SUMMARY OF SOME COMPUTATIONAL AIDS FOR OBTAINING A FORMAL SEMANTIC DESCRIPTION OF ENGLISH

by John Olney , Carter Revard, and Paul Ziff

The work reported herein was supported by contract F1962867C0004, Information Processing Techniques, with the Electronic Systems Division, Air force Systems Command, for the Advanced Research Projects Agency Information Processing Techniques Office.

System Development Corporation
2500 Colorado Avenue
Santa Monica , California 90406

Some Computational Aids for Obtaining a Formal Semantic Description of English

by

John Olney, Carter Revard, and Paul Ziff

System Development Corporation

2500 Colorado Avenue

Santa Monica, California 90406

The goal of a formal semantic description of a language may be characterized approximately as follows: to specify a set of lexical entries and a system of rules sufficient to provide appropriate readings for both individual sentences and multi-sentence connected discourses in the language, given appropriate structural descriptions of the sentences and discourses. (It cannot be required that only appropriate readings be derivable from the description, since semantic disambiguation often requires extralinguistic information. However, one would expect that the appropriateness of the readings derived will be correlated with their provision for inferences semantically warranted by the text segments in question.) As a contribution toward developing a formal semantic description of English (henceforth abbreviated as FSDE), we have transcribed two English dictionaries into machine-usable form and are developing programs for processing the information contained therein. The dictionaries are Webster's Seventh New Collegiate Dictionary (W7) and The New Merriam-Webster Pocket Dictionary (MPD), both published by G. & C. Merriam Co. Apart from pictorial illustrations and tables, the transcript of W7 is complete, even to the extent of preserving virtually all typographic distinctions made in the original.

We make the simplifying assumption that the word and phrase senses to be accounted for by the lexical entries and rules of an FSDE include all the senses described for entries in W7. Our routines for processing the W7 and MPD transcripts are designed to aid in the discovery of:

(a) rules for obtaining certain of the senses described for W7 entries from other senses described for the same entries or from senses described for other W7 entries from which the first were derived morphologically; and

(b) semantic components and rules for combining them to yield specifications of senses that cannot conveniently be obtained by the rules referred to in (a) above.

A first approximation to the set of lexical entries of an FSDE could be specified when the rules and semantic components referred to in (a) and (b) have been discovered; their discovery will of course require many, many man-years of effort and also close collaboration with researchers in English syntax. The basic contribution of our dictionary processing routines to this enterprise will be to permit semiautomatic generation of lexical fields for the main entries in W7. One type of lexical field which we plan to generate would display all words and phrases that are W7 entries and have been derived with the addition of a given affix or have been derived from a given base word by any morphological process. Appended to the field would be a printout listing the senses given in W7 for the words and phrases in the field. We have already constructed such derivational fields manually for a considerable number of affixes. By providing syntactic feature characterizations of the usages of the entries in each field, we have often been able to discern which element (if any) of the sense shift(s) accompanying the addition of the affix corresponds to a shift in syntactic function. In general, we anticipate that the derivational fields will prove quite useful as a computational aid for discovering the rules referred to in (a) above.

We also plan to generate semiautomatically a type of lexical field closely resembling what has traditionally been called a semantic field, viz., to construct for a given sense described at some entry in W7 a labeled network in which the nodes are other senses described in W7 that are closely related semantically to the given sense, and the labels indicate the type of relationship, e.g., synonymy, antonymy, etc.

A major objective of our current research is to determine, for the types of uses of the fields that we now foresee, exactly which relationships between word senses should be considered constitutive of semantic fields, and to what extent these relationships are signaled by machine-recognizable characteristics of W7 sense descriptions. Our results to date indicate that: (1) the semantic field output for a given word sense in the form of such network should be based on occurrences of the word in that sense, and occurrences of the nonfunction words used to describe that sense, in other sense descriptions provided in W7; and (2) a variety of other clues involving the syntax of defining phrases, stylistic regularities observed in W7, derivational morphology, the etymological information provided in W7, etc., should be taken into account before the field is output.

As part of our stylistic study of W7, we are working out conceptual analyses of the meanings of nonfunction words and phrases that are used very frequently in W7 definitions, starting with those chiefly used in defining derived senses. We will use these analyses to develop precise terminology for describing semantic relationships.

Probably the most frequent and serious errors in our mechanical tracing of fields will result from establishing links (in the network) between word senses whose descriptions in W7 contain a pair of word occurrences that are

spelled the same, are functioning as the same part of speech, but are used in radically different senses. Clues based on stylistic regularities in W7 should help in disambiguating words occurring in sense descriptions, but the availability of MPD in machine-usable form should help still more. Although MPD describes only one-sixth as many senses as W7, nonetheless, for almost all word occurrences in W7 sense descriptions it defines the senses appropriate for those occurrences. Moreover, each MPD sense description is usually such a straightforward abbreviation of one or more W7 sense descriptions that our routines will generally be able to determine which W7 sense description(s) it abbreviates. Words occurring in W7 sense descriptions can then be disambiguated semantically to the extent of provisionally eliminating senses described for them in W7 which are not abbreviated in MPD. No doubt the tracing of semantic fields by our routines will still have to be monitored closely by the user, at least initially.

Edward Bendix has already shown in his recent dissertation how a plausible approximation to specifying semantic components of each of a group of words closely related semantically can be obtained by applying a series of semantic tests to all pairs in the group. Accordingly, we anticipate that our semiautomatically generated semantic fields will be used mainly to aid in discovering the semantic components and rules referred to in (b) above --more particularly for discovering components of words which, unlike kinship and color terms, do not belong to domains of the English vocabulary that are obviously well-structured.

As a further contribution toward discovering the elements of (b), we are attempting to arrive at a gross approximation to a set of semantic primitives for English by roughly the following procedure: Where A is

the set of words occurring in W7 sense descriptions, find B, the set of words occurring in the sense descriptions of the members of A, and print out all pairs like 'infantile paralysis' and 'poliomyelitis', each of which occurs (as a crossreference) in the sense description of the other. Wherever possible eliminate one member of every such pair from B, then find C, the set of words occurring in the sense descriptions of the remainder of B, and iterate the process until the set most recently found is not significantly smaller than the set found immediately before it. The members of the smallest set found will then be scanned to see whether any further eliminations can be made. Preliminary results obtained by this procedure will be presented, together with the results of our stylistic analysis of W7 and our study of derivational processes. Applications of our routines for processing the W7 and MPD transcripts to lexicography, and to tasks in computational linguistics not directly concerned with semantics, will be discussed briefly.