Alfred Hoppe
LIMAS II, Bonn

Communicative Grammar and Machine-Assisted Text Contents Analysis

Hoppe, A.: **Communicative grammar and machine-assisted text contents analysis.**
In: Int. Classif. 11 (1984) No. 1, p. 9–12, 7 refs.

None of the endeavors undertaken so far in the field of text contents analysis has proved satisfactory, since semantic implications were, in the end, merely superimposed onto traditional grammar, e.e. onto descriptions of the form domain of language. The approach of LIMAS, however, uses a semantic syntax which can be abstracted to such a degree as to permit it of being converted into an algorithm and made computer operable. For this the Communicative Grammar was introduced. Its procedures are outlined, its present state and possible applications in information banks, retrieval, translation, language courses and through the use of microprocessors are described. (I.C.)

1. The overall situation in machine-assisted language processing

a) Language and computer

The usual LDP procedures (LDP = linguistic data processing) start out from computer logic, which is based on the "And/Or scheme" and the binary system of numbers. But since language is a mental product of man, who does not think along binary lines only, it employs a logic – namely the logic of man – which is more differentiated than and differently structured from that of physical processes and which (or the frequent lack of which) cannot be completely grasped by computer logic, even granting that this logic may still be expanded or developed further.

This becomes particularly clear when one does not see language exclusively as a sequence of characters but also as a sequence of contents elements and their linguistically-described relations, hence when one strives for an analysis of text contents. For particularly the linguistic relations of the contents often fail to correspond to the relations of the forms in which they are expressed – and even more frequently they do not correspond to the logical relations, let alone to the scientifically defined relations, of the things in the world.

b) Priority of language

From this it should not be concluded (as Bar Hillel did) that as a matter of principle there can be no contents-oriented, machine-assisted analyses of texts, nor any machine-assisted translation procedures, but only that the traditional knowledge of language and its semantics were not sufficient to describe language, based on this state of the art, in computer-operable fashion. This means that the linguistic system, being a process of the human art of formulation, must have priority over the computer system in the sense that e.g. an economic system is first of all described the way it is (factual analysis of a process) and only then is made computer-operable by suitable processing of the description arrived at.

As far back as 1966 American experts had declared – in the so-called ALPAC memorandum – that satisfactory machine-processing of language and texts will be impossible failing an expansion of the knowledge of linguistic contents and concentration of research on linguistics rather than on hardware and software development in LDP.

c) The sacred cow of traditional grammar and the computer approach

In the past 20 years a number of "new" grammars have come into existence, culminating more or less in Noam Chomski's approach and in the euphoric hope that it would prove possible to have the computer itself develop a computer-compatible grammar.

None of the endeavors undertaken so far in the field of text contents analysis has proved satisfactory (as has already become public knowledge down to the technical journalistic level, see Computer Magazin 2/82), since semantic implications were, in the end, merely superimposed by them onto traditional grammar, i.e. onto descriptions of the form domain of language (Ispra 1966). Therefore they have been unable – despite all attempts at refinement – to overcome the limits necessarily inherent in the inventorying of grammatical forms.

The temptation consisted and consists in the fact that the world of forms appears to be much more readily transferable to computer logic than the "logic" of the contents of language. Nor can this logic be generally equated with the "logic" of the natural sciences or with a scientific system of the relations among the things in the world.

2. The new approach. The concept of the LIMAS Research Group.

Supplementarily to this development, another research approach was embarked upon as far back as 1964 with the establishment of the LIMAS (*L*inguistik and ***ma***schinelle *S*prachverarbeitung = Linguistics and machine processing of language) research group. The concept underlying this approach was favored by the US Government and financed by the German Federal Government until it was stopped by the German Research Minister in 1976. However, the founder of the LIMAS Research Group has since then continued to pursue this approach until the present state of the art was reached.

2.1 The premises of the LIMAS approach

Premise No. 1: The communication contents

Man does not primarily formulate his thoughts for some such purpose as finding grammatical forms for his communication contents, such as subject, predicate, object.

Nor are these the things he wishes to communicate; he does not even need to know them. What he is interested in, rather, is: linguistically expressing, and recognizing in the forms thus expressed, the *contents* he has in mind. But to do this, not even the contents of words will suffice such as they are offered today by dictionaries and/or "concept network systems" and by data and information banks. No, language conveys its essential information only by means of its own interrelated system of word contents, hence by a semantic syntax of communication contents. They are the controlling agents of man's formulation process.

Premise No. 1 therefore means: A contents-oriented, machine-assisted language processing process must first of all succeed in describing, by means of the semantic syntax, the linguistic formulation process as performed in the mind of thinking man, following which it must duplicate it, formalize it and describe it in computer-operable terms.

Premise No. 2: The laws of language and those of the machine

The description of the linguistic contents must be of such nature as to do full justice, on the one hand, to the laws of language (formulation process of linguistic contents) and to permit of adaptation, on the other hand, to the laws of machine-assisted processing. This however, is not possibe on the level of the multibillion variety of forms of expression but only on a higher abstraction level of contents and their mutual relations.

Premise No. 3: Semantic syntax

Just as there is a grammar of the world of linguistic forms – without which no one would be able to speak or understand in linguistically correct fashion – so there must also be a grammar of elements of linguistic contents – without which no on would be able to speak or understand language with meaningful contents. Therefore, in addition to the formal syntax, this semantic syntax needs to be elaborated.

Only this syntax can be abstracted to such a degree – i.e. represented in the form of abstract classes in such fashion – as to permit it of being converted into an algorithm and made computer-operable.

2.2 The Communicative Grammar procedure

a) Classification

Through classification of the contents elements, the semantic syntax and the syntactic-linguistic relations pertaining to it are represented by a system of codes, lists, matrices, tables and parameters. The codes are added to the words in the dictionary. They represent man's knowledge about the semantic syntax of language. They provide the access to the working instruments mentioned. These, because of the neighborhood functions, already comprise the contents factors of neighboring words belonging to one and the same contents complex and needed for the formation of it. Therefore only a rudimentary formal analysis of the forms of expression is necessary.

By means of the semantic syntax, every group of words belonging together, and every sentence, whether individually or as part of a text, as well as every concept occurring in this connection can be analyzed with respect to its relation to other concepts of the sentence, of the text or of the subject field concerned, and the result can be formalized.

The semantic syntax is a grammar of communication contents (Communicative Grammar, or CG (in German: KG)), which not only comprises the word contents (word semantics) but also the relations – understood contents-wise – of these word contents among one another.

Example: "*The X company* produces/manufactures/fabricates/makes/develops/tests/tries out/builds/erects/renews/completes/(re)constructs/elaborates/evolves/forms/shapes/renovates/contrives/prepares/works out a *text processing machine*." If the "X company" is to be recognized from the text as the *producer* and the "machine" as *its product*, then the common meaning of the 20 verbs that may occur in the text must be recognized. The company's role as producer is not shown in the dictionary and is not part of the word contents of "company" but rather of the contents of the sentence. This its role is not expressed by e.g. "subject/object" but rather by the semantic syntax, i.e. by the meaning of one of the verbs mentioned in the sentence. Other verbs than those mentioned (e.g. "sells") establish a different relation between "company" and "machine". The content syntax thus is a different one. The verb, therefore, is the word defining the relation between the words "company" and "machine".

The same relation is also expressed by e.g. "machine manufacturing company" or "the manufacture of machines by the X company" or "the companies manufacturing machines" or "the X company's machine production".

All words which assign such roles (meanings, contents) – they may also be roles of place, time, cause, effect, condition, association, change of state, etc. – to other concepts are assigned a code corresponding to the roles, a code which stands for the roles these words assign to the words in their syntactic neighborhood. That means: there is only one code for those 20 verbs and for all nouns, adverbs, connectives, prepositions and adjectives assigning the same role to their syntactically connected neighborhood words. Instead of the words "company" and "machine", all words which can take their place in the syntactic neighborhood of the code word (:"manufacturer, nation, firm, Germany", or "food, consumer goods, hosiery, energy") are assigned the same role in this syntactic unit represented by the code as the words "company" and "machine".

As a result, the analysis becomes independent of the form of expression. The syntactic, contents-based connection remains the same, however. The number of forms of expression is unlimited.

b) Three stages of analytic results

The formulation of analytic results takes place in three stages which are determined at the same time. The first stage assigns the concept a still diffuse, contents-based relation to the other concept (e.g. agens, patiens, occurrence relation, occurrence purpose); the second stage indicates the specific role (e.g. donans, donare, donare addressee, donare object), while at the third stage the specifications of the second stage are combined into classes of higher generalization and abstraction. There are about 30 such classes on this third level.

The specific roles of the second stage, e.g. DONARE, ACCEPTARE, RAPERE (= to take away, to rob), are abstracted on the third level from their contents variants and combined into VARIARE-ASSOCIATION, so that it is possible to say what something is

– and was – associated with, while at the second stage it is said in what manner the association relation has been changed. A distinction is made between VARIARE and INVARIARE-ASSOCIATION.

At which of the three stages the analysis result is printed out depends on the preference of the user. The printout form is an expression of the natural language.

The meanings attached to a word in the contents complex, which at all three stages go beyond the lexical meaning of the word, are called "semantic roles". They are at least as important communication contents as the lexical contents of the words.

c) *Classification systems*

Classification is required for this procedure. The classes obtained (see above!) are added to the words concerned in a compact code. As a result, the second and third stages of the semantic syntax can comprise millions and millions of forms of expression.

The lexical contents of the nouns are classified according to their contents reference possibilities within the contents complexes. (Examples: PERSON, ANIMALIUM, VEGETATIVUM, CONCRETE, CONCRETE-MOVED, EMOTION, VALUE, et al.). This gives rise to some 30 noun classes, whose codes are added to the words in the dictionary.

The words bringing about the connections are classified in like fashion. Through their codes, and through the procedure using the noun classes occurring within the contents complex, their semantic roles are determined and added as analysis result in the three stages. Each of these classes comprises from 2 to 500 or more words.

Within the semantic-syntactic complex, the relations from the contents classes to the connective class words are frequently established by verbs in sentences. Examples: the horseman gave his horse to the wounded man (DONARE); the wounded man accepted the horse from the horseman (ACCEPTARE); the wounded man stole the horse from the horseman (RAPERE). "Giving", "accepting" or "taking away" "something" is something only a PERSON, a PERSON/CORPORATE BODY or an ANIMALIUM can do to a PERSON or an ANIMALIUM. The "something" can belong to numerous noun contents classes, while "horseman" in the three sentences on one occasion plays the semantic role of the giving agent (DONANS), on another occasion that of the acceptance direction reference (ACCEPTARE reference), and on the third occasion that of the turning-away reference (RAPERE reference). At the same time he was in all three cases the first association carrier, with the wounded man being the second such carrier. All three verbs express a VARIARE ASSOCIATION.

The generalized third stage therefore comprises far more forms of expression than the second stage. If all substituents of the possible classes of this contents complex and their permutations are counted along, one arrives at hundreds and hundreds of millions of possible forms of expression with the same semantic roles in their contents complex.

d) *Detachment of the contents complexes from the forms of expression*

Each contents complex is presented detached (abstracted) from the form of expression in which it appears (sentence conglomerate, sentence, group of words, composite word). It is thus representable in all languages in forms of expression germane to them and forms part of any human thinking.

e) *Metalingua*

Like the traditional form syntax, the semantic syntax is presented to a high degree in Latin terminology, so that it may be applied to all languages having the appropriate thought contents. It constitutes a metalingua in simple form.

f) *Synergetic action cycles*

The contents complexes reveal themselves as synergetic action cycles. They are order and relation structures organizing themselves out of the neighborhood function of specific contents factors (classes).

There are no more than five basic models for these action cycles. They also occur alongside one another in combinations, so that, like the factors of a single action cycle, they can become themselves, according to the same model, factors of a new, superimposed action cycle. Considerably simplifying operative possibilities, they permit a binary procedure structured according to "And" and "Or" functions.

g) *Communicative grammar*

All factors for the description of the word classes, the contents factor complexes, the relations and the action cycles are – as contents elements or complexes – named according to their linguistic contents, thus becoming objects of communication. Therefore this description of the semantic syntax of language is called "Communicative Grammar" (CG).

3. Current status of the Communicative Grammar

Available in completely finished form, the system of the CG has been published in book form and in numerous technical articles highlighting various application aspects (for documentation, classification, translation, etc.). The LIMAS research group has conducted several machine-assisted model tests for text analysis and for partial translations into English.

Thus e.g. a 4000-word text (= 24 sentences) was subjected within 60 seconds, in a non-economical procedure, to a contents analysis followed by querying in arbitrary language, as was demonstrated in 1975/76 already via telescreen before a committee of the German Federal Ministry for Research and Technology. 12 000 punched cards were involved.

4. Applications

a) *Information bank*

With the results of the text analysis it is possible to build up, with machine assistance, an information bank whose data comprise not only the classified words (descriptors), word contents classes and concepts, but also their semantic roles as obtained from the analytic procedure as well as their contents-based relations to other word contents and concepts which in any given case are automatically determined from the text and either automatically or manually added to the lexicon words.

b) Retrieval

The formulation of the queries is completely independent of language. Being processed according to the same procedure as employed by the text analysis, they thus directly receive those addresses to the bank which refer to the query contents and which again consist of the same codes. The information output can occur, in the form desired in the given case, at any of the three abstraction stages mentioned. It can be formulated in any clear text form desired.

c) Translation

Owing to the metalingua formulation of the analysis results the printout of the desired information can occur in all those languages which have been appropriately processed according to the CG system and are connected to the information bank. As a result, querying is also possible from and in another language.

Since the procedure and the CG are reversible as a matter of principle (like human speech and understanding, too, are based on only *one* grammar, *one* procedure), the metalingua formula makes it possible for any form of expression representing its contents complex to be generated in any of the languages connected and to be printed out in another language. In a more expanded form, the CG guarantees a true-to-meaning translation into any language, including languages not belonging to the Indo-European family of languages.

d) Microprocessors

Since the Communicative Grammar in its binary and process-oriented structure – including the reversibility of the process paths – follows the human formulation and understanding process on its paths from the linguistic contents to their forms of expression and vice versa, it is possible to present these process paths in flow diagrams. This involves, however, certain properties of the conjunctions and disjunctions at the knots as discussed in (4), page 130 et seqq.

From this, integrated circuits are built up which can be reduced to 5–6 models. The linking of such circuits again and again gives rise to identically or similarly functioning simulators of the – forever recurrent – complexes of the morphologic and semantic syntax, from the nominal word groups to the occurrence complexes to the sentence plans and all their modalities, e.g. down to those into which the partial complexes initially to be regarded as valid can be imbedded (e.g. the DONARE complex into the modalities: necessity, surmise, possibility, rumor, indirect speech, permission, desire, et al. Assuming a number of 10,000 full verbs, this alone would produce 1,440,000 different forms of expression that may be collected.

These complex, integrated circuits and their interlinkings in the formulation process of language may be "wired up", including their different logical elements at the knots. They replace in their function the analysis and synthesis programming and, if transferred to microchips, operate as much faster between their lexicon inputs and their result outputs (the latter in metalingua form) as a microchip is faster than the corresponding software. In this process, the implication of the structures of the intralingual semantic syntax ensures that cumbersome decision operations such as a purely morphologically structured system would require are largely dispensed with.

On page 2 of this issue of IC a draft design (already in extreme reduction) is given of the verbal formulation process to be used on a microprocessor.

e) Language instruction

From the beginning of its development, the system has been applied with increasing success for some 60 semesters by now at Bonn University in German language courses for non-German students of from 10–15 different mother tongues from Europe, Asia and Africa. At present there is also a demand at the secondary school level for a new German school grammar, for the formalistic systems in use there have for more than 30 years proved incapable of ensuring satisfactory linguistic instruction in the mother tongue (i.e. German).

The continued theoretical development of the CG is being handled in the LIMAS II research group in Bonn.

References:

(1) Hoppe, A.: Semantische Steuerungen im Prozeß der Formulierung sprachlicher Formen. In: Festschrift Helmut Gipper. Amsterdam 1979

(2) Hoppe, A.: dtv Wörterbuch der deutschen Sprache. Hrsg. G. Wahrig, 1978. Buchbesprechung in: Int. Classif. 5 (1978) No. 3, S. 179–180

(3) Hoppe, A.: Klassifikation innersprachlicher, semantischer Komplexe. In: Kooperation in der Klassifikation. Proc. 2. Fachtagung, Ges. f. Klassif. Frankfurt-Höchst, 6.–7. 4. 1978. Frankfurt: INDEKS Verl. 1978. = Studien zur Klassifikation, Bd. 2, p. 47–59

(4) Hoppe, A.: Die semantische Syntax der GESCHEHEN-KOMPLEXE: In: Kommunikative Grammatik, Teil I, Bonn: F. Dümmler 1981.

(5) Hoppe, A.: Vorsprachliche Konzeption semantischer Komplexe des Geschehens und deren Wortung. In: Peuser, G., Winter, S. (Hrsg.): Angewandte Sprachwissenschaft. Festschrift f. Günther Kandler. Bonn: Bouvier Verl. 1981.

(6) Hoppe, A.: Die synergetische Funktion begrifflicher Klassen, dargestellt an ihren sprachlichen Zusammenhängen. In: Numerische und Nicht-numerische Klassifikation. Proc. 5. Fachtagung d. Ges. f. Klassif., Hofgeismar, 7.–10. April 1981. Frankfurt: INDEKS Verl. 1982. = Studien zur Klassifikation, Bd. 10, p. 166–183

(7) Hoppe, A.: Die Selbstorganisation semantischer Strukturen. In: Festschrift Joh. Knobloch, Bonn 1983.

Dr. Alfred Hoppe, LIMAS II
August Bier-Str. 20, D-5300 Bonn 1