Computers and Translation 2 (1987)

Semantic Space

Walter A. Sedelow, Jr., and Sally Yeates Sedelow

Walter A. Sedelow, Jr. is Professor of Computer Science, Univ. of Arkansas at Little Rock; and Adjunct Professor of Electronics and Instrumentation, Graduate Institute of Technology, Univ. of Arkansas. Sally Yeates Sedelow is Professor of Computer Science and Adjunct Professor of English, Univ. of Arkansas at Little Rock; and Adjunct Professor of Electronics and Instrumentation, Graduate Institute of Technology, Univ. of Arkansas.

In an earlier article, 'The Lexicon in the Background' (Sedelow & Sedelow 1986) for this journal, we provided detailed description of the conversion, for our research, of *Roget's International Thesaurus*, 3rd Edition (1962), into computer-accessible form. In addition a formal mathematical model of thesauri initially devised by Robert Bryan (1973) was presented, along with a summary of research by Dale (1979) and Patrick (1985) based on that model.

In this article we attempt to provide the theoretical and intellectual rationale for the approach we have taken to the study of natural-language semantics, placing it within the context of work which, in some instances, may be seen as representing an alternative to our approach and, in others, may be seen as complementary; while in still others, the scholarly and scientific knowledge cited has clearly been foundational to our own structuring of research on semantics. At this stage, we certainly make no claims for the architectural distinction of the edifice which we hope is now taking shape, but we have conducted sufficient empirical research (Dale 1979; Patrick 1985; S. Sedelow 1969; S. Sedelow 1985; W. Sedelow 1985; Warfel 1972) on both the explicit structure (as defined by Roget) and the implicit structure (as included in the Bryan model) of the *Thesaurus* as to have confidence in its reliability as a reasonably good guide to semantic relationships as they are embodied in the use of English.

As noted at the close of 'The Lexicon in the Background':

> 'In the interest...both of forwarding the cause of better machine translation -- prospectively in the oral mode as well as with reference to written text -- and also, more profoundly, of contributing to the development of linguistics, we are moving on with new research efforts into the lexical and semantic structures of dictionaries and thesauri, in addition to looking into the lexical and semantic properties of, hopefully, 'randomly' chosen bodies of unrestricted text. The objective is to build a rich and a useful integrated body of science as to knowledge representations for a substantial and comprehensive subset of the English language.' (Sedelow & Sedelow 1986:80-1)

It is ever more apparent that natural language computing is coming to include a definable sub-specialty of computational lexicology, which is being fostered in part by the growing awareness (with a far greater prospective awarenesss in the future) of the need for effective *applied* computational lexicology. To take but one exemplary phase of the need to mobilize computational lexicological knowledge, engineering and technology (especially high technology but also conventional technology) are a powerful case in point, e.g. the growth of communicational demand between, say, Chinese users of Western and (particularly) American technology and U.S., or at least English-speaking, opposite numbers moves forward apace. A conservative current estimate gives us fifty million learners of English in China; and with all due regard for the intensities and motivations of those students of English, it is not unfair to indicate that most or at least many of those users are going to find that their facility in English is

sufficient for effective scientific and technical communication only if there is supplementation -- and, practically speaking, that means in the computer modality -- of their understanding with semantic prosthetics (and occasionally syntactic prosthetics) to clarify exact meanings. For some types of engineering and product/service adoption by the East from the West, it is especially crucial that fine-grained and operational understanding be of a high level of quality. One need only think of petrochemical processing plants and various sorts of medically related technologies to see how crucial it is to avoid ambiguities, especially operational ambiguities, with reference to these bodies of knowledge.

Our own research efforts are focused, in part, on the lexeme and concentrate on associational lexemic structures. We are taking a formalized and mathematical approach to the strength of semantic associations among lexemes through the study of clustering patterns at the lower levels of the *Thesaurus* hierarchy. We have already accomplished a great deal in the way of understanding, i.e. graph-theoretically speaking, the connectivities of the *Thesaurus*, and we will use some of the strength metrics which we have been exploring to create a flexible representation of those lexemic connectivities. That formalized modeling -- where *formalized* is used in the sense that was established by us in the contrasting of *formal* and *formalized* (see Sedelow & Sedelow 1979; 1983) -- will give us a mode of knowledge representation appropriate to accounting for our understanding of the language which has been poured into the *Thesaurus* hierarchy of Roget, and of the 'world' contained, if only (in part) by implication, in that linguistic mold which we call the *Thesaurus*.

The model-relative and model-theoretic character of everyday natural language recently has been highlighted in such work as that of Sowa (1984), where issues of incommensurability across discourse domains are both explicit and implicit. The studies of Robert Hingers, a former student of ours at Kansas, are at the moment serving to bring together crucial aspects of our understanding of the dialectic in the debate over the nature of mapping and commensurability (im)possibilities across lexical and speech domains, even intra-lingually, and even within a scientific discipline. Efforts by Lowe, including his recent article in the *International Journal of Man-Machine Studies*, and others provide yet another route -- in that instance, a route utilizing the work of Stephen Toulmin -- toward some measure of formalization at least of the semantics in scholarly transactions and within a scholarly *traditio*. For a more rigorous appreciation of what can be done on that front -- including what can be done utilizing certain sorts of computerized graphics -- attention should be especially directed to the work of John Bristow Smith (1983) of the University of North Carolina at Chapel Hill, as represented in his chapter on 'Computer Criticism'.

It must be evident that both implicitly and explicitly our approach to the work that we have undertaken and are undertaking, as well as our sense of its significance, is in no small measure governed by a feeling for the great importance of general, non-domain-specific approaches to the building of intelligent systems, including such natural language computing intelligent systems as are used in conversational computing, and in other forms of AI capability, that one generally finds in consultant and expert systems. In computer science there has been an implicit bifurcation of orientation -- some might be inclined to say, loosely, of paradigms -- with reference to the appropriate scale of the domain for computer-based natural language interfaces, etc. One orientation is well represented by the work of such a Carnegie-Mellon-style approach as is taken by Edward Feigenbaum (Barr, Feigenbaum, & Cohen 1981). Feigenbaum and others of that school tend to study extremely constrained semantic domains (within a single language) which frequently are constricted even further by being devised in an almost Weberian way ('means-to-end rationality', Max Weber) in order to accomplish specific procedural goals. In *The*

Handbook of Artificial Intelligence, the authors say that 'in AI a representation of knowledge is a combination of data structures and interpretative procedures that, if used in the right way in a program, will lead to "knowledgable" behavior' (Barr, Feigenbaum, & Cohen 1981). It would be hard to imagine a more restricted approach; but it should be emphasized that in this instance the use of the term *restricted* is not meant to be value-laden, but only to call attention to the scale of domain. Within the MIT computer science community, Patrick Winston is a leading exponent of a similar emphasis on the desirability of a concentration on only limited domains of discourse.

There is, to be sure, a well-established general tradition in science that it is sometimes wise to proceed from a simpler, more restricted case to more complex cases, e.g. to start with the hydrogen atom. But as we well know, the (sometimes tacit) choice of scale, especially smallness of scale, may create its very own problems which can only be fruitlessly attacked within its limitations. One thinks, for instance, of the problems that would emerge from teaching anatomy exclusively on the basis of a single cell as the largest basic unit, with all other structures construed as merely combinations of cells; with such an approach, even physiology would be far more difficult than it need be. Or imagine doing a macro-economics with only micro-economic models; the models would then be a part of the problem rather than a part of the solution (what they doubtless would be passed off as).

It seems abundantly clear that for all the short term gains with work -- albeit heavily pump-primed (*contra* Rosenblatt's Perceptron) -- which has given us very micro-world (if not toy) applications, there is in parallel an extremely attractive set of opportunities available in the area of computational linguistics (broadly construed) and knowledge systemics (W. Sedelow 1968), where the unit of analysis is a fuller subset of language than what has been attended to in the more massively financed AI undertakings to date. One must be extremely careful not to confuse the lack of success of some previous general approaches -- such as the earlier phases, at least, of The General Problem Solver -- with any necessary limitation on that sort of approach, nor *a fortiori*, with what is possible by using certain current techniques. At the very least, it may now be possible to do what may have been previously impossible.

The extension of our current studies into the semantic structure of dictionaries, thesauri, and large bodies of internally consistent -- at least loosely semantically-consistent -- text is, among other things, a commitment to that more comprehensive approach. Taking the more comprehensive approach is, of course, encouraged more by another dictum of scientific methodology and tradition than that urging the treatment of the simpler case first, a positive sanction deriving from the importance of dealing with the general case whenever possible. The history of mathematics can be a magnificent text for those who would preach homilies on the merit of going for the more abstract, the more comprehensive solution. It may be that, in part, we are seeing within one aspect of the history of computer science, a kind of replay of some of the contrast in research strategies as between the natural sciences and the mathematical sciences. And for those who mistake engineering for science, the resultant trivialization (of a computer science) is disastrous.

Within the domain defined by this approach to the building of natural language computing AI systems (especially those which are conversationally expert or consultant), the choice of the *Thesaurus* as one type of structure to better understand is, if not dictated, at least indicated by certain properties which thesauri have as sources for the formation of computerized knowledge representations. Thesauri have the merit of being already semi-formalized (semi-formal in accordance with the definitions of formality and of formalization in Sedelow & Sedelow 1979; 1983). They also, by implication at least, embrace substantial subsets of any given natural language as a whole, inasmuch as they have the interesting property of serving an infinite

multiplicity of functions as knowledge bases for people. And it should never be forgotten that thesauri are culturally validated (for all the difficulties of defining the term *cultural validation*) in that a wide language community can read and understand them in whole or in part, either with or without extensive overt understanding or utilization of their formalized internal structure. Similarly, almost no one, including perhaps most of the staff of a dictionary-making firm (or those temporarily organized for dictionary construction), understands from a mathematical perspective the structure of the dictionaries they refer to, or even the dictionaries they are building; and we know as a matter of historical fact in conjunction with our earlier explorations on behalf of the holder of the rights to *Roget's Thesaurus*, that for it the same holds in spades.

But even computer scientists have been constrained -- and, in a sense needlessly so -- in their understanding of the formal properties of thesauri and of dictionaries. That deficiency was a function of the lack of adequate theoretical models with which to examine those structures. It is tempting to point out that some part of the slowness in developing appropriately rigorous models is to be attributed to a narrow definition of the nature and scope of mathematics, even on the part of some who are practitioners within the mathematical sciences. Thus, the relationship of the structures of language to the structures of mathematics more narrowly construed has not been perceived in terms of its revealing only special cases of properties of structure in human symbolization. As to that perception, the level of sophistication attained slightly more than a century ago by Gottlob Frege was not truly regained until Alonzo Church formulated his Lambda Calculus; similarly, Babbage's level of sophistication in computer design was not regained until the fourth and fifth decades of this century.

In part as a function of the Bryan model and of our experience with it, it is now possible to do more with the study of general semantic structures (e.g. dictionaries), particularly thesaural structures. Such work may be seen as, at least in principle, a necessary prelude to higher quality mechanical translation -- owing to the prerequisite necessity of coping in general-form way with the mapping function across the semantic spaces created by different natural languages. One may say in passing that there are rich theoretical harvests to be garnered from these studies in comparative semantic space creation, ranging over issues posed or implied by Vico to issues posed or implied by Chomsky. It may also be mentioned as having technical implications (as distinct from the more global, theoretical, and even ideological implications apropos of Vico and Chomsky), that such research is potentially transferable to the study of synonym dictionaries and other sorts of word-finders, as well as the generation of thesauri from dictionaries.

An instance of the way in which the next chapter in this research story will include some paragraphs linking back to what has been written previously by members of this research group is in the utilization of further work on the study of semantic distance, through establishing the 'costs' of traversal of branches of a tree among various leaf nodes. The lesser the cost, the greater the putative semantic relatedness; but, although in a general way there is mileage, so to speak, to be gotten out of those sorts of traversal studies, a note of caution is in order. Our previous research establishes that such tree traversal costs are not consistent, nor are they always reliable indicators of word association, even though in many instances they can be so used. Primarily, the problem is that while two given entries which are near to each other in the tree are generally likely to be related, it does not follow that entries which are far removed are not so related. More than that, such (far-spaced) entries in the explicit *Thesaurus* structure may even be identical; certainly, they may be members of a semantic equivalency class. We have found that entire semi-colon groups that are near duplicates can be located in quite widely spaced portions of the same thesaural tree.

It has become ever more forcefully evident to us that an implicit biasing in favor of determinacy and categoriality have impeded the building of useful models for the study of natural language function. That model of lawfulness which in the study of the natural realm historically has led to the notion of physical 'laws' owes its origin in no small measure not only to the tradition of Roman law, but also to the peculiarities of its so-called 'reception' in the Western world, particularly in the latter Middle Ages (Gilmore 1942). Hermeneutics applied to scientific texts, as has been done by so many including (if only implicitly) Stephen Jay Gould (1987), reveals to us how important the assumptive entailment content of received verbal traditions can be in structuring subsequent scientific work, even far beyond the level of conscious awareness -- a result of precisely the implicit associational structure which our research is concerned to disclose. We desperately need to overcome the (tacit) associations of determinacy and categoriality, and in order to do so we need, in effect, to deconstruct some of the implication of structure itself; it is a category to be hermeneutically explored. Practically speaking, what such an approach may be taken to mean can be the application of the noble work undertaken by Lotfi Zadeh and his associates as to 'Fuzzy Sets' (e.g. Zadeh et al. 1975). It may be, though, that an even more powerful approach (in its relevance for our work) is latent in research led by Zdisław Pawlak at the Polish Academy of Sciences in Warsaw, work being extended in this country through the efforts of Jerzy Grzymala-Busse at Kansas. The Pawlak studies are being accomplished under the rubric 'Rough Sets'.

Somewhat ironically perhaps, in light of the attention we are giving to the significance of a shift in string scale (an *upward* shift in size or length of language elements to be examined), we are presently finding that a shift downward in scale within the Roget hierarchy is indicated for the effective understanding (through the examination of alternate modes of chaining, see Sedelow & Sedelow 1986:79-80, *re* Archie Patrick's work employing the Bryan model) of associational relationships within the hierarchy. But, on the other hand, that kind of downward shift is very consistent with our concern to transcend tacit implications of linguistic or symbolic structures which have been carried along from an earlier period in intellectual history or dialectic. In this instance, by moving the focus of attention with the Roget hierarchy downward in scale, we obviate some of the difficulties which may be latent in the Aristotelian and Enlightenment sort of scheme for structuring knowledge which Roget used -- no doubt somewhat unconsciously -- in building the upper levels of the hierarchy. And those upper levels of the hierarchy are perhaps more open to question -- even questions as to the degree of cultural validation -- than the more fine-grained or lower-level structures of that same hierarchy. There may even have been some transference effects of an undesirable sort, if users assumed that the unattractiveness (for them anyway) of the upper levels of the Roget typology implied -- albeit logically unnecessarily -- difficulties at a lower level. That process could have obscured information within the hierarchy which otherwise might have been turned by them to good account.

In a development well alluded to in the work of Sowa cited earlier, there has been much recent interest in the utilization of semantic nets in conjunction with efforts to use certain sorts of graph-like structures for capturing and holding semantic information, as in expert systems. There are opportunities for us to contribute to the development of semantic net approaches to these special purpose -- even domain-specific -- AI knowledge representations, through a reconsideration of the implications of our experience with the Bryan model. Also, perhaps, opportunities exist for enhancing our own studies through tapping into the reservoir of experience with a semantic net approach, now to be freshly re-applied in conjunction with the study of the *Thesaurus*, rather than with (more *ad hoc*, less comprehensive) attempts to render machinable the semantic content of natural language text or oral statement. Part of the very considerable

latent power in the Bryan model is its utilization of an abstract relationality for getting at semantic dimensions; lexemic meaning, then, is interpreted within the Bryan model as entirely a matter of differential associations. As an abstract object instantiated, the *Roget Thesaurus* is a very sparse subset of the power set of all lexemes.

A purely combinatoric model for discussing intellectual developments (which might better be thought of as symbolic element regroupings forming, over time, trajectories) has its attractions. The work of Bernard Williams (of Ergosyst, Inc.) on the theory of simultaneous independent inventions and the explication of that phenomenon is also apropos and related. Williams' analysis of the simultaneous independent invention of various components of the contemporary computer can be seen as an exemplification of the approach which earlier he had developed more abstractly with reference to the general process of simultaneous independent invention. Both of his studies (a Master of Arts Thesis and a Doctoral Dissertation, Department of History, University of Kansas), at their varying levels of abstraction, may be seen as special cases within the larger framework of combinatoric relationality for which Bryan's model for the *Thesaurus* may be understood as another type case.

With that kind of relational model, then, the whole co-occurrence pattern of any two lexemes in any of several possible sets of clusters contains crucial information as to their semantic association; hence, we may construe it as a collection of nested and/or intersecting clusters, where cluster membership may be taken to imply semantic relatedness of some type. These co-occurrence relationships among lexemes are, within this model, defined wholly in terms of the formal structure of the *Thesaurus* or, alternatively put, exclusively in terms of their mutuality of relationship. We can thus see the *Thesaurus* as incorporating probability constraints for a more abstract word game of the type that Wittgenstein came to see as a proper general characterization of arguments or intellectual developments, or even conversations. You might think of the specifics of individual Wittgensteinian word games as no more than the realization of a certain subset of possibilities implied by the larger word game net which is embodied in a thesaurus. This wholly formal definition enables us to link into the tradition (at least from Frege forward and, more tenuously, from the writings of the medieval Raymond Lull) of powerful formality -- formality in the mathematical sense, rather than in some of its looser 'cultural' meanings -- which we find Frege taking to the limit in his founding of mathematical logic a century ago, and for which there is the more immediate precedent of Alonzo Church and the Lambda Calculus, especially as it was developed by Richard Montague and is being further developed by those in the Montague tradition (such as Robert Bryan, through his doctoral dissertation at the University of Kansas). As with all sorts of powerful approaches, it is a great gain when one is able to avoid many of the pitfalls which, in a sense, may otherwise be unavoidable, by not getting into unnecessary commonsensical semantical specifications that often partake of being a part of the problem -- perhaps surprisingly to Sowa -- rather than of its solution.

Bearing in mind these concepts which inform the direction in which we are taking our new research, one can also retrospectively see more clearly the significance of what already has been done. The general heuristic of gaining advantage by going to more abstract representations, which has been so characteristic in the mathematical sciences, where the elimination of unnecessary assumption has been so crucial (as with the celebrated instance of Riemannian geometry), here works repeatedly. In yet another way of seeing how our research directions are related to the most fundamental vectors in the history of computer science, one should note that we are looking at this *Thesaurus* as graphically depictable, with the quality of being conceptualized as a Boolean matrix within which the words constitute rows and the categories,

columns (ranks and files). In that classic mode of address, *a cell is true* when the word associated with that cell intersects the associated, orthogonal category -- that is to say, when an entry which is a member of the word defined by the cell is also a member of the category defined by the cell; so that a walk through the *Thesaurus* connectivity structures involves starting at a *true* cell and traveling on a row or column to another *true* cell. The discussion in 'The Lexicon in the Background' of alternate modes of chaining, and their respective pay-offs, provides specific illustrations of itineraries among true cells.

Just as co-occurrence has proved to be a crucial concept in citation indexing research (and in what it in turn reveals about the structure of scientific dialectic or argumentation, or scientific conversation or communication, or *traditio*), so here, too, strong links have that same sort of character. In the terms defined above as to the nature of a semantic molecule, strong links exist when there are two or more pairs of true cells on the vectors representing the molecules and also on the orthogonal vectors. Christopher Gunn devised two exemplary representations of such strong connectivity: a text representation and a matrix representation; we here instance those representations in figures 1 and 2 to illustrate in a more graphical way the above observations.

(Labeled arcs indicate degree of overlap (non-normalized))

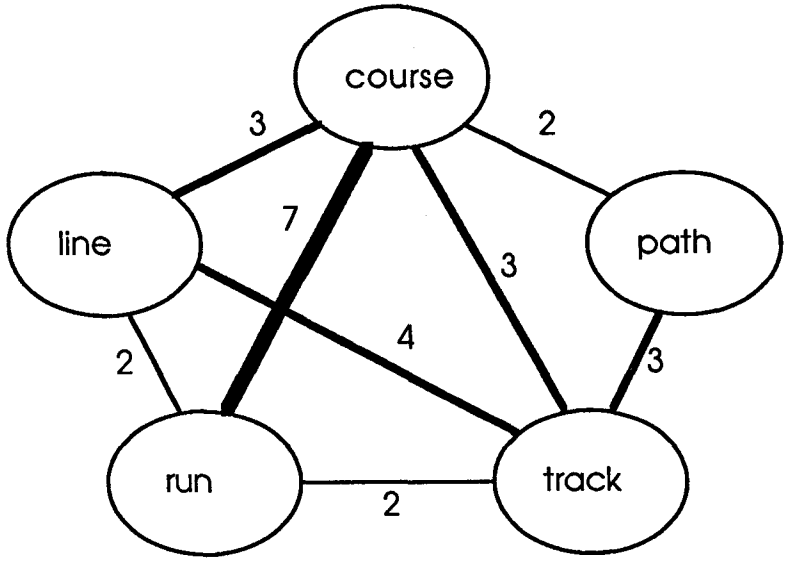

Figure 1. Strong linkages in 657.2.1 (way)

To:

	course	path	track	run	line
course	--			.78	.33
path		--	.25		
track		.42	--		.44
run	.47			--	
line	.20		.33		--

From:

Figure 2. Normalized bidirectional strong linkages (d*) in 657.2.1 (way). (Values < .2 are blanked; diagonal is undefined)

Looking very directly toward the future, we see promise in the contention of Christopher Gunn that, although it would be not inexpensive computationally, an option to consider would involve a thoroughgoing reclustering intra-thesaurally. Then, new molecules (in Bryan's sense of molecule) are created for all strongly connected components of existing molecules, whilst there is a discarding of all those entries which do not participate in linkages that are strong. And, apropos categories, there is a relaxation technique available under which such higher order clusters as paragraphs or sections may be re-partitioned into new semicolon groups that are based on the strong connectivity in the remainder of the *Thesaurus*, which for these purposes is held constant. Now, both these word and category reclusterings can be done on an iterative basis. Christopher Gunn expects that for words the result of such a reclustering could be a separation of the original word clusters into homographs and differing senses of text strings which are equivalent. As to categories, while the expected result is less evident, it does appear, from preliminary explorations, that the strength of c-links is affected significantly by the frequent presence of words which are general and imprecise, such as the verbs *set, rest*, etc.

At the 1985 annual meeting of the American Society for Information Science, an impassioned set of pleas was made -- and can be heard on the tapes for the appropriate session -- both by university faculty and by commercial R&D personnel (as from the Institute for Scientific Information, in Philadelphia) with reference to the utilization of models employing the concept *semantic primitives*. It is at least an interesting fact that such general terms as cited in the preceding paragraph appear to correlate well with concept classes which certain other researchers have identified as exemplificatory of 'semantic primitives'. We would like to identify such primitives or, in Pawlakian terms, their 'rough equivalents', and a metric apropos c-links would be useful to that end and is a high priority task for the months ahead.

If one uses the old philological categories of case, then one may notice that our prospective research will tend to stress attention to verbs, i.e. attention to the Roget subset comprising all the verbs in the *Thesaurus*. While there are arguments for attending to various cases which have occurred in the history of even contemporary linguistics, and while we think there is a special importance to the category, *verb*, it is a fact that we also intend to extend our model to the full *Thesaurus* once we have, with reference to verbs, developed our work sufficiently. It would also be our hope to derive sparse connectivity matrices for all words, representing all immediate (i.e. hop count of 1) linkages.

It may seem intuitively obvious that the way to go with combinatoric analysis of the sort we

have in mind, both with reference to the understanding of properties of the *Thesaurus* and with reference to the specification of characteristics of particular long language strings, is to stipulate the properties of any given matrix (in, thus, an alternative knowledge representation) as a set of constraints on the comprehensive combinatoric, rather than attempting (inductively) to build up a set of allowed combinations. To make at this time one significant point only, the feasibility of comparisons -- as for example of semantic space utilization differentials across languages, in each instance as a matter of a restriction on the general combinatoric differentiated by language -- is thus (i.e. using the constraints approach) facilitated for interlingual research as might be used in machine translation, general semantics, etc. But, of course, there are threats of combinatoric explosiveness that must be coped with, even though yesterday's realistic fears are today's comparatively inexpensive procedures, owing to the happy decline in prices for computer hardware components.

We noted above the proposed expansion from a discussion of verbs alone to an analysis of connectivity matrices for all immediate linkages for all words, categories, and entries that are marked by strong connectivity. The computation of paths of length greater than one has led to the derivation of a cost of travel function. That function may be used as a preliminary model for multi-link or multi-hop traversals. Even with current improvements in hardware and systems, the exhaustive computation of all shortest paths through the connectivity structure is not immediately practicable. For the verbs alone, it would take approximately a 20k x 20k matrix representation for only those verb entries which are strongly connected. Fortunately a heuristic algorithm or algorithms of a strength-first type are available to make the task simpler. With such algorithms we will have entries into full matrices which represent the total connectivity among words, categories, and entries; and those matrices can be seen as a definable subset of the total possible set of matrices if the full combinatoric realization were not constrained. Presumably, we will use entry data for much further work of this sort, since the entry matrix can be extended to include all the information in the word matrices and category matrices. Either fully-filled or partially-filled matrices can be construed as knowledge representations in network form for the semantic dimensions in the *Thesaurus* . Imagination and skill alone constrain the forms into which the entry matrices may be mapped. Types of desired manipulability for those matrices strongly influence the choice of knowledge representation schema. Given our objectives, we wish to be able, of course, to derive global models from such entry matrices of semantic structure for the *Thesaurus* . And the analysis of the mapping functions needed for movement from one thesaurus into a thesaurus of another language will be crucial both for its theoretical and its applied results.

One attractive approach is a graph theoretic treatment of the matrix, in which the cells represent cells in a digraph; that is useful apropos sparse matrices with paths of link equal to one. There has been a hypothesizing to the effect that the filled matrix of all strongest paths of length greater than or equal to some number and of strength greater than or equal to some number will support a more powerful dimensional analysis.

In this paper we have indicated some of the intellectual antecedents apropos our research effort. At this point, we reach back to the influential work of Charles Osgood, of approximately three decades ago. It will be remembered that, interestingly enough, a dialectic between approaches favored by Osgood and those of Chomsky contributed a great deal to the origination of a powerful, if now somewhat faded, approach to many issues in contemporary linguistics. Our interest in Osgood is apropos the similarity of the modeling we propose of the semantic space of a language -- say of English, first, and perhaps Chinese next -- to the n-dimensional eigensystem techniques which Osgood and his associates used to derive the semantic differential measures.

Although we see similarities between some of our first pass techniques and the approach taken by Osgood, there are certain difficulties deriving from his implicitly Euclidean spatial structure, n-dimensional though it be, which was a product of a factor model made popular especially by the work of J. P. Guilford and his numerous doctoral students. It is unlikely that we will be able to persist with the constraint under which they operated which required the input connectivity matrices to be symmetric and positive-definite, so that eigenvectors and eigenvalues are guaranteed real.

Just as we clearly see our indebtedness to the great Frege and major figures following him for one aspect of our work, here, in prospect, we see it likely that another major figure in the mathematics originating in Germany during this last 100 years will prove to be of real significance to us. More specifically, we have the higher generality spatial model using infinite-dimensional Hilbert space to fall back on. We do not really expect that the characteristics of the semantic space of English or any other language will be amenable to the requirements of factor analysis, but it does seem probable that a Hilbertian model will suffice; it does make possible the interpretation of complex eigensystems. Especially interesting to natural language computer scientists is the fact that many well-established applications of this type of Hilbert mathematics (as in quantum mechanics, for example) are indicative of the utility of it for the treatment of connectivity properties which are probabilistic or fuzzy. Again, we see how effective work in natural language semantics may be much influenced by the utilization of non-categorical methodologies which allow for the more effective manipulation of fuzziness and probability. As mathematics itself experiences a greater involvement with not only classic Zadehian fuzziness, but also with Zadeh's progression from probability theory to possibility theory, it may transpire that for the analysis of mathematical semantics (as in analyzing programming language semantics) the utilization of some form of the Hilbertian approach which we are proposing for natural language semantic space delineation will prove broadly significant. If or when that happens, we will see further how any given ('natural') language may be construed not only in some form of parallelism with mathematical languages of various sorts (whether algebraic or programming languages, or formal languages), especially where out of the same culture, but also how we may effectively regard natural language as a degenerate form of mathematical symbolization (rather than seeing mathematical language as a special and refined form of natural language). Reflecting on the power latent in these approaches, we suspect that in the near future there may be a substantial breaking-out onto higher ground with reference to the effectiveness of studies of semantics of whatever sort, that is, whether with reference to the semantics of natural language or with more fully constructed languages (one might say more fully Viconian languages, where by Viconian one means that they are/were the product of comparatively overt human consciousness).

But however fast and far we are able to move with reference to the understanding of the comprehensive properties of semantic spaces as differentially developed by particular languages through time, there is even less question as to the speed and distance we will be able to cover with reference to the development of models of local features in the *Thesaurus* structure using these approaches. And, thus, for a given language we expect to be able to derive significant understanding of field structure properties for particular sub-spaces of its total semantic space, as also more adequately to understand the strength and limitations of that concept of semantic primitive objects or its analog in superordinate cluster objects, not to mention an understanding of case relationships among the members of lexemic populations.

We have mentioned above our interest in the more organized comparison of thesauri and of dictionaries, and, more particularly, the comparison of abstract models appropriate to each. Our

research also will explore ways in which the formally used techniques in the process of defining dictionaries are like and unlike those in the making of entries for thesauri; further, we will explore the extent to which a thesaurus may complement a dictionary; more especially, we will be interested in seeing how dictionary definitions may be enhanced (against some criteria of operational precision) through connectivity relationships for some of the terms which have been derived from the analysis of connectivity properties for those same terms in some sub-space within a thesaurus.

In the course of other research as to discourse analysis, we have found the work of Halliday and Hasan very useful (1976). We have employed Halliday and Hasan as a donnée, while others have worked from different bases, in the developing of linking structures across more particularistic representations of semantic content. The concepts related to semantic collocations as presented by Halliday and Hasan are also much to the point in studying thesaural structure. We look to the generation of semantic grammars capable of differentiating variations in senses and meanings as a function of context and word choice. Text-understanding and text-generation systems will be used to test the capabilities of the analytical apparatus derived from these latter studies.

Acknowledgments

Significant segments of the early research on thesauri were supported by ONR Project NR348-005. Current research is supported by NSF Grant GA410.

REFERENCES

Barr, Avron, E. A. Feigenbaum, and Paul R. Cohen. 1981. The handbook of artificial intelligence. 3 vols. Los Altos: Kaufman.

Bryan, Robert M. 1973. Abstract thesauri and graph theory applications to thesaurus research. Automated language analysis, ed. by Sally Yeates Sedelow. Lawrence: University of Kansas, Departments of Computer Science and Linguistics.

Dale, Claudia J. 1979. Investigations of nouns, Roget's international thesaurus. Master of Science thesis. Lawrence: University of Kansas, Department of Computer Science.

Gilmore, Myron P. 1942. Argument from Roman law in political thought 1200-1600. Cambridge: Harvard University Press.

Gould, Stephen J. 1987. Time's arrow, time's cycle: Myth and metaphor in the discovery of geological time. Cambridge: Harvard University Press.

Halliday, M. A. K., and Ruqaiya Hasan. 1976. Cohesion in English. London: Longman.

Patrick, Archibald B. 1985. An exploration of an abstract thesaurus. Master of Science thesis. Lawrence: University of Kansas, Department of Computer Science.

Roget's international thesaurus. 1962. New York: Thomas Y. Crowell.

Sedelow, Sally Yeates. 1985. Computational literary thematic analysis: The possibility of a general solution, Proceedings of the 48th ASIS Annual Meeting, ed. by Carol Parkhurst, 22:359-62.

------. 1969. PREFIX. Automated language analysis, ed. by Sally Yeates Sedelow. Chapel Hill: University of North Carolina, Departments of English and Computer and Information Science.

------, and Walter A. Sedelow, Jr. 1969. Categories and procedures for content analysis in the humanities. The analysis of communication content, ed. by George Gerbner. New York: John Wiley & Sons.

------,------. 1986. The Lexicon in the background. Computers and Translation 1:2.73-81.

Sedelow, Walter A. 1985. Semantics for humanities applications: Context and significance of semantic 'stores', Proceedings of the 48th ASIS Annual Meeting, ed. by Carol Parkhurst, 22:363-6.

------. 1968. History as language. Computer Studies in the Humanities and Verbal Behavior 1:4.

------, and Sally Yeates Sedelow. 1983. Computers in language research: Formalization in literary and discourse analysis. Berlin: Mouton.

------, ------. 1979 (eds.). Computers in language research: Formal methods. The Hague: Mouton.

Smith, John Bristow. 1983. Computer criticism. Computers in language research: Formalization in literary and discourse analysis, ed. by Walter Sedelow and Sally Yeates Sedelow. Berlin: Mouton.

Sowa, John F. 1984. Conceptual structures: Information processing in mind and machine. Reading, Mass.: Addison-Wesley.

Warfel, Sam. 1972. The value of a thesaurus for prefix identification. Automated language analysis, ed. by Sally Yeates Sedelow. Lawrence: University of Kansas, Departments of Computer Science and Linguistics.

Zadeh, Lotfi, et al. 1975. Fuzzy sets and their applications to cognitive and decision processes. New York: Academic Press.