TOWARDS A CORE VOCABULARY FOR A NATURAL LANGUAGE SYSTEM

Hubert Lehmann
IBM Deutschland GmbH, Scientific Center
Institute for Knowledge Based Systems
Wilckensstr. 1a
D-6900 Heidelberg, Germany
Email: LEH at DHDIBM1.BITNET

ABSTRACT

The desire to construct robust and portable natural language systems has led to research on how a core vocabulary for such systems can be defined. Statistical methods and semantic criteria for doing this are discussed and compared. Currently it does not seem possible to precisely define the notion of core vocabulary, but it is argued that workable criteria can nevertheless be found. Finally it is emphasized that the implementation of a core vocabulary must be seen as a long-range research program rather than as a short-term goal.

Motivation

Reasearch on natural language processing systems today strives for the construction of **robust** and **portable** systems.[1] A system is robust, if it can handle a large variety of user inputs without giving up or producing unexpected results. A system is portable in the sense intended here, if it is not geared to a single subject domain, but can be ported with a reasonable effort to a variety of subject domains. It is common understanding that there exists a central fragment of a language which 1. is required for dealing with virtually any subject domain, and 2. is invariant with respect to meaning and use accross subject domains. It is of course a non-trivial empirical question whether such a central fragment really exists, and if so, to say what it is, but a number of researchers seem to share the assumption that it does (cf. e.g. Alshawi et al. (1988)). Any robust and portable system would then have to handle this core fragment.

In this paper I am concerned with a second – related – assumption, namely that there exists a **core vocabulary** which is needed for handling any subject domain. This assumption is also shared by many researchers, and it underlies the production of basic vocabularies for language learning such as Oehler (1980). Usually the authors claim that their word lists are based on statistical investigations, but they also emphasize that they did not slavishly stick to the statistics but used additional criteria such as "usage value", "availability", "familiarity", or "learnability" without ever saying how these are established.[2]

I will address the following questions:

1. How can the intuitive notion of **core vocabulary** be properly defined?
2. How can statistical methods be employed to define a core vocabulary and how do they relate to semantic criteria?
3. What semantic criteria can be found to define a core vocabulary?

Definitions of a core vocabulary

There are several ways to define **core vocabulary**, I can think of the following three:

1. The core vocabulary consists of the *n* most frequent words of a language.
2. The core vocabulary is that vocabulary which is common to all native speakers of a language.
3. The **semantic** core vocabulary consists of those words which suffice to define all of the remaining vocabulary of a language.

The first two definitions call for statistical methods which shall be discussed in the next section, and the third one obviously requires a semantic approach which shall be discussed in section "Semantic criteria".

Statistical methods

Frequency counts have well established the basic properties of the frequency distribution for text corpora. Thus in Kucera and Francis (1967) we get coverage figures like this for their complete corpus of about 1 million tokens:

```
  10 most frequent words:    24.26 %
 100 most frequent words:    47.43 %
1000 most frequent words:    68.86 %
```

---

[1] The research described here has been conducted in the context of the LILOG project (Herzog et al., 1986). It has profited from intensive discussions with R. Mayer. Much of the underlying statistical work on text corpora is due to U. Bandara and G. Walch from the speech recognition project SPRING (Wothke et al., 1989).

[2] Our investigations are based on German, but for ease of reference also some English examples are given.

These figures vary only slightly with corpus size, and also for German similar values are observed. However, while coverage figures are rather stable with respect to the n most frequent words of a corpus, what are the n most frequent words may vary widely with corpora or subcorpora. Two parameters responsible for this variation are obvious:

1. Subject matter and
2. Communicative function.

Thus in the "Kultur" section of a newspaper which we have analyzed we see that words like *Musik, Theater, Regisseur*, etc. occur with a drastically higher frequency than in the other sections, which of course can be attributed to subject matter. But personal pronouns, in particular 1st and 2nd person pronouns, also show a much higher frequency, and this can hardly be attributed to subject matter, rather to different communicative functions of feuilletonistic writing and say economic news.

All of this relates of course to the much discussed issue of what constitutes a representatitve corpus for statistical linguistic analysis. Since specific subject matters and communicative functions vary in importance for different speakers of a language, it will be difficult if not impossible to eliminate arbitrariness. Rather, a definition of representative corpus must take into account the research goals pursued.

For a natural language system which is supposed to analyze and generate texts, to engage in dialogues with users, and which is to acquire knowledge from the analysis of definitions and rules formulated in natural language, one needs a corpus of texts where all these aspects are sufficiently represented. We were able to draw upon a variety of corpora none of which would show all the features required, but the combination of them seems to be quite reasonable.

We compared the following five word lists:

1. Oehler (1980): Grundwortschatz consisting of 2247 words,
2. Erk (1972): scientific texts from 34 disciplines, 1283 words with frequency ≥ 20,
3. Pregel/Rickheit (1987): texts by primary school children, 593 words with frequency ≥ 20,
4. SPRING-corpus of newspaper texts, 2733 most frequent words,
5. DUDEN (1989): definitions for words beginning with *a*, 2693 words with frequency ≥ 4.

From these, word lists B_n were formed consisting of those words occurring in at least n of the original word lists ($1 \leq n \leq 5$). The lengths of these lists are B_1: 5409, B_2 : 2248, B_3: 1215, B_4: 565, and B_5 : 116.

The size of B_5 shows that a really common core of a variety of texts may be extremely small, the successive loosening of restrictions used here allows for a balanced extension of this very small core. The list B_3 was chosen as the **statistical core vocabulary** serving as a base for applying semantic criteria, because the overall core vocabulary was envisaged to have a size of approx. 1500 words. Inspection shows that many intuitively basic words and very few idiosyncratic words are contained due to the method of intersecting the word lists. Hence, B_3 seems quite reasonable.

Semantic criteria

If one takes the n most frequent words of any frequency count one will no doubt discover that these words will not exhibit a **linguistic closure** in the sense that natural sentences can be formed with all and only the words in the set. Further one will see that semantic relations will be incomplete. Thus one finds in Oehler (1980) which is based on the old Kaeding count that *weiblich (female)* occurs but not its antonym *männlich (male)*. For a core vocabulary to be set up for a natural language system, I think, one must strive for linguistic closure, since otherwise, one ends up with words one cannot use. This means that you cannot base the core vocabulary on frequency counts alone.

Furthermore, one cannot expect that one will have just the vocabulary needed to formulate definitions for the words in the list chosen. To avoid circularity, one will have to accept that certain words cannot be defined within the vocabulary, but one will also have to accept that for some words less than complete definitions can be given. Because of this lack of definability, a **semantic core vocabulary** can only be understood as an approximative notion geared towards "the best one can do". What one can hope to do, is to define

1. taxonomic relations,
2. "selectional restrictions" or **constraints on semantic compatibility**, and
3. meaning rules of arbitrary complexity (including classical definitions).

I propose to formulate all of these types of rules in natural language for B_3 trying to stay within at least the vocabulary of B_1 , to add the words used in the formulations to the original set, and continue until one cannot think of further rules. I claim that one can achieve a fixed point from where on no new words are added to the set, and that at this moment one has reached a rather good approximation to a semantic core vocabulary.

There is undoubtedly a relationship between frequency and semantic relevance: since taxonomic relations are often exemplified by anaphoric references, since semantic compatibility constraints lead to the co-occurrence of ap-

propriate words, and since other more complex semantic relationships are bound to be exhibited in the various threads of discourse, one has all reason to expect a certain amount of congruence between frequency counts and the semantic core vocabulary as defined above.

The work on formulating taxonomic relations, semantic constraints and other meaning rules is underway, and since it will involve all of the vocabulary, linguistic closure will be achieved at the same time.

As an example, take a taxonomic rule for *Arm* which is in B_5

Jeder(B_3) Arm ist Teil(B_4) eines Körpers(B_3).
(Every arm is part of a body.)

The word *Körperteil (body part)* is only available in B_1 and was therefore not used, or instead of *Teil* one could also have used *Glied*(B_3, *member*), but then the rule would not have covered arms of machines or rivers. This highlights a big problem in the natural language formulation of meaning rules: how is ambiguity dealt with? Space does not permit a full discussion here, therefore suffice it to say that it is one of our research goals to formulate meaning rules which specify criteria for disambiguation.

Linguistic description

The preceding discussion has concentrated on how to establish a core vocabulary. Now a few brief remarks shall follow on how the words of the core vocabulary can be linguistically described.

The morphology of languages such as German is well understood and has been coded for an extended vocabulary in the lexical database of the LEX project (Barnett et al., 1986). This database also contains detailed syntactic information, in particular on government patterns.

It is the description of the semantic (and pragmatic) properties of many words one would obviously want to include in a core vocabulary that will confront us with huge unsolved theoretical problems. Be it modal verbs or propositional attitudes, sentence adverbs or "abstract" nouns of various kinds. Investigations on some individual words have generated heaps of literature, for others it seems that people have not even dared to look at them. Does this make the enterprise of implementing a core vocabulary a futile one? I think not. I think the implementation of a core vocabulary should be seen as a long-range research goal for both computational and theoretical linguistics, and furthermore that natural language systems provide a good environment for doing experiments in semantics, because they encourage an integrated treatment of linguistic phenomena.

Conclusions

Our research on establishing a core vocabulary for German in the framework of the LILOG project has revealed that currently no absolute definition can be given, but ways have been shown how to arrive at a working definition with respect to the objectives of natural language systems. It has been shown that both statistical methods and semantic criteria can, and I think, have to contribute to the establishment of a core vocabulary.

The linguistic description and thus the implementation of a core vocabulary depends heavily on progress in theoretical linguistics, in particular in semantics and pragmatics, but I want to stress that focussing on a core vocabulary is a fruitful way to direct linguistic research, which can be supported by the need for integrated treatments in natural language systems.

References

Alshawi, H., D. M. Carter, J. van Eijck, R. C. Moore, D. B. Moran, F. C. N. Pereira, and A. G. Smith (1988): "Research Programme in Natural Language Processing – Annual Report", *Nattie Project Document NA-16*, Cambridge: SRI International.

Barnett, B., H. Lehmann, M. Zoeppritz (1986): "A word database for natural language processing", *Proceedings 11th International Conference on Computational Linguistics COLING86 August 25th to 29th, 1986, Bonn, Federal Republic of Germany*. 435-440.

Erk, H. (1972): *Zur Lexik wissenschaftlicher Fachtexte*, München: Hueber.

Herzog, O. et al. (1986): "LILOG – Linguistic and Logic Methods for the Computational Understanding of German", *LILOG-Report 1b*, Stuttgart: IBM Deutschland.

Kucera, H., W. N. Francis (1967): *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Oehler, H. (1980): *KLETT Grund- und Aufbauwortschatz Deutsch*. Stuttgart: Klett.

Pregel, D., G. Rickheit (1987): *Der Wortschatz im Grundschulalter*. Hildesheim: Olms.

Wothke, K., U. Bandara, J. Kempf, E. Keppel, K. Mohr, G. Walch (1989): "The SPRING Speech Recognition System for German", in: *Proceedings of Eurospeech '89. Vol. 2*, 9-12.