AUTOMATIC CONSTRUCTION OF A KNOWLEDGE BASE BY ANALYSING TEXTS IN NATURAL LANGUAGE

Werner FREY, Uwe REYLE and Christian ROHRER

Dep. of Linguistics, University of Stuttgart

ABSTRACT

We present a system which translates sentences from a subset of German into a database. This database will function as the basis for a question-answering-system.

The system is applied to a complete text and not to isolated sentences. As an intermediate stage between the German text and the database we use the Discourse Representation Structures (DRS) invented by Hans Kamp. Kamp's system has been chosen because it handles intrasentential and intersentential relations uniformly. Within Kamp's system one can account for certain types of anaphoric relations for which no other linguistic theory has provided a solution.

The input to our system is analysed by a parser which is based on lexical functional grammar. This is the first attempt to combine research on discourse representation with lexical functional grammar with the help of the formalism of Definite Clause Grammar.

For the construction of the database out of the DRS's, two solutions arc proposed. First, a translation of the DRS's into a set of PROLOG clauses enriched with some additional deductive principles. Second, the formulation of inference rules which operate directly on the DRS.

So far we have implemented the following components: parser of German, translation rules which map syntactic trees into DRS's and rules which translate DRS's into PROLOG-clauses.

I. The Fragment of German and the Parser

Our parser is based on the formalism of lexical functional grammar (LFG). The implementation of the LFG-parser itself is described in (Reyle/ Frey, 1983).

We have chosen LFG for four reasons:

a) LFG is based on research in theoretical and in computational linguistics.
b) LFG has already been applied to a variety of languages.

This report describes work done in the Department of Linguistics at the University of Stuttgart. It was supported in part by the German Science Foundation project Ro 245/12.

c) LFG does not require transformational rules.
d) The output of the parse (called 'f-structure') constitutes in form and content an excellent input to our semantic component.

Our fragment of German comprises the following constructions: relative clauses, conditional clauses, universally and existentially quantified noun phrases, sentence and constituent negation, personal pronouns and definite descriptions.

II. Semantic representation

A. Introduction

As semantic representation we use the discourse representation theory of Hans Kamp (Kamp, 1981 a)

Let D = $S_1, \ldots, S_n$ be a discourse of the fragment. A discourse representation structure (DRS) for D is constructed by successively parsing the sentences and translating the resulting f-structures into a DRS. The translation of an f-structure of a sentence S_i proceeds through the application of certain DRS-construction rules which operate, so to speak, 'from the outside in' on the f-values of the grammatical functions and features contained in the f-structure of S_i. Thus the order in which the rules are to be applied is imposed by the hierarchy of the f-structures. It is important to note that the construction of the DRS for D at the stage of processing the sentence S_i is done with respect to the DRS constructed for the sentences $S_1, \ldots, S_{i-1}$.

A DRS for a discourse D consists of a partially ordered set of D(iscourse) R(epresentations) which in turn consist of:

(i) a set of discourse referents: discourse individuals, discourse states, discourse events and discourse times.
(ii) a set of atomic conditions of the following types: Let u and v be discourse individuals.
  (a) α(u), α(u,v), where α is an I(ntransitive) V(erb) or T(ransitive) V(erb) respectively.
  (b) α(u), u = β, where α is a C(ommon) N(oun) and β a P(roper) N(oun).
  (c) conditions handling temporal relations, which we will ignore in this paper (see Kamp, 1981 b).

B. Translation of f-structures into DRS's

The following natural language constructions of the fragment trigger via their encoding in the

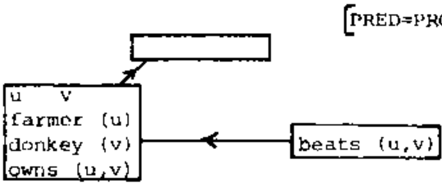
the f-structure the construction of the DRS in the following way: Universal quantifiers, conditional sentences and negation introduce subordinate DR's. Quantified common nouns and proper nouns introduce discourse referents. Pronouns do not introduce new discourse referents but must refer to a discourse referent which has already been introduced. Relative pronouns pick up the reference introduced by the nounphrase which dominates the relative clause. Personal pronouns and definite descriptions pick up a discourse referent which was introduced before and which is accessible in the partial order of DR's.

For the construction of the DRS out of the output of our LFG-parser (for a detailed description of the parsing mechanism see Reyle/ Frey, 1983) consider the following example:

(1) Every farmer who owns a donkey beats it.

```
(1')
PRED='beat< [SPEC=every
             PRED='farmer
             NUM=sing.
             ADJ= [PRED='own<RPRO, [SPEC=a
                                    PRED='donkey
                                    NUM=sing.   ]>]]
                                   [PRED=PRO]>
```

(1'')



The names of the grammatical [unctions and features correspond to procedures in our program. In (1') this means that we will have the following procedures: PRED, SPEC, ADJ, NUM. The SPEC procedure builds up the DR configuration and introduces discourse referents. Furthermore it is responsible for the representation of quantifier scope ambiguities.(Scope ambiguities cannot be represented in f-strur:tures (for good reasons). Scope ambiguity is a semantic phenomenon and must be handled in the semantic representation.) The PRED procedure introduces atomic conditions on the discourse referents. ADJ guarantees the correct semantic relationship between the relative clause and the common noun and NUM specifies the discourse referents to be discourse individuals or discourse sets. Crucial for the program is the ability to build up structures which are not fully instantiated and leave unspecified parts as variables (see Pereira, Warren, 1980).

C. Truth conditions for DRS's

The DRS-construction rules are such that they allow for a uniform treatment of indefinite descriptions in arbitrary contexts. Consider the sentences:

(2) A farmer owns a donkey.

(3) If a farmer owns a donkey, he beats it.

In (2) the indefinite phrase 'a donkey' is to be interpreted as existentially quantified, whereas in (3) it has a universally quantified reading. By the DRS-rules indefinite descriptions are interpreted as referential terms: they always introduce a discourse referent.

The resulting DRS's are:

```
(2')  | u    v
      | farmer  (u)
      | donkey  (v) 1
      | [Owns  (u rY l
```

(3') the same as (l'') above.

Whether the discourse referent introduced by an indefinite description implies existence or not, depends on its position in the partial ordering of the DR's. Exactly those referents introduced in the main DR's imply existence; all other referents are to be'interpreted as universally quantified, if they occur in the antecedent-DR of a pair of sub-DR's, or as existentially quantified, if they occur in the consequent-DR. Of course the interpretation of the latter, i. e., the existentially quantified referents depends on the interprctation of all universally quantified referents accessible from them. Thus the role of the partial ordering of the DR's is twofold. First it forms the base for the definition of truth--conditions on the DRS. Second it defines an accessibility relation '-<—', which restricts the set of possible antecedents of a pronoun or definite description to the accessible ones. The truth conditions of a DRS are necessarily recursive. Recall that each DR consists of a set of individuals and atomic conditions, i. e., it can be considered as a partial model. A partial function f from a DR m into a model M is called a proper embedding, if it is a homomorphism of in into M-with respect to the- atomic conditions of m. Suppose 1 satisfies the predecessor of a sub-DR m of k - if there is one; if not let 1 be the empty function. Then g is said to satisfy m, if it is a proper embedding of m that extends 1, and if m dominates two sub-DR's m_1 and m_2 then additionally every k that extends g and satisfies m_1 must be extendable to a h that, satisfies m_2,. Let us call the union of the 'highest' DR's of a"DRS k the main DR of k. Then k is true in M, if there is a proper embedding of the main DR of k into M (for an exact, formulation of the truth conditions see (Kamp, 1981 a)).

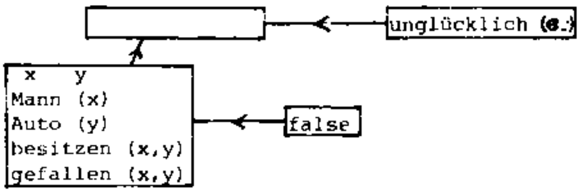
Two remarks, are in order. First, the truth--definition of a DRS of a discourse D is defined for the whole discourse and not for the conjunction of the individual sentence's that occur in D. This is necessary in order to account for intersentenfial dependencies like anaphorical relations. Second, in spite of the fact that we do not. want to check the truth of a discourse D in a given model the exact-model theoretic interpretation of the DRS's is crucial for our aims to build up an (intentional) database out of D. The reason is that in order to get correct answers to questions abaout such a database, containing facts and rules, the structures upon which it will be built up must *be* logically transparent.

D. Anaphoric pronouns and definite description

In analysing a pronoun one has to look for a suitable discourse referent which is to be substituted for the pronoun. The set of candidates is restricted by the partial ordering of DR's. Only discourse referents, which are accessible from the DR under construction, are possible antecedents (see example (Consider however sentence (4):

(4) Wenn kein Mann ein Auto besitzt, das ihm gefallt, dann 1st er unglucklich.
(If no man owns a car he likes,then he is unhappy

The corresponding DRS looks like:



(We represent negation by means of a DR false for which no proper embedding exists.)
It is not possible to find an antecedent for the second pronoun in example (4). Therefore a text which consists only of this sentence is not wellformed.

Another problem which should be tackled intersententially are definite descriptions. The sentence "the mad hatter hates Alice" is only used correctly if a hatter was mentioned in the preceding discourse and if it is clear to the reader that there is only one mad hatter. Therefore, in order to analyse definite descriptions we have to take into account the preceding discourse and we have to check whether the existence and uniqueness conditions are fulfilled. For the DRS Theory this means that we have to look for a discourse referent introduced in the preceding DRS for which the description holds and we have to check whether this description holds for one referent only. Our algorithm proceeds as follows: First an intermediate stage is built up where the definite description is replaced by a variable over discourse referents. The attributes of the definite description are given to a search procedure. This procedure succeeds if one and only one match with a subset of the attributes of an already introduced accessible discourse referent is possible. A text is coherent only if all its variables over discourse referents can be resolved.

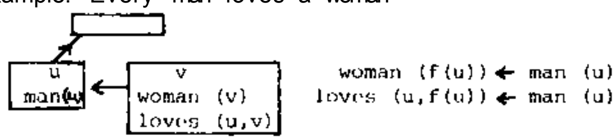
In our approach, contrary to others, it is not necessary to expand the representation of the sentence containing the definite description with clauses stating the existence and uniqueness conditions.

111. From DRS to database

From the model theoretic interpretation of the DRS it follows that the atomic conditions contained in the main DR express the facts asserted in the discourse. The information contained in the sub-DR's express the rules.
In order to answer queries about the content of the discourse represented by the DRS we consider two possibilities:
(I) translating the DRS into a PROLOG - database
(II) formulating inference rules which operate directly on the DRS.
Both lead to a logical database.
The extensional part (i.e. the facts) and the intensional part (i.e. the rules viewed as integrity constraints) are expressed in the same language.

*Surely this is a special case. The attributes of the definite description could be hyponomious to the ones with which the match works. Furthermore the required discourse referent does not have to appear in the DRS representing the text; it could appear in a presupposed DRS which represents world knowledge.

ad (I): For the translation of the DRS's into PROLOG-clauscs we developped an algorithm which operates directly on the DRS's. Thus it is not necessary to translate the DRS's first into predicate logic.
If only the consequent-DR*s of a DRS introduce a new pair of sub-DR's the DRS is equivalent to a set. of Horn-clauses. The translation procedure replaces the discourse referents introduced in the main DR and in the consequent-DR's by their corresponding Skolem functions. All other discourse referents are replaced by variables.
If the consequent-DR's contain more than one atomic formula, each of them gives rise to a separate Horn-clause.
Example: Every man loves a woman



DRS PROLOG-clauses

For arbitrarily branching DR's which are not equivalent to Horn-clauses the algorithm is more complicated. In this case it is not convenient to translate into clausal form, because the procedural interpretation of the resulting clauses depends on the Skolem function (appearing in the antecedents) the extension of which is not represented in the programme. The clausal forms for the sentence
Jede Frau, die jeden Mann liebt, den sie kennt, ist uberfordert.
(Every woman who loves every man she knows is stressed).
are:

iiberfordert (x)<--Frau (x) , not (Mann (f(x))
iiberfordert (x) <- Frau (x) , not (kennen (x,f(x)))
iiberforderf (x) <--Frau (x) , lieben (x,f(x))

To get a programme we have to translate into

iiberfordert (x) <- Frau (x) , not (Mann (y) ^
kennen (x,y) ^ not (lieben (x,y)))

(with 'not' interpreted as 'negation as failure').
Hence one has to take into account special PROLOG requirements and has to build up new representations. Furthermore the programme has to be enriched with special 'meta'-procedures to render adequate answers (e.g. a rule yielding the contraposition of a clause).
ad (II): Because of the problems mentioned under (I) we want to investigate the possibility of using the DRS itself - together with inference rules operating on it - as database. A set of inference rules has been proposed by Hans Kamp.

REFERENCES

Kamp, H., "A Theory of Truth and Semantic Representation". In: J.A. Groenendijk (ed.), Formal Methods in the Study of Natural Language I., 1981

Kamp, H., "Evenements, Representations discursives et References temporelle". In: Languages 64, 1981

Reyle,U./Frey,W., "A Prolog Implementation of Lexical Functional Grammar". In: this volume

Warren, D.H.D./ Pereira, F.C.N., "Definite Clause Grammars for Language Analysis". In: Artificial Intelligence 13, 1980, 213-278