BOEL VICTORIA BØGGILD-ANDERSEN

Valence Frames Used for Syntactic Disambiguation in the EUROTRA-DK Model

Abstract

The EEC Machine Translation Programme EUROTRA is a multilingual, transfer-based, module-structured machine translation project. The result of the analysis, the interface structure, is based on a dependency grammar combined with a frame theory. The valency frames, specified in the lexicon, enable the grammar to analyse or generate the sentences. If information about the syntactical structure of the slot fillers is added to the lexicon, certain erroneous analyses may be discarded exclusively on a syntactical basis, and complex transfer may in some cases be avoided. Where semantic and syntactical differences are related, problems of ambiguity may be solved as well. This will be exemplified, and the frame theory will be explained. The paper concentrates on the valency of verbs; according to the EUROTRA theory the verb is the governor of a sentence.

1 The EUROTRA Model

The structure of the system as a whole is as shown in figure 1.

Units at the different levels:

EBL: Signs, codes.

EMS: Morphemes.

ECS: Words, syntactical categories, phrasal categories.

> **STRUCTURE:** Constituent structure: indicates the natural sequence of the sentence constituents by means of syntactical categories and subcategories.

ERS: Syntactical functions (surface syntactical relations): Governor, subject, object etc., modifiers.

Fig.1. The modular structure of the EUROTRA system.

```
   ANALYSIS  MODULE:                     SYNHESIS MODULE:
   levels:                               levels:

   -------------------------------       -----------------------
  ! EBL: Normalized text          !     !         EBL           !
  !      (sign analysis)          !     !                       !
  !-----   -----------------------!     !------        ---------!
         T-RULES                               T-RULES

   ------   -----------------------      ------        ---------
  ! EMS: Morphological structure  !     !         EMS           !
  !      (morphological analysis) !     !                       !
  !-----   -----------------------!     !------        ---------!
         T-RULES                               T-RULES

   -----    -----------------------      ------        ----------
  ! ECS: Constituent structure    !     !         ECS           !
  !      (surface syntax)         !     !                       !
  !-----   -----------------------!     !------        ---------!
         T-RULES                               T-RULES

   -----    -----------------------      ------        ---------
  ! ERS: Relational structure     !     !         ERS           !
  !      (surface syntactical     !     !                       !
  !       relations)              !     !                       !
  !-----   -----------------------!     !------        ---------!
         T-RULES                               T-RULES

   -----    -----------------------      ------        ---------
  ! IS:  Interface structure      !     !         IS            !
  !     (deep syntax with some    !     !                       !
  !      semantic features)       !     !                       !
  !-----   -----------------------!     !------        ---------!
             !                                  ^
             !                                  !
   ---------V-----------------------------------!-----------
  !                                                         !
  !           T R A N S F E R M O D U L E S                 !
  !---------------------------------------------------------!
```

STRUCTURE: Dependence structure: canonic sequence. Indicates the surface syntactical relations, determines agreements and percolates features. Sequence of units: governor, subj, obj/obl_ag/attr_subj/ pobj/obl_go/obl_loc/compl, iobj(obj2)/pobj/compl/obl_go/obl_loc/ attr_obj, modifiers. (Abbreviations are explained in fig.4.; the source is the EUROTRA Reference Manual version 5.0, 1989. A later version was distributed after the symposion.)

IS: Deep syntactical relations; governors, arguments and modifiers with some semantic features.

STRUCTURE: Dependence structure: canonic sequence. Indicates the depth syntactical relations. Elevates certain simple entities by percolating them to the relevant mother node as feature bundles. Sequence: governor, argument1, argument2, argument3, argument4, modifiers.

Each level consists of two components, a lexicon and a grammar. The system identifies its units at a certain level by using the lexicon, where at the same time the information about the various lexical entries, necessary for the grammatical rules at each level to function, is drawn.

Fig.2. The structure of a level.

```
 -------------------------------------------------
!                         !                        !
!         LEXICON         !        GRAMMAR         !
! identification of units !  combination of units  !
!_________________________!________________________!
```

Between the levels the T-RULES ensure the correct transformation of the output from one level to the form necessary as input for the following level.

As it will appear, the sign analysis occurs at the EBL level, the morphological analysis at the EMS level and the syntactical analysis at the three following levels. Some semantical information is included in the IS level. We shall concentrate on the last two syntactical levels, the ERS and the IS levels, and we shall pose the question: What information must be present at the lexical entries at these levels for the grammar and thus the system to function? In order to determine this, we must look more closely at the task of these two levels in the total analysis process.

2 The ERS Theory

On many points the ERS theory is in agreement with the theory behind the f-structure in LFG (Lexical Functional Grammar). Our grammar at the ERS level is a dependence grammar. A dependence structure consists of a governing lexical unit (GOVERNOR) and (possibly) a number of sentence members (DEPENDENTS), which presuppose the presence of the GOVERNOR. We distinguish between two types of DEPENDENTS:

Fig.3. Dependents at the ERS level.

1: COMPLEMENTS, which fill out a place in the frame
of the governor, i.e. are frame bound or valency bound:
The GOVERNOR requires their presence.

2: MODIFIERS, which also presuppose the presence of the GOVERNOR but do not fill out a place in its frame, are not required by the GOVERNOR.

At ERS the complements of the main verb are constituted by the following syntactical functions, from which, however, the Danish implementation differs on certain points:

Fig.4. Verbal complements at the ERS level.

SUBJ.
Example: DA: PETER sover.
ENG: PETER sleeps

OBL_AG (the subject in passive clauses).
Example: DA: Suppe spises AF MANGE.
ENG: Soup is eaten BY A LOT OF PEOPLE.

OBJ (the indirect object, if this is not a sentence).
Example: DA: Peter spiser SUPPE.
ENG: Petes eats SOUP.

ATTR_SUBJ (the subject complement if this is not a sentence).
Example: DA: Problemet er ULOESELIGT.
ENG: The problem is INSOLUBLE.

ATTR_OBJ (the object complement if this is not a sentence).
Example: DA: Det kalder jeg EN OVERDRIVELSE
ENG: That's what I call AN OVERSTATEMENT.

POBJ (frame bound prepositional phrase i.e. prepositional phrase governed by the main verb and requiring a particular preposition).
Example: DA: Jeg haaber PAA EN FORANDRING.
ENG: I am hoping FOR A CHANGE.

OBL_LOC (obligatory prepositional phrase denoting place).
Example: DA: Peter bor I SLAGELSE.
ENG: Peter lives IN MANCHESTER.

```
OBL_GO     (obligatory prepositional phrase
denoting direction).
Example: DA:  Peter tog TIL PARIS.
         ENG: Peter went TO PARIS.

COMP (clause-shaped complement).
Example: DA:  Peter har lovet, AT HAN NOK SKAL KOMME.
         DA:  Peter hoerte HANS KOMME.
         ENG: Peter has promised, THAT HE WILL BE
              THERE.
         ENG: Peter heard HANS COMING.

OBJ2 (the direct object, if this is not a
sentence).
Example: DA:  Jeg skylder HAM en tjeneste.
         ENG: I owe HIM a favour.
```

For the sake of clarity I have chosen sentences which do not bear much resemblance to the constructions that we usually work with in texts from the Commission.

From LFG we have also taken over the so-called ”Principle of Completeness and Coherence“, which can be formulated as follows:

> A structure must contain ALL the complements required by its governor AND NO OTHERS.

From this principle follows:

> Complements can only be described as obligatory. Empty elements (empty nodes) must be inserted in a number of cases, for instance in passive clauses where the logical subject (the agent) is not present, and in infinite constructions without an explicit subject directly connected to the infinite verb form.

The ERS guidelines are of a directive character, and not obligatory; in the Danish implementation we have differed from them for example by not including the category COMP, which, as we saw it, was defined not by its syntactical function, but exclusively by its being clause-shaped. In the Danish implementation we have simply allowed that subjects as well as objects may also consist of substantival sub-clauses or infinitives. This simplifies the mapping between the ERS and IS levels, and besides it agrees more with our linguistic intuition, also because precisely the same applies in Danish to words and phrases governed by a preposition. (for example in a so-called POBJ); in this position we may also find both nouns, noun clauses and infinitives.

3 The IS Theory

The next level, the IS level, is, however, actually legislative. It is here that the decorated tree structures, which constitute the starting point of the synthesis AFTER the transfer process, are formulated. Since, as it is well known, EUROTRA is a multi-lingual translation system, it is necessary to model the IS level in such a way that, using a common feature theory, it describes the linguistic features that are relevant for translation purposes and which are common to the languages in question, in a way that is compatible with all nine EEC languages. To formulate this theory and the common feature theory is a difficult task — some might even say an impossible one — but it is also a challenge, because no theories existed that might be transferred to a multi-lingual, transfer-based machine translation system. The IS theory has been formulated in a cooperation between linguists in EUROTRA with constant feedback from the various language groups, and it rests on the following main principles:

1. IS is primarily a syntactical theory.

2. The starting-point of the description is English, which functions as a kind of meta-language.

3. The IS theory consists of a dependence grammar with a sunken governor, combined with a frame theory.

4. The theory must satisfy the following requirements:

    (a) The description must be adequate; it must, as far as possible, disambiguate polysemantic surface structures.

    (b) The description must be calculable; it must be formalized so as to permit a computer to calculate the relevant phenomena.

As mentioned above, the theory must be able to describe the linguistic features, relevant for translation purposes and common to the nine EEC languages; hence all non-significant differences are neutralized. This applies to the individual language (for example the difference between the active and the passive voice) as well as to differences between the EEC languages. An example of a difference specific to a particular language is the difference between noun clause types (infinitive constructions or that sentences). If one wants to specify this difference in the analysis, it must be done on the underlying level, the ERS level. It is a monolingual matter, whether a verb requires finite or non-finite clausal complements. Some verbs do not take clausal complements at all, others take special types, and some (the support verbs) require deverbal nouns as objects, while the equivalents in other languages may not have the same restrictions. I shall show an example indicating these differences in the next section.

The neutralization of differences, specific to a particular language, is done in the following way:

> As mentioned earlier, already at the ERS level a "euroversal", canonic sequence of sentence members is determined. Thus word orders, specific to a particular language, are neutralized. At the IS level this sequence describes a small, defined, number of depth syntactical relations between the members.
>
> Certain sub-systems (tense, aspect, modality, etc.) are removed from the actual structural representation and re-coded, attributed to the overall sentence complex by calculation.
>
> Certain surface phenomena are removed from the tree structure and are represented instead in the overall sentence complex as features, if they are relevant for the translation.

As has been mentioned, the IS structure is also a dependence structure, including a governor and two types of dependence relations. These are:

> ARGUMENTS, of which a maximum of four may occur. An argument number may only be indicated, if the preceding one also occurs in the frame of the governor
>
> MODIFIERS, which do not occur in the frame of the governor.
>
> Thus, the maximum completion of a sentence, in which the main verb is always regarded as the governor, is as follows (the Kleene star indicates zero, one, or more occurences of the subsequent member):
> S = GOV, ARG1, ARG2, ARG3, ARG4, *MOD.

The relation between the complements of the ERS level and the arguments of the IS level can be schematically described as follows:

```
SUBJ ------------- ARG1
OBL_AG ----------- ARG1

OBJ -------------- ARG2
ATTR_SUBJ -------- ARG2

COMP ------------- ARG2 eller ARG3
POBJ ------------- ARG2 eller ARG3
OBL_GO ----------- ARG2 eller ARG3
OBL_LOC ---------- ARG2 eller ARG3

ATTR_OBJ --------- ARG3
OBJ2 ------------- ARG3

Frame-bound  arguments (PP's)  not  otherwise
indexed ---------- ARG4
```

So, both at the ERS level and at the IS level it is necessary to specify the valence structure of the lexical units in the level specific lexicons. And where can information about the valence of Danish words be obtained? Generally speaking, possibilities are few; no actual valence dictionaries for Danish exist, as they do for German for example. The Danish EUROTRA group has to work out these dictionaries themselves.

4 The Coding of Verbs in the Lexicon

In the Danish implementation we indicate the valence of the words at the ERS level already in the lexicon for the ECS level. We shall see how this can be done in the case of the verbs, taking a concrete example.

The Danish verb BEMAERKE (notice, remark) is mono-transitive. the lexical entry to this verb may for instance look like this in the ECS dictionary:

```
'bemaerke_v1' = cat=v, ers_frame=f20, ctrl=no, dalu='bemaerke',
reflex=no, vfeat=nstat, auxlu='have', t=no, term=xx0.
```

The formula ers_frame=f20 refers to the sentence rule that describes sentences with mono-transitive verbs. There are, however, different meanings of this verb. If we follow the definitions in Nudansk Ordbog, which, in Denmark, comes closest to the medium-sized, monolingual dictionary used for the project, we can make a division into these entries, where only the definitions differ. The examples are coded in the format used for the Lemma dictionary, where information necessary for the different levels are gathered under the relevant entries:

```
'bemaerke_v1' = cat=v, scat=mainv, level=zero, dalu='bemaerke',
darno=v1, ers_frame=f20, dapform1=no, dapform2=no, dapform3=no,
dapform4=no, daisframe=arg12, reflex=no, daparg1=no, daparg2=no,
daparg3=no, daparg4=no, auxlu='have', vfeat=nstat, flex_type=fx1,
dcons=no, oc=yes, infl=root, term=xx0.
%% Coder: boel  16-Jun-89
%% Source: experiment
%% DEF: iagttage, laegge maerke til
%% Comments:
%% Examples:Ingen bemaerkede hans fravaerelse. NDO.
```

```
'bemaerke_v2' = cat=v, scat=mainv, level=zero, dalu='bemaerke',
darno=v2, ers_frame=f20, dapform1=no, dapform2=no, dapform3=no,
dapform4=no, daisframe=arg12, reflex=no, daparg1=no, daparg2=no,
daparg3=no, daparg4=no, auxlu='have', vfeat=nstat, flex_type=fx1,
dcons=no, oc=yes, infl=root, term=xx0.
%% Coder: boel  16-Jun-89
%% Source: experiment
%% DEF: udtale, ytre
%% Comments: Ambiguous example in the NDO.
%% Examples: Han bemaerkede, at han var forhindret. NDO
```

Here, we use f20 in both cases: The verb only takes an object as a complement. Dalu means Danish lexical unit, darno refers to Danish reading number. DEF means definition. Information preceded by %% is not relevant for the grammar.

According to this, there must be two different entries in the lexicon, where the coding is identical, but the definitions differ. Hence, an analysis of the sentence:

DA: Kommissionen har bemaerket en rimelig udvikling inden for erhvervslivet.

ENG: The Commision has noticed a reasonable development in industry. (bemaerke=notice, sense 1, bemaerke_v1 above)

produces two identical results at the ERS level, both of this form:

Fig.5. ERS object with no object differentiation.

```
                          cat=s
                            !
  ------------------------------------------
 cat=v           cat=np                 cat=np
 dalu=bemaerke   sf=subj                sf=obj
 sf=gov             !                       !
                    !                       !
                 cat=n           -------------------------
            dalu=Kommissionen   cat=n    cat=pp     cat=ap
                 sf=gov         sf=gov   sf=pobj    sf=mod
                           dalu=udvikling     !       !
                                              !       !
                             -------------------   cat=adj
                         cat=p           cat=np    sf=gov
                    dalu=inden_for       sf=compl  dalu=
                         sf=gov               !    rimelig
                                              !
                                          cat=n
                                  dalu=erhvervsliv
                                          sf=gov
```

If, however, we supplement our lexical entries with the information that BEMAERKE in sense 1 may have certain types of sentence objects (an NP, an at-clause or an interrogative clause), while in sense 2 (bemaerke_v2 above) the word only takes at-clauses or pronouns as the object, the entries will look like this:

```
'bemaerke_v1' = cat=v, scat=mainv, level=zero, dalu='bemaerke',
darno=v1, ers_frame=f244, dapform1=no, dapform2=no, dapform3=no,
dapform4=no, daisframe=arg12, reflex=no, daparg1=no, daparg2=no,
daparg3=no, daparg4=no, auxlu='have', vfeat=nstat, flex_type=fx1,
dcons=no, oc=yes, infl=root, term=xx0.
```

```
%% Coder: boel  16-Jun-89
%% Source: experiment
%% DEF: iagttage, laegge maerke til
%% Comments:
%% Examples:Ingen bemaerkede hans fravaerelse. NDO.
```

f244: The verb only takes an object as complement. The object is: an NP OR a nominal at-clause OR an interrogative clause.

```
'bemaerke_v2' = cat=v, scat=mainv, level=zero, dalu='bemaerke',
darno=v2, ers_frame=f262, dapform1=no, dapform2=no, dapform3=no,
dapform4=no, daisframe=arg12, reflex=no, daparg1=no, daparg2=no,
daparg3=no, daparg4=no, auxlu='have', vfeat=nstat, flex_type=fx1,
dcons=no, oc=yes, infl=root, term=xx0.
%% Coder: boel  16-Jun-89
%% Source: experiment
%% DEF: udtale, ytre
%% Comments: Ambiguous example in the NDO.
%% Examples: Han bemaerkede, at han var forhindret. NDO
```

f262: The verb only takes an object as complement. The object is: a nominal at-clause OR a pronoun.

Furthermore, we change and sub-divide our grammar rules on the basis of this information. As a result, only one analysis, using the first entry of BEMAERKE, is possible, and the number of analyses are reduced. We are thus able to discard certain erroneous analyses exclusively on a syntactical basis, because it turns out, that semantic and syntactical differences may be connected. And this must happen at the ERS level, this being the level where a distinction is still made between different object types and sub-clause types. At the IS level this specific distinction is neutralized.

The following transfer rules will ensure the correct translation:

```
1: cat=v, dalu=bemaerke, darno=v1 =>
   cat=v, enlu=notice, enrno=v1.

2: cat=v, dalu=bemaerke, darno=v2 =>
   cat=v, enlu=remark, enrno=v2.
```

5 Conclusion

As claimed above, problems of semantic differences may in some cases be related to syntactical differences. In these cases, more systematic use of the syntactical information may solve some, if certainly not all, semantic problems of machine translation.

We have also made it theoretically possible to GENERATE or produce the correct sentence structure for a sentence translated into Danish, precisely by including information about sentence structure in the lexicon, contained in the more specific indication of valence. What forms the basis of the translation into Danish is a tree structure, where the individual words make up the leaves on the tree. In the transfer process, the words of the source language are exchanged with the equivalent words of the target language. In case of sentential complements, finite or non-finite, the tree does not specify which sentence type makes up the object of the sentence. The Danish lexicon will contain the information about the syntactical combinations of the different verbal entries, which make possible the establishment of a more specific sentence structure. In these cases we avoid having to work out word-specific rules of complex transfer; the problem can be solved monolingually in a more general way.

The fact remains that we lack dictionaries in Danish that allow us to draw information about the combination possibilities of Danish words to an extent that suits our purpose. One of the many tasks that the Danish EUROTRA group faces is to produce such dictionaries. I hope to get the possibility to experiment with drawing this information from the Gyldendal dictionary: "Dansk Sprogbrug“ by Erik Bruun. Here we find examples of the use of Danish words and a typology comparable to a rough valency description. We shall have to complete this information and transform it to a formalism suited for this special purpose.

References

Bresnan, Joan (ed): *The Mental Representation of Grammatical Relations.* MIT press. Cambridge and London 1982.

Boeggild-Andersen, Boel Victoria: *Forslag til udvidelse af verbalkodningen i den monolingvale ordbog.* Intern Rapport. 15. sept. 1988. EUROTRA-DK, Copenhagen.

Boeggild-Andersen, Boel Victoria: *Diderichsens saetningsskema anvendt ved 'parsing' i EUROTRA.* Lecture from the Danish EUROTRA seminar at the University of Copenhagen 1988. In press.

Boeggild-Andersen, Boel Victoria: *Verbale komplementer og argumenter i EUROTRA-teorien og deres adskillelse fra modificerende led.* Lecture from the Danish EUROTRA seminar at the University of Copenhagen 1989.

Boeggild-Andersen, Boel, Hanne Fersoe, Lise D. Johansen and Patrizia Paggio: *Sentential complements and non-finite clauses.* PO-22 report, EUROTRA 1989.

Bruun, Erik: *Dansk Sprogbrug.* Gyldendal. Copenhagen 1978.

Nudansk ordbog. Politikens forlag. Copenhagen 1989.

The EUROTRA reference manual 5.0, draft version. 1989.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik: *A Grammar of Contemporary English.* Longman. London 1972, 1986.