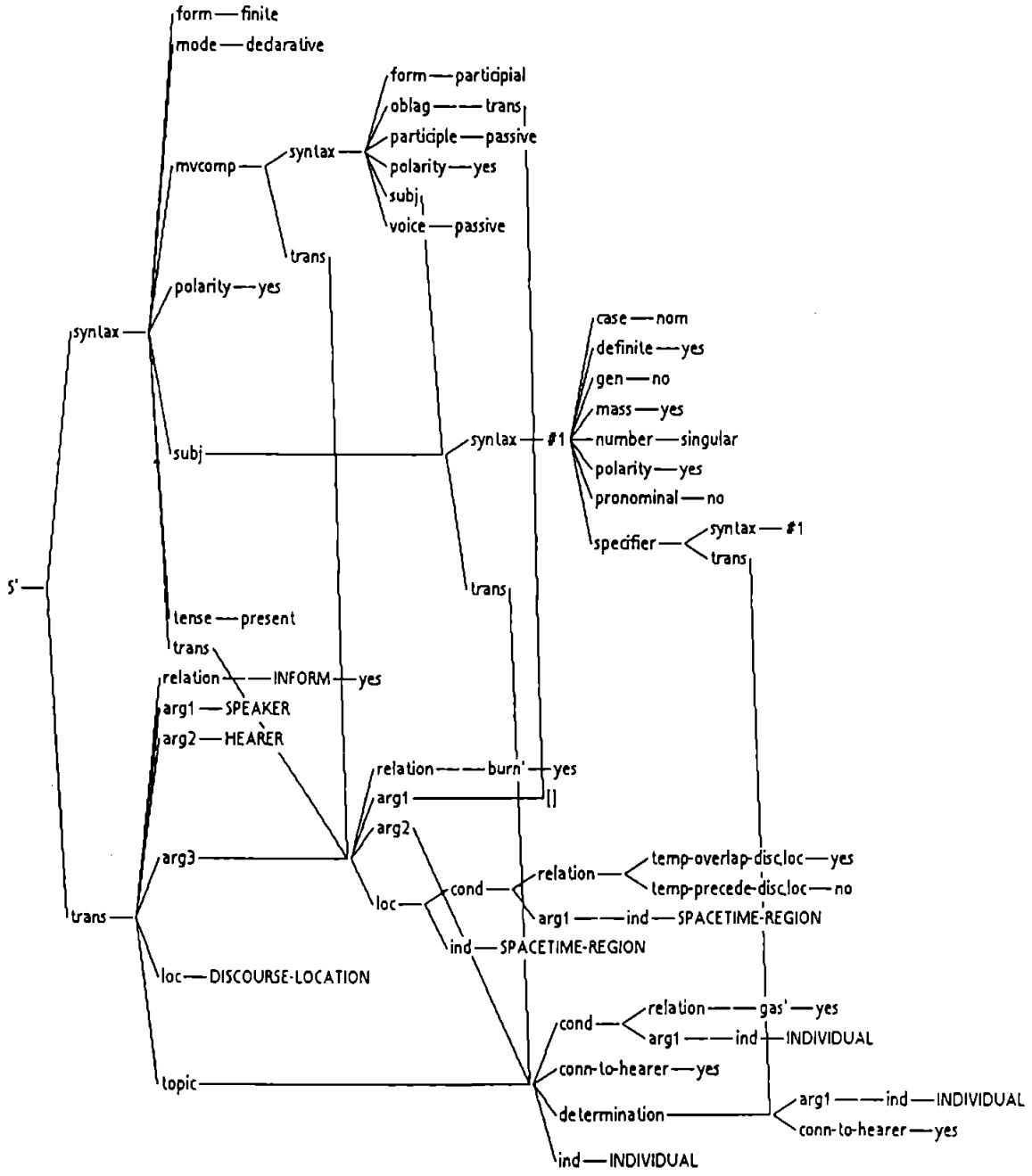

FIG 6 : PONS, simplified feature structure representing *The gas is burned.*

Next, terminal nodes are substituted with target word stems, and inflected word forms are found. **Mode 3** is used in all instances where 1 and 2 are insufficent. In mode 3 interlingual translation is carried out: the semantic representation of the source text functions as an interlingua expression. This representation is contained in the *trans*-part of the feature structure.

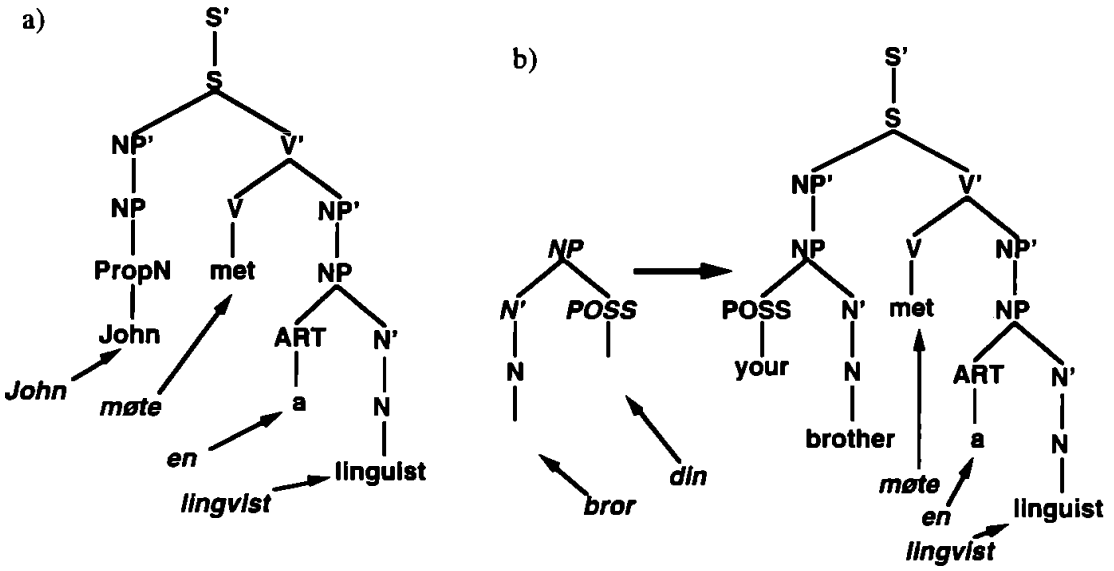

FIG 7 : a) PONS, mode 1: word-for-word correspondence:
John met a linguist. –> John møtte en lingvist.
b) PONS, mode 2: rule-to-rule correspondence:
Your brother met a linguist. –> Broren din møtte en lingvist.

The *trans*-structure is a situation schema: the notion of a situation schema has its origin in Situation Semantics (Barwise and Perry 1983, Fenstad et al. 1987) where situation schemata are used to represent the semantic relations contained in linguistic expressions. A situation schema consists of a set of attributes and values, where attributes designate types of roles in a fact and values refer to role fillers. A situation schema representing a sentence contains not only the propositional content, or the described situation, of that sentence. It contains also grammaticalized information about the utterance situation. To achieve translational equivalence the situation schema must include the information that is necessary to construct a target sentence that will express the same propositional content and have the same pragmatic function as the source sentence. To generate a target sentence from a situation schema the system must extract from the target grammar word stems and rules which express the semantic relations contained in the situation schema. Next, the full feature structures associated with these rules and stems are unified into the situation schema, extending this to a feature structure containing both syntactic and semantic information. To determine word order the syntactic rules of the target grammar are processed to build the constituent trees which are compatible with the feature structure.

The systems compared

In situation schemata in PONS linguistic meaning expressed by the source sentence is coded in attribute-value pairs neutral between source and target language. Translation via situation schema is based on the idea that two expressions from two different languages are translational equivalents if they are represented by the same situation schema. The situation schemata in PONS are declarative descriptions stating which expressions of source and target language that, at least according to the system, are translational equivalents. A situation schema is a representation neutral between analysis and synthesis, and also neutral with regard to direction of translation. Thus, the relation that holds between source and target expression is bidirectional and declarative.

Since PONS is a purely declarative system, the same syntactic rules in a grammar may be used for analysis as for generation. This is due to the fact that both analysis and generation are related to the same kind of representation, namely the feature structure where syntactic and semantic properties are interrelated, but contained in separate modules.

In PONS no language pair specific information is used in interlingual mode. Neither is any language specific information about how semantic relations are linked to syntactic functions contained in the situation schema. Accordingly, generation in mode 3 requires a fair amount of syntactic processing. To avoid inefficiency, grammars must be written with care, so that the generation algorithm does not build a number of trees representing different rules but identical strings.

As opposed to the situation schema in PONS, the transfer dag in Toolkit is language pair specific and dependent on the direction of translation. It follows from this that the transfer dag is not neutral between analysis and generation and may only be used for the purpose of generation. To generate a string from a transfer dag and to analyse a string to produce an f-structure cannot be reversible operations when execution of transfer rules transforms the source dag. The relation between source and target expression is unidirectional and irreversible.

As a consequence of the transfer strategy and the somewhat procedural character of the system, Toolkit needs separate rules for analysis and generation. Functional ordering rules specify how syntactic functions contained in the transfer dag are projected onto constituents of the target sentence. Moreover, a particular transfer entry specifies in what way semantic roles are linked to syntactic functions in the target language. Considerations of efficiency lies behind the use of separate rules for generation. Both transfer and generation rules are designed to keep the amount of work done during generation at a minimum. A result of this is

that it is not necessary to build parse trees during generation. It should, however, be mentioned that a subset of the rules found in the Toolkit system are in fact neutral between analysis and generation. But analysis rules as well as generation rules employ structure-building operations and are therefore of a procedural kind. It is a question whether it is easy enough to keep track of effects that result from applying and modifying the different kinds of rules in the Toolkit system. This pertains to analysis, transfer and generation rules.

Acknowledgements

The author is indebted to Helge Dyvik, Torbjørn Nordgård, Magnar Brekke and Roald Skarsten for their valuable help, information, criticism and encouragement during the preparation of this paper.

References

Barwise, J. and J. Perry. 1983. *Situations and Attitudes.* The MIT Press, Cambridge, Massachusetts.

Bresnan, J. (ed.). 1982. *The Mental Representation of Grammatical Relations.* The MIT Press, Cambridge, Massachusetts.

Dyvik, H. 1990. *The PONS Project: Features of a Translation System.* SKRIFTSERIE Nr. 39 SERIE B, University of Bergen, Department of Linguistics and Phonetics.

Executive Communication Systems, Inc. 1985. *Machine Translation Toolkit. Utilities User's Manual.* Provo, Utah.

Executive Communication Systems, Inc. 1985. *Machine Translation Toolkit. LECS User's Manual.* Provo, Utah.

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

Karttunen, L. 1986. *D-PATR – A Development Environment for Unification-Based Grammars.* REPORT No. CSLI-86-61, Center for the Study of Language and Information, Stanford University.