ASCOF -- A Modular Multilevel System for French-German Translation

Axel Biewer, Christian Féneyrol, Johannes Ritzke, Erwin Stegentritt

Universität des Saarlandes
D-6600 Saarbrücken
West Germany

This paper is an overview of ASCOF, a modular multilevel system for French-German translation. In ASCOF, the classical divisions of the translation process (analysis, transfer, synthesis) have been adopted.

The analysis is realized by three phases: (1) the morphological analysis, (2) the identification of non-complex syntactic phrases and the macrostructure of the sentence, and (3) the determination of the structure of complex syntactic phrases and the syntactic functions, in which syntactic and semantic criteria are used. Semantic criteria are stored in a semantic network. The syntax-oriented parts of the system interact with this semantic network during the identification of the syntactic functions. The lexical transfer operates on the standardized output tree of the analysis. The structural transfer and the syntactic synthesis are achieved by transformational grammars; the morphological synthesis, at least, generates the word form of the target language (German).

1 Project History and Status

ASCOF (Analysis and Synthesis of French by means of COMSKEE) is a computer system for the processing of natural language with the purpose of translating written French texts into German texts. This system has been under development since 1981 at the University of the Saarland at Saarbrücken, West Germany (Project C of SFB 100). At present (1984-1985), the research team, which has drawn upon the experience and findings of previous studies, consists of six members.

[SFB 100 is a research group in which different linguistic and computer science-oriented projects cooperate. The SFB 100 was founded in 1973 and is financed by the DFG (German Research Foundation). Other projects of the SFB 100 have been developing the programming language COMSKEE and the systems SUSY and SUSY II. ASCOF is an independent system especially conceived for French and German. General descriptions of ASCOF are given in Féneyrol, Ritzke, and Stegentritt (1984) and Stegentritt (1983). Detailed descriptions of the various problems and their solutions are discussed in Féneyrol and Stegentritt (1982) and in Ritzke (1982).]

2 Application Environment

The system is programmed in COMSKEE (Computing and String Keeping Language; cf. Mueller-von Brochowski et al. (1981), Messerschmidt (1984).) For the computer scientist, COMSKEE is a procedural (imperative) format-free, block-oriented programming language such as ALGOL and Pascal, yet comprising some of the qualities of functional languages (such as LISP or PROLOG).

For the linguist, COMSKEE is a powerful device especially due to its dynamic data types – dictionary, set, sentence, and string – and its dynamic operations – such as positional and contextual substring access and assignment.

The system runs on a SIEMENS 7561 under the system BS 2000. ASCOF has been conceived as a completely automatic translation system. As yet, we have been less concerned with end-user application than with fundamental research. For this reason, we have focused primarily upon linguistic and computer science problems, rather than upon processing speed and the like.

0362-613X/85/020137-154$03.00

3 General Translation Approach

3.1

In ASCOF the "classical" divisions (cf. Vauquois 1975) have been adopted: analysis, transfer, and synthesis. The result of the sentence analysis is represented as a standardized tree structure, which then serves as input for the transfer and synthesis of the target language.

3.2

The ASCOF analysis takes place in three steps based on different grammar and algorithm types. The morphological analysis PHASE I is carried out by an algorithm that realizes actually a mere pattern matching; in PHASE II context-free grammars identify non-complex syntactic phrases and the macrostructure of the sentence. A reduction in the homographies of word classes is simultaneously achieved for the complete sentence. PHASE III determines the syntactic functions within the sentence, using syntactic and semantic criteria, and carries out the semantic disambiguation of lexemes. This phase of analysis is performed by algorithms similar to ATN, representing an interactive system. [The term **interactive system** might be problematic in this context as this term often denotes components interacting with the user. Here we are concerned with the process communication between different components.] Consequently, the ASCOF analysis does not constitute a one-pass parser but a system of parsers (cf. **Figure 1a**). The strategy applied resembles that of cascaded ATNs (Woods 1980) and was chosen for the following reason: the complexity and length of the sentences to be analyzed require – for reasons of efficiency – parsing strategies appropriate to the different problems, that is, context-free grammars for PHASE II, which works exclusively with syntactic information, and formalisms similar to ATN for PHASE III, where syntactic and semantic information is combined. A similar combination of syntax and semantics often occurs in modern parsers of various orientations, e.g., in the determinism parser put forward by Marcus (1980, chap. 10).

This approach not only allows a step-by-step realization of the test phases required for the development but also provides the user with alternative options for the output owing to the different depths of analysis in PHASE II and III.

3.3

Beyond the phase of analysis, ASCOF includes a phase of transfer and synthesis, where the words of the source language are exchanged for those of the target language and where simultaneously structures are altered in the tree structure if necessary. The changes of structure are carried out by a transformational grammar. The grammar operates on trees; grammar and algorithm are separate from each other and the algorithm interprets the externally stored rules of the grammar. [The documentation of the transformational component is put forward by Reding (1985); for the discussion of transformation grammars in machine translation, cf. Vauquois (1975), Boitet, Guillaume and Quézel-Ambrunaz (1982), and Huckert (1979).]

On the leaves of the output tree, produced by the syntactic synthesis, a further algorithm operates, which interprets a set of morphological rules in order to generate the correct word forms of the target language. The transfer and synthesis components of ASCOF are shown in **Figure 1b.**

The separation of grammar and algorithms allows the application of the above-mentioned components in other languages as well, provided that the grammar is replaced.

4 Linguistic Techniques and Computational Realization

4.1

The most sophisticated phase within ASCOF concerns analysis (French); the synthesizing phase (German) has not yet been developed to such an elaborate extent.

This paper consequently concentrates on the description of the analyzing phase. Much space is devoted to segment analysis, the interaction of the complement analysis and the analysis of complex noun phrases, which is discussed in detail and illustrated by examples.

4.2

PHASE I of the analysis consists of the sentence/text input and the morphological analysis. Each word form is assigned the set of possible categories as well as the morpho-syntactic information. A full form and a stem dictionary (both approximately 47,000 entries) as well as a suffix dictionary (inflectional suffixes) are available. Unknown word forms undergo a derivational analysis (based on Stegentritt 1978).

4.3

Within ASCOF's PHASE II, we can distinguish two different parts, both realized by means of context-free grammars; in the first part, categorial (word class) ambiguities are resolved through a CFG working on the string of word classes issued from PHASE I. By applying this grammar, we obtain one – or possibly more – string(s) of unequivocal word classes. [A special word class problem occurring in French, concerning the ambiguity of (-ant) forms, is discussed in Féneyrol (1982).] As the consulted grammar itself represents a rudimentary syntactic analysis, here we arrive at an initial decomposition of the sentence into so-called simple syntactic units, such as nominal, prepositional, adjectival, adverbial, and verb units; (coordinating and subordinating) conjunctions, introducing words (relative pronouns and others), and commas form proper units, too.

For a sentence such as

(1) la directive du Conseil du 20 juillet, qui, à l'article 4, prévoit une augmentation du prix du maïs de 3%, touche les régions du sud et les pays africains qui dépendent de ce produit d'importation.

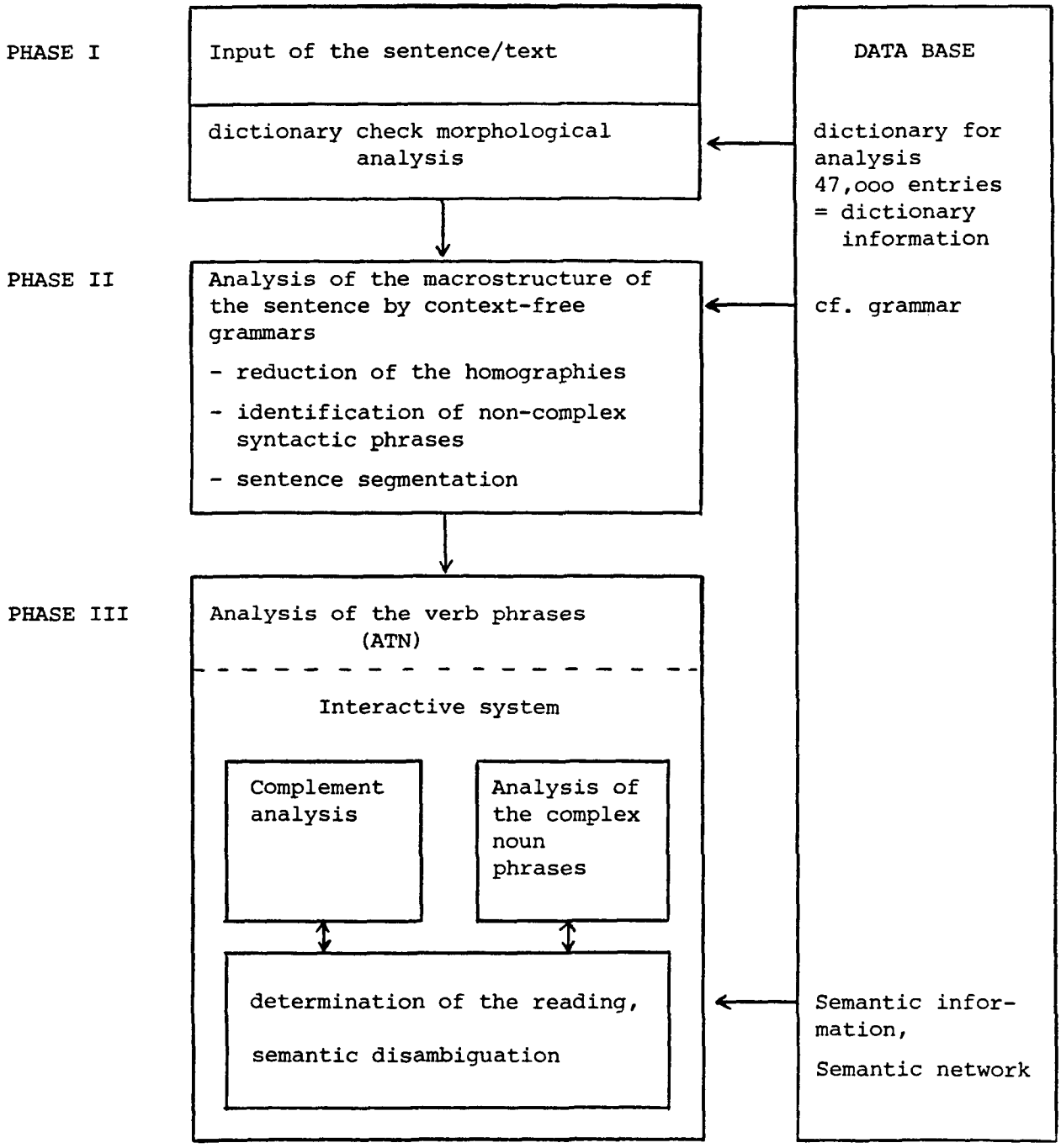

Figure 1a. Analysis components in ASCOF.

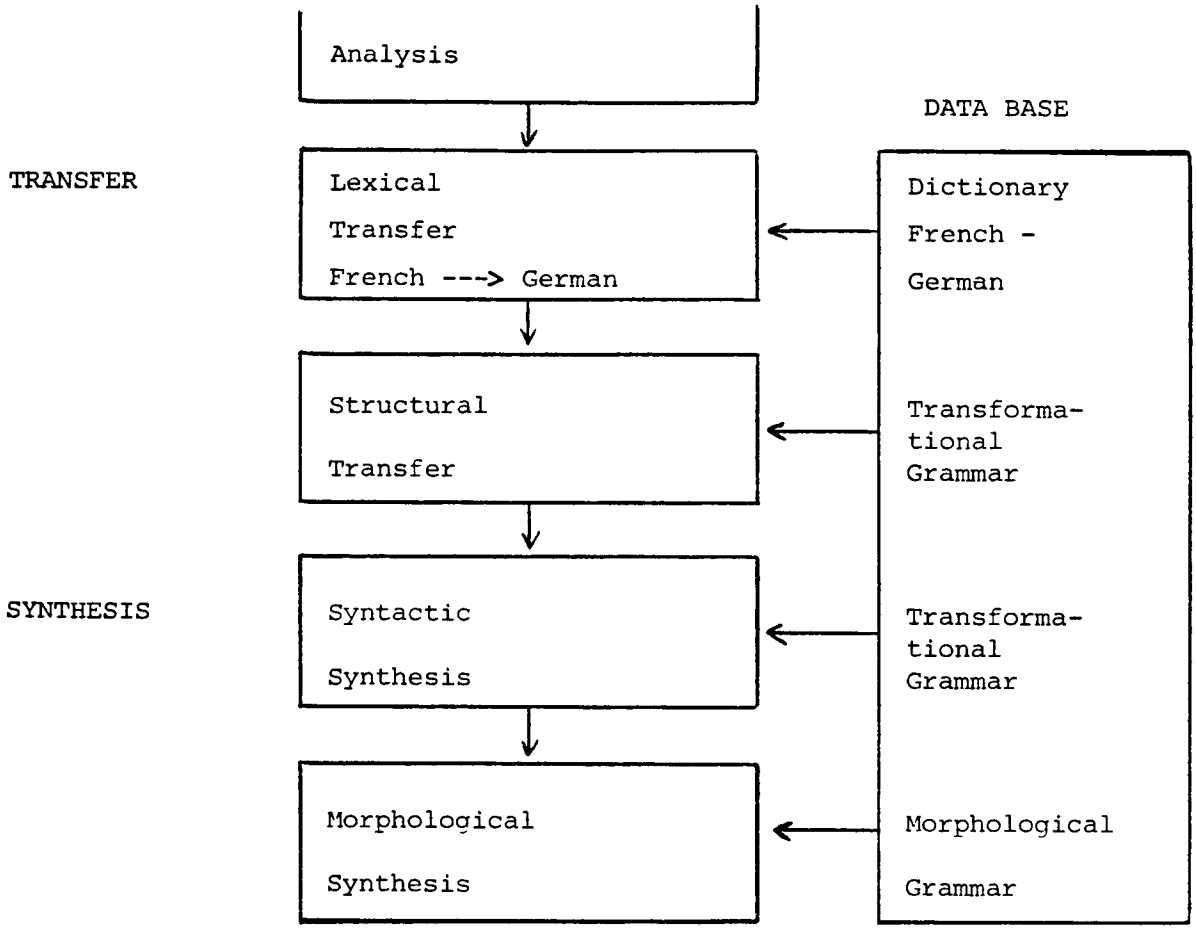

Figure 1b. Transfer and synthesis.

the result of this first part of PHASE II will be as follows:

(2) nnog pg pg com rel com pg com verb nnog pg pg pg com verb nnog pg coord nnog adj rel verb pg pg eos

[The abbreviation "eos" means "end of sentence", "nnog" means "nominal noun phrase", "rel" means "relative word"; the other abbreviations need no comment.] This string of simple syntactic units serves as input for the second part of PHASE II, concerned with the segmentation of the sentence, that is, its decomposition into main and/or subordinate clause(s); the following shows in more detail how this part of ASCOF works. [Some of the problems appearing in the segmentation of French sentences as well as a description of another type of segmentation module used earlier are described in Féneyrol (1983).]

The operation of segmentation itself takes place in two steps, each of which is realized by a CFG; in the first step, we apply the CFG IDENTIFICATION to the string under (2) in order to "transform" the sequence of simple syntactic units into so-called Segmentation Units (SU) [we use the following conventions: "( )" for facultativity, " | " for alternatives, and "*" for repetition]:

IDENTIFICATION: (TV1, NTV1, S1, R1)

TV1 :nnog; com; rel; verb; coord; adj; pg; eos

NTV1 :SUNNOG; SUN; SUCOM; SUCOORD; SURELS; SUVERB; SUEND

R1 1 : S1 → (SUNNOG | SUN | SUCOM | SUCOORD | SURELS | SUVERB)* SUEND
2 : SUNNOG → nnog (SUN)*
3 : SUN → (pg | adj)*
4 : SUCOM → com
5 : SUCOORD → coord
6 : SURELS → rel (SUNNOG | SUN | SUCOM | SUCOORD)* SUVERB
7 : SUVERB → verb
8 : SUEND → eos

Taking a look at some rules of IDENTIFICATION, one should note that, for example, in rule 2 SUNNOG is built up without taking account of the internal structure of the complex noun phrase; the actual relations inside the SUNNOG, *l'augmentation du prix du maïs de 3%* – in fact: *(l'augmentation (du prix (du maïs)) (de 3%))* – are not of importance in this part of ASCOF; it is only necessary to identify the complex noun phrase as SUNNOG. [The accurate description of complex nominal phrases is

the main task of NOMAL; for this section, see 4.4.3.] Rule 6 builds up the relative clause, the minimal inventory consisting of relative word (rel) and verbal phrase (verb) with the bracketed facultative elements between them.

The result of IDENTIFICATION for sentence (1) via structure (2) is

(3) SUNNOG SUCOM SURELS SUNNOG SUCOM SUVERB SUNNOG SUCOORD SUNNOG SURELS SUN SUEND

These Segmentation Units obtained from IDENTIFICATION themselves constitute the terminal vocabulary of a second CFG GROUPING. [Both CFGs (IDENTIFICATION and GROUPING) here only contain the categories occurring in the example (1).] Its task consists of moving from the terminal Segmentation Units to the axiom TXS (Text Sentence) by way of a left-to-right, bottom-up parsing, providing a decomposition of the sentence into clauses and indicating their mutual relations.

List of abbreviations:

TXS	Text Sentence
CS	Complex Sentence
S	Sentence
NOV	NOGC + VERBAL
NOGC	Complex Noun Group
SSBCL	Sequence of Subordinate Clause(s)
CSBCLC	(Comma) SBCL (Comma)
SBCL	Subordinate Clause
POST	Postverbal domain
NEUTC	Complex Neuter Unit
NEUT	Neuter Unit (/= NOG)

GROUPING: (TV1, NTV1, TXS, R1)

TV1 :SUNNOG; SUCOM; SURELS; SUVERB; SUCOORD; SUN; SUEND

NTV1 :CS; S; NOV; NOGC; VERBAL; POST; NEUTC; NEUT; SSBCL; CSBCLC; SBCL

```
R1  1 : TXS    →  CS SUEND
    2 : CS     →  S ( SUCOM S | SUCOORD S )*
    3 : S      →  ( SSBCL ) ( SUCOM ) NOV
    4 : NOV    →  NOGC VERBAL
    5 : NOGC   →  NOG ( SSBCL )
    6 : NOG    →  SUNNOG ( SUCOM SUNNOG
                         | SUCOORD SUNNOG )*
    7 : VERBAL →  SUVERB ( POST )
    8 : POST   →  NOGC | NEUTC
    9 : NEUTC  →  NEUT ( SSBCL )
   10 : NEUT   →  SUN ( SUCOM SUN | SUCOORD
                                    SUN )*
   11 : SSBCL  →  CSBCLC | SBCL ( SUCOM SBCL
                         | SUCOORD SBCL | SUCOM
                                    VERBAL )*
   12 : CSBCLC →  ( SUCOM ) SUBCL ( SUCOM )
   13 : SBCL   →  SURELS ( POST )
```

As a single correct result of GROUPING, we obtain for our example the structure represented in **Figure 2.**

It can be easily demonstrated that some other derivation attempts will not succeed; thus, for instance, if we try to apply rule 2 of GROUPING in order to arrive at a coordination of main clauses through the conjunction *et*, the second presumed main clause will not be completable since a verb phrase is missing (in contrast to a sequence such as . . . *et les pays africains* (relative clause) *souffriront de la famine.*). For another case of an ineffectual attempt, we can note rule 11, which would initiate a coordinate Sequence of Subordinate Clauses (SSBCL); here this would lead to a missing verb phrase belonging to the first SUNNOG of the sentence.

Each correct augmentation result – such as the one in Figure 2 – forms the basis upon which the subsequent PHASE III of ASCOF will then operate clause by clause. [We are at present working on the combination of the different distinct parts (word class disambiguation/ decomposition into simple syntactic units; IDENTIFICATION and GROUPING); this conception will ultimately lead to a single CFG achieving PHASE II of ASCOF.]

4.4

4.4.1

The task of analyzing verb sequences, the first step of PHASE III (cf. Figure 1a), is to group together isolated verb elements (finite verbs, participle I, participle II, infinitive) within a segment to assign a structural description to these phrases (e.g., to determine VOICE and TENSE) and – ultimately – to interpret those phrases as nodes of a tree structure.

The grammar of the verbal analysis is conceived as a two-step ATN (cf. Biewer 1985).

The transitions between the different states of the ATN are guided primarily by a subclassification of verb elements (e.g., participle II of *avoir*). The conception of the ATN enables the processing of an unlimited number of infinitive phrases; the integration of these infinitive phrases into the tree structure is governed by categorial information – recorded in dictionaries – as well as by tests concerning the syntactic context. Nevertheless, the interpretation of non-complex infinitive phrases as nodes of a dependency tree does not exclude ambiguities that cannot be resolved at this stage of the verbal analysis. Accordingly, it is a set of hypothetical structures, gained by purely syntactic and surface-related data, that forms the output of the verbal analysis.

4.4.2

When the analysis of verb phrases is completed, the sentence is structured in such a way that parts of main and subordinate clauses and their interrelations are identified. Furthermore, non-complex syntactic (one-nuclear) noun and prepositional phrases as well as verb sequences are determined. An interactive component operates on this input performing (1) the complement analysis, (2) the analysis of complex (multinuclear) noun and prepositional phrases, and (3) the disambiguation of

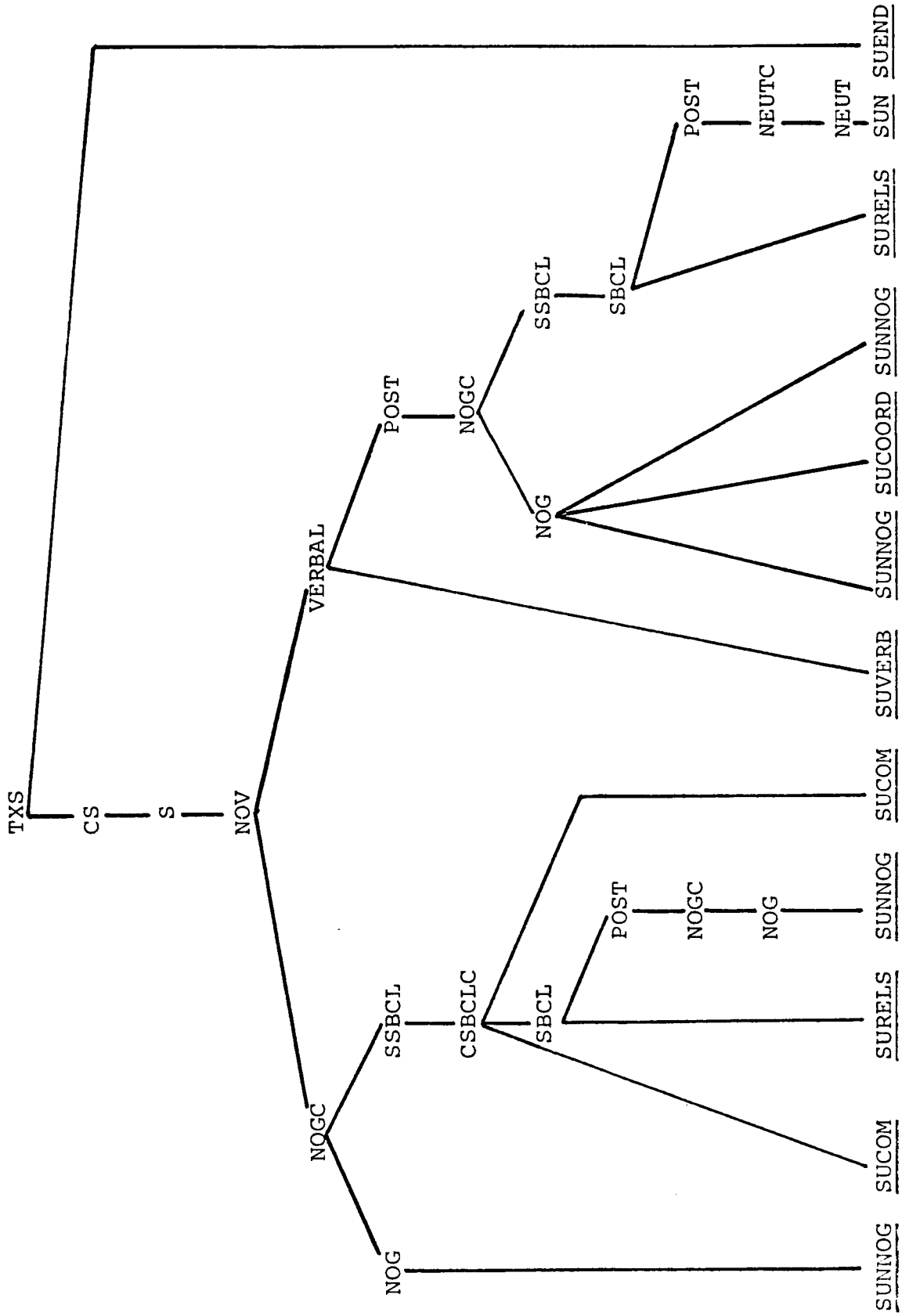

Figure 2. Tree-structure of the segmented phrase.

lexical items, all according to syntactic *and* semantic criteria.

In contrast to many other systems of machine translation or linguistic data processing, semantics and syntax are here equally treated, neither having priority over the other. Another characteristic feature is the interaction between each step of analysis: the complement analysis determines the head of a complement (SUB, K1, K2 in **Figure 3**), and the internal structure of the complement itself (hypotactic and paratactic relations) is determined by the complex nominal analysis (NOMAL), the findings of which are then returned to complement analysis (cf. Figure 3). This interactive procedure is guided by the syntactic-semantic features of the dominating verb. The required syntactic features have already been calculated in the previous analysis; in contrast, the semantic information is memorized in the data base of the semantic network and must be taken from there (cf. 4.4.4).

ASCOF analysis offers, as output, syntactic-functional structure trees representing a synthesis of constituents and dependencies.

In the complement analysis, the dictionary entries of the verb – the verb being the central node of the sentence – determine those noun or prepositional phrases that belong to the verbal frame. The remaining phrases are parts of complex noun or prepositional phrases or of adverbials. For the preverbal and postverbal fields, complement analysis operates separately, reflecting the dissimilar structure of both fields. The component SUB, operating in the preverbal field, defines subjects of non-dominated verbs as well as possibly occurring pronominal complements; dominated verbs (e.g., causative constructions) are analyzed by another component, not described in this paper (INFSUB, cf. Stegentritt (1982, 1984)). In the postverbal field, the components K1 and K2 are activated for the complement analysis. In every case, only the first constituent of the complement is determined. Then the component NOMAL (cf. 4.4.3) is called and investigates the structure of the identified complement.

The component determining the subject (SUB) is bipartite; in addition to the preverbal pronominal complements, the first part also defines the subject, provided that it is a pronominal subject. If there is no pronominal subject, part two defines a nominal subject. In the first part, the analysis progresses from right to left, starting at the finite verb/auxiliary. All preverbial elements are processed subsequently, until a pronominal subject is found.

Figure 4 represents the first phase of SUB as an ATN. The different paths correspond to the possible distributions of syntactic phrases in the preverbal field of the French language. If the verb is directly preceded by a comma, it is highly probable that this comma indicates the end of an insertion, such as in sentence (4). In this case, the analysis is continued up to the next comma, with no consideration of the insertion. The edges and states of the ATN that had been activated are listed in (4a).

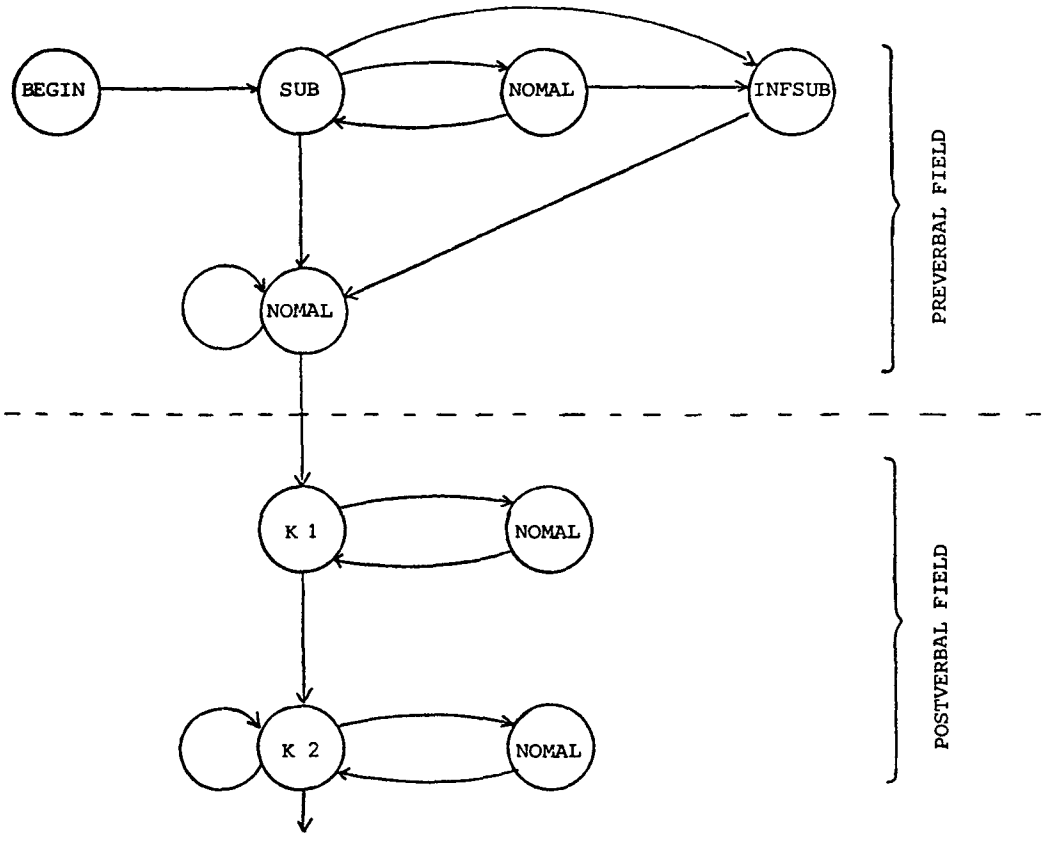

Figure 3. Interaction of different components in Phase III.

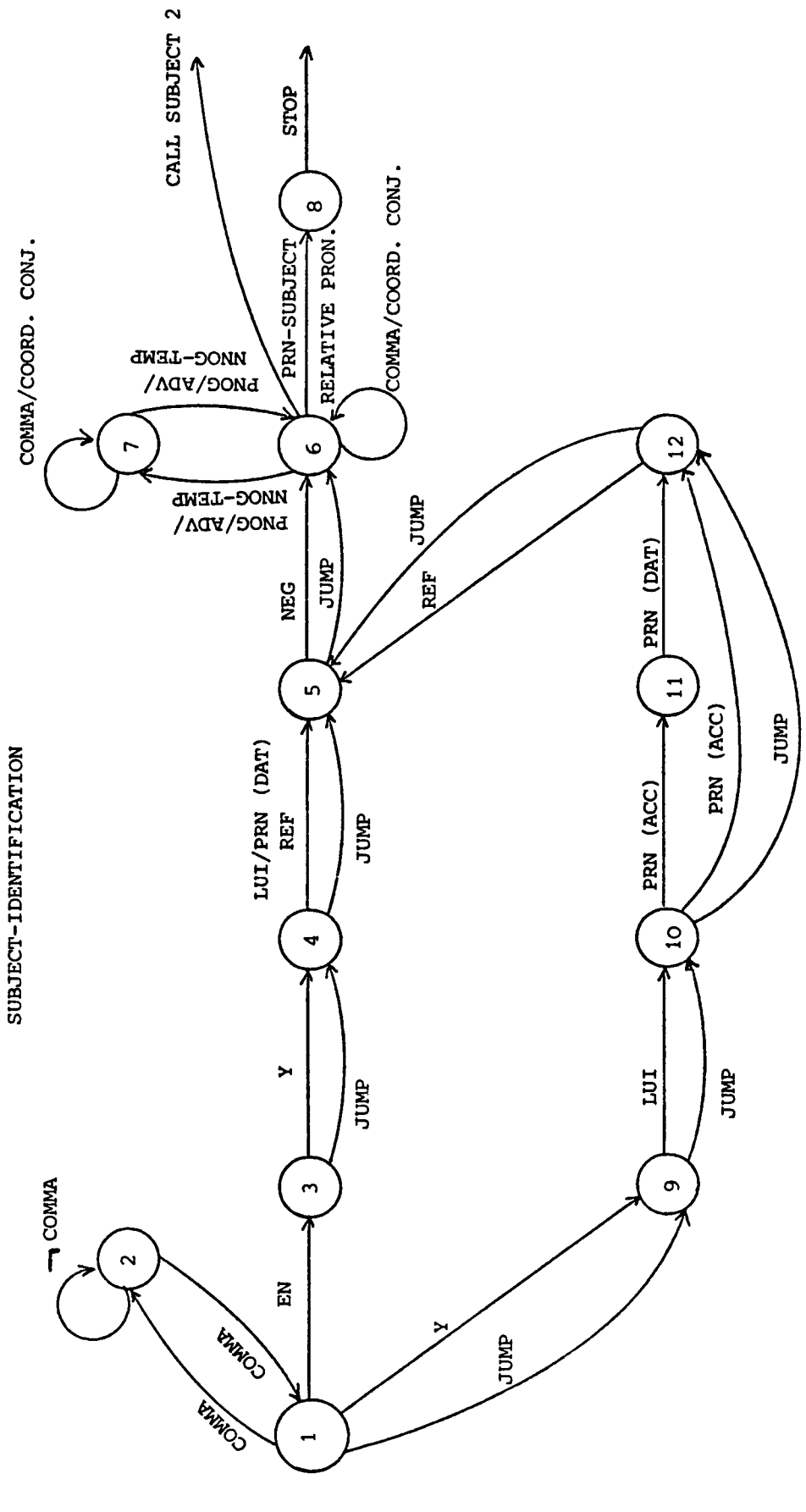

Figure 4. ATN grammar of the preverbal field.

(4) . . . qui, à l'article 4, prévoit . . .

(4a) 1 – COMMA – 2 – (not COMMA)*4 – 2 – COMMA – 1

In the first phase, combinations of clitics as noted in (5) - (7) may be identified.

(5) il ne le lui propose pas

(5a) 1 – JUMP – 9 – LUI – 10 – PRN(acc) – 12 – JUMP – 5 – NEG – 6 – PRN-Subject – 8 – STOP

(6) il me le donne

(6a) 1 – JUMP – 9 JUMP – 10 PRN(acc) – 11 – PRN(dat) – 12 – JUMP – 5 JUMP – 6 PRN-Subject – 8 – STOP

(7) il m'en parle

(7a) 1 – EN – 3 – JUMP – PRN(dat) – 5 – JUMP – 6 – PRN-Subject – 8 – STOP

If the network comes to the end of a clause or meets a configuration not covered by the rules, the second phase of analysis – starting at the beginning of the clause in the reverse direction of analysis – is called to search for the first nominal phrase (i.e., the leftmost). This is, however, a mere candidate for subject function, the form of a syntactic phrase (NNOG) not yet indicating the actual function of this phrase. As an illustration, compare (8).

(8) Après les changements de cours pour le froment, *l'avoine* et la seigle, la directive (. . .) touche (. . .)

As the set of rules in this second phase supplies the first noun phrase (NNOG, = *l'avoine* in (8)) as a subject, the phrase in question must be checked as to whether or not it is part of a coordination of prepositional phrases in which the preposition can be deleted. This subject-control mechanism presupposes that the coordinating conjunctions *et* or *ou* always indicate the end of a coordination. It follows that the subject candidate cannot be the subject if the following structure is given:

(9) CANDIDATE-COORD-NNOG-COMMA-NNOG-VERB

In this structure, however, the final preverbal NNOG (in (8), *la directive*) is the correct subject. In structure (10):

(10) CANDIDATE-COORD-NNOG-VERB

the subject candidate is confirmed. The head of the subject having been identified, phase NOMAL is called to examine a possibly complex internal structure of the subject and further potentially complex nominal sentence elements in the preverbal field – as, for example, the adverbial insertion *à l'article 4* in sentence (4) (cf. 4.4.3).

The complement analysis in the postverbal field is carried out by another pair of components (K1 and K2), also cooperating with NOMAL. K1 is the less sophisticated complex, identifying nothing but direct complements – which it recognizes merely by form – and prepositional complements obligatorily demanded by the verb. Yet because of the various functions prepositional phrases may assume, prepositional complements at this stage are to be identified only if no more than one PG within the segment is a potential candidate. Applied to our example (1)

(1) la directive du Conseil du 20 juillet, qui, à l'article 4, prévoit une augmentation du prix du maïs de 3%, touche les régions du sud et les pays africains qui dépendent de ce produit d'importation.

the results are:

- for *prévoir: une augmentation* as a direct object
- for *toucher: les régions* as a direct object.

Again, the determination of the coordinative relationship between *régions* and *pays* is given within NOMAL (cf. 4.4.3).

Yet, whenever the prepositional complement is optional or when more than one candidate for an obligatory prepositional complement is available, semantic criteria must be observed *in addition* to formal and syntactic criteria in order to correctly determine the complements.

The semantic restrictions that must be respected when a certain phrase is attributed the function of a complement are denoted in the verb entry in the dictionary. According to these (semantic) restrictions, the algorithm searches for the required criteria, namely among the nuclear lexemes of the potential complement phrase under consideration. The semantic information for these lexemes, however, is not stored with their dictionary entries (e.g., nouns) themselves, but is represented by a semantic network (cf. 4.4.4).

The strategy used by ASCOF in K2 shall be illustrated by the three verbs in sentence (1).

Prévoir does not demand the attribution of a further complement, the verbal frame being saturated as soon as *une augmentation* is identified as a direct complement. The verb *toucher* raises other problems; *toucher* may occur

a. without any complement: *le tireur d'élite a touché.*
b. with a direct complement: *il touche ma main.*
c. with an *à*-complement: *le chercheur touche au but.*
d. with a direct complement and an optional *de*- complement: *il m'a touché du doigt.*

As to sentence (1), the direct complement *les régions* having been located, it must still be examined whether one of the following *de*-phrases is a complement; that is, whether verbal frame (b) or (d) is valid. A *de*-phrase, however, can for instance be the complement of *toucher* if the noun of this phrase is semantically labeled as "part of the body". Consequently, the semantic value of the nuclear lexemes of the *de*-phrase following *les régions* must be examined in this respect. The results obtained through consultation of the semantic network show that the nuclear lexeme *(sud)* of this phrase does not satisfy the required condition. It follows that verbal frame (b) is actualized in this sentence. Simultaneously, this analysis defines the reading of the verb.

As to the last verb of sentence (1), *dépendre,* no direct complement can be found. Thus the reading 'to be dependent on' versus the reading 'to take down' is immediately determined. This reading of *dépendre* demands an obligatory *de*-valency. Yet no restrictions can be formulated for the semantic quality of the nuclear lexeme of the *de*-phrase. This implies that only formal and syntactic criteria can be considered. In the case of the first phase, *de ce produit,* the formal and syntactic criteria hold true (a preposition introducing a phrase, an article) and thus allow this phrase to be identified as a complement and handed over as a head-phrase to the analysis of the complex noun phrases. A second run of K2 checks the following phrase, *d'importation,* which is rejected by those criteria identifying this phrase as a compound in NOMAL. The criteria are: no proper noun, number = singular, no article, no attribute. One single attribution is performed accordingly.

An interesting variant of sentence (1) is:

(11) ... dépendent de l'effet de l'importation.

where both phrases may function as a complement. If the second phrase *(de l'importation)* is identified as a complement, the first postverbal phrase can only be attributed the function of an adverbial. In a concluding control phase for those phrases scheduled as adverbials, the structure variant of *de l'effet* as an adverbial and *de l'importation* as a complement can be erased, *effet* not being apt to assume the function of an adverbial in this phrase, in contrast to prepositional phrases, such as *de cette manière* or *de cette façon.*

4.4.3

The task of the analysis of complex noun phrases (multinuclear noun- and prepositional phrases) is to identify their boundaries and to describe their internal structures, i.e., to determine the syntactic-functional relations between the various parts of the complex. [For more details, cf. Ritzke (1985).]

The following syntactic-functional relations are defined for the analysis:

- head-function of the central phrase of a complex *(la maison du père)*
- paratactic relations of phrases depending on the head:
  coordination (CO): the coordinated phrase has the same syntactic function as the phrase with which it is coordinated *(l'étude de la détermination et de la classification).*
  apposition (AP): a paratactic relation on the level of syntactic analysis *(du froment tendre, plante comestible qui ...)*
- hypotactic relations of phrases depending on the head:
  prepositional complement/object (PO): signalizes a very close relation between two phrases; the preposition of the second phrase is mostly synsemantic; the PO and the inflectional case of the two languages (French, German) frequently correspond to each other *(la voix de son maître – die Stimme seines Herrn).*
  [For more details concerning the function of prepositions and their analysis, cf. Ritzke (1981).]
  prepositional complement/part of a compound (PC): implies an even closer unity of the two phrases; PC and the governing phrase, in contrast to other structural units, represent primary units often functioning as a one-nucleus phrase; this is of great importance to the language pair French and German, the complex phrase in French being equivalent to a German one-word compound (e.g., *la notation de base – Basisnotation)*
  prepositional attribute (PA): signals a relatively free relation to position and introductory preposition, which is always autosemantic; for this reason, the PA may assume many different semantic values (e.g., temporal or local attribute, etc.) *un groupe simple à l'intérieur du syntagme).*

Basic elements for the analysis of complex noun phrases are syntactic phrases having only one nucleus. They may occur in the form of noun phrases, prepositional phrases, or pronouns functioning as nouns.

Every one-nucleus noun – or prepositional – phrase of these basis elements is directly relevant to the analysis of complex noun phrases as a potential element of a complex phrase. Every non-nominal phrase (e.g., verb phrase, etc.) is indirectly relevant as an indicator of the boundary of the complex.

As for French, the identification of the left boundary of a complex noun phrase presents no problem since the phrase furthest left of a complex is its central phrase (its head) in the majority of cases. The head represents the syntactic function of the whole complex on the sentence level and at the same time marks the boundary of the left side. Further phrases are located on the right side of the head, the analysis being directed from left to right.

The head of a complex noun phrase is identified and determined in syntax and function by means of the analysis of verb complements. The complement analysis forms an interacting system with the analysis of nominal complexes – as explained above. The potential head of a potential complex noun phrase, identified by complement analysis, is a signal for the call of the complex nominal analysis.

The search for further phrases on the right side of the head begins at the head. Here, two phrases are checked to identify a boundary. These phrases consist of the head on the left and a dependent phrase on the right which is connected with this head. The syntactic function is simultaneously identified. Thereupon, complement analysis is called again to determine further complements or non-complements, respectively.

Non-complements (adverbials) may also have a complex internal structure. Accordingly, after the analysis of complex-structured complements, non-complements are analyzed as to their possibly complex structure. The preverbal field is investigated first. If, for example, the complex structure of a subject has been identified there, a further complex nominal analysis is undertaken

for the first phrase of the sentence – provided that this phrase is identified as adverbial and that other phrases follow in the field between its position and the head of the subject. This is done in order to clarify whether this phrase is a central phrase of an adverbial complex.

A similar approach is applied for the postverbal field: first the complements, then the non-complements, are analyzed as to their possibly complex internal structures.

In principle every phrase can be related to another phrase in its direct neighborhood. This phrase again may dominate a subsequent phrase, etc., so that highly recursive structures may appear within a nominal complex (right-branches cascades).

At the same time two related phrases may be disconnected by other phrases, when several phrases of different syntactical relations are dependent upon the same phrases. In addition to the potentially recursive structures, discontinuous relations must be detected as well.

The algorithm NOMAL has been conceived especially for the solution of those recursions and discontinuities within the complex nominal syntax (cf. **Figure 5**).

NOMAL is a procedural (sub)system similar to an ATN. Each syntactic function has its own procedure. The procedures are called successively by the main program whenever a pair of phrases – the head and the actual phrase – is examined. The identification of the relation that is – or is not – established between both phrases is governed by rules containing conditions for the head and the actual phrase. Each procedure is provided by a proper set of rules; the monitoring of the rule conditions is carried out by test operations on edges.

If the testing of a rule attains a positive result – that is, if the conditions in this rule for head and actual phrase hold true – the phrases are unified and their relation interpreted according to the procedure where the rule was found. Valid conditions in the procedure PO, for example, have the effect that the relation between the phrases is identified as a prepositional complement/object. The approach is analogous for the remaining procedures.

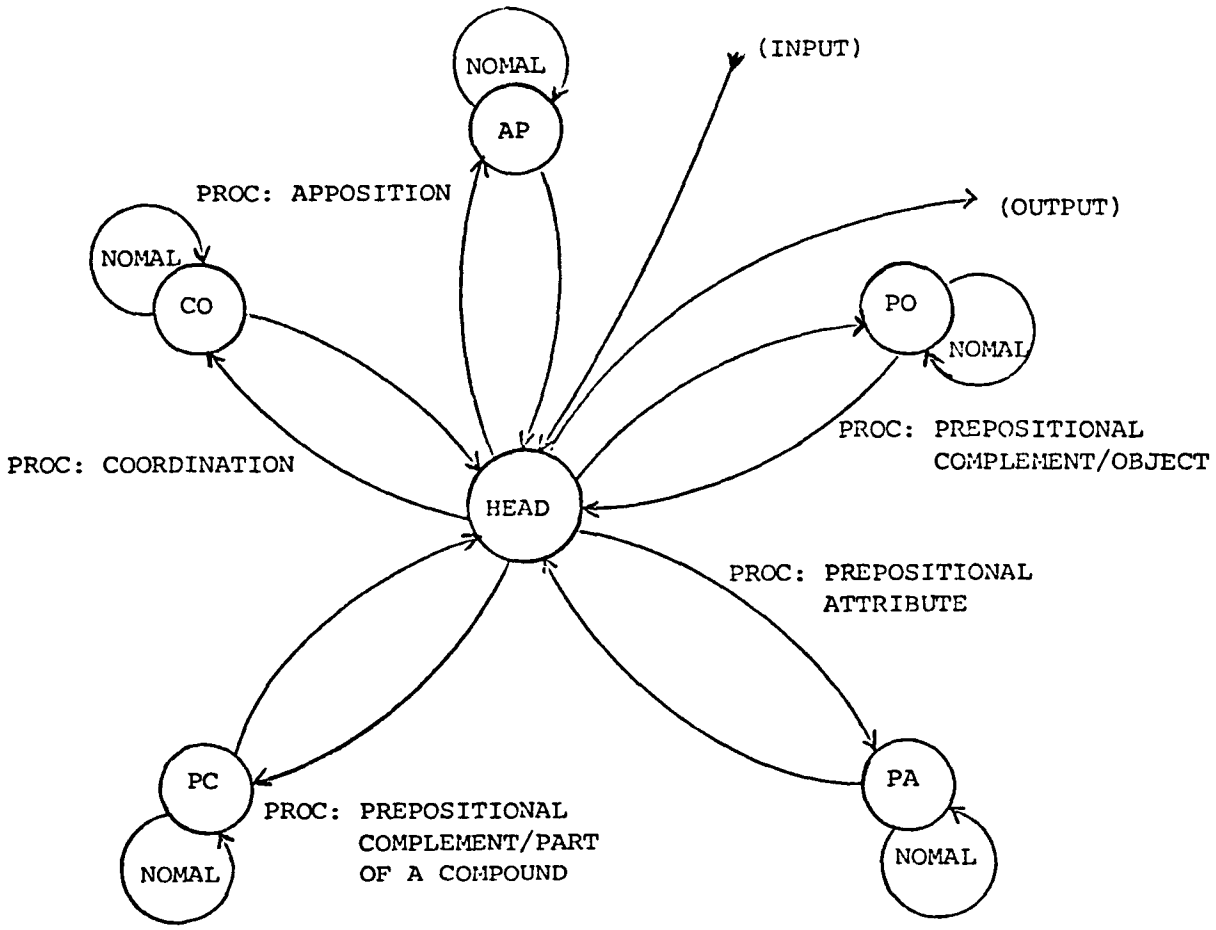

Figure 5. The subsystem NOMAL.

The actual phrase having thus been attributed a function, it becomes the head itself in the next run; the actual phrase will then be the subsequent phrase. Analogously, further positive results gradually move the actual phrase to the head of the next run; the head is deferred to the right, thus allowing the correct analysis of recursive cascades.

If none of the procedures attains a positive result, the new run is begun with another head, that is, with the phrase to the left of the head last tested. If there is still no positive result, the head keeps moving analogously to the left, until all the possibilities have been tried. The actual phrase is identical for all runs, thus allowing the analysis of relations between two discontinuous phrases.

In every case, a boundary is reached when none of the runs offers a positive result. The former actual phrase then no longer belongs to the complex but rather indicates its boundary.

The conditions denoted in the rules consist of morpho-syntactic and semantic information. Morpho-syntactic information (e.g., an article, the form of an introductory preposition) can be read directly, the previous step of the analysis having made it available. In contrast, the semantic information must be deduced from the semantic network, which serves as a data base for the syntactic-functional analysis (cf. 4.4.4).

The importance that evaluating rule conditions holds for the analysis shall be demonstrated in the following by the complex noun phrases in example (1).

The nominal complex *Les régions du sud et les pays africains:*

Formal indicators can be checked at this point for the determination of the paratactic relation: in our example, the coordinating conjunction *et*. Commas and the sequences of comma and conjunction may be evaluated as well. Yet the mere checking of such indicators is generally not sufficient for determining the proper parts of a coordination.

In order to determine the correct parts, the semantics of the nouns must also be checked, that is, the closeness of the semantic relation between the candidates in question must be checked in the example. This is accomplished by retrieving operations within the semantic network using the nominal lexemes of the different phrases. The common generic term (geographic place), which is reached by *région* and *pays*, satisfy the two conditions: coordination indicator *et* and greatest semantic closeness. They are parts of the coordination, the phrase *du sud* is recognized to be a prepositional complement/object of the first phrase by the procedure PO. The result is the following structure:

((les régions (du sud)(PO)) (les pay africains)(CO))

The nominal complex *la directive du Conseil du 20 juillet:*

For hypotactic relations, too, the examination of semantic criteria is necessary over and above that of the syntactic conditions. The search for a certain preposition at the beginning of a clause – in our example, *de* as an indicator of object references – must be performed by examining the semantic closeness of the nuclei (nouns) of the phrases. In the given example, the direct time indication, which suggests an attributive relation, is thus identified. In the network, a temporal attribution is closer to the information of 'act of legislation' (connected with the lexeme *directive*) than to 'institution' (connected with the actual reading of the lexeme *Conseil*). Accordingly, valid conditions exist for the relation of prepositional complement/object between *la directive* and *du Conseil* and for the relation prepositional attribute between *la directive* and *du 20 juillet.* The last relation is discontinuous. The result of the analysis is the following structure:

((la directive (du Conseil)(PO))(du 20 juillet)(PA))

The nominal complex *une augmentation du prix du maïs de 3%:*

The last non-complex phrase, *de 3%,* of this complex phrase holds a – likewise discontinuous – relation to the first phrase, *une augmentation,* whereas the remaining *de*-phrases form a cascade and each depends upon its preceding phrase (the relation of prepositional complement/object).

The phrase *de 3%*, an indication of quantity, is more closely related to the action, *augmentation* (action/state of affairs: increase), by the corresponding edge configuration in the network than to *prix* (notion of a measure) or to *maïs* (plant, agricultural product). Accordingly, a relation is established between *de 3%* and *une augmentation*, disregarding the interposed phrases. The result is the following structure:

((une augmentation((du prix (du maïs)(PO))(PO))
(de 3%)(PA)))

The nominal complex *de ce produit d'importation:*

In order to determine the relation between the two phrases, formal criteria may be more readily applied than in the previous cases. The second phrase is characterized by the fact that neither a determinative nor an attributive element is contained or follows. Moreover, the nominal nucleus neither is in the plural nor is it a proper name. The phrase is introduced by the preposition *de*. All the characteristics are rule conditions for an actual phrase in the procedure prepositional complement/part of a compound. These conditions being fulfilled and no other candidate being present, the relation between the phrases may be determined in this particular case without semantic examinations. The result is the bracketing:

(de ce produit (d'importation)(PC))

As it is difficult in many cases, mainly in the postverbal field, to determine a function with certainty after a single run of the analysis of complex noun phrases, decisions are made only if there is no doubt (e.g., if there is

only one potential candidate or if candidates can be excluded because of semantic incompatibility, etc.). Otherwise a solution must be sought through repeated interaction with its complement analysis, which must then be given priority. Yet, given the basic alternative of the PP-attachment (complement, adverbial, part of a complex noun phrase), a reliable solution cannot always be found and several solutions may be admitted in such doubtful cases.

4.4.4

Conceding that structural descriptions within machine translation – functioning as the input for a transferring component and thus defining a sort of "interlingua" – cannot be established without semantic information, the appropriate form for representing this information as well as its integration into the analysis process appears to be controversial.

With regard to the various ways of integrating semantic information, and above all to the proper moment of its integration, there are two extremely different points of view. On the one hand, syntax may be accorded absolute priority, in such a way that all syntactic and structural descriptions of a given input sentence are generated and only then semantically interpreted and filtered according to the semantic information. On the other hand, the analyzing algorithm can be guided by semantic information, the "level" of syntactic relations only being consulted as a secondary control mechanism.

Most machine translation systems favor the first point of view, partly owing to the history of this science. Yet then, of course, the problem with highly complex sentences is how to deal with syntactic ambiguities. If, for example, an input sentence is attributed a number of tree structure by the syntax-driven analysis, part of the syntactic relations will necessarily be represented redundantly; if, on the other hand, the input sentence is attributed a chart or chart-like structural description by the syntactic analysis, much more time must be sacrificed to the semantic component – as a consequence of the more complex input structure.

Semantically driven parsers, whose equivalents can mainly be found in language-oriented AI research, guarantee a highly efficient parsing, but are extremely dependent upon their discourse area, due to their considerably limited "expectancy", and generally operate on non-complex sentences, linguistically speaking.

In order to isolate correct structural descriptions as soon as possible, cross-connections between syntax and semantics are established by the ASCOF system (cf. 4.4.2). This is to say that semantic information is made available for the identification of syntactic-functional relations and that, vice versa, syntactic and functional relations can be used for semantic interpretation (disambiguation).

The flow of information between syntax and semantic in ASCOF analysis is therefore not based on a phase model (e.g., the sequential operations of separate modules). The semantic information in the ASCOF system is stored in a semantic network. (For different conceptions of semantic networks, cf., above all, Findler (1979).)

The implementation of the network structure is based on the COMSKEE data structure "dictionary", which is a string-indexed array (of variable length) over strings (of variable length) that can be held externally. "Of variable length" means that a string can be of any size so that its length is only limited by the memory space of the machine.

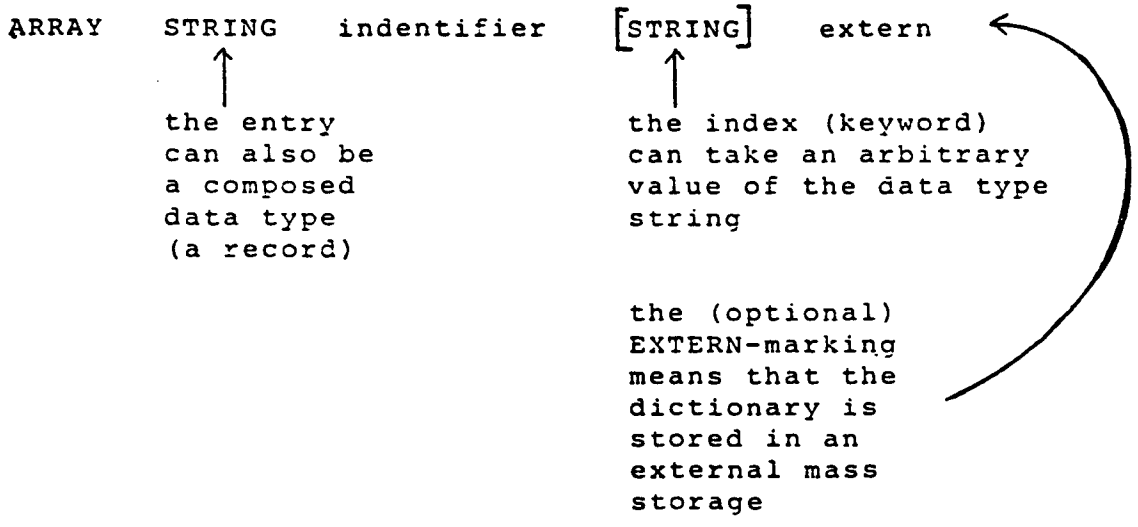

Figure 6. Declaration of the COMSKEE data structure "dictionary".

A module that interprets this structure as a network operates on the "dictionary", as shown in **Figure 7a.** [Figures 7a–7c. Structures of the dictionary entry.]

7a.

index	entry
household utensil	ISA : *artifact*
artifact	ISA : concrete object

A part of the entry can be the index (keyword) of another entry. Thus it is not necessary to store explicitly the information that a person has a head; rather this information can be deduced *on demand* by following an ISA path, which means in our implementation, interpreting an array element such as:

7b.

index	entry
A	ISA: B,

as a pointer from A to B labeled by ISA, as in

7c. **A** —ISA—> **B**

All directly addressable nodes in the network represent basic forms that point to semantically unambiguous readings following denoting edges. Consequently, they represent the FORM-CONTENT relation. For example, the basic lexical form *cuisinière* points in the semantic readings of 'female cook' and 'oven', following the denoting edges, as **Figure 8** shows:

As also shown in Figure 8, several basic lexical forms may refer to the same semantic reading when following the denoting edges. The relation defined by denoting edges thus takes into account the linguistic phenomena of synonymy and polysemy.

ISA edges, which define a special implicit relation, are important for two reasons. First, they allow the defining of a sort of semantic distance between linguistic units, and second, they economically administer information deducible by inheritance paths.

Semantic distance in the linguistic sense may be of importance for the correct identification of coordinative structures. The complex noun phrase in

(12) La vente des réfrigérateurs et des cuisinières

may – on a purely syntactic level – be attributed two structures (*des* can be interpreted as a preposition as well as an indefinite article), cf. **Figures 9a** and **9b:**

Based on the information as illustrated in Figure 8, it is now possible to carry out a lexeme disambiguation of *cuisiniere* (namely, as (cuisinière__2)) as well as to give structure 9a priority over structure 9b, since (cuisinière__2) is semantically more closely related to *réfrigérateur* than (cuisinière__1) and, furthermore, since (cuisinière__1) and (cuisinière__2) are more closely related to the node 'réfrigérateur' than to 'vente'. Because of the relations expressed in Figure 8, structure 9c can be assigned to sentence 12:

As made clear by this example, syntactic-functional and semantic decisions are made almost simultaneously in ASCOF. A system admitting a two-phase model for the informational flow between syntax and semantics would not only have to organize the ambiguity of structures in sentence (12) but – in order to undertake a lexeme disambiguation of *cuisinière* – would also have to construct exactly those syntactic relations directly available in an interactive system like ASCOF

For lexeme disambiguation as well as for the identification of syntactic-functional relations, linguistics units that may function as predicates are of major importance. These are primarily adjectives and verbs. Each "verbal node" following denoting edges refers to a set of verbal frames whose complement slots are described by syntactic and semantic conditions.

In detail, each complement slot is attributed the following information.

- A complex of conditions that must be fulfilled by a syntactic phrase in order to function as the corresponding complement:
  - syntactic conditions
  - semantic conditions (to be fulfilled by the head of the phrase in question)
    e.g., for *toucher* in 4.4.2, the formal-syntactic conditions for the prepositional object are: preposition = *de*; the semantic condition for the relevant head is : ISA: part of the body.
  - the necessary occurrence of the argument
  - the possibility of rejecting competitive verbal frames

In the following, the procedure and the efficiency of the semantic network within the syntactic-functional analysis shall be explained with the aid of a representative example.

The French verb *fumer,* for instance, is assigned three verbal frames; they correspond to the readings of 'to smoke a cigarette, to give off smoke, to smoke a trout', cf. **Figure 10.**

As in Figure 10, a network relation between entity nodes and predicate nodes is established according to semantic conditions defining the restrictions for the different complement slots. In Figure 10, this allows the interpretation of *fumer* as 'to give off smoke' and the rejection of the other interpretations (cf. the action commands within verbal frames). This is to say that once *cheminée* is identified as a subject, there is no need to look for a direct object. Furthermore, Figure 10 demonstrates that "semantic markers" alone do *not* guarantee correct semantic interpretations. It is by no means sufficient to know that *poisson* is a concrete object in order to interpret *fumer* in

(13) La cuisinière fume du poisson

in the sense of 'to smoke-dry'; because the meaning of (fumer__3) 'to smoke-dry' and (fumer__2) 'to smoke' –

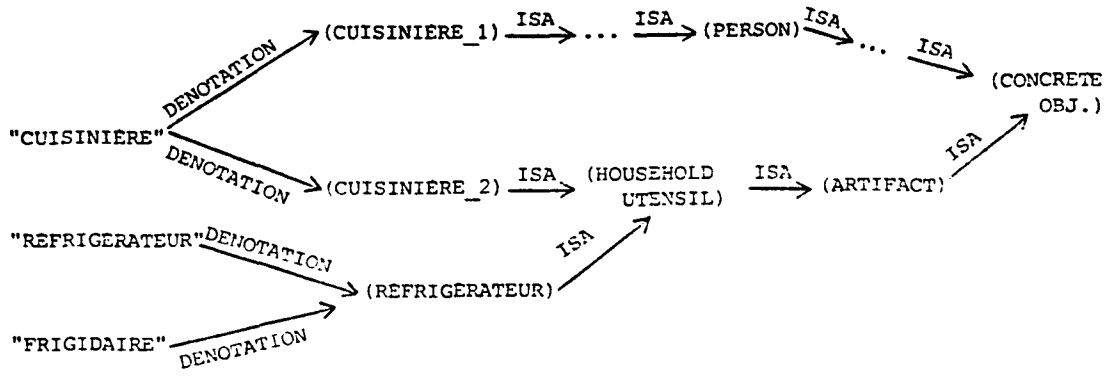

Figure 8. Example of the network structure.

Figures 9a-9b. Possible phrase-structures of (12).

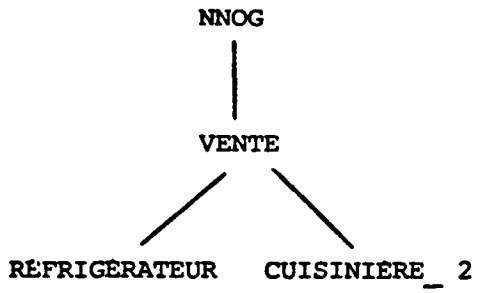

Figure 9c. Correct interpretation of (12).

as in (14) – is not discernible by the semantic marker /concrete/.

(14) La cuisinière fume une cigarette

In a componential model, every animal would have to be marked /can be smoked (dried)/, whereas in the network model this information may be deduced by following a "function" arc from 'animal' to 'viande'. Moreover, with the complement identification, semantic conditions attributed to the different complement slots may be considered for lexeme disambiguation. For example, the reading 'oven' for the noun *cuisinière* in (13) and (14) can be rejected on grounds of the semantic conditions attributed to the subject slots.

At the end of this part of the analysis, the syntactic-functional structure of the sentence is produced (i.e., main and sub-/coordinate clauses, the verb predicate and its complements, adverbials, as well as the internal structure of noun and prepositional phrases are identified). The analysis yields no interlingual structures but rather structures more closely related to the source language. These are represented by canonical trees. At the same time, due to the interaction between syntax and semantics, many of the lexical ambiguities are resolved. Before starting ASCOF's final two phases of transfer and synthesis, any remaining ambiguities must be cleared up especially by means of further semantic information.

4.5

4.5.1

The phases of the system subsequent to the analysis include a transformation component (transfer and syntactic synthesis) and a morphological synthesis. In the transfer phase, the source language lemmata are first replaced by their target language equivalents. Here it is not sufficient to merely exchange lexemes since target language differences of a lexical nature arising from syntax and/or semantics need to be borne in mind. Instead, the replacement process must take into consideration the information contained in the analysis tree, for example, the syntactic and semantic verb frame actualization. Subsequent to the lexical transfer, the analysis structures are transferred to the structures appropriate to the target language (syntactic synthesis). This transfer component is implemented as a **tree transformation algorithm** that interprets externally stored rules (transformation instructions). Linguistic data and algorithm are thus strictly separated from one another so that this component may also be used for other language pairs. The transformation algorithm runs through the analysis tree in preorder, tests for each node whether a package of rules exists for the given node label and – provided that the conditions of a rule are fulfilled – carries out the instructions that refer to some few elementary operations. The conditions and the transformation instructions can refer to both subtrees and the attribute-value pairs associated with the node. (For example, the inheritance of the target language gender of the head of a NOG to attribute and determiner complex.)

4.5.2

The input for the morphological synthesis is the labeled tree taken from the syntactic synthesis. In contrast to the syntactic synthesis, the morphological synthesis operates only locally, in other words, the pre-terminal nodes are examined and processed isolated from each other. Tree transformations are thus no longer carried out. The basic forms of the lexemes and the morpho-syntactic information from the pre-terminal nodes serve as keys that call the appropriate rule of the morphological generative grammar. Grammar and algorithm are separate. The grammar for the morphological synthesis of German is able to generate German word forms, provided that the necessary information supplied from all of the preceding phases is complete and accurate.

5 Future Prospects

The ASCOF system has been implemented in the sections described in the present paper and has in part already been tested. Here we have a version that covers syntactic structures and vocabulary based on texts (EC bulletins) taken from the agricultural sphere. The vocabulary of this sphere is completely covered on the morpho-syntactic level. As yet, owing to the complexity of the semantics, the semantic data bases (semantic network) has been developed only as a prototype. The improvement of the quality of the translations performed by ASCOF is basically dependent upon the development and elaboration of the semantic networks.

References

Bates, Madeleine 1978 The Theory and Practice of Augmented Transition Network Grammars. In Bolc, Leonard, Ed., *Natural Language Communication with Computers.* Springer, Berlin, Heidelberg, New Yor, W. Germany/USA: 191-259.

Biewer, Axel 1985 Ein ATN zur Verbalgruppenanalyse des Französischen. In Figge, Udo L. 1985: 13-26.

Boitet, Christian; Guillaume, Pierre; and Quézel-Ambrunaz, Maurice 1982 Implementation and Conversational Environment of ARIANE 78.4. An Integrated System for Automated Translation and Human Revision. In Fix et al. 1985: 225-236.

Féneyrol, Christian 1982 La forme en (-ant): un problème de l'analyse automatique du français. In Fix et al. 1985: 237-262.

Féneyrol, Christian 1983 La segmentation automatique de la phrase dans le cadre de l'analyse du français. In *Actes du Colloque International Informatique et Sciences Humaines. 1981. LASLA. Université de Liège.* Liège, Belgium: 353-368.

Féneyrol, Christian; Ritzke, Johannes; and Stegentritt, Erwin 1984 Esquisse d'une analyse fonctionnelle du français (ASCOF). In *Actes de la XIe Conférence Internationale de l'ALLC.* Louvain-la-Neuve. Forthcoming.

Féneyrol, Christian and Stegentritt, Erwin 1982 Komplementanalyse in komplexen Sätzen des Französischen. In Stegentritt, Erwin, Ed., *Maschinelle Sprachverarbeitung 1981. Vortäge auf der 12. Jahrestagung der GAL, Mainz 1981. Sektion Maschinelle Sprachverarbeitung. Sprachwissenschaft – Computerlinguistik 8.* AQ, Dudweiler, W. Germany: 55-68.

Figge, Udo L., Ed. 1985 *Romanistik und Datenverarbeitung 1983. Akten des Deutschen Romanistentages, Berlin 1983. Sektion Romanis-*

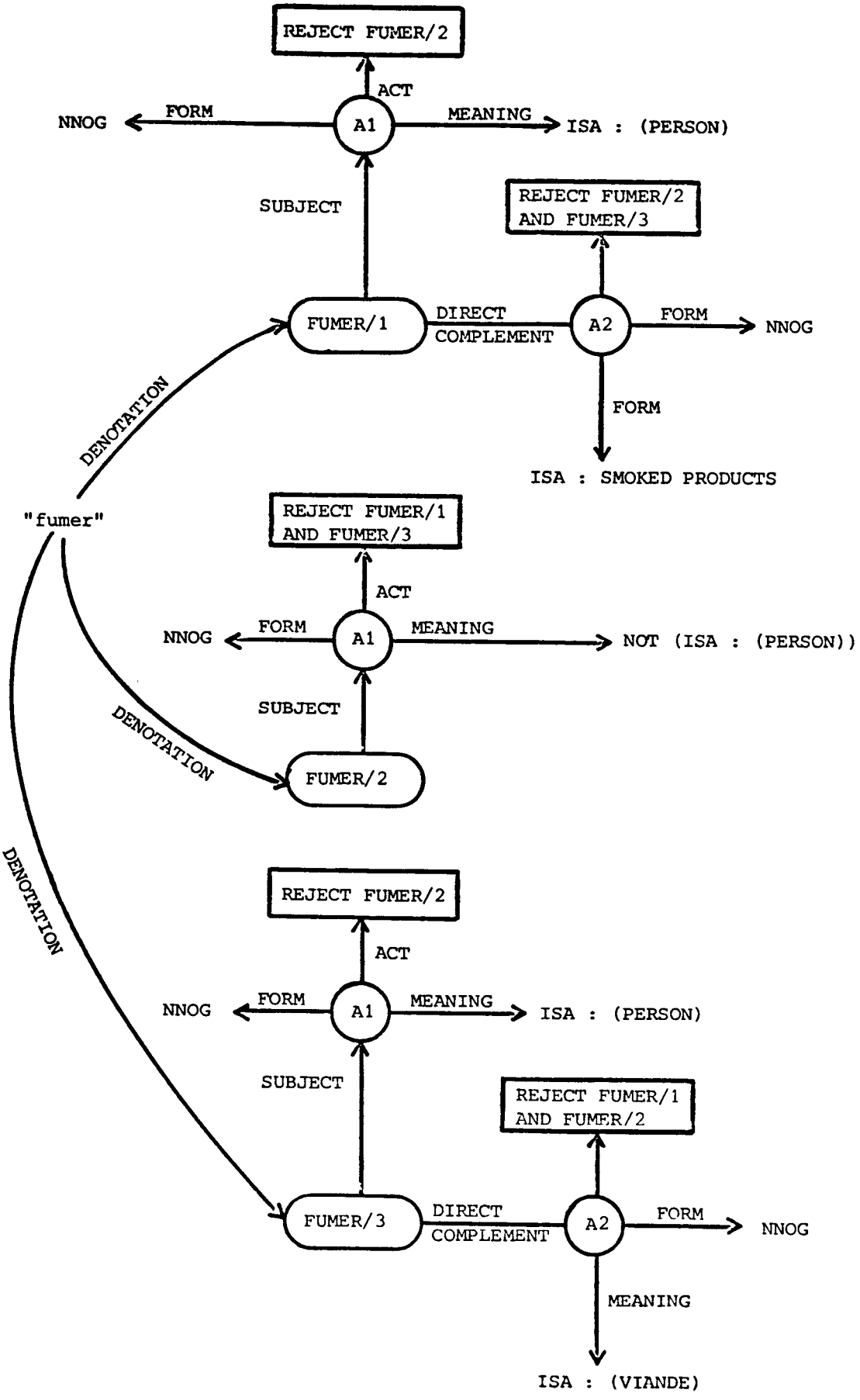

Figure 10. Verbal frame of the verb *fumer*.

tik und Datenverarbeitung. Sprachwissenschaft – Computerlinguistik 11. AQ, Dudweiler, W. Germany.

Findler, Nicholas 1979 *Associative Networks, Representation and Use of Knowledge by Computers.* Academic Press, New York, San Francisco, London, USA/Great Britain.

Fix, Hans; Rothkegel, Annely; and Stegentritt, Erwin, Eds. 1985 *Sprachen und Computer, Festschrift zum 75. Geburtstag von Hans Eggers. Sprachwissenschaft – Computerlinguistik 9.* AQ, Dudweiler, W. Germany.

Huckert, Edgar 1979 *Automatische Synthese des Französischen aus einer logischen Basis. Sprachwissenschaft – Computerlinguistik 2.* AQ, Dudweiler, W. Germany.

Marcus, Mitchell P. 1981 *A Theory of Syntactic Recognition for Natural Language.* MIT Press, Cambridge, London, USA/Great Britain.

Messerschmidt, Jan 1984 *Linguistische Datenverarbeitung mit COMSKEE.* Teubner, Stuttgart, W. Germany.

Mueller-v. Brochowski et al. 1981 *The Programming Language COMSKEE, 2nd Revised Report. Linguistische Arbeiten, Neue Folge,* 4. SFB 100, Saarbrücken, W. Germany.

Reding, Helene 1985 *Transformationskomponente für Baumstrukturen. Linguistische Arbeiten, Neue Folge,* 11. SFB 100, Saarbrücken, W. Germany.

Ritzke, Johannes 1981 Problème de l'analyse automatique des prépositions du français. In Schwarze, Christoph, Ed., *Analyse des prépositions. IIIe Colloque franco-allemand de linguistique théorique du 2 au 4 février 1981 à Constance.* Niemeyer, Tübingen, W. Germany: 139-157.

Ritzke, Johannes 1982 Zur automatischen Analyse adverbialer Relationen. In Fix et al. 1985: 287-307.

Ritzke, Johannes 1985 Automatische Analyse des Französischen: Komplexe Nominalstrukturen. In Figge, Udo L. 1985: 37-54.

Stegentritt, Erwin 1978 *MORPHO II B. Automatische derivationelle Analyse des Französischen. Sprachwissenschaft – Computerlinguistik 1.* AQ, Dudweiler, W. Germany.

Stegentritt, Erwin 1982 LAISSER + Infinitiv – Konstruktionen in der automatischen Analyse des Französischen. In Fix et al. 1985: 309-327.

Stegentritt, Erwin 1983 Aperçu d'une analyse automatique du français. In Cignoni, Laura and Peters, Carol, Eds., *Linguistica Computazionale, Vol. III. Supplement Actes du VIIe Congrès de l'ALLC, Pisa 1982:* 243-250.

Stegentritt, Erwin 1984 Automatische Subjektidentifizierung in französischen Infinitivsätzen des Typs FAIRE + Infinitiv. In Figge, Udo L., Ed., *Romanistik und Datenverarbeitung 1981. Akten der 10. Sektion des Deutschen Romanistentages, Regensberg 1981. Sprachwissenschaft – Computerlinguistik 7.* AQ, Dudweiler, W. Germany: 22-38.

Vauquois, Bernard 1975 *La traduction automatique à Grenoble.* Dunod, Paris, France.

Woods, William A. 1980 Cascaded ATN Grammars. *AJCL* 6(1): 1-12.