CONCEPT ASSOCIATION

SALLY YEATES SEDELOW
Department of Computer Science
University of Arkansas at Little Rock
Little Rock, AR 72204

ABSTRACT

The complement to decomposition in scientific research is composition. In human language computing, composition is achieved by way of semantic association and the generation of strings of entities. That generation of strings takes place progressively: e.g., strings of symbols (words), strings of strings (sentences), strings of strings of strings (paragraphs), etc. The mathematical (topological, graph-theoretic) analysis of Roget's Thesaurus (1962) has opened a door onto a broad vista of potential achievements in such areas as artificial intelligence and expert systems, through the analysis of concept association, or concept composition.

For the purposes of natural language (e.g., English) concept association, words are salient. But the importance of words as bearers of concepts has come to be recognized, within recent linguistic paradigms, as something of a "spin-off" from efforts to describe the structure resulting from combining symbols and word strings into sentences. As Chomsky (1965) described sentence structures, syntactic part-of-speech designations and rules for acceptable syntax were primary; semantics comprised interpretations mapped onto the syntax. But Chomsky popularized the tree-graph representation of sentences, intending it as a description. Nonetheless, as he showed (still arguing that it was a description), once a descriptive tree is available, it is possible not only to ascend the tree (e.g., from the word, table, to the part-of-speech, noun, to the composite category, noun-phrase [in combination with, for example, an article], to the composite category, sentence [in combination with a verb phrase]), but it is also possible to start at the top (sentence) and generate the rest of the tree by a series of rewrite rules (e.g., S → NP + VP). One can hypothesize that once the idea of generating sentences gains power, one begins to consider what in fact seems to happen, or be uppermost in our minds, when engaged in such generating. Most of us do not think first of such strings as Determiner-Noun-Verb-Determiner-Noun but rather of some concept(s), or information to be conveyed in the form of words or collections of words. Not surprisingly, some of Chomsky's students became known for their work in generative semantics; more recently still, linguists schooled within the transformational paradigm are speaking of lexical-relation grammars. Thus, within a paradigm growing out of the earlier structuralist paradigm, with emphases in both cases upon the syntax of strings of strings (sentences), there is now a strong emphasis upon the lexicon, upon the words, and their meaning, strung together into sentences.

With the paragraph (strings of strings of strings), we enter the realm of discourse analysis — of trying to explain why we perceive a paragraph, or a chapter, or a book as unified, or 'holding together' as a single piece of text. For the English language, one of the more rigorous analytical systems was developed by Halliday and Hasan (1979), who explored and listed the kinds of "ties" that bind texts together. Although some of the general categories are at least meta-syntactic (e.g., the category of reference, especially pronominal reference), in fact the emphasis is upon the lexicon, upon words which are actually present, as is the case in most of the categories employed, or are assumed to be present, as in elided segments that repeat some aspect of earlier content. Clearly, a pronoun such as "he" will refer to some specified single male elsewhere in the discourse; hence, the two instances will refer to the same individual and, in the act of doing so, bind the text together. Under the category of "substitution," words such as "thing" are made concrete by other words (as with "This is the thing....) and, of course, there are many types of ties comprising words that are closely related semantically, i.e., they can be used in very similar contexts.

Within the field of artificial intelligence, and very notably artificial intelligence as embodied in expert systems, there is a great need for good natural language interfaces, sometimes only as "front-ends" but sometimes comprising the backbone of a given system. For the development of sophisticated tutoring systems or computer-based consultants or any interactive system for which the preferred mode of interaction is a natural language such as English, significant progress can only derive from much more conceptually agile (and therefore more wide-ranging as to domain) systems than are available today. The research discussed here focuses upon large general-purpose lexicons, notably thesauri and more particularly, Roget's International Thesaurus, Third Edition (1962), which have the great advantage of being culturally-validated — that is they have been used by, in the case of Roget's, English-language speakers for many decades and are thus descriptive of the language as it is used, in part because they are also prescriptive. This research is directed toward building a foundation for intelligent systems that can range as widely as the semantic space for an entire language permits them to. Further, it has long been evident that approaches to natural language knowledge representation are so laborious as to daunt entire generations of graduate students likely to be involved in building those representations — hence, given the approach advocated by many of building conceptual structures "from scratch," the possibility of intelligent and expert systems able to work within any but very restricted domains is remote. It seems desirable, therefore, to redouble our efforts to use, at least as a basis for a domain-transcendent natural language knowledge representation, the very large lexicons already available which people find adequate for many semantic discriminations.

After examination of a number of lexicons, our research focused in upon Roget's, which has both an explicit and implicit structure, both deriving from concept association. The explicit structure is hierarchical (N.B., that the Roget's used here is not set up like a dictionary) with seven or eight levels, depending upon how they are specified. At the top are eight very general categories and at the bottom of the hierarchy are the individual words, grouped together with other, closely semantically associated, words in a semicolon group (i.e., the boundaries are formed by semicolons). An example of such a group would be: inspiration, inhalation, indraft or indraught, inflow, inrush, suffiation, insufflation, afflation, afflatus. The implicit structure depends upon the multilocality of a number of words in the Thesaurus — a given word may appear in a number of different places in the Thesaurus, thus providing the means for rulefully traversing the Thesaurus cross-hierarchically. Bryan (1973) has produced an elegant mathematical model of the Thesaurus which, among other properties, defines aspects of the implicit structure, including chains, stars, and neighborhoods.

Two questions immediately come to mind concerning a resource like Roget's: 1. Is it really a reliable guide to English semantic space as defined by cultural usage? and 2. Can it provide the types of relationships needed in computer-based natural language implementation?

In answer to the first question, empirical research suggests that the Thesaurus can be accurately regarded as the skeleton for English-speaking society's collective associative memory. Briefly, the ways in which this conclusion has been tested are: A. the determination of when the initial characters in an English word are functioning as a prefix;

our working hypothesis was that when a possible prefixed word form and its possible stem occur close together within the Thesaurus, they probably can be considered as a stem and a prefixed form; (This assumption was borne out in a large majority of cases, while at the same time the Thesaurus correctly showed, e.g., that the words "vent" and "prevent" cannot be grouped together as stem and prefixed form. Thus, as a measure of 'semantic distance' between English words, the Thesaurus is useful in a satisfying way in this test [Warfel, 1971-2; Sedelow, 1969; Sedelow, 1985a; Sedelow, 1985b, 1988]); B. early experiments with content-analytic programs using the Thesaurus showed that it provides semantic clustering conformal with experience and expectations as to usage patterns (Sedelow and Sedelow, 1969); C. starting from a low level (indicated by semi-colon boundaries, hence called semicolon groups) with the syntactic subset comprising verbs, navigation of 'chains' based on the topological model (Bryan, 1973; Patrick, 1985; Sedelow and Sedelow, 1986) produced distinctions among homographs — a very important achievement, given the pervasiveness of ambiguity in multi-domain natural language knowledge-bases, user queries, and system responses, and thus problems involving disambiguation; D. a distribution of the so-called Chinese simplicia, as categorized by Karlgren (1923), against categories in Roget's showed semantic gaps conformal with observations made more 'anecdotally' by scholars comparing aspects of Chinese and English (e.g., book and private conversation of Bloom [1981]); E. research exploring the interaction between the Thesaurus and abstracts of articles in the 1985 SCAMC (Symposium on Computer Applications in Medical Care) Proceedings which produces a conceptual overview of the abstracts based on intersections between textual context and thesaurus concepts — for which the results are quite satisfactory (Brady, 1981); F. a distribution of the UNIX Spelling Dictionary against terms occurring in the Thesaurus shows a very high correlation with the grouping of entries in the Thesaurus as to semi-colon group, paragraph, category, etc. A distribution of the Oxford Advanced Learner's Dictionary against the Thesaurus also has produced a very high correlation; G. inasmuch as the sentence "Time flies like an arrow" is a classic in discussions of ambiguity, it is worth noting that the Thesaurus, again interacting with the text of the sence (Brady, 1989), produces the reading that seems often to come to mind first, i.e., the speed with which time goes by.

These studies of the Thesaurus — studies that range over many different kinds of problems/applications and many types of text — suggest that it is appropriate to conclude that the initial working assumption as to the potential usefulness of a culturally validated resource such as Roget's has been substantiated with specific reference to Roget's. We do not claim that the Thesaurus is 'perfect' — we can ourselves generate examples of desirable modifications and additions that are well grounded theoretically — but it has performed very well in a number of tests over rather wide-ranging discourse domains.

The answer to the second question — Can the Thesaurus provide the types of relationships needed in computer-based natural language implementation? — must be brief. It is our hope that the associative structure of the Thesaurus will lend itself to the kinds of approaches suggested by genetic computing (Walbridge, 1989) and natural computation (Richards, 1988) thus obviating the need for some of the structures currently being used for natural language interface programs. Insofar as additional relationships are desirable, some of them can be derived from the Thesaurus as it stands. For example, the hierarchical structure will specify IS-A (taxonomic) relationships and can help deal with the issue of "inheritance" (as in frames). Insofar as stereotypes (e.g., user expectations in a given context), in their more general sense, are culturally-prescribed and induced, the networks of associated terms in the Thesaurus ought to provide good indices to stereotypes; at least as they function to elicit knowledge of the user,one or two nodes in the net should serve to "haul in" sets of related terms which might either exist in the user's current knowledge base, and thus help characterize the user, or which could be tutorially aded, and thus help inform the user. These are domains within which we now are proposing to do research, along with a number of others which time and space prevent describing. As is evident, though, concept association is the key to our approach as it is, we believe, the key to natural language communication and understanding.

LITERATURE CITED

BLOOM, ALFRED H. 1981. Linguistic shaping of thought: a study in the impact of language on thinking in China and the West. New Jersey: Laurence Erlbaum. 128 pp.

BRADY, JOHN. 1988. ICSS (Interlingual Communication Support System) and a Wittgensteinian language game. Proceedings, European Studies Conference, University of Nebraska/Omaha. pp. 20-27.

BRYAN, ROBERT. 1973. Abstract thesauri and graph theory applications in thesaurus research, in S. Sedelow, ed., Automated language analysis, 1972-1973. University of Kansas Departments of Computer Science and Linguistics, Lawrence. pp. 45-89.

CHOMSKY, NOAM. 1965. Aspects of the theory of syntax. The M.I.T. Press, Cambridge. 251 pp.

HALLIDAY, M.A.K. and RUQAIYA HASAN. 1979. Cohesion in English. London: Longmon Group Limited. 374 pp.

KARLGREN, BERNARD. Analytic dictionary of Chinese and Sino-Japanese. 1923. (Reprinted by Dover, 1974 and by Taipei, Ch'eng-wen Publishing Company, 1966.) 448 pp.

PATRICK, ARCHIBALD. 1985. An exploration of an abstract thesaurus instantiation. M.S. Thesis, Computer Science Department, University of Kansas, Lawrence. 122 pp.

RICHARDS, WHITMAN, ed. 1988. Natural computation. The M.I.T. Press, Cambridge. 561 pp.

ROGET'S INTERNATIONAL THESAURUS, third edition. 1962. New York: Thomas Y. Corwell. 1258 pp.

SEDELOW, SALLY YEATES. 1969. Prefix, in S. Sedelow, ed., Automated language analysis, 1968-1969. University of North Carolina/Chapel Hill, Departments of English and Computer & Information Science. pp. 12-25.

SEDELOW, SALLY YEATES. 1985a. Computational lexicography. Computers and the Humanities. 19:2, 97-101.

SEDELOW, SALLY YEATES. 1985b. Computational literary thematic analysis: the possibility of a general solution. Proceedings of the 48th ASIS annual meeting. 22:359-362.

SEDELOW, SALLY YEATES with DONNA WEIR MOONEY. 1988. Knowledge retrieval from domain-transcendent expert systems: II. research results. Proceedings of the 51st annual meeting of the American Society of Information Science. 25:209-212.

SEDELOW, SALLY YEATES and WALTER A. SEDELOW, JR. 1969. Categories and procedures for content analysis in the humanities, in George Gerbner, ed., The analysis of communication content. New York: John Wiley & Sons, Inc. pp.487-499.

SEDELOW, SALLY YEATES and WALTER A. SEDELOW, JR. 1986. Thesaural knowledge representation. Proceedings, Waterloo University (Ontario) Conference on Lexicology. pp. 29-43.

WALBRIDGE, CHARLES. 1989. "Genetic algorithms: what computers can learn from Darwin," Technology Review, M.I.T. January, pp. 47-53.

WARFEL, SAM. 1972. "The value of a thesaurus for prefix identification," in S. Sedelow, ed., Automated Language Analysis, 1971-1972. University of Kansas Departments of Computer Science and Linguistics, Lawrence. pp. 31-49.