On the Form of Lexical Entries and their Use in the Construction of Discourse Representation Structures*

Hans Kamp and Antje Roßdeutscher
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Keplerstr. 17
D-7000 Stuttgart 1

1. Introduction

The central task of natural language semantics is to articulate how meaning is determined by form. From a computational perspective this task divides into two. On the one hand a computationally viable semantics should describe the formal aspects of interpretation: it should describe how the meaning of a natural language expression or text can be culled from its recognizable form - in other words, how a recognizer of form, that is, a (syntactic) parser, can be paired with an algorithm which converts the structures that the parser computes into specifications of content. On the other hand, the theory should account for language generation, i.e., for the conversion of content specifications into actual sentences or texts. Here we only consider the problem of interpretation.

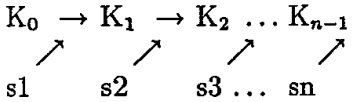
Traditionally the problem of how form determines meaning is regarded as capable of factorization into two separate problems, a lexical and a structural problem. The lexical problem is that of determining the meaning of what is recognizable by its form as a particular word or lexical item. One way of solving this type of problem in daily life is to look up the given word in a lexicon or dictionary - indeed, this is what most dictionaries are designed for. The second problem is that of determining how the meanings of the individual words of which a complex expression (such as in particular a sentence or text) is made up jointly produce the meaning of the expression (the sentence, text, or whatever) as a whole.

Syntactic structure is essential to the second problem. For it is primarily the way in which the words occurring in a complex fit together syntactically which determines how their meanings combine into the meaning of the entire expression. For long formal semantics of natural language was almost exclusively concerned with this second problem—i.e., with analyzing how the syntax of a complex expression determines its meaning as composed from the words contained in it. The analysis of the meanings of individual words was left to the lexicographers and lexical theorists, while formal semantics itself made do with very general assumptions about what sorts of things the meanings of lexical items are. This method can go a surprisingly long way towards providing insight into the form-meaning relation of complex expressions. Nevertheless it is evident that the semantic description of an individual sentence or text which differentiates it from others with parallel syntactic structure, but built from different words, will have to supply more than a purely schematic account of the particular lexical items the sentence or text contains. A semantics able to provide this kind of differentiation must combine the insights of formal semantics with those of a theoretically informed lexicography.

*The research reported in this paper is being supported by the DFG in the context of Sonderforschungsbereich 340. We thank Nicholas Asher, Franz Beil and many other colleagues in the IMS for suggestions, criticism and other forms of valuable assistance.

The very project of computational semantics - that of associating specifications of content, or semantic representations, with syntactic structures - derives its point from the conviction (which goes back at least as far as Frege) that syntactic structure tends to conceal semantic and logical structure. In particular, accounting for the logical relations between different sentences or texts in terms of their syntactic structure directly tends to be very awkward and can as a rule be accomplished much more perspicuously using derived structures, usually referred to as semantic representations or logical forms. There is also another, arguably more compelling reason why syntactic structures cannot serve as representations of meaning. According to generally accepted views of syntax the unit of syntactic representation is simply the sequence consisting of the unrelated syntactic structures of the sentences which compose it. In contrast, the unit of semantic analysis is not the individual sentence but the coherent discourse or text: the meaning of a coherent text usually resides partly in certain connections between the sentences of which it consists, and thus transcends the meanings which those sentences can be recognized as having on their own. A sequence of unrelated syntactic structures is unable to capture those connections.

The approach towards semantics on which the present paper is based was developed with the specific aim of accounting for the intersentential connections between sentences that distinguish coherent texts from mere concatenations of sentences. This approach - Discourse Representation Theory, or DRT for short - analyzes how the meaning of the i-th sentence s_i of a text $s_1,\ldots,s_n$ is determined by the syntactic structure of s_i together with the context K_{i-1}, which in turn is determined, partly or wholly, by the meanings of the preceding sentences $s_1,\ldots,s_{i-1}$. In fact, the ways in which s_i is semantically connected with the antecedent part of the text often make it impossible for its meaning to be represented in isolation, and the only possibility of capturing it lies in showing how the meaning representation K is updated through incorporation of the information which s_i contributes. Thus we arrive at the following text interpretation schema:



K_n captures the meaning of the entire text $s_1, \ldots, s_n$. The representations K_i are called *Discourse Representation Structures* or *DRS*s.

Let us look at an example to see how this might work in some detail. Consider the following three sentence "text":

(1) Der Tourist erkrankte an Typhus. Nach drei Wochen war er wieder gesund. Ein Arzt aus Izmir hat ihn geheilt.

(The tourist came down with typhoid. After three weeks he was well again. A doctor from Izmir has cured him.)

We will assume for simplicity that the starting representation K is empty. So the first sentence is to be interpreted and represented free of context. The processing of the sentence (in fact this is true for the great majority of sentences; it is true in particular for the two remaining sentences of (1)) consists in introducing representatives - so-called *discourse referents* - for the arguments of the verb. On the one hand these are inserted into the argument positions of the verb; on the other hand they act as arguments to conditions which express the descriptive information contained in the argument phrases themselves. A quite distinct assumption that is crucial to the representation of (1) concerns the referential arguments of verbs: we assume that the verb of a sentence has, in addition to the arguments expressed by actual argument phrases, also, as an implicit argument, an event (or state of affairs, as the case may be); this is an event of the type defined by the verb together with its other argument(s). The verb's tense determines, sometimes in cooperation with temporal adverbs, prepositional phrases or subordinate clauses, where this event or state of affairs is located in time, relative to the utterance time of the represented discourse or text and/or with respect to other times

or events that are already present in the representation. Thus the first sentence of (1) is analyzed as saying "There exists an event before the utterance time n (n stands for "now") which is an event of the tourist coming down with typhoid". In the notation customary for DRSs this analysis is represented as follows:

(2)

n e y z
der Tourist(y) Typhus(z)
e < n e: y erkranken an z

(The condition "e: y erkranken an z" expresses that e is an event consisting of y coming down with z.)

Similarly the second sentence is analyzed as saying that there is a state of affairs s before n which is located three weeks after the event e and which consists in the tourist being healthy. Note that in order to arrive at this analysis the second sentence has to be related to the (DRS of the) first sentence in several ways. First, a connection must be made between the pronoun *er* and the tourist mentioned in the first sentence, or—this is how anaphora is viewed in DRT—between *er* and the discourse referent y that represents the tourist in (2). Second, the prepositional phrase *nach drei Wochen* has to be interpreted as denoting a time three weeks after event e. (There is also a third connection, which has to do with the adverb *wieder*. We will ignore this connection for now, but will come back to it below.) The update of (2) with this new information is as in (3):

(3)

n e y z s u
der Tourist(y) Typhus(z) u = y
e < n e: y erkranken an z
s < n $\lvert[e,s]\rvert_{weeks} = 3$ s: u gesund sein

("$\lvert[e,s]\rvert_{weeks} = 3$" expresses the fact that the temporal distance between e and s measured in weeks is equal to 3.)

The effect of updating (3) with the information contained in the third sentence is given in (4):

(4)

n e y z s u e′ x v w
der Tourist(y) Typhus(z) u = y v = y
Izmir(w) Arzt aus w(x)
e < n e: y erkranken an z
s < n $\lvert[e,s]\rvert_{weeks} = 3$ s: u gesund sein
e′ < n e′ < s e′: x heilen v

It should be noted that the total meaning which (4) represents cannot be analyzed as a conjunction of three propositions expressed by the three sentences of (1). To see this, one has to know that a DRS is to be read as an assertion to the effect that there exist entities corresponding to its discourse referents (displayed in the top line) which satisfy all the conditions that are specified for those dicourse referents in the lines below. Thus (4) means that there was a time before the utterance time n at which the given tourist came down with typhoid and a later time at which he was healthy again and that there was a doctor from Izmir who cured the tourist. In other words, (4) amounts to an existentially quantified conjunction with conjuncts coming from the different sentences of (1). It is a fact well-known from symbolic logic that such an existential formula is normally not equivalent to a conjunction of formulas that correspond to the individual conjuncts.

DRSs such as (2) - (4) suffer from a fundamental deficiency. In fact, this is generally true of DRSs as they have been constructed until recently (see for instance Kamp (1981), Kamp and Reyle (1991)).

The deficiency is this: the DRS-conditions "e: y erkranken an z", "s: u gesund sein" or "Arzt(x)" of (4) are treated as unanalyzed predicates, unrelated to any lexicon or lexical theory to account for the specifics of their meanings or the semantical and logical relations between them. DRSs of this kind reflect the perspective of formal semantics as sketched above, and are deficient for the reason given there. It was to remedy this lacuna that the work we will outline in the next two sections has been undertaken.

2. The syntax and semantics of lexical entries

The lexical entries which we will discuss concern a small number of related words, such as *heilen, gesunden, gesund, erkranken*. Some of these are found in our sample discourse (1). According to the analysis we will offer all these words relate to a central concept which we will denote as HEILEN. This concept is expressed directly in an intransitive sentence such as (5):

(5) Der Fuß heilte.

Our lexicon has an entry for this intransitive use of the verb *heilen* which looks as follows:

(6)

heil-en	$\{\langle\theta 1, f1\ \rangle\}$
	Theme1
HEILEN	SEL RESTR:
(e, y_{th1}, z_{th2})	bodily part

Heilen can also be used transitively. In fact, it has several transitive uses which our approach must distinguish. One of them is exemplified by the sentence

(7) Ein Arzt aus Izmir hat den Touristen vom Typhus geheilt.
(A physician from Izmir has cured the tourist of typhoid.)

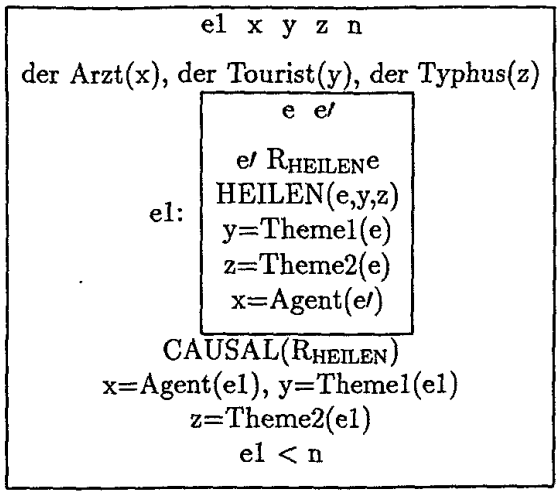
The lexical entry corresponding to (7) is given in (8)

(8)

heil-en	$\{d\langle\theta 1, f1\rangle$	$\langle\theta 2, f2\rangle\}$	von+DAT
	Agent	Theme1	Theme2
CAUS(HEILEN)	SEL RESTR:	SEL RESTR:	SEL RESTR:
$(ec, x_{ag}, y_{th1}, z_{th2})$	capable of intention to cure	organism or bodily part	ailment or disease

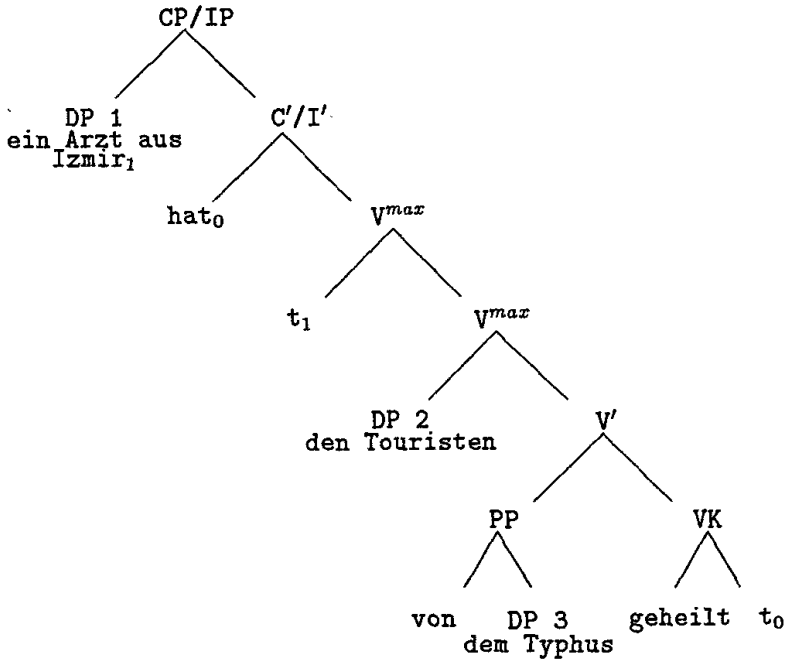
Like the vast majority of entries in the lexicon we are developing, (6) and (8) consist of a syntactic component (the top tier) and a conceptual one (the bottom tier). The conceptual component is often shared, in part or in whole, with several other entries. It is this component which contains the information that should flow into the DRSs which represent sentences or texts containing the lexical items in question. The syntactic component of an entry is specific to that entry and serves to identify it as corresponding to particular uses of the particular words. The introduction of lexical information into the DRS involves replacing a lexical item occurring in some minimal DRS condition—such as e.g. the condition "e*l*: x heilen v" from (4)—by the conceptual component of the relevant lexical entry, while substituting discourse referents for some or all of the argument symbols occurring in that component. For instance the DRS for (7) will not contain the word *heilen*, but rather the conceptual structure the verb expresses and so will look like this:

(9)



Let us have a closer look at how (9) is obtained. In section 1 we implied that DRSs are constructed from syntactic structures. More specifically, we assume that these syntactic structures are the ones postulated in Frey and Tappe (1991). (The theory which postulates these structures - an interpretation of X-bar-Theory for German - is being developed in a project which, like our own, belongs to a larger research effort - the SFB 340: Sprachtheoretische Grundlagen für die Computerlinguistik, located at the Universities of Stuttgart and Tübingen.) According to the theory of Frey and Tappe the syntactic structure of (7) has the following form:

(10)



In syntactic trees of this sort the arguments of the verb are readily identified by their positions: They are those phrases whose base positions are subordinate to the maximal projection (the Vmax node) of the verb. In order to replace the verb *heilen* by the corresponding conceptual component in the DRS built from (10), the parser, which has already constructed (10), must match the verb arguments of (10) with those of a suitable lexical entry for this verb. In the present instance the parser will select (8) as the correct entry, because it can recognize the arguments specified by that entry as matching those found in (10). (We omit the details) The match enables the DRS

construction procedure to replace the lexical verb in the DRS with the conceptual component of (8) while substituting for the symbols filling the argument positions of that component in (8) the discourse referents that represent the matching arguments of (10).

In the present case this does not yet yield the DRS given in (9), which has a complex event structure in lieu of the condition CAUS(HEILEN)(e1, x,y,z) which we find in (8). To obtain (9) as specified, the construction procedure must appeal to an *axiom* belonging to the theory which provides the logical foundation for the lexicon; we will refer to this theory as LT (for "Lexical Theory"). The axiom in question spells out the principal implication of the assumption that transitive *heilen* denotes the causative of the concept HEILEN:

(11) $(\text{CAUS(HEILEN)})(\text{ec},x_{ag},y_{th1},z_{th2}) \Longleftrightarrow \text{ec}:$ $\dfrac{\boxed{\begin{array}{c} e\prime\ \ e \\ e\prime\ R_{\text{HEILEN}}\ e \\ \text{Agent}(e\prime)=x_{ag} \\ \text{HEILEN}(e,y_{th1},z_{th2}) \end{array}}}{\text{CAUSAL}(R_{\text{HEILEN}})}$

This axiom expresses the notion that the subject of *heilen* as it is used in (7) does something —i.e. is involved in some event e'— which causes an event e instantiating the concept HEILEN (as we find it for instance in (6)) and that the other two arguments of the verb are the two participants of e. However, transitive *heilen*, like its English counterpart *cure*, involves more than just causing someone to get better. Lexicalized causatives – i.e. causative concepts that are expressed by a single verb rather than by the compound "cause + intransitive verb" – involve not just any causal relation, but one of a special "direct" and/or "intended" kind. It has proved extremely difficult to articulate what distinguishes the causal relations which are implied by lexicalized causatives from causal relations in general. We will assume that the relations we need are uniquely determined by the concepts C and accordingly represent them as R_C. The fact that R_C is causal, i.e. satisfies the predicate CAUSAL, implies that any events related by it are ipso facto causally related. This is expressed by another axiom of LT:

(12) $\boxed{\begin{array}{c} \text{e1}\ \ \text{e2} \\ \text{CAUSAL(R)} \\ \text{e1 R e2} \end{array}} \Longrightarrow \text{e1 CAUSE e2}$

A further fact about the causative of HEILEN is that any one of its instances ec entails the occurrence of a process e instantiating the concept HEILEN:

(13) $(\text{CAUS(HEILEN)})(\text{ec},x_{ag},y_{th1},z_{th2}) \Longrightarrow \text{HEILEN}(e,y_{th1},z_{th2})$

3. Inferences on DRSs

There is one aspect of the lexical entries (6) and (8) which we have not yet explained. This is the use of specifications such as *Agent*, *Theme1* and *Theme2*. These terms denote *thematic roles*. Thematic roles are functions from the events or states that instantiate the concepts expressed by verbs to certain participants of those events or states. These functions turn out to be useful in capturing certain generalizations over clusters of concepts. For instance, the concept HEILEN is one of a number of process concepts C all of which involve a participant that has a particular property RES(C) once the process is completed which it did not have when the process started. For the case of HEILEN this implication is captured by the following pair of axioms:

(14)

e HEILEN(e)	⇒	s1 s1:¬ (RES(HEILEN))(Theme1(e))) s1)(e	
e HEILEN(e)	⇒	s2 s2: (RES(HEILEN))(Theme1(e))) e)(s2	

("s1)(e" is to be read as "s1 abuts e".)

Another feature of HEILEN is that its two themes—the patient and the ailment—are "separated" at the time the process is complete. In this respect HEILEN resembles a number of other lexicalized concepts, among them concepts of cleaning, where the separation between Theme1, the thing cleaned, and Theme2, that of which the thing is cleaned, admits of a more literal interpretation. For concepts C belonging to this class we denote the associated relation as SEP(C). The relation between C and SEP(C) can also be captured in the form of a pair of axioms of LT. We only give the special instances for the case where C is HEILEN

(15)

e HEILEN(e)	⇒	s1 s1: ¬ (SEP(HEILEN)) (Theme1(e),Theme2(e)) s1)(e	
e HEILEN(e)	⇒	s2 s2: (SEP(HEILEN)) (Theme1(e),Theme2(e)) e)(s2	

(As (15) shows, the language of LT must be able to represent polyadic as well as monadic predicates)

To conclude, we will illustrate how DRSs can be used, in combination with the lexicon, to verify certain inferential relations between sentences and texts. As an example consider our original text (1):

(1) Der Tourist erkrankte an Typhus. Nach drei Wochen war er wieder gesund. Ein Arzt aus Izmir hat ihn geheilt.

Intuitively (1) entails

(16) Der Tourist gesundete vom Typhus.

(17) Der Arzt hat den Touristen vom Typhus geheilt.

In order to verify the implication to (16) we need lexical entries for some of the other words occurring in (1). First the entry for *erkranken*.

(18)

erkrank-en	{⟨θ1,f1⟩ }	an+DAT
	Theme1	Theme2
(ANT(HEILEN))	SEL RESTR:	SEL RESTR:
(e,y_{th1},z_{th2})	organism	disease

Erkranken is the antonym of the verb *gesunden*, which instantiates the concept HEILEN(e, y, z), as illustrated by (16). This relation of antynomy is one which we find between many pairs of lexical items. In case those lexical items denote processes, the relation has the - obvious - logical properties given in

(19) If C is a process concept such that s: RES(C)(theme$_i$(e)),
then ANT(C) is a process concept such that s: ¬RES(C)(theme$_i$(e))

Since *gesunden* directly expresses the concept HEILEN, its antonym *erkranken* expresses the concept ANT(HEILEN). Instantiating (19) to this concept gives:

(20)

e (ANT(HEILEN))(e)	⇒	s1 s1: (RES(HEILEN))(Theme1(e))) s1)(e

e (ANT(HEILEN))(e)	⇒	s2 s2: ¬(RES(HEILEN))(Theme1(e))) e)(s2

A second set of implications involving ANT(HEILEN) derives from the separation axioms (15) for HEILEN. Since ANT(HEILEN) is the reverse of HEILEN, its result state is one in which the two themes, patient and ailment, are no longer "separate":

(21)

e (ANT(HEILEN))(e)	⇒	s1 s1: (SEP(HEILEN)) (Theme1(e),Theme2(e)) s1)(e

e (ANT(HEILEN))(e)	⇒	s2 s2: ¬(SEP(HEILEN)) (Theme1(e),Theme2(e)) e)(s2

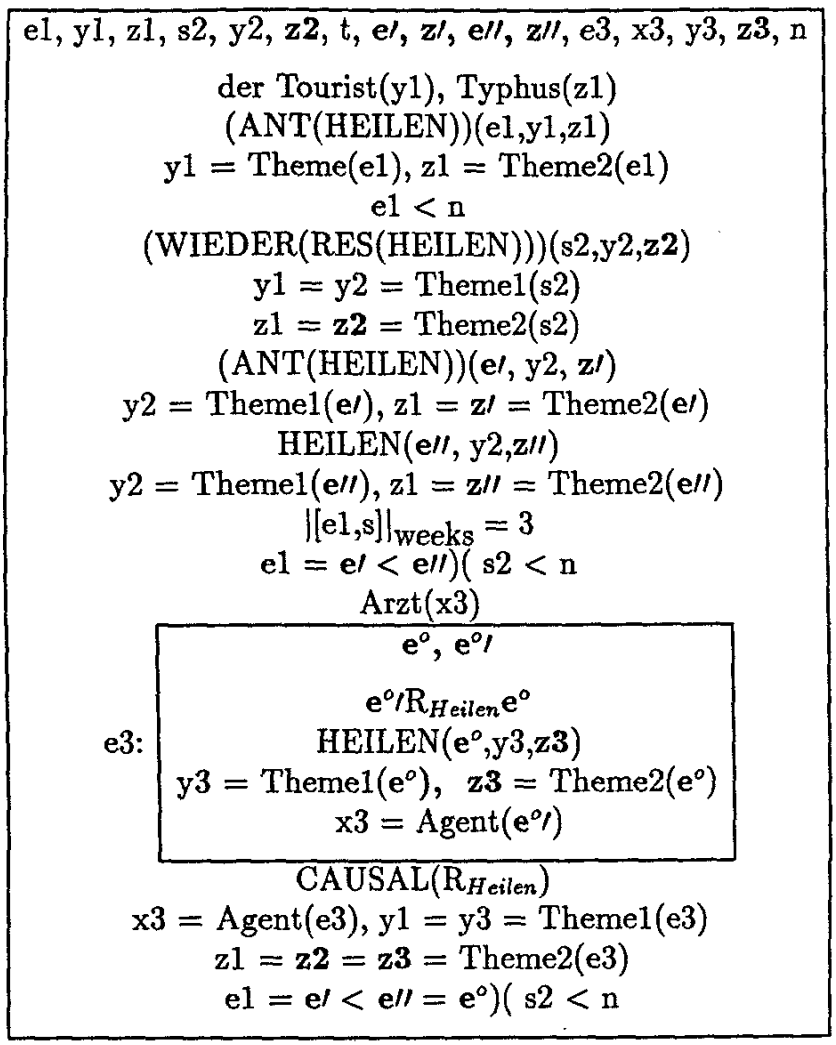
With the help of (18) we can construct a representation for the first sentence of (1). To extend this DRS to a representation of the second sentence we also need an entry for the adjective *gesund*. As *gesund* denotes the state which results from a process of HEILEN (though it need not necessarily arise in this way), its entry can be represented in the following form:

(22)

gesund	{Θ}
(RES(HEILEN)) (y_{th1})	Theme1 SEL RESTR organism or bodily part

To process the second sentence of (1) we need to know something about the word *wieder*. All that is needed in the present context is this: When *wieder* is part of a clause describing a state of affairs s, it entails that a state s/ answering to the same description obtained at some earlier time t_1 and that at some intermediate time t_2 a state s// obtained which did not answer the description. In the present case the entailment is that our tourist mentioned in the first sentence, y1, had been healthy at some time in the past and that since then he had been ill. These implications are supported by the DRS for the first sentence of (1) (which acts as context for the interpretation of the second sentence). For this DRS entails, via the implications given in (20), that y1 was healthy before he caught typhoid and then was not healthy once he had caught it. In the light of what has been said it should now be clear how the second sentence can be processed. As the third sentence presents no further problems, we give the DRS for the entire three sentence text, without showing the intermediate stages:

(23)



(The boldface discourse referents are ones that belong to the concepts but are not explicitly mentioned in the sentences. See Kamp and Rodeutscher (1991) for details.)

Note that (23) supports the inference that between the time of the completion of the process e1 described by the first sentence and the state s2 described in the second y1 changed from a state of non-health into one of health. Moreover, (21) tells us that the first of these two states is one in which y1 is ill with typhoid. On the assumption of the existence of such a change, i. e., of the existence of a process that leads from the first state to the second, a discourse referent e// with the appropriate conditions has been added to the DRS. The thus expanded DRS contains a smaller DRS, representing the sentence (16), as a part.

(16) Der Tourist gesundete vom Typhus.

The inference to (17) involves a further step, in which e// is identified with e°, the process that is part of the event of the doctor curing the patient. It should be noted that this last identification doesn't strictly follow—it would in principle have been possible for y1 to have suffered simultaneously from another disease and for the doctor to have cured him of that one, while he overcame the typhoid on his own account. For lack of space we cannot go into the nature of such default inferences here.

References

[Frey,Werner / Tappe, Thilo (1991)] "Zur Interpretation der X-bar-Theorie und zur Syntax des Mittelfeldes. Grundlagen eines GB-Fragments" AIMS-Reports, Stuttgart to appear.

[Kamp, Hans (1981)] "A Theory of Truth and Semantic Representation", in: Jeroen Groenendijk, Theo M. V. Janssen and Martin Stokhof (eds.), *Formal Methods in the Study of Language,* Vol.I, Mathematisch Centrum, Amsterdam, reprinted in J. Groenendijk, T. Janssen and M. Stokhof (eds.), (1984), *Truth, Interpretation and Information,* Foris, Dordrecht, pp. 1–41.

[Kamp, Hans / Reyle, Uwe (1991)] *From Logic to discourse.* Vol.I, to appear: Kluwer, Dordrecht 1991

[Kamp, Hans/ Roßdeutscher, Antje (1991)] "Remarks on Lexical Structure and DRS-Construction", AIMS-Reports, Stuttgart 1991