[From: *Human translation, machine translation*. Papers from 10th annual conference on computational linguistics, Odense, Denmark, 22-23 November 1979, ed. Suzanne Hanon and Viggo Hjørnager Pedersen.]

Structuring linguistic information for machine translation

The EUROTRA interface structure

By Bente Maegaard and Hanne Ruus

One of the basic principles within the EEC is that all important documents must be available in all six Community languages. The amount of documents to be translated is continually growing and the cost of the translation services rapidly increasing. To cope with this problem the Commission some years ago turned to machine translation for assistance. It purchased an American system, SYSTRAN, which translates between some Community language pairs (English-French and vice versa, and English-Italian). SYSTRAN is a system for pre-translation and its results are not quite satisfactory. This fact, combined with the ever growing demand for translations made the Commission arrange *a* meeting in Luxembourg in February 1978.

The participants in this meeting were researchers from universities and research centres in the field of computational linguistics and machine translation in the Community countries. As a result of this meeting a team was set up with the aim of investigating the possibilities of coordinating efforts within the field in Europe, using the existing European know-how and experience.

This coordination group has made a proposal for a EUROpean TRAnslation system, EUROTRA. This system is multilingual in its conception, and is meant to translate between the six present Community languages, while being extensible to other languages (Portuguese, Greek and Spanish will soon become Community languages). We (Hanne Ruus and Bente Maegaard) have participated for Denmark in the work on the project description since September 1978.

The other universities represented in the coordination group are Leuwen (Belgium), Manchester/Essex (England), Grenoble (France), Saarbrücken (Germany), Pisa (Italy) and Delft (Netherlands). Margaret King, England, is the chairperson of the group.

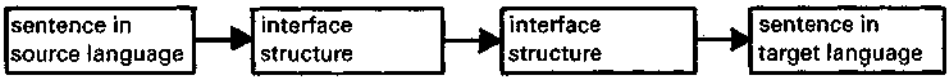
Apart from the multilinguistic aspects already mentioned, the most important characteristic of the project is that the system is going to be developed separately in the various Community countries: for each language there will be a group working in its own country.

The translation process is divided into three modules: analysis, transfer and generation, two of these, analysis and generation, being monolingual and consequently being developed entirely by one language group. The third module, transfer, is bilingual and must be developed jointly by pairs of language groups.

It seems evident that the members of the language groups developing the monolingual modules and cooperating with other groups on transfer, must be native speakers of the language they treat. It would also be possible to get native speakers of the Community languages to work in groups in a centralized organisation, e.g. in Luxembourg. The reason for emphasizing the decentralization of the system and thereby of the organisation is partly that computational linguistics and machine translation

centres and university institutes in all the Community countries will benefit from research carried out locally, and partly that all the specialists in the field can hardly be persuaded to move to Luxembourg.

The EUROTRA translation process can be illustrated in the following way:



analysis transfer generation

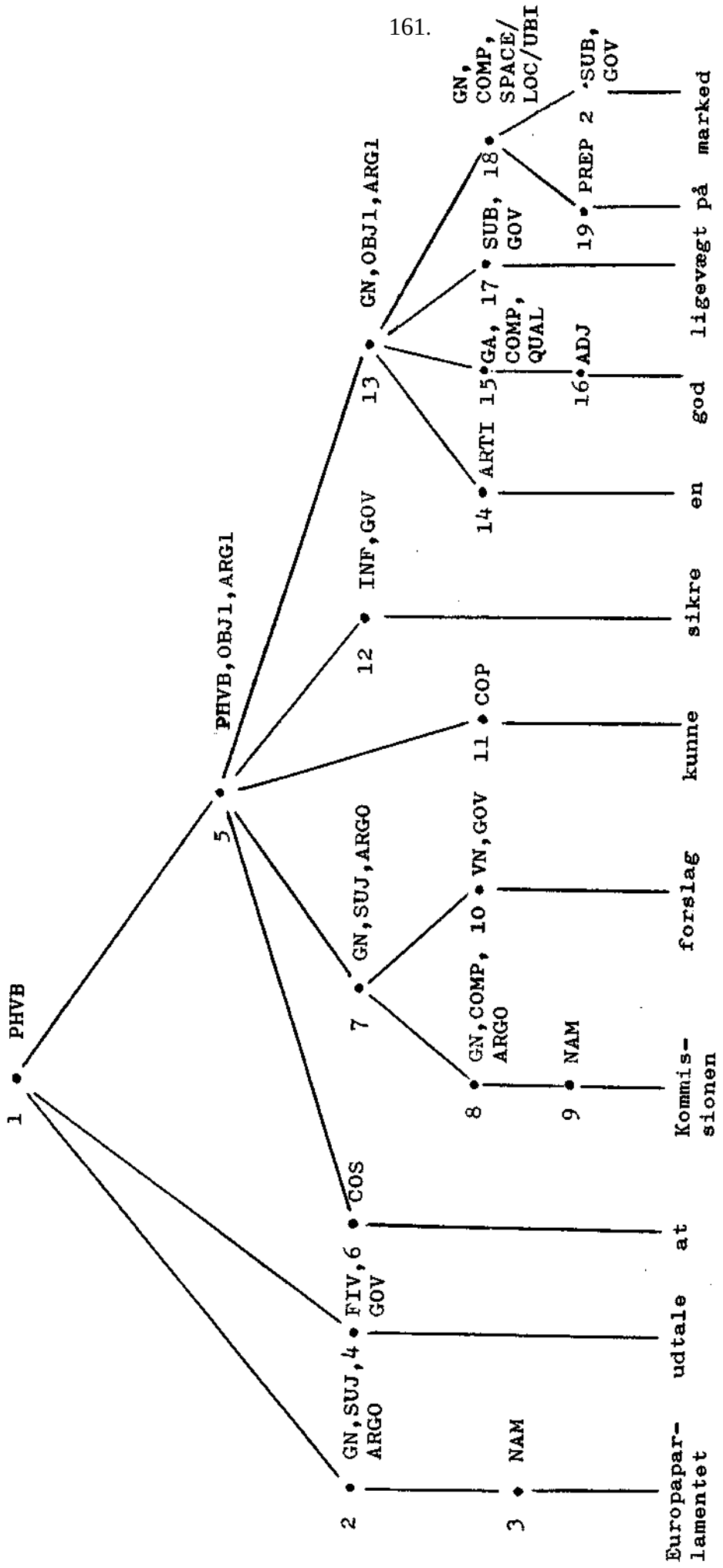
Fig. 1.

For an English text to be translated into Danish, the analysis module to be used will have been developed by the English group, the transfer module by the English and the Danish group jointly, and the generation module by the Danish group. The analysis module and the generation module are independent of the target and source languages, respectively, i.e. the same English analysis module will be used for translating into all the other languages. The transfer module on the contrary will be specific for each pair of languages, i.e. for 6 languages there will be 6 x 5 = 30 different transfer modules. For reasons of economy it is therefore essential that the transfer phase be restricted to an absolute minimum: transfer deals solely with problems the solution of which is necessarily based on bilingual information.

To benefit from results of research carried out in the European centres and universities experimenting with various grammar models and formalisms for computational analysis of natural language it is essential to allow free choice of grammar model and strategy within each module of the system.

To make sure that the work carried out by the different language groups can be conjoined into one working system the output from each module is strictly defined. Each language group has to produce interface structures of the kind described below i.e. dependency trees. This means that the result of the analysis must be described using the dependency formalism, which has proved to be efficient for this purpose. But the fact that the dependency formalism is used to express the result of the analysis does not impose any restrictions on the choice of grammar type or strategy. The limits within which such a choice is possible will be set by the common software, which is not yet fully specified, one of the basic principles in the design of the software being that it must allow for different analysis strategies and different types of grammar.

As it is emphasized in the papers of Johnson and Keil in this volume, the main obstacle in the history of machine translation is ambiguity. Most ambiguities can be resolved by a



Europaparlamentet udtalte at Kommissionens forslag kunne sikre en bedre ligevægt på markedet.

Fig. 2.

deeper linguistic analysis e.g. a morphological ambiguity such as Danish *taler* finite verb (Eng. *speaks)* or form of a noun (Eng. *speaker)* can be resolved in the syntactic analysis, similarly syntactic ambiguities can be resolved by looking at semantic information. For this reason it has been decided to put information from different levels of linguistic analysis into the EUROTRA interface structure.

The interface structures consist of dependency trees. In the trees the nodes are decorated with labels giving information at three different levels of analysis: the morpho-syntactic level, the syntactic function level and the logico-semantic level. These three levels correspond roughly to the three phases in the history of machine translation outlined in Rod Johnson's paper.

The labels at the morpho-syntactic level indicate formal properties of the constituents such as the word class (e.g. node 3 NAM, proper noun), the morphological class (e.g. FIV, finite verb, node 4), the formal characteristics of a complex unit such as PHVB, i.e. constituents containing a verbal core (e.g. node 5), or GN, noun group (e.g. node 13). The node numbers refer to the nodes of the tree in fig. 2. At the syntactic function level the labels indicate the syntactic function of the constituents such as subject SUJ, direct object OBJ1, indirect object OBJ2 and adverbials COMP.

As mentioned before, the interface structure is a dependency tree. This means that among a set of sister nodes one is seen as governing the others, this node is labeled GOV and the relations between the GOV-node and any of its sisters are indicated in the labels of the sister nodes. Any Gov-node directly dominates a leaf in the tree. The leaves contain references to the lexical units of the words in the input text.

As all deep level relations can be read off the labels, it is possible to retain the word order of the text in the order of the leaves of the tree. This feature is useful when translating between closely related languages, as the word order of the input text need only be changed, when the surface word order rules of the target language do not permit the word order of the source language.

Fig. 2 is the interface structure for the Danish sentence *Europaparlamentet udtalte, at Kommissionens forslag kunne sikre en bedre ligevægt på markedet* (Eng. *The European Parliament estimated that the Commission's proposal would ensure a sounder balance on the market).* The labels on the nodes are written in the following order: morpho-syntactic class, syntactic function, logico-semantic relation. The indication of morpho-syntactic class is obligatory.

The nodes 1,2,3,4 contain the analysis of *Europaparlamentet udtalte.* Node 2 dominates a noun group consisting of one proper noun (NAM, node 3), it is the subject of the sentence (SUJ) and the logico-semantic deep subject (ARG0) of the GOV-node (node 4), which has the morpho-syntactic label FIV, finite verb. Node 5 dominates a subordinate clause (PHVB), which is the direct object in the main clause (OBJ1) and is related to the GOV-node as ARG1, deep object.

Information about the inflected forms of the words in the sentence is stored in morphological variables for tense, number, person etc., these variables being attached to the appropriate nodes in the tree. So the information that the surface form *bedre* is the

comparative form of the lexical unit *god* is stored in a variable at node 16.

Node 8 shows the difference between the logico-semantic subject and the syntactic subject. *Kommissionens forslag* (Eng. *the proposal of the Commission)* is a nominalization of the sentence *Kommissionen foreslår* (Eng. *the Commission proposes),* the structure of which would be as follows

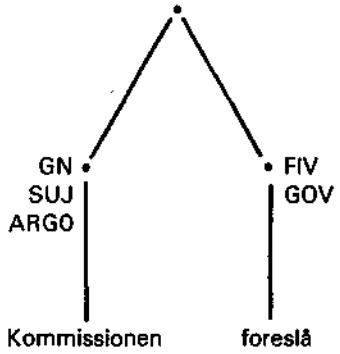


Fig. 3.

and therefore the structure of the noun group is

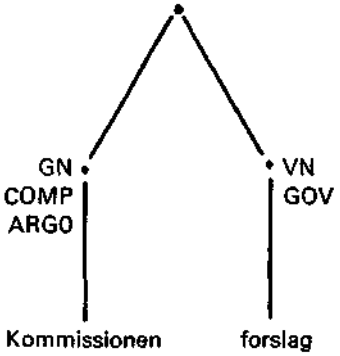


Fig. 4.

In the nominalization the logico-semantic relation is the same, but the syntactic function

changes.

To provide for the correct analysis of the noun group it is necessary to have access to the information that *forslag* is a verbal noun (VN) and to information about what kind of ARG0 it combines with.

The logico-semantic level does not use only labels of the form ARG0, ARG1, etc., but also labels that describe functions like the deep cases of Fillmore. Examples of this are the SPACE/LOC/UBI of node 18 and the QUAL of node 15. SPACE/LOC/UBI is used for an adverbial phrase which indicates a place where something takes place. In the system of logico-semantic labels there is a series of labels indicating spatial relations, and a parallel one indicating temporal relations. The label QUAL is used for adjectives as attributes, for relative clauses, etc.

The last particular comment on the labels in the tree is that the modal verb *kunne* is labeled as an auxiliary, COP (node 11). The choice of this label is not based on thorough investigations of modals and their behaviour in translation. This question will be considered in the future and is on the wait list for discussion in the coordination group.

A very important general characteristic of the interface structure is the flatness of the tree: the number of nodes and of levels is restricted. Some of the advantages of doing this are that one needs less storage capacity and gets shorter search paths in the tree. If we compare the basic structure of this tree to the equivalent structure of an ordinary IC (Immediate Constituents) or context-free grammar, we see that the dependency structure simply consists of three branches from the root and consequently

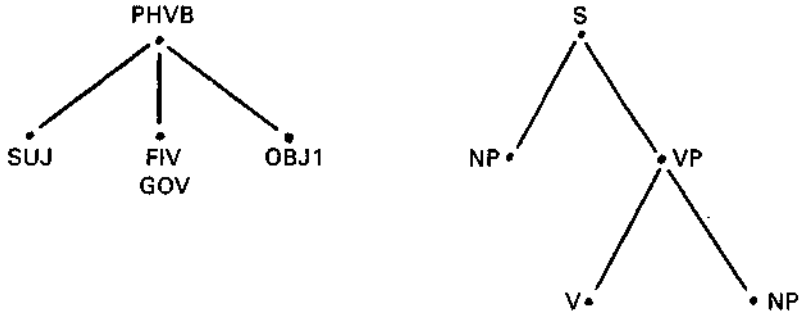
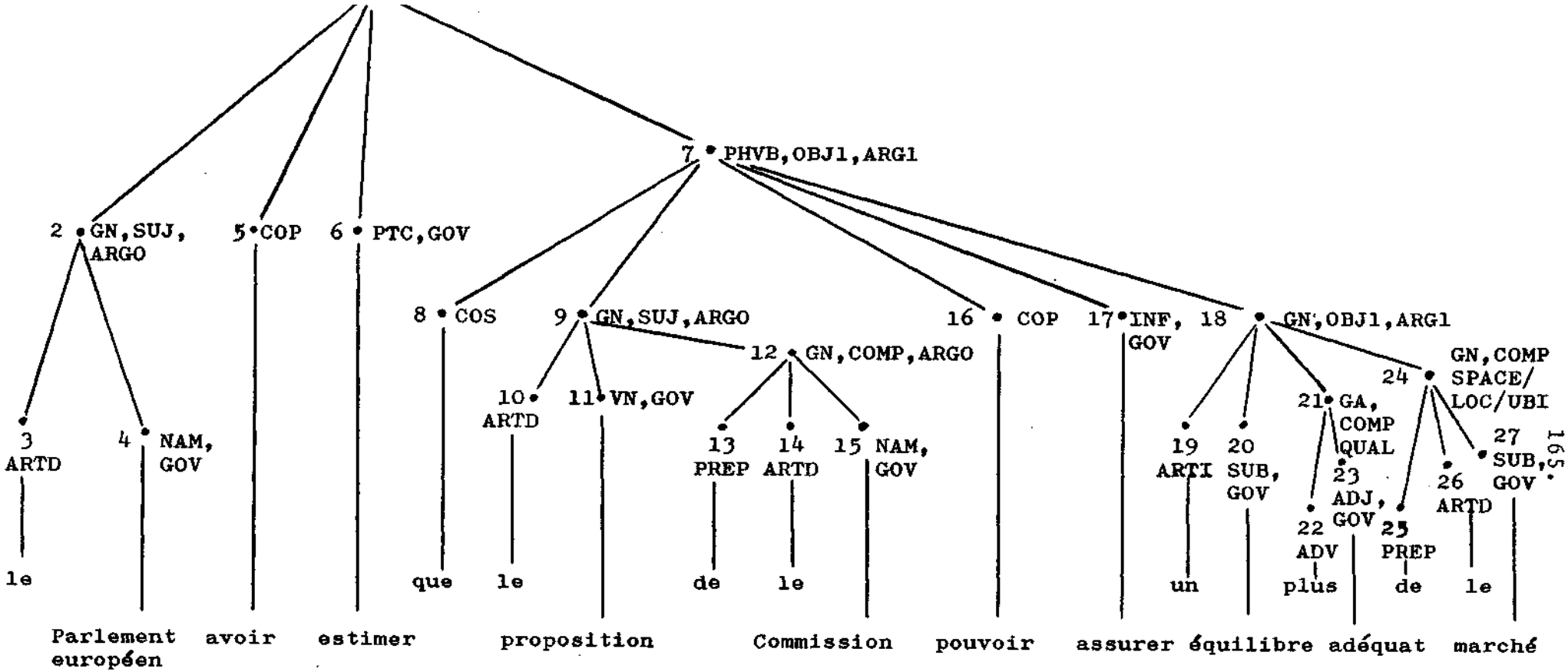


Fig. 5.

has got only two levels and four nodes, whereas the context-free tree has got three levels and five nodes. Although the flat tree is simpler than the other one it does not contain less information, - on the contrary. Another example of the adequacy of the flat trees is seen in the structure of the noun group (GN), *en bedre ligevægt* (the sub-tree



Le Parlement européen a estimé que les propositions de la Commission pouvaient assurer un équilibre plus adéquat du marché.

Fig. 6.

dominated by node 13).

Short of displaying acceptable translations produced by the system the best way to demonstrate the usefulness of the EUROTRA features described so far is by working through an example going from the interface structure of some source language into the corresponding sentence of a target language.

As Danish is our mother tongue we feel best qualified to predict features in the Danish generation module, so the example will outline the way from a French interface structure through transfer and generation into an equivalent Danish sentence.

Fig. 6 shows the interface structure for the French sentence *Le Parlement européen a estimé que les propositions de la Commission pouvaient assurer un équilibre plus adéquat du marché.*

The fact that the content of the sentence is fairly similar to that of the Danish sentence of fig. 2 should not lead to the conclusion that the translation will necessarily result in this Danish sentence.

The transfer phase consists of lexical transfer whereby the French lexical units on the leaves of the tree are substituted by appropriate Danish lexical units. It is obviously simple to interchange *Parlement européen* with *Europaparlamentet* on the leaf of node 4 and equally simple to substitute *ligevægt* for *équilibre* on the leaf of node 20. But as soon as a French lexical unit corresponds to several Danish lexical units, the correct equivalent must be found by looking at the information in the labels on the nodes. *estimer* of node 6 e.g. can be translated into Danish as *agte* or *mene* and *assurer* of node 17 can mean *forsikre* or *sikre.* In the case of *estimer* morpho-syntactic and syntactic function labels are used: *mene* is chosen because of the morpho-syntactic class PHVB of the OBJ1 governed by *estimer.* In the case of *assurer* the choice is based on a logico-semantic relation and a semantic feature in one of the lexical units of the relevant constituent. The translation *sikre* is chosen because the lexical unit of the governing node 11 of the ARG0 of *assurer* has the feature non-human in its semantic description.

After lexical transfer the French lexical units representing the content words of the sentence have been substituted by Danish lexical units as illustrated in fig. 7.

We shall now briefly touch on the generation phase and describe some of the mechanisms which are necessary to generate a Danish sentence from the interface structure resulting from the transfer phase. In the generation phase, the constituents of the sentences and of the phrases must be rearranged to get the correct surface order in the target language. Until this moment the interface structure has shown the structure of the source sentence and consequently the surface word order of the source sentence. Only now, at this very late stage in the process, it is necessary to make structural changes in the tree. In the present example the order of the nodes 19, 20, 21 must be changed so that *un équilibre plus adéquat* will map into *en mere passende ligevægt.*

The choice of verbal tense is also target language specific. The information necessary to make the correct choice will come partly from the grammar of the target language and partly from the labels in the interface structure. The interface structure resulting

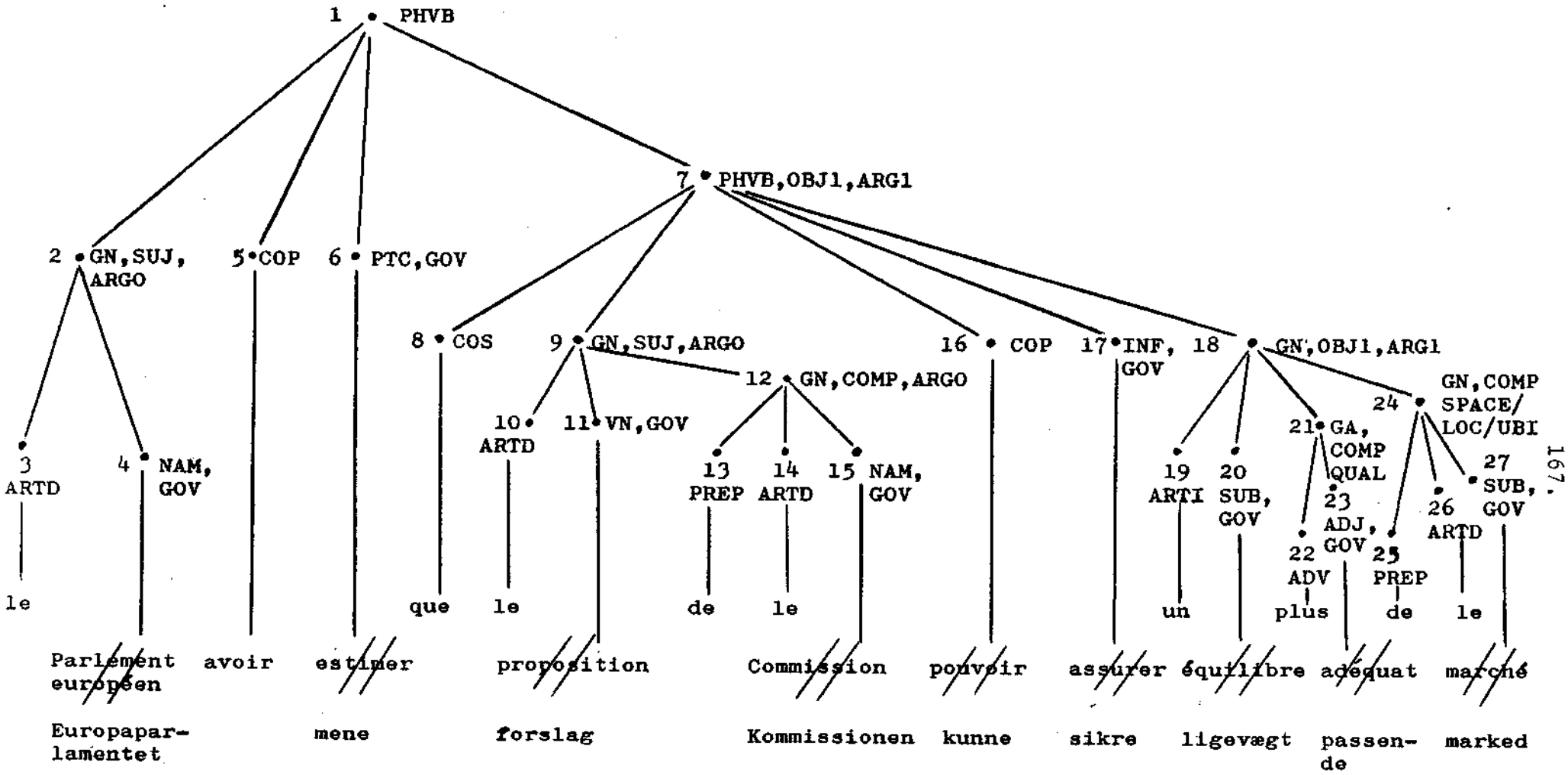

1 PHVB
2 GN,SUJ, ARGO
5 COP
6 PTC,GOV
7 PHVB,OBJ1,ARG1
3 ARTD
4 NAM, GOV
8 COS
9 GN,SUJ,ARGO
16 COP
17 INF, GOV
18 GN,OBJ1,ARG1
10 ARTD
11 VN,GOV
12 GN,COMP,ARGO
13 PREP
14 ARTD
15 NAM, GOV
19 ARTI
20 SUB, GOV
21 GA, COMP QUAL
22 ADV
23 ADJ, GOV
24 GN,COMP SPACE/ LOC/UBI
25 PREP
26 ARTD
27 SUB, GOV
le
Parlement européen
avoir
estimer
que
le
proposition
de
le
Commission
pouvoir
assurer
un
équilibre
plus
adéquat
de
le
marché
Europaparlamentet
mene
forslag
Kommissionen
kunne
sikre
ligevægt
passende
marked

Fig. 7.

from the French analysis must therefore provide information which enables the Danish generation module to choose the past tense (præteritum), *mente,* for the compound French verbal tense (passé composé), *a estimé,* in node 5 and 6.

Function words are normally not translated during transfer, as they are often dependent exclusively on the target language. In order to choose the correct word or expression, the information resulting from the linguistic analysis, as it is stored in the labels, is combined with information from the target language dictionary, e.g. information concerning the possible constructions of a verb. In the example considered, the French *que* in the leaf of node 8 should be interchanged with *at* because it introduces a subordinate clause which is the object (OBJ1) of the verb *mene.*

The preposition *de* of the noun group dominated by node 24, *du marché,* must be translated by *på.* This decision is made by combining the logico-semantic label SPACE/LOC/UBI with dictionary information concerning the lexical unit *marked.*

One of the last tasks to be carried out in the generation phase is the application of morphological rules: the correct inflected forms of nouns, adjectives etc. must be constructed according to the information in the labels (indicating definite/indefinite form, number, etc.), and furthermore the generation module must take care of concord within phrases and sentences.

As a more complicated example of morphological generation we can consider the nodes dominated by node 12 in the French tree, *de la Commission.* The expression *de la Commission* is structurally equivalent to the above-mentioned *du marché,* but whereas *du marché* was translated by *på markedet, de la Commission* must be translated by a genitive, *Kommissionens.* It is possible to make the correct decision because of the logico-semantic labels: some rule will state that a noun group (GN) which is ARG0 for a verbal noun (VN) is a genitive form in Danish.

By working through this example we have demonstrated that generation as well as transfer makes use of labels of all three levels. The result of the generation will be *Europaparlamentet mente, at Kommissionens forslag kunne sikre en mere passende ligevægt på markedet.*

The examples of interface structures have shown what kinds of linguistic information the system will be working on, but very little has been said about the sources of this information.

There will mainly be two sources of information: grammars and dictionaries. These two kinds of linguistic data will be kept separate from the processes or algorithms using them. The common software will provide tools for writing grammars and dictionaries and for expressing strategies of analysis. The software will facilitate the incorporation of existing grammars into the system. It will e.g. be relatively easy for us to reformulate the morphological analysis of Danish developed for DANwORD to make it conform to the EUROTRA conventions.

One of the main tasks of the Danish group will be to construct an analysis grammar, which will produce structures like fig. 2 for most Danish sentences, and to test it out.

Another main task will be constructing the generation grammar which will finally produce the sentence in Danish. Generation may be seen as the reverse of analysis, and many rules from the analysis grammar can be reversed for use in generation, but as we have seen generation raises problems differing from those of analysis, if only because the interface structure reflects the analysis of a sentence in a foreign language.

Other tasks will be constructing pairs of transfer modules from and into Danish in collaboration with other language groups, but the really heavy task will be the construction of the Danish dictionary with all the necessary information. The lexical units in the dictionary will have information attached about morphology, derivation, syntactic valency and semantic characteristics. The amount of information for any given lexical unit will to a certain extent depend on how ambiguous it is. Information about word frequencies like that compiled in the DANwORD project will also be stored in the dictionary, as it can be foreseen that even the fairly powerful linguistic tools envisaged in EUROTRA will not always be sufficient to choose the correct analysis for a given sentence. In such cases statistical information about the words in the sentence can play an important part in guiding the system to make an acceptable choice.

From our point of view the most interesting prospect of the EUROTRA project is the possibility of testing out different algorithmic grammars and hereby discovering to what extent the linguistic analysis required for acceptable machine translations of non-fiction texts can be formulated algorithmically. Moreover this project offers the opportunity of testing out the adequacy of different linguistic models for the description of Danish, hereby obtaining new insights into the structure and mechanisms of our mother tongue.

References.

Margaret King: A New Attempt at Machine Translation, in *Actes du Colloque de Saint Maximin, 1980,* IRIA.

Margaret King: EUROTRA: A European Machine Translation System, in *Lebende Sprachen* (forthcoming).

The DANwORD project is described in several issues of SAML, the periodical of the Institute of Applied and Mathematical Linguistics, University of Copenhagen.