Harald H.Zimmermann
Universität des Saarlandes, Saarbriicken, Germany

Language and Language Technology

To Alfred Hoppe on his 80th Birthday

Zimmermann, H.H.: **Language and Language Technology.** To Alfred Hoppe on his 80th birthday.
Int.Classif. 18(1991)No.4, p.196-199, 6 refs.
In Language Technology, theoretical concepts have to be matched with practical limits and possibilities. It is explained that there are low-level fields like spelling check, automatic hyphenation or free text indexing where word based techniques can play a considerable part even without integrating syntactic or semantic analysis. On the other hand it is claimed that, without the integration of syntactico-semantic features and world knowledge, higher sophisticated tools like machine translation (not to speak of speech recognition) will not reach a functionally acceptable level. The work and the semantically based theory of A. Hoppe is compared with these considerations and it is shown that his conception as well as the practical results can be of extreme usability for high quality language technology development. (Author)

1. Introduction

If one is dealing with language problems, one will often perform Sisyphian labor. Hardly does one believe to have advanced a step, when new phenomena crop up which, if not rendering the entire structure questionable, at least manifest its limitations.

In my life I have met three outstanding scholars who, each for his own part and in his own field, have tried to describe language and language formalisms and make them useful for practical purposes as well. They are - to put it in somewhat general terms, - the lexicologist and lexicographer Gerhard WAHRIG, the syntactician and philologist Hans EGGERS, and the semantician and cybernetician Alfred HOPPE.

Refusing, as all of them have, to content themselves exclusively with abstract theories and the development of formal description systems, they have instead linked up theory with empiricism and with hard, systematic work on the material. And they all were or are aware that man's linguistic competence is more than can be mapped today onto even the most up-to-date machines. Let me quote, as one standing for all, A.Hoppe. In the chapter "Language and Machine" of his latest book "Theory of Semantic Syntax: Firm Combinations" we read, as an introduction:

"If one wishes to entrust the understanding of a language and thus the process of thinking to a machine, then the procedure (...) is co-determined by the design and the mode of operation of this machine. Even the very first semantic step of the ladder is unattainable for it. The person competent in language, on the other hand, is enabled by this competence to move effortlessly from step to step up and down the ladder." (1, p.124)

I may be permitted - in all modesty and with great respect for these scholars to seek their company with the following considerations. In so doing I will - in accordance with my teaching and research and development activity in the field of information and language technology - primarily take an *engineer's point of view*. This engineer is confronted with the general question whether - and if so: within what limits - there exist possibilities for so integrating 'language' (still used quite vaguely here) into language-technological processes that it becomes possible to machine-evaluate or machine proceed utterances of human language (as found - in written or spoken form - on the surface).

In the following I wish to exclude the vast, but ultimately 'unintelligent' field of the physical storage and transport of linguistic utterances, such as e.g. digital language tansfer and digital language storage systems.

What engineering tasks in which language processing plays a part are to mention here? Without systematizing it here any further, I will mention a few fields in which there have been application-oriented developments for several years:

(1) electronic word processing,
(2) man/machine communication (question & answer systems),
(3) machine 'understanding' of spoken language,
(4) automatic indexing (up to contents analysis),
(5) machine and machine-supported translation of language.

I do not deny, in so doing, the *interaction*, repeatedly described by A.Hoppe of the most varied linguistic factors, especially not the central importance of semantics (as influenced by Hoppe) for the linguistic understanding process. This applies in particular to the generative part, i.e. to the case that the machine performs more or less independently the generation of linguistic utterances.

I am likewise aware that in particular the syntacto-semantic characteristics as systematically developed by Hoppe in his "Communicative Grammar", i.e. the theory of a semantically dominated syntax, play a central part here. Similar considerations are found in the works by M.and G.Gross in the French language.

The engineer's starting point, in contrast, is somewhat different. His problem consists in finding for his 'clients' practically effective problem solutions for a specific field. The problem ultimately confronting him is (only): are there in this field any solutions *at all* in which semantics - or viewed more broadly: linguistic and world-oriented knowledge - can be exploited only *partially*? For, as Hoppe rightly says at the end of the aforementioned chapter with respect to information technology: it must "not forget that its clients also think and talk".

2. Dictionary and morphology

The (electronic) dictionary is regarded in the following as the machine's store of knowledge. The knowledge required for language processing (today) gets into this store via (human) linguistic experts. In the present connection it plays only a subordinated role how this dictionary is organized technically. However, relational databank systems for the storage and consistent updating of data are available, together with expert system parts containing in particular rules for the derivation of characteristics. In Hoppe's concept, too, the dictionary plays an important part, as here, among other things, the *semantic roles* of words are listed (cf. 'Theory', 2, p.108).

If in the dictionary field one adheres to an *open* concept, e.g. with the possibility of adding - or possibly also modifying - any desired characteristics one will in my opinion be prepared for all imaginable applications. But this does not mean at the same time that from the very start all possible applications must or can be considered. For in practice this is a costs and marketing question. The language engineer will primarily let himself be guided by whether and in how far specific solutions can be attained at reasonable costs. From the above fields I select three examples:

a) Automatic hyphenation and spelling corrections in word processing;
b) Dictionary-based synonym-provision and translation aids;
c) Automatic (word-oriented) indexing

In all cases it is first of all important that the client should not be 'disappointed' *for volume reasons*. Thus an approach as used e.g. by Knuth for hyphenation in English texts (System TEX): 'Hyphenate only known word forms' has practically no value in German (as a strongly composing language). Automatic hyphenation and aid in spelling is meaningful in German only when more than 100 000 word stems are stored and more-over a morphological flexion analysis as well as derivation and decomposition procedures are applied. Only then the system will not come to a halt at linguistically *correct* words, since they are unknown to the system: The problem of as complete as possible *morphological identification* must first be solved before restrictions to it can be dealt with. Interestingly, the current printed dictionaries in German do not indicate any syllabication marks, although they are important for recognizing faulty word compositions.

The possibility to admit also in composition topical formations as linguistically correct (Kanzlerreise = Chancelor's trip, Buchüberreichung = book presentation, ...) leads necessarily, however, to overidentifications, unless further criteria (but which ones?) can be used for blocking: the typographical errors "Waldkauf" (= forest purchase) instead of "Waldlauf" (forest run) or "Maustür" (mouse door) instead of "Haustür" (house door) belong in here (and appear at least identifiable within certain limits), but what about the sentence "gib mir *seinen* Brief zurück" (give me back *his* letter), instead of "... *meinen* Brief..." ("... *my* letter...")? Investigations on error identification on word basis have shown that some 95% of all errors are of such nature that they can be reliably spotted, with the finding of the remaining 5% being left (today) to human intervention.

Things are different with *automatic hyphenation*: With good procedures it is practically possible in German already on the word level to attain qualities comparable to any intellectual hyphenation. "Weak points" are presented primarily at linking points, where, e.g., the suffix "er" and the prefix "er" collide (Druck-er-zeugnis = either printing product or printing certificate) and in the (rare) cases of differing hyphenation possibilities (best known example Wach/stube (guard room) and Wachs/tube (wax tube).

Where electronic (lexical) translation aids are made available the printed book is first of all replaced by the electronic dictionary. The strategy is comparable here to the use of the printed dictionary. The advantage of electronic procedures is evident at two points: There is no (or hardly any) need for a user any longer to know the alphabet and the aid is available to him *during writing* more or less by pressing a button (3).

For differentiation purposes one will provide aids and set characteristics. It will be most interesting to examine to what extent Hoppe's formal classifications can be resorted to as an external basis. I consciously make a difference here: On the system side the form characteristics can surely be found but all experience indicates this will be of little use to the lay user. So a bridge needs to be built from the "system's view" to the "user's view" (a quite customary procedure in information/database technology). Possibilities presenting themselves are: Replacement of the characteristics by prototypical examples (GETR/DONS - "Vater"; GEZL/DONM - "Auto") or the automatic generating of an example (schenken_1: Fritz schenkt Paul ein Auto) (donate_1: Fritz donates a car to Paul), etc.

Something particularly to be wished is the improvement of searches in (bibliographic as well as textual) databases or their depth analysis with language technology methods. Here two major points are to be noted: In the long run it is economically not feasible, to index linguistically in depth the "big" databases already during their construction phase (as, e.g., DPA, JURIS, PATDPA in Germany; Chemical Abstracts internationally). At present practically all text data bases are searched on a word-form basis (in the free text mode), and the user is partially required to perform downright acrobatics (in so-called *truncation*). Undoubtedly very helpful is the availability of automatic truncation aids in which the system automatically makes available the possible stems and - if going beyond a single word class - also pseudo stems. Initially sufficient for this purpose is a reduction algorithm which in comparison with the identification procedure as used in spelling control also supplies references to (basic) stems. In view of the unreliability of the original material possible over indexations (example: Schraubenmutter/ - muttern? / mütter?) hardly carry any weight, on the contrary: a differentiation a priori (e.g. Bank = finance institute or seat) would not be appreciated by a database, as it lacks corresponding differentiations.

This limitation does not apply when e.g. a *consulting system* is being developed for proposing to a database user suitable terms for a search in a database: here, Hoppe's categorizations (e.g. on differentiation of meaning) might play an important role.

3. Syntax and Semanto-Syntax

Since Fillmore, at the latest, "neutral" (as to meaning) syntax analysis in Chomsky's sense has been internationally discarded. As Hoppe's early systematic works had already gone in the same direction, in line with the fact that, following the lead taken by Weisgerber, language phenomena have for some time been approached from an integral point of view.

At this point I would rather not go further into Hoppe's step model, recommending instead the reading of Hoppe's new publication. All present-day procedures and approaches, particularly in the field of machine translation (including the classical systems SYSTRAN and METAL as well as EUROTRA), proceed meanwhile from the recognition that language analysis and synthesis need a *semantic component*. Many of the existing differences are rather to be found in the field of analysis and synthesis *strategy:* while Hoppe assigns semantics a "controlling function", with EUROTRA it is a "level" alongside surface syntax, while in SYSTRAN its primary role is the *disambiguating function*, regarded as essential by Hoppe as well.

While the Hoppe graph and the so-called interaction/network system constitute, in my eyes, an interesting *representation form*, they do not solve, e.g., the *parsing* problem (at best, a language generator might be built), as one finds in the linguistic expression forms a confused mass of (surface) ambiguities which first of all must be disentangled. Undoubtedly interesting here is the idea of taking the concept system itself (or more precisely: the system parts corresponding with the expression forms = "concept words") as basis of the analysis rather than - as generally customary - the syntactic structure (cf. "Theorie", p.114).

That there are interdependencies between the various "levels" (if such an analogy should be admissible at all) is evident already from the simplest examples. The sentences "he *seed* the woman in the garden", or "he too late came, he not was admitted" remain intelligible. In the case of a formally correct sentence using non-sensible words it cannot be necessarily decided whether the sentence is semantically correct (which it may be after having been "translated" into sensible words. The sentence: "this mouse eats the cat" (theoretically) even causes the semantic syntax to fail.

Nevertheless it must be retained: Without (consistent) application of a semantically oriented syntax *superior* systems of computerized language processing, particularly of machine translation, must fail. Their specific efficiency becomes evident especially in socalled "univocalization" of former syntactically ambiguous structures (potential alternatives), or, what is at least as important: the disambiguation of *word* meanings. In the interest of "vindicating" existing translation procedures it should be noted, however, that it took almost a generation to lay in some partial fields, the foundations for a system of the Semantic Syntax, which - in addition - must first prove itself in practice. In any event, such a procedure will only then become effective, if it is realized *on a larger than sentence scale* (using either a paragraph or the entire text as its context level) and, in addition, also the problem of pronominal reference is considered (and solved along). Such a semantic analysis on the level of the *sentence context* - as Saarbrücken investigations during the eighties have shown - fails because of the ambiguity of the pronouns (meaning that there is too little univocalization/disambiguation).

4. Language and World Knowledge

The question where "linguistic" knowledge (= language system related knowledge) and where "world knowledge", that is "subject related knowledge" begins, is - from the language engineers point of view - a "philosophical" topic (4). Here too, I would like to give an example (referring at the same time to the above mentioned example of the cat and the mouse): If in a question-answering system (the example comes from PLIDIS, an earlier development by the Institut für Deutsche Sprache, Mannheim) the question is asked: "by how many points did VW rise yesterday?" (at least), the following data are necessary for answering:

- current date (of the day)
- VW = VW share
- value of the share from the preceding day and the day before
- rise = increase in value/change in value (the share might also have decreased in value)
- point = numerical value in whole numbers

At least the following operations (rules) must be applied:

- yesterday = current date minus one (data knowledge) ...
- recall of values from data base
- mathematical comparison operation

If one moves inside a small "world" (stock exchange information, weather report, schedule of events; any number of such "worlds" may be imagined), certain functions become important which - in systems of general language, appear at best rudimentarily but which likewise have strong effects on "understanding" (and are even absolutely necessary for the answers).

5. Summing-up and Perspective

Language technology - if it wants to really aid or facilitate the work of its clients/users - must make use of *all* available means supplied by linguists, psychologists, experts in specific fields or computer scientists. Higher valued systems (e.g. for machine translation) need (however) a strong, semantically based syntax.

Not just of today, but ever since the research work by the LIMAS group, the works of Alfred Hoppe have formed an important element of these developments (5). Because of their high degree of formalization they are, in addition, particularly suited for being used in application systems, e.g. for text analysis or machine translation. With the two parts of the "Semantic Syntax" (1981/1991) material of versatile applicability is now available.

Nevertheless the fact should not be underestimated that it is a long (and moreover expensive) way from a theoretically well-founded description - despite the relatively broad material basis - until practical application. The greatest chances for practical application of this work are to be found, in my opinion, in the field of machine translation and - as occasionally practiced by Hoppe himself - in the (newly developing) field of so-called teachware, i.e., of computer-assisted learning.

We are gratefulfor the fact that now not only Alfred Hoppe's conceptual structure is complete, but that also a broad material basis for specific applications has now been created. To the best of my abilities I will try to contribute to making his concepts and models directly and indirectly effective. To the jubilar I wish - from the bottom of my heart - further creative energy and health.

References

(1)Eggers, Hans et al.: Elektronische Syntaxanalyse der deutschen Gegenwartssprache. Tübingen 1969.
(2)Gross, Maurice: Grammaire transformationelle. Syntaxe du verbe. Paris 1968.
(3)Hoppe, A.: Grundzüge der Kommunikativen Grammatik, Teil 1: Die semantische Syntax der Geschehen-Komplexe. Bonn 1981.
(4)Hoppe, A.: Grundzüge der Kommunikativen Grammatik, Teil 2: Theorie der semantischen Syntax: Feste Fügungen. Bonn 1991.
(5)Luckhardt, H.-D., Zimmermann, H.H.: Computergestützte Übersetzung - Praktische Anwendungen und angewandte Forschung. Saarbrücken 1991.
(6)Panyr, J., Zimmermann, H.H.: Information Retrieval: Überblick über aktive Systeme und Entwicklungstendenzen. In: Batori, I., Lenders, W., Putschke, W. (Eds.): Comput.Linguistics. Berlin/New York 1989. p.696-708