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

A MULTILAYERED APPROACH TO THE HANDLING OF WORD FORMATION

Wolfgang Hoeppner

Research Unit for Information Science and Artificial Intelligence
University of Hamburg, Germanisches Seminar
Mittelweg 179
D-2000 Hamburg 13
Fed. Rep. of Germany

The treatment of word formations has until recently been a neglected topic in natural language AI research. This paper proposes a multilayered approach to word formation which treats derivatives and compounds on several different levels of processing within a natural language dialogue system. Analysis and generation strategies being developed for the dialogue system HAM-ANS are described. Identification of word formations, semantic interpretation, and evaluation in the context of a dialogue are the main levels of analysis on which the system successively attempts to infer the implicit relations between word formation components. Generation of word formations is viewed as a process comparable to the generation of elliptical utterances.

INTRODUCTION

Any linguistic theory has to account for word formation as a way of expressing complex relations, facts or situations, and nearly every theoretical approach contains at least suggestions as to how to handle word formation. Not until recently has word formation become a topic in natural language processing within the framework of artificial intelligence (cf. FININ 1980, McDONALD 1981).

It is argued here that similar attention should be paid to word formation as has already been paid, for example, to sentence structure. This is justified not only because such phenomena as derivatives and compounds obviously do occur in natural language, but rather because natural language AI systems to a large extent already contain the sort of knowledge needed to understand word formation, and therefore seem to be well suited for investigations in this field (cf. SAMLOWSKI 1975). Having been discarded in early days of AI research as too tedious and expensive a task (CERCONE 1974) the analysis of word formation, especially compounding, seems to be a major way to increase linguistic coverage and to reduce vocabulary errors, one of the most frequent sorts of errors in natural language systems (see THOMPSON 1980)

Generally speaking, the trouble with word formation is that in contrast to sentence structure the relations between constituents are not overtly marked in word formations. In addition, there are seldom explicit clues indicating whether a given word is lexicalized or analyzable, or how to interpret the latter ones. Furthermore, derived and compound words incorporate ambiguities on several different linguistic and cognitive levels. Therefore, it is a challenging task for natural language AI research to study how a system can identify, understand, and make use of word formation.

ANALYSIS OF WORD FORMATION IN HAM-ANS

The approach to the handling of word formation described in this paper is part of the development of the natural language system HAM-ANS (Hamburg Application-Oriented Natural Language System). HAM-ANS, which is based on the earlier system HAM-RPM (see v.HAHN et al. 1980) provides natural language access to other software systems. While a natural language interface to a large relational data-base system dealing with fishery data is currently being designed, the other two major application areas of HAM-ANS - a hotel-reservation system and a motion analysis system dealing with a street crossing (see JAMESON et al. 1980, MARBURGER et al. 1981) - are studied and further developed in an implemented version which covers the complete natural language dialogue. Analysis and generation of word formation has been integrated into the system in the context of these two domains of discourse and the examples given below are taken from dialogues about these domains.

The main idea in our approach is that derivatives and compounds cannot be treated appropriately in a separable component whose output is a semantic interpretation (FININ 1980) or a paraphrase (BORGIDA 1975, McDONALD/HAYES-ROTH 1978). Instead, the question of how, or even whether, a word formation is to be analyzed should be decidable on different levels of processing covering the meaning of word constituents, their interpretation in the context of utterances, and their interpretation in situational context.

It is quite a tradition in theoretical linguistics to discriminate between compounding and derivation as the two basic means of word formation. It is not our concern here to add new arguments in favour of or against such a simple distinction; rather, we use it to delimit the broad field of word formation and to indicate the linguistic data our approach is designed to capture. Leaving aside the differentiation between lexicalized and analyzable words for the moment the overall research objective is to handle those words which can be first segmented into semantically meaningful units and then interpreted by making use of knowledge sources and the inferential capacity of a natural language AI system.

A common characterization of derived words is that they are formed by combining a free morphemic part with a bound one. The order of these morphemic parts yields the discrimination between prefixation and suffixation on purely structural grounds. There is, however, also a semantic difference between two types of derivation: It appears to be rather difficult to determine the meanings of prefixes and their semantic relation to the free morphemic part of the word in a general way. We will therefore exclude prefixes from our treatment of word formation within HAM-ANS for the time being and concentrate on derivation by means of suffixes and on composition.

IDENTIFICATION OF DERIVATIVES AND COMPOUNDS

In both English and German derivatives are written as one word delimited by blanks or punctuation marks. Compound words in these languages are represented differently. A German compound is written as one string of letters, the segmental units of an English one are clearly indicated by a blank or a hyphen. This orthographic difference incorporates a difference in the problems of how to identify compounds in both languages. The ambiguity between English compounds and syntactic constructions, attribution in 'woman doctor', has to be handled within the syntactic analysis (cf. MARCUS 1980) and does not occur in German because of its graphemic representation of compounds. On the other hand a system analyzing German compounds has to identify the meaningful segments as a first subtask; several approaches in the area of computational lingustics have dealt with the problem of identifying segments in isolated compounds (cf. v.HAHN, FISCHER 1975, SCHOTT 1978). These systems rely heavily on graphemic and morphemic rules, the latter using additional lexical information. Characteristically the analysis of isolated compounds will at best produce more than one segmentation, as e.g. for the word STAUBECKEN which should be segmented in either STAU-BECKEN (reservoir) or in STAUB-ECKEN (dusty corners), but the determination of the intended meaning lies beyond the scope of these approaches.

In the system HAM-ANS the starting point of word-formation analysis is contained in the lexical analysis component, its main task being the reduction of inflected word forms and providing lexical information for the subsequent syntactic analysis. Whenever a word is not contained in the lexicon, the system removes possible inflectional suffixes before trying to recognize it as a derivative or a compound. Only if this attempt fails will the user be asked for information about the word. Employing the contents of the system's lexicon is certainly a simple way to define lexicalized formations. This sharp distinction between lexicalized and analyzable words, as used in the current implementation, does not do full justice to observable degrees of lexicalization; therefore it will yield to an improved conception. The segmentation of words not contained in the lexicon makes use of a table of derivative suffixes, a set of graphemic restrictions and the definitions of basic lexical items stored in the lexicon. Graphemic restrictions incorporate rules for the reduction of vowel mutation often cooccurring with suffixation and for the detection of juncture morphemes.

In a first step, derivative suffixes are recognized by comparing final segments of the word under inspection with the entries of the suffix table. The analysis of derivatives in HAM-ANS is to a large extent based on work done for different purposes in the area of computational linguistics (HOEPPNER 1980), major deviations being the extensive use of a lexicon and a smaller selection of productive suffixes. Apart from the literal form of the suffixes the entries of the table contain information about gender (for nominal suffixes), part of speech of the derivative and the basic form being derived and expressions of the system's semantic representation language SURF, which later on is integrated into the semantic representation of the whole word. The lexicon serves as a device for ascertaining that the remaining part is a lexical unit known or accessible to the system.

Having identified a derivative suffix and thus determined the word to be a derivative, the remaining part, however, can recursively turn out to be an analyzable formation, say a compound. So a second step (in the processing of a nonderived word the first step) is the attempt to split the word into two components both of which have to be ultimately transformable into canonical forms, for example by removing vowel mutation or analyzing a derivated part in the way described above. Search in the lexicon is performed by constructing a hypothetical first constituent and looking for the most similar lexicon entry. This yields the second constituent as the remainder which by consulting the lexicon leads to a revision of the initial hypothetical assumption or confirms it.

In principle these two steps in identifying the structure of compounds and derivatives should interact recursively to allow for the handling of multiple compounding and derivation (for restrictions on multiple derivation see HOEPPNER 1980). In HAM-ANS the analytical capacity at the moment is restricted to compounds with two parts and to singular derivation. This limitation is not so much determined by the identification process but rather by the state of elaboration of those processes which relate and integrate the semantic interpretation of a word formation into the knowledge already available to the system.

After the system has successfully segmented an initially unknown word, the result of the identification is a structure containing the identified parts together with those grammatical features which in the course of further processing will guide the construction of a semantic interpretation and which provide grammatical information for the whole word. To illustrate this resulting lexical structure, an example for the word 'STRASSENFEGER' (street cleaner) is given in Fig. 1, indicating also the origin of the associated grammatical features (the features and their values are given here in English).

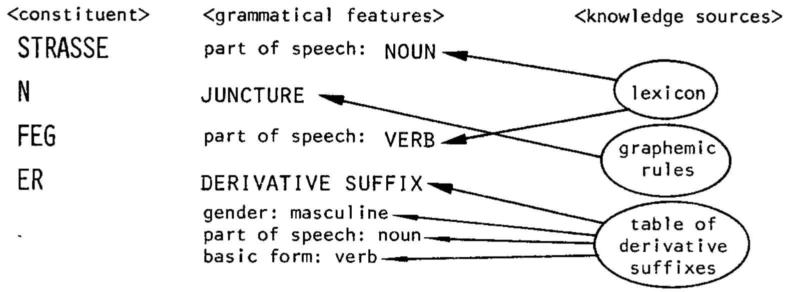

Figure 1: Lexical representation of STRASSENFEGER

SEMANTIC INTERPRETATION OF DERIVATIVES AND COMPOUNDS

So that the system needn't analyze an unknown word each time it occurs in an utterance, the information gathered so far could be stored in lexical memory, as is done with explicit information given by the user about unanalyzable words. The goal of word-formation analysis, however, is not completed with the segmentation of words and the assignment of features to their parts. A more important step is to relate structural knowledge about derivatives and compounds to conceptual knowledge and to transform lexical structures into semantic structures. The logic-oriented representation language SURF (see JAMESON et al. 1980) is the device in HAM-ANS which expresses semantic relations between parts of utterances and likewise between lexically analyzed words. An interpretation process has accordingly been implemented which maps lexical representations of analyzed words onto expressions of SURF having the same type as that constructed by the parser for simple words of the same class. For example, a compound noun is represented by a 'description' in the same way as a simple noun in a noun phrase would leave the parser. The only difference is that the representation of a compound contains explicit relations between its constituents. An example interpretation of the German compound STUHLBEIN (chair leg) is given in fig. 2, the letter T in the last line standing for the whole-part relation in the system's conceptual semantic network.

```
(d-o: AND
      (lambda: x1 (af-a: ISA x1 BEIN))
      (d-o: AND
            (lambda: x2 (af-a: ISA x2 STUHL))
            (lambda: x3 (af-a: T x2 x1))))
```

Figure 2: Semantic interpretation of STUHLBEIN

The representation of the simple noun BEIN (leg) would correspond to the first argument of the outermost conjunction, which is likewise a 'description'.

Let's now take a closer look at the way the transformation of a lexical representation into a SURF representation is achieved. As mentioned above the table of derivative suffixes includes one or more SURF-expressions for each suffix. The expression provided for the suffix -ER, in STRASSENFEGER, together with a verb stem leads to a case-frame instantiation with the agent being a male person and an objective case to be filled either by a genitive attribute or a compound constituent as in this example.

Compounds require a more interesting transformation process to discover relations between their parts. Analyzing the lexical representation, different inference strategies are selected depending on the parts of speech of the constituents. For instance, a compound consisting of two adjectives activates processes trying to establish a coordination of the two concepts (e.g. DUNKELBRAUN (dark brown)). The transformation of nominal compounds applies the system's inferential capacity to detect possible links between the two concepts in the conceptual semantic network. In addition to the part-of relation the following relations are inspected and used for the semantic representation in SURF:

- physical object and its material, e.g. HOLZTISCH (wooden table)
- property of an object, e.g. HAARFARBE (colour of hair)
- physical object in its preferred location, e.g. COUCHTISCH (couch table)
- combination of physical objects, e.g. RADIOWECKER (clock radio).

Finally, compounds with a verbal element are transformed by trying to fit the remaining constituents into the slots of the verb's case frame. The example STRASSENFEGER is represented as the instantiated case frame of FEGEN (to sweep) with the noun STRASSE (street) filling the objective slot.

At this stage the lexical representation and the semantic interpretation of a compound or a derivative are stored in the system's lexical memory for several reasons:

- to eliminate the need for repetition of the whole analysis each time the word occurs,
- to form the basis for analogy-driven resolution of word formations,
- to enable the system to use understandable words while generating utterances from semantic representations (see below).

An example of a semantic interpretation which is still ambiguous at this processing stage is the one for BILDERRAHMEN (picture frame), which besides a semantic representation expressing a whole-part relation would, by reference to the case frame of the German verb RAHMEN (to frame), be interpreted as an object-verb nominalization (the framing of pictures). Once the appropriate semantic interpretations have been inferred, processing continues with the parsing of the entire input utterance. The ATN grammar of HAM-ANS treats compounds and derivatives in the same way as other words of the same class except that they are more frequently ambiguous, so that the parser more often has to use knowledge of case frame restrictions or attribution congruency to select an appropriate reading.

EVALUATION OF WORD FORMATION IN DIALOGUE CONTEXT

Trying to evaluate entire utterances is the ultimate processing phase for accepting, reinterpreting or rejecting interpretations of word formations in HAM-ANS. Suppose the client in the hotel-reservation situation already knows about a desk in the room being offered and asks whether the desk chair is a comfortable one. Having interpreted the compound SCHREIBTISCHSTUHL (desk chair) as a chair which is conceptually located to a desk, the system would try to identify a referent with this property. According to the system's intentions (cf. JAMESON/WAHLSTER 1982) it might reject the existentially presupposed interpretation of 'desk chair', or it might find an appropriate referent and accept the interpretation. A third possibility, which is particular plausible in this communicative setting, is to take any chair in the neighbourhood of the desk and set it up as the object referred to. A similar case is the relaxation of one part of an additive compound, e.g. to agree to an object's property stated as 'dark brown', even if only 'brown' is a proper attribute according to the system's extensional knowledge and no contrary information relating to the object's brightness is available.

It should be emphasized that the task of analyzing word formation in an ongoing dialogue is not finished when the system is able to interpret a compound or derivative in utterances or even has given a satisfactory reply. The knowledge gained through the analysis and the commitments to the interpretation chosen have to be integrated into the knowledge sources. At present two consequences are associated with a successful analysis: First, the conceptual and referential semantic networks are updated to allow for subsequent reference. Second, the knowledge sources representing the partner's assumptions about the domain of discourse are updated accordingly (cf. JAMESON/WAHLSTER 1982).
The worst case conceivable appears to be misspelling in a way which allows the word

to be acceptable on structural and semantic grounds but doesn't make sense in the context of the utterance and the dialogue. These cases, however, seem to be rare.

GENERATION OF WORD FORMATIONS

The unbalanced relation between the analytic capabilities of natural language AI systems and their generative capabilities is found in the area of word formation as well. In HAM-ANS, research in word formation generation has started with two approaches. The first is a rather simple one: By analyzing compounds and derivatives, the system has created a semantic interpretation in terms of the language SURF and kept it in lexical memory. The basis for answer generation is a structure of the language SURF, which makes it possible to check agreement between certain parts of the answer and entries in lexical memory. An example of this method is the substitution of the description LAMPE AUF DEM SCHREIBTISCH (lamp on the desk) by SCHREIBTISCHLAMPE (desk lamp).

The second approach enables the system to make use of word formations in its own utterances without having previously analyzed a corresponding word. For those parts of utterances which might be verbalized using word formations, e.g. modified nouns or coordinate attributes, a set of patterns is provided which, by means of a matching process bind relation identifiers and canonical word forms. These are handed to a generation component whose task it is to decide on derivation, compounding or no word formation at all, and to yield morphologically correct junctures.

ACKNOWLEDGEMENT: Research on HAM-ANS is currently being supported by the German Ministry for Research and Technology under contract 08IT1503 8.

REFERENCES

[1] BORGIDA, A. T.: Topics in the Understanding of English Sentences by Computer. Dept. of Comp. Sc., University of Toronto, Technical Rep.78, Febr.1975

[2] CERCONE, N.: Computer Analysis of English Word Formation. Dept. of Comp. Sc., University of Alberta, Technical Report TR74-6, April 1974

[3] FININ, T. W.: The Semantic Interpretation of Compound Nominals. Coordinated Science Laboratory, University of Illinois, Report T-96, June 1980

[4] v.HAHN, W./FISCHER, H.: Ueber die Leistung von Morphologisierungsalgorithmen bei Substantiven. ZDL, Beiheft 13, 1975, 130-150

[5] v.HAHN, W./HOEPPNER, W./JAMESON, A./WAHLSTER, W.: The Anatomy of the Natural Language Dialogue System HAM-RPM. In: L.Bolc (ed.): Natural Language Based Computer Systems. Muenchen/London: Hanser/Macmillan 1980, 119-253

[6] HOEPPNER, W.: Derivative Wortbildung der deutschen Gegenwartssprache und ihre algorithmische Analyse. Tuebingen: Narr 1980

[7] JAMESON, A./HOEPPNER, W./WAHLSTER,W.: The Natural Language System HAM-RPM as a Hotel Manager: Some Representational Prerequisites. In: R.Wilhelm (ed.): GI-10. Jahrestagung, Saarbruecken; Berlin: Springer 1980, 459-473

[8] JAMESON, A./WAHLSTER, W.: User Modelling in Anaphora Generation: Ellipsis and Definite Description. To appear in: Proc. of the ECAI-82, Orsay 1982

[9] MARBURGER, H./NEUMANN, B./ NOVAK, H.-J.: Natural Language Dialogue about Motion in an Automatically Analysed Traffic Scene. In: Proc. of the 7th IJCAI, Vancouver 1981, 49-51

[10] MARCUS, M. P.: A Theory of Syntactic Recognition for Natural Language. Cambridge, Mass.: MIT Press 1980

[11] McDONALD, D.: COMPOUND: A Program that Understands Noun Compounds. In: Proc. of the 7th IJCAI, Vancouver 1981, 1061

[12] McDONALD, D./ HAYES-ROTH, F.: Inferential Searches of Knowledge Networks as an Approach to Extensible Language-Understanding Systems. In: D.A.Waterman, F.Hayes-Roth (eds.): Pattern-Directed Inference Systems. N.Y. 1978, 431-453

[13] SAMLOWSKI, W.: Deutsche Nominalkomposita in einem Sprachverstehensprogramm. In: G.Veenker (ed.): Zweites Treffen der GI-Fachgruppe Kuenstliche Intelligenz. Abt. Informatik, Univ. Dortmund, Bericht 13/75, 1975, 90-109

[14] SCHOTT, G.: Automatische Kompositazerlegung mit einem Minimalwörterbuch zur Informationsgewinnung aus beliebigen Fachtexten. In: F.Wingert (ed.): Klartextverarbeitung. Berlin: Springer 1978, 32-43

[15] THOMPSON, B. H.: Linguistic Analysis of Natural Language Communication with Computers. In: Proc. of the 8th COLING, Tokyo 1980, 190-201