COLING 82, J. Horecký (ed.)
North-Holland Publishing Company
© Academia, 1982

PARSING GERMAN

Ingeborg Steinacker, Harald Trost
Department of Medical Cybernetics
University of Vienna

The first part of this paper is dedicated to an overview of the parser of the system VIE-LANG (Viennese Language Understanding System). The parser is a production system which uses an interleaved method that combines syntax and semantics. It parses directly into the internal representation of the system, without producing an intermediate syntactic structure. The last part discusses the relationship between some special features of the German language, and properties of the parser that originate in the language.

GENERAL APPROACH

A sentence is parsed word per word, from left to right. The parser is largely a data-driven production system. Productions involve the use of syntactic and semantic information at all major stages of the process. Noun phrases, for example, are recognized by an ATN which verifies the result of syntactic analysis semantically. It returns semantically valid NPs only. The parser belongs to the class of semantic parsers as suggested by [1], [4], [7]. It has two main sources of information: one is a semantic net, which propagates the information about selectional restrictions, the other is the parsing-lexicon, which for each word contains different senses associated with the information necessary to distinguish one sense from the others. Information includes syntactic features of the sentence (infinitive, surface-cases of dependent noun phrases ....), semantic restrictions and words that occur together with the input-word.

The productions make use of a correspondence between syntactic information in the sentence and the roles of the net (see chapter internal representation for an explanation of roles). Productions are used not only for generating the internal representation of constituents but also as expectations that guide the analysis of the rest of the sentence.

The generation of the internal structure corresponding to the sentence is centered around the verb. Since the representation of other constituents can be initiated independently of the verb, the parser builds a semantic structure immediately after a constituent is recognized. These structures are stored in a list, until the main verb of the sentence has been found. Then the parser tries to fill the case-slots of the verb with the given structures. The semantic categories of the structures have to be matched against the value restrictions of the roles of the verb.

INTERNAL REPRESENTATION

The source of semantic information is a Structural Inheritance Net [2]. This net formalism has the advantage of being epistemologically clear and explicit. SI-Nets are based on a strict discrimination between few structural components, and their content (what is represented). Real world knowledge is represented in the form of concepts and roles. Roles explain relationships between concepts. A concept is defined by its attributes which consist of two parts: the role and the value restriction. The value restriction is a concept which defines the range of possible fillers for the attribute, the role defines the function of a filler with regard to the concept being defined. Role-filler concepts can be regarded as semantic categories.

Generic concepts are organized in a hierarchy of super- and subconcepts. A subconcept inherits the attributes of the superconcept. If a concept has more than one superconcept it inherits the combined set of attributes. When processing an input individuals of the addressed concepts are instantiated. These individuals constitute the episodic layer of the net.

A word sense addresses either a concept or the attribute of a concept. If an input word relates to a concept, as most nouns and verbs do, that concept is instantiated. If it corresponds to a role both the concept and the attribute are instantiated, i. e. the generic concept, the role defining the attribute and the value restriction. Most adjectives and most prepositions are mapped into roles (size, colour, location, time, ...) but also some nouns (e.g. father is the role of a person in the concept family).

The net is structured in a way that facilitates the incorporation of results gained in linguistics: attributes of actions are defined in a way corresponding to cases of a case grammar. This can best be illustrated by an example: Actions are represented as net-concepts, e.g. DO. The concept DO is defined by attributes with roles like AGENT, OBJECT, GOAL, RESULT, that are restricted by adequate role-filler concepts. By defining attributes in this way a correspondence between surface cases in a sentence and roles of the net can easily be established.

THE PARSING-LEXICON

In the parsing-lexicon each word-sense is associated with productions. These productions reflect the correspondence between surface cases of the sentence and semantic cases within the net. The number of tests in a production correlates to the number of senses of a word. By executing these tests the parser gains the information necessary to choose the correct reading of a word. Tests check the syntactic and the semantic context in which an input word is found. Sometimes morphological information and the occurrence of certain words have to be taken into consideration as well. The range of tests reflects our general approach to parsing: combining syntax and semantics at all stages of the parsing process [8].

Depending on the stage of the process the failure of a test is interpreted in two ways. If the end of the sentence has been encountered the result is taken as false, if parsing is in progress the test is repeated at later stages of the process.

Actions associated with the tests mostly deal with semantic structure-building procedures. Some actions are used to control the parsing process. Usually the semantic structure for a constituent of the sentence is built after the constituent is recognized but actions can delay the creation of net-structures. The reasons for such a delay are explained in the following chapter.

A verb-sense is recognized by taking into consideration the syntactic surroundings of the verb and the semantic categories that match the selectional restrictions defined by the verb. After a verb-sense has been chosen expectations are built up regarding missing constituents. The occurrence of certain surface-structures also leads to the formation of expectations. Therefore tests that are associated with verbs first check the surface structure of the sentence (cases, prepositions...). The constituents that satisfy these syntactic tests have to fulfill semantic selectional restrictions. After having passed these tests, actions create the semantic representation for the verb and fill its roles with the selected constituents.

Unless an entry in the lexicon includes a test regarding subject and object of a sentence the following default actions are executed automatically: the subject of a sentence is mapped onto the AGENT and the object (accusative) is mapped onto the OBJECT of the action.

A Detailed Example

The two senses of 'gehen' in the following example can be disambiguated by using the entries in the parsing-lexicon listed below (parts of the entry which are irrelevant to the example are left out). These sample entries include important kinds of tests and actions.

(1) 'Ich gehe in den Park.' (I walk into the garden.)
(2) 'Der Bus geht nach Wien.' (The bus is bound for Vienna.)

```
gehen1 (move along)
C((CASE NOM) AND (RESTRICTION ANIMATE)) ->
    A(CRI LOCOMOTION)
gehen2 (bound for)
C((CASE NOM) AND (RESTRICTION PUBL.-TRANSPORT.)) ->
    A(CRI (PUBL.-TRANSPORT))
C(PLOC) ->
    A (CRV(+,DESTINATION,*)).
```

In the example the '+' parameter is an individual of the concept PUBL.-TRANSPORT, the '*' parameter is the location expressed by the prepositional phrase, namely Vienna. The nounphrase 'Ich' (I) fulfills the restriction ANIMATE, because speakers are always interpreted as humans.

Surface-tests:

Case-tests search for an NP of the surface-case indicated by the second parameter. If an NP is found that satisfies the condition, the tests that are connected by AND or OR to the case-test are executed. The constituent of the sentence which satisfies the tests is referred to with an asterix '*' in the associated action(s).

The test PLOC refers to a prepositional phrase that indicates some location. It is a test which uses syntactic and semantic information.

Restriction-tests:

These semantic tests are used to check selectional restrictions. They are often used in combination with syntactic tests. If both tests are met by a constituent this is a significant indicator, that the correct interpretation has been selected.

Structure-building Actions:

The action CRI(concept) creates an individual of the concept. The action CRV(pl,p2,p3) individuates an attribute of the concept pl. The concept pl, the role p2 and the concept P3 as value-restriction are instantiated. If pl or p3 are addressed by '+' the parameter refers to the first concept that was individuated when processing this particular entry in the parsing-lexicon. A '*'-parameter refers to the semantic representation of the constituent which satisfies the first test of the production.

SPECIAL FEATURES OF GERMAN

Morphological Ambiguities

We believe that making use of the interaction between syntax and semantics has many advantages over a strictly sequential approach to parsing. Introducing semantic information helps to resolve some ambiguities at an early stage of the analysis and thus to avoid unnecessary backtracking. Typically, morphological ambiguities can be resolved by such an interaction.

The German language is rich in inflectional forms, therefore the morphological component often comes up with more than one possible stem for an input word. These stems usually belong to different categories of words, e.g. 'meinen' can be interpreted as a verb (to suppose) or it can be reduced to the possessive pronoun 'mein' (my). Syntax restricts the type of a constituent, which is expected at a given point in the analysis. Usually it is sufficient to use syntactic information to disambiguate morphological ambiguities of this kind.

If a word is reduced to two different stems of the same category of words, selectional restrictions in the semantic net are used to choose one stem. The parsing-lexicon relates surface cases to semantic restrictions of the attributes of the action. In most cases this informaton is sufficient for disambiguation.

The inflected form 'gehoert' is reduced to the two verbs 'hoeren' (to hear) and 'gehoeren' (to belong to).

(3) Dieses Buch gehoert mir. (This is my book.)
(4) Hast du dieses Geraeusch gehoert?
(Did you hear that noise?)

In (3) the subject of the sentence has to be a 'POSSESSIBLE OBJECT', in (4) the object of has to be a subconcept of 'SOUND'. A violation of selectional restrictions. is a clear indicator that the wrong interpretation of the verb has been chosen.

Disconnected Constituents

Another characteristic feature of the German language is the verb second phenomenon. In German a verb can occupy three different positions within a sentence: the first in questions and commands, the second in main clauses, and the last in subordinate clauses. Compound predicates are divided into two parts. The auxiliary or the modal verb hold the place of the verb, and the rest of the predicate is put at the end of the sentence. One has to deal with a two-piece predicate whenever compound tenses are used, in structures involving the infinitive etc.

For a parser that uses a traditional approach of sequential syntactic and semantic processing these features cause extensive backtracking. The method of combinig syntactic and semantic analysis does not avoid backtracking completely but it makes re-interpretation easier. This claim is supported in the following paragraph using the example of a compound predicate.

(5) Mein Bruder hat das Buch, von dem du mir erzaehlt hast,
schon gelesen.
(My brother already read the book, which you told me about.)

In (5) the object and a relative clause separate the two parts of the predicate. One possible reading of the verb 'haben' is to possess. The object 'das Buch' satisfies the semantic restriction 'POSSESSIBLE-OBJECT', therefore 'hat' is taken as the predicate and a possess relation is established between the representations for subject and object. When the past participle 'gelesen' is encountered at the end of the sentence this decision has to be revised in favour of the compound predicate 'hat gelesen'.

The possess relation which was established has to be replaced by the concept that is addressed by 'lesen, namely 'INFORMATION-TRANSFER'. The semantic representations of the object book and the relative clause are not afflicted by this change. Book also fits into the hierarchy of 'INFORMATION-SOURCE' and therefore satisfies the selectional restrictions for the object of 'INFORMATION-TRANSFER' also.

Separable prefixes also add to the problem of finding the right verb. Syntactically verbadjuncts are particles, that are part of the verb. In some tenses a verbadjunct becomes separated from the verb and is put at the end of the clause. Verbadjuncts can specify the verb, but sometimes they change its sense completely (aufhoeren = to stop, hoeren = to hear).

(6) Das Kind hoert nach einer Stunde endlich zu weinen auf.
(After an hour the child finally stops crying.)

Such features either cause delay in the construction of the internal representation for a sentence, or they result in backtracking because the correct meaning of the verb becomes apparant at the end of the sentence.

CONCLUSION

The structure of the German language adds some difficulties to the general problem of parsing natural language. Flexible word-order and multiple sources for ambiguities led us to choose a data-driven approach. Syntactic and semantic information are used for disambiguation of existing structures and for expectations that control processing of new input.

Since backtracking is inevitable in some cases we tried to make it as efficient as possible. The integration of syntax and semantics facilitates backtracking to a large degree because semantic representations for all constituents are built independently. If backtracking occurs e.g. after having selected a wrong verb-sense, the parser has to destroy the existing semantic representation and replace it with the one indicated by the new verb. The slots of the instantiation of the concept for the new verb have to be filled with the already existing structures instead of having to start the parse all over again.

ACKNOWLEDGEMENTS

This research is part of the project 'Development of a NLU System with regard to Medical Applications' (supervision R.Trappl), which is supported by the Austrian 'Fonds zur Foerderung der wissenschaftlichen Forschung', grant #4158.

REFERENCES

[1] Boguraev, B.K., Automatic Resolution of Linguistic Ambiguities, University of Cambridge, (1979).
[2] Brachmann, R.J., A Structural Paradigm for Representing Knowledge (Bolt Beranek and Newman, 1978).
[3] Buchberger, E., Steinacker, I., Trappl, R., Trost, H., Leinfellner, E., A NLU System for Medical Applications, SIGART 79 (1982), 146-147.
[4] Gershman, A.V., Knowledge-Based Parsing, Yale Univ., RR-156,(1979).
[5] Oden, G., On the Use of Semantic Constraints in Guiding Syntactic Analysis, Univ. of Wisconsin, WR-3, (1978).
[6] Palmer, M., A Case for Rule-driven Semantic Processing, in: Proceedings of the 19th Ann. Meeting of the ACL, Stanford, (1981).
[7] Riesbeck, C.K. and Schank, R.C., Comprehension by Computer: Expectation-based Analysis of Sentences in Context, Yale University, RR-78, (1976).
[8] Steinacker, I., Parsing between Syntax and Semantics, Automatische Sprachverarbeitung 81, Potsdam, (1981).
[9] Trost, H. and Steinacker, I., The Role of Roles, Some Aspects of World Knowledge Representation, in: Proc. of the 7th Int'l. Joint Conf. on Artificial Intelligence, Vancouver, (1981).
[10]Wilks, Y., Processing Case, Am. J. of Computational Linguistics 4,(1976).