Analysis of "natural" language discourse*

by SALLY YEATES SEDELOW
The National Science Foundation
Washington, D.C.
and
The University of Kansas at Lawrence

ABSTRACT

Referential linkage in extended language strings (multiple-sentence, paragraph, etc.) is of great interest to computer scientists, linguists, and literary scholars concerned with the analysis of discourse. In all three disciplines, semantic relationships are central to approaches to inter-sentential, inter-paragraph, and inter-supraparagraph linkings. This paper compares and contrasts some of the directions taken in current research and explores the possible utility of a general-purpose thesaurus for the construction of semantic frames of reference. The utility of such a thesaurus for measuring semantic distance between terms and thus establishing possible linkages is suggested by an experiment concerning word prefixation.

Discourse analysis is now being undertaken by research scientists in a number of disciplines. Its meaning varies from field to field and even from scientist to scientist; traditionally, it has implied the structural analysis of a relatively large number of consecutive natural language sentences, paragraphs, chapters, or larger units. (It is not restricted to the spoken word, although that is one meaning of discourse.)

As a former professor of English literature with a major interest in the analysis of language-strings of the length of Hamlet, or Paradise Lost, I have watched the growing interest in discourse of my newer colleagues in computer science and in linguistics with both genuine excitement and, it must be confessed, a frequent feeling of déjà vu. Thus, extended debate on the subject of "frames" at a recent workshop on theoretical issues in natural language processing sounded very like discussions over the years by literary scholars and students on topics such as "frame of reference" and "point of view." I feel that my new colleagues are, in a sense, just beginning to learn to talk but, on the other hand, perhaps in time their talk will be couched in more precisely used terms than those employed in analogous literary conversations. I certainly hope so, since such precision was for me a prime motivation in having become associated with computer scientists and linguists.

Although discourse analysis has been a primary concern of literary scholars for many years, for both linguists and computer scientists it does represent an exciting new concern. The reasons for computer science coming to this study rather late are quite obvious, but that linguists should be so slow to arrive on the scene may seem to some rather surprising. For the sake of clarification, it might be noted that for a number of decades literature and linguistics (at least as linguistics is practiced in this country) have been very much separated from each other. Linguistics departments have sometimes provided, as a service, courses on the English language, but the heart of the discipline has been elsewhere. The "elsewhere" has been a world of micro-events, such as the basic units of sound, or the basic units of grammatical structure, which were long considered to be solely syntactic. Given this concern with microunits, the outer bounds considered necessary for adequate study of such units were provided by the sentence. As I have already noted, literary scholars find it necessary and desirable to concern themselves with texts of lengths greater than a sentence; thus, the work of linguists has not been of overwhelming interest to literary scholars, and vice versa. Further, there has been relatively little of the interdisciplinary dialogue which might have pushed linguists somewhat sooner toward looking at the types of language phenomena of interest to literary scholars.

More recently, linguists have been forced, perhaps partially through their own need for new approaches to the study of language but also because of efforts to use computers for language-dependent tasks—such as machine translation—to try to place their microunits within systems which, in turn, have proved to be parts of ever larger systems. Early efforts to use the computer to provide translations produced generally unsatisfactory results partly because semantic systems had been essentially ignored by linguists. Attempts to apply the computer to other areas entailing heavy

* The views expressed in this paper are those of the author, and do not necessarily reflect any policies of the National Science Foundation.
The SRI grammar, described by Jane Robinson, can also handle elliptical references between sentences, as in the following sequence: "What is the length of the surface displacement of the Lafayette? ... What is its draft? and "What is the length of the Lafayette? ... The Ethan Allen?" In this sequence of four sentences we have an example of anaphoric reference through the use of "its" in the second sentence and of elliptical anaphoric reference in the fourth sentence which, in its entirety, would read "What is the length of the Ethan Allen?".

Let me offer one more example of current speculation by a computer scientist about an aspect of small scale discourse analysis; in this case, the concern is with linkages among single sentences and an individual's "store" of earlier verbal and sensory experience. This example is the beginning of a fable, as told by Wallace Chafe, a linguist at Berkeley, and retold by Marvin Minsky:

There was once a Wolf who saw a Lamb drinking at a river and wanted an excuse to eat it. For that purpose, even though he himself was upstream, he accused the Lamb of stirring up the water and keeping him from drinking . . .

Minsky says that to understand this fable one must realize that the wolf is lying. To understand "even though" one must realize that the contamination in question doesn't move upstream. This realization in turn requires us to understand (among other things) the word "upstream," itself. Minsky then devotes several pages to indicating how his framework might help one understand some of these terms as well as the more extended meaning of these two sentences. By frame, Minsky means a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party. . . . For visual scene analysis the different frames of a system describe the scene from different viewpoints, and the transformations between one frame and another represent the effects of moving from place to place. For nonvisual kinds of frames, the differences between the frames of a system can represent actions, cause-effect relations or changes in metaphorical viewpoint.

A little later, I want to return to this example to indicate how some of my research can help to get at "frames of reference" which, in turn, will help to delineate discourse structures where the structure depends upon the perception of semantic relationships.

For macro discourse analysis, the most relevant current work in computer science is content analytic and I will recur to it briefly later. When we turn to linguistics, we find a number of linguists—notably Robert Longacre and Joseph Grimes and students associ-
ated with them—who have become involved in the analysis of macro discourse ranging up to short narratives in length. In *The Thread of Discourse,* Joseph Grimes provides a broad overview of these discourse analytic techniques, a number of which would be useful for discourses of any length. Many of the approaches to discourse analysis described by Grimes depend on too much preliminary background for use in this paper, but a brief description of two of the approaches would seem worthwhile in order to provide some sense of linguists' orientations toward this new area.

One very simple technique is span analysis. An identification span consists of a series of identifications of the same participant in a narrative—a series in which no identification is stronger than the one before it. Strength of identification is defined as a ranking that goes from proper names, to explicit descriptors like "the last speaker at the symposium," to common nouns like "the speaker," to nouns used generically like "the fellow," to pronouns like "him," to reference without identification. When a shift occurs from a weak form of identification to a stronger form, the current span is terminated and a new one begins. The spans can be shown graphically simply by taking a piece of paper, listing vertically from 1 to n the number of clauses, or phrases, and drawing vertical lines to show the length of the spans. Such spans can be used for many other characteristics of discourse such as setting, time, and place, and can reveal patterns which graphically illustrate the differences in style between one work and another, between one author and another, and so on.

I would like to mention one other discourse analytic technique described by Grimes and, in fact developed by his students, Mary Ruth Wise and Ivan Lowe. This approach depends upon the roles taken by participants in a narrative. Here, roles are described by such terms as agent, experiencer, instrument, and other analogous categories drawn from case grammars, which are now of great research interest to both linguists and computer scientists. This approach to discourse analysis studies the shifting of roles among the participants in a discourse. The perception of shift is based upon relative rankings among cases such as agent, experiencer, and so on. In this system, agent is the highest ranking role and a case called essive, which is rather analogous to the existential state, is the lowest ranking.

One type of role shift is a reversal, which involves an exchange of relative ranking between two participants in a narrative. In some simple narratives, reversal is the only shift which occurs. The text tends to start out with one character as agent and the next as, say, goal, then reverses so that the second character is agent and the first has a role of lower rank. A second reversal brings the characters back into the original orientation; in some of the simple folk narratives studied by Grimes and his students, this second reversal signals the beginning of a new paragraph.

Another kind of role shift is described as a switch, in which the second and third participants undergo reversal of rank. In a number of narratives studied by Grimes and his students, the regular progression of events is carried by single operations such as a single reversal or a single switch. Whenever there are composite operations, such as a switch and reversal, or reversal and switch, there is a surprise, an interruption or a point in the narrative where things go wrong.

Grimes tries, with varying degrees of success, to provide a text-analytic operational significance to such categories as cohesion, theme, setting, background, evaluation, and many others, including a range of types of linkages between sentences. Almost all, if not all, the categories for discourse analysis he explores depend upon understanding word meanings, phrasal meanings, clausal meanings, paragraph meanings, and so on. Meaning—semantics—is the key to discourse analysis and therein lies a challenge for any efforts to use the computer for discourse analysis.

My own research on computer approaches to discourse analysis has been partly built upon training as a literary scholar. Thus I have been concerned with texts of extended lengths and I have wanted to find methodologies, procedures, and/or algorithms suitable for use with a broad range of texts. By contrast, for the most part computer scientists concerned with discourse have tended to constrain the universe of discourse with which they deal to a very small segment of reality. The reasons for such constraints are obvious when you consider the complicated problems which must be dealt with in even a very modest universe; but I have chosen not to follow this path because my own preferences and training make it un congenial and, especially, because as a scientist I strongly prefer general solutions for natural language problems to *ad hoc* solutions.

When I began working with computers I decided to concentrate upon structures of meaning because those are of major importance to discourse which, as I have stated before, is of central importance to most literary scholars. At that time (thirteen years ago) other people were working on parsers and I planned simply to borrow a parser from someone else when the need arose. It is doubtless a splendid testimonial to man's ingenuity as a user of language that the general-purpose parser I need is still not available, although there has been excellent research on this problem and there are parsers to which I would now like access.

My approach was to develop a set of content analytic programs which I used to look for literary themes. Themes comprise semantically-related words and form

---

*Much of the research described on the following pages has been supported by the Information Systems Branch, Office of Naval Research.*
part of the structure of a discourse. For example, part of the structuring form of *Hamlet* is provided by a network of associations involving disease, madness, and decay, just as part of the structure of *Paradise Lost* involves a semantic network of words associated with variety and variation. I have written and talked about this research elsewhere and I will not discuss it further here. But I do want to talk about an outcome of that research—that is, further research, on the nature of thesauri.

Thesauri now come in many forms. In the context of information retrieval, a thesaurus most often consists of a list of terms, sometimes divided into categories, relating to some specialty area. The kind of thesaurus I have in mind is not one of these special-purpose word lists, but rather a general-purpose reference work showing semantic relationships among words in a whole language. Specifically, I have in mind the English-language thesaurus, *Roget's International Thesaurus*. Although no such general reference work is "perfect," or even nearly perfect, it does have the strong advantage of validation through time and across a culture. I argue that large corpora such as *Roget's Thesaurus* or the Merriam-Webster's dictionary should be investigated for their potential utility in dealing with semantics for all kinds of computer-based tasks. They also lend themselves to segmental exfoliation when a finer semantic mesh is needed for a specific purpose. Such potential sources of semantic information have been largely ignored by those seeking to use the computer to capture meaning in natural language. Notable exceptions to this rule are John Olney at SDC and his associates who have put *Webster's Collegiate Dictionary* into computer-accessible form; among those using this version of the dictionary are Robert Simmons and Robert Amsler who, in a recent paper state that "from a linguistic point of view . . . the proper source of semantic information is a large corpus of ordinary usage of the language."

One of the reasons these general purpose corpora haven't been more extensively employed is that they are difficult to use. Another is that they are difficult to get into a useable form. Part of my own research effort in recent years has been directed toward putting *Roget's International Thesaurus* into computer-accessible form. The *Thesaurus* is now in computer-accessible form although the editing hasn't been completed.

Editing is a major problem because the *Thesaurus* makes many assumptions concerning the ability of the human reader to fill in elisions, supply context, and otherwise rely on the enormous store of information in the human brain. An example of a frequently occurring type of elision is "jump up or off" in which the jump for "jump off" is omitted. An example of the dependence upon the human reader to fill in or at least understand what a possible extension of items in the *Thesaurus* might be is the frequent use of etc. to indicate the continuation of lists. Examples of such lists are "Pan-American, Pan-Pacific, Pan-Hellenic, etc.", (multiply by 5, etc.)", "hourly, daily, etc.", and so on. These and other analogous problems which must be dealt with in order to use the *Thesaurus* in automated systems have been discussed elsewhere.16,17

Our reason for putting the *Thesaurus* into computer-accessible form was to use it for content analytic applications. In order to put it to such use, it is desirable to have a clear understanding of the structure of the *Thesaurus*; approaches to the study of the structure have been explored by a graduate student, Robert Bryan, and are described in research reports for this project (available through the Computer Science Department at the University of Kansas). A major potential utility of the *Thesaurus* not only for content analysis but for many other efforts to use the computer to understand natural language is as a guide to measures of semantic distance. That is, the number of nodes one must pass through in the *Thesaurus* to get from one word to another, or from one concept to another, can be taken as measures of semantic distance. Again, the graph theoretic approaches outlined by Bryan are intended to facilitate research on this question. It should be borne in mind that since any given term may appear at several different nodes in the graph, semantic space in the *Thesaurus* is much more complex than the organizing tree-structure for the *Thesaurus* content might lead one to suppose.

The viability of the *Thesaurus* as a potential source of information concerning semantic distance was a central issue in an experiment we undertook with reference to prefixation. Our experiment was to see whether the *Thesaurus* might be used for determining when a particular string of alphabetic characters is serving as a prefix. The experiment is described in some detail by Sam Warfel in one of our research reports. As you may know, the *Thesaurus* is organized hierarchically, with the basic text consisting of 1040 semantic categories each with a number and a label, e.g., "854. lack of feelings." Each of these numbered categories is divided syntactically and semantically. For the purposes of this experiment the syntactic categories were ignored, because a preliminary investigation suggested that syntactic categories added no useful information toward the determination of prefixed words. (An obvious reason for this impression is that syntactic category membership is most often indicated by suffixes.)

To clarify the following discussion, it may be worthwhile to describe briefly the general structure of the hierarchy in the *Thesaurus* so that ensuing references to "levels" will be understandable. The 1040 categories are related at a higher level in the *Thesaurus* by the "Synopsis of Categories," which is not part of the basic text but is presented as an outline following the Preface. In this synopsis the *Thesaurus* is divided into eight classes (e.g., Class Six: Intellect). Each
class is divided into several labeled sub-classes indicated by Roman numerals (e.g., I. Intellectual Faculties and Properties) and each sub-class is divided into labeled sub-sub-classes designated by capital letters (e.g., L. Conformity to Fact). Each sub-sub-class is divided into several of the 1040 categories (e.g., 515. Truth) which are numbered consecutively throughout the text. Each of the 1040 categories is subdivided into numbered ‘paragraphs’ of words; for example, there are twenty-two such divisions under 515. Truth, so that there are subcategories 515.1–515.22. These ‘paragraphs’ are, in turn, subdivided into semi-colon groups; that is, within a paragraph those words most closely related are grouped and delimited by semi-colons, e.g.,

accuracy, correctness, rightness;
extactness, exactitude;
presciseness, precision;

Thus, in the following discussion of the value of semantic distance measures for determining prefixation, the levels referenced are as follows:

Level 1 e.g., Class Six: Intellect
Level 2 e.g., I. Intellectual Faculties and Properties
Level 3 e.g., L. Conformity to Fact
Level 4 e.g., 515. Truth
Level 5 e.g., 515.3
Level 6 e.g., accuracy, correctness, rightness;

Our assumption was that words occurring either in the same place in the hierarchy or very near each other in the hierarchy are more likely to be related than those further away. Therefore, it seemed possible that words having the same root and differing only by prefix would tend to fall together in the Thesaurus, (i.e., to be only short distances apart). In general, our sampling bore out our hypothesis. That is, words such as joy and enjoy (863), accurate and accuracy (891), courage and encourage (891), occur in the same fourth level categories in the Thesaurus hierarchy. On the other hand, words such as vent and prevent are shown not to belong together because prevent does not occur in either the same category or any of the categories related to categories associated with the unprefixed root vent.

To gain a further sense of the utility of the Thesaurus for this purpose, we compared the use of the Thesaurus with that of our prefix recognizing program (which is operational and which is based on ad hoc decisions on all the words in the Random House dictionary). We used the Thesaurus for the comparison of thirty-eight word pairs which had been turned up during the course of a computer run of the PREFIX program. PREFIX produced ten word pairs which struck me as viable pairings. Had the Thesaurus been used to examine those same word pairs, eight of them would have been linked together. The word pairs are shown below, with those bearing the * indicating the pairings also produced by the Thesaurus:

integration-disintegration
* able-able
* courage-encourage
* danger-endanger
* doubted-undoubted

Of the ten pairs on the above list, the grouping of moralize and demoralize is at least questionable. The Thesaurus' judgment here may be preferable to my own when I was earlier looking at the results of the PREFIX run.

The PREFIX program grouped together three word pairs which should not have been grouped. The Thesaurus identified none of these words as appropriately paired:

cent-recent pare-prepares tribute-distribution

The PREFIX program produced twenty-five pairings which I judged to be helpful in some contexts (some of them very rare and esoteric) but not in others. Those pairs in the list below which would have been linked by the Thesaurus are again shown with an *:

* compass-compassible
* conceivable-inconceivable
* cover-discover
* fend-defend
* fluence-influence
* joy-enjoy
* junction-conjunction
* ligation-obligation
* mode-outmode
* ply-apply
* promise-compromise
* refutable-irrefutable

As should be apparent from the examples above, as well as, hopefully, from any sampling others might make of the Thesaurus, while the Thesaurus is not perfect for this task, nonetheless, it is rather good. Some problems could be taken care of if inconsistencies in the organization of the Thesaurus were eliminated. For example, although most negation and reversal relationships are expressed in adjacent categories under the same third level headings, there are exceptions to that organizational principle. Thus, although there is a category at the fourth level in the hierarchy labeled “disintegration” there is no comparable category labeled “integration,” despite the fact that there are categories labeled “order” and “disorder” as well as “continuity” and “discontinuity.”

The same inconsistency is apparent with talkativeness and untalkativeness where the third level categories containing the words are quite far removed...
from each other, unlike other positive-negative category pairs:

(Level 2) III. Communication of Ideas

(Level 3) B. Modes of Communication

M. Uncommunicativeness; Secrecy

552. Communication 611. Uncommunicativeness

552.4 611.2

communicativeness, taciturnity, untalkativeness . . . ;
talkativeness . . . ;

A different problem arises with the pairs birth-rebirth and born-reborn where the first word in each pair occurs in fourth level categories related to either "beginning" or "physical birth" while the second word in each pair occurs only in fourth level categories related to the religious experience of conversion. Therefore, the words in the pairs are not judged by the program to be prefixed since the metaphorical relationship between the two uses of birth is not shown in the Thesaurus.

In order to achieve comparability in these measures of semantic distance provided by the Thesaurus, organizational inconsistencies in the Thesaurus such as revealed by this experiment would need to be coped with either through cross referencing or shifts in the structure of the Thesaurus. Metaphorical relationships will be more difficult to capture, although it is the case that the Thesaurus, unlike a dictionary, is rather strong on making some metaphorical relationships explicit. This facet of the Thesaurus is useful, as we shall see in an experiment with the Thesaurus and the few sentences from the fable about the wolf and the lamb cited earlier.

For the sake of convenience, the initial sentences of the fable I cited earlier are repeated:

There was once a Wolf who saw a Lamb drinking at a river and wanted an excuse to eat it. For that purpose, even though he himself was upstream, he accused the Lamb of stirring up the water and keeping him from drinking. . . .

The experiment I want to undertake here is first to use the Thesaurus to see what kind of contextual "frame" it can provide to facilitate understanding of the fable. When the Thesaurus fails, I will suggest the use of a parser or dictionary, either separately or in combination with each other or with the Thesaurus. Let me say, prefatorially, that this experiment, as is the case with Minsky's article on frames, ignores many difficulties that would in fact arise were one using the Thesaurus, as well as, for that matter, the dictionary and parser to cope with the fable. The goal in this next little exercise is to see whether there is any point in pursuing further the use of the Thesaurus for such a task.

Our assumption—a strong one—will be that the computer has no information about any of the words in these sentences and that, in fact, the Thesaurus must be used to provide the context.

You will observe that the first sentence of the fable is: "There was once a Wolf who saw a Lamb drinking at a river and wanted an excuse to eat it." A lookup of "river" in Roget's will locate "river" under "running water" and it will be linked with "stream" as well as with "drinking water." Thus the notion of river as drinkable emerges. The word "wolf" appears in a listing of animals. "Lamb" does not appear in that list but it is linked to "sheep" which, in turn, appears as an animal. "Sheep" is also linked to "meat" which is linked in the Thesaurus index to "meat" which in turn occurs with "feed" and "dine." Given metaphors' pleasing property of frequently having some relationship to (symbolic) reality, it would probably be useful to find in the Thesaurus the use of "wolf" to mean "devour," which in turn occurs in the index under "eat up." Thus, as it happens, the Thesaurus associates wolf with eating and on the basis of this association it might be possible for a computer program to assume that the creature who "wanted an excuse to eat it" is the wolf. The lamb is not particularly associated with drinking through any references in the Thesaurus. Rather, a syntactic parser might be expected to work out the relationship between the lamb and drinking in the phrase "lamb drinking at a river."

The chief problem in this sentence is the identification of "it." To repeat the sentence: "There was once a wolf who saw a lamb drinking at a river and wanted an excuse to eat it." "It" could refer either to the lamb or to the river. The only route to disambiguation that looks possible to me is at the syntactic level, which might point up a parallelism between the wolf seeing an object—the lamb—and eating it. In both cases, "it" is the object of an action by the wolf and possibly through this parallelism "it" might be identified with lamb. I find no information in the Thesaurus that would enable a program to perform this disambiguation. It is possible to deduce from the Thesaurus that a lamb can be eaten if you follow a somewhat circuitous path which, under the listing of animals, provides the word "flesh" in connection with "horse flesh" and under "eating" provides the word "flesh eater." You will remember that "sheep" appeared under "animals" and that the word "lamb" was linked with "sheep." Unfortunately, there is no information in the Thesaurus which would suggest that a river cannot be eaten.
To the contrary, the words "drinking" and "drink" occur under the general heading of "eating," so one might conclude that the "lamb," for example, could just as well be said to be eating a river as to be drinking it. The preposition "at" in the phrase "lamb drinking at the river" might be helpful here, but it is difficult for me to see how it could solve this problem.

With reference to the word "excuse" in this sentence: the wolf "wanted an excuse to eat it"—the Thesaurus lists "excuse" along with "guise," "mark," "pretext" and "false pretence" and it also links "false pretender" with the expression "wolf in sheep's clothing." Syntactic analysis might be necessary to establish that it is the wolf, not the sheep, which is linked to "false pretender" in that group but, given such analysis, "excuse" is linked to "wolf."

The portion of the second sentence we have is: "For that purpose, even though he, himself, was upstream, he accused the lamb of stirring up the water and keeping him from drinking. . . ." Syntactic analysis should establish that "he, himself" is the subject of the second sentence and rules governing anaphoric reference would identify the subject, "he, himself," with the subject, "wolf," of the first sentence.

The meaning of "upstream" and its connotations for this little narrative are quite difficult to get at. As noted earlier, the Thesaurus is structured so that, frequently, categories having opposite meanings are juxtaposed. For example, Ascent, which contains the word "upstream" is next to Descent, which contains the word, "downstream." It would be possible to get from one category to the other by searching on the word "stream." If one looked up "upstream" in Webster's Seventh Collegiate Dictionary, the definition "at or toward the source of a stream" turns out not to be terribly helpful. On the other hand the definition of "downstream" is "in the direction of the flow of the stream" and flow is associated with movement both in the dictionary and in Roget's Thesaurus. The index of the Thesaurus provides many clues that "move" and "motion" entail a change of position; for example, one finds the phrases "move back" and "move forward."

The dictionary, as you remember, associates movement or flow with directionality and "downstream" is shown by the thesaurus structure to be the opposite of "upstream." Further, the "even though" in this sentence implies some condition—opposite or contrary to—not consonant with one's "frame" involving "upstream."

At this point, the computer-based information processing programs might be in a position to try to produce some version of the representation suggested by Minsky in his discussion of this small segment of discourse. It might be desirable to represent the relative positions of the wolf and the lamb vis-à-vis the direction of flow of the river; and it may, indeed, be necessary to build into the computer program some sort of "primitive sense perception" to show that disturbing the water at the lamb's location doesn't affect the water at the wolf's location. This kind of "sense-perceived" knowledge, which forms the basis for much exploratory work in computer science, does not depend, initially, upon word associations but rather upon visual experience. (It may even depend also on a stored analytical model of physical processes.) Much knowledge, of course, has an ultimate dependence upon visual experience but this particular perception is difficult to track down to its final meaning through either thesaural associations or dictionary definitions. Given some such primitive representation of the factual situation it would then be desirable to try to contrast the verbal statement after "accuse" with the realistic representation produced by the "even though."

One can see how it might be possible to combine a parser, a thesaurus, and a dictionary, as well as, perhaps, sensory primitives to conclude that the wolf intends to pick a fight with a lamb, but I am not certain that it will be clear whether the wolf wants to eat the lamb or the river—although the suggested syntactic clues might provide some probability that the wolf would like to eat the lamb. If the syntactic information doesn't provide a satisfactory resolution, then one envisions having to provide some primitive sense perception based on the size of a river relative to that of a wolf or on what animals, perhaps even wolves, have been seen to eat, or some combination of those perceptions. The role of Roget's and the dictionary, obviously, would be to try to reduce markedly the number of such perceptions one would have to build into a system and to capture, instead, many such perceptions through patterns of word association which presumably reflect one's sense perceptions (and [implicit] reality models).

It is certainly clear to me, as I'm sure you will be able to guess, that an effort to get down to cases and actually program a computer to make its way through a thesaurus and dictionary, draw the appropriate inferences, and combine those inferences with syntactic and sensory information is going to be exceedingly complicated. But it needs repeated, strong emphasis that we require general, not ad hoc, solutions to these problems. I argue that one should be exploring the possible utility of these general-purpose reference works with an eye toward revising and adapting them to the needs of specific discourse analyses for the very pragmatic reason that no one seems to have the patience to construct from scratch a thesaurus or dictionary that deals with a very large segment of "reality" and, at the same time, is specifically designed for a set of "language understanding" computer programs. Further, there is at present no consensus as to what theories or procedures the computer programs should embody; therefore, there is no consensus as to how words and their semantic relationships should be represented. It is also the case that general-purpose reference works have some claim to cultural valida-
tion; such is not the case for ad hoc thesauri or dictionaries for special-purpose programs.

In summary, one can surmise, I think, how one might build upon some of the types of analyses I have suggested to use the computer to analyze discourse. For small scale, or micro discourse, syntactic parsers will be valuable, as will dictionaries and thesauri for their guides to meaning. Representation of at least visual perception "primitives" probably cannot be circumvented. I've said nothing about sound patterns, but research directed toward enabling the computer to move from the written word to spoken "output" as well as to enable the computer to understand speech is currently in progress and such research clearly has relevance for the study of patterns of sound in discourse. For macro discourse, all the procedures relevant for small-scale discourse are germane. In addition, the kinds of thematic analyses made possible by my programs are relevant for extended discourse. It is then possible to make visible patterns of thematic occurrence within a text by graphically portraying the themes at the locations in which they occur. John B. Smith of Penn State, once a student of mine, used some of my programs and some of his own to portray occurrence patterns of images in Joyce's Portrait of the Artist—and discovered that those moments Joyce had described as epiphanal were graphically obvious because of the coincidence of major imagery patterns at those points in the Portrait. Spans of setting, of identification, and of other categories to which we have alluded will eventually be amenable to computer analysis if we are able to deal with the many thorny issues related to perception of meaning that I have illustrated through examples in this paper.

The importance of computer-based discourse analysis for the many natural language applications in computer-based systems cannot be overemphasized. The long-awaited breakthroughs in information retrieval and in other application areas, such as computer-assisted instruction, which are dependent upon information retrieval (broadly defined) must await an increased algorithmic and computational capability for the analysis and generation of extended discourse. It behooves us to explore every possible avenue to full or partial solution of the many complex problems which must be solved in order to achieve computer-based discourse analysis. The implications are immense—both for the development of the non-numeric aspects of computer science and for the next major stage in the application of computers to the solution of human problems and to aiding with society's work-a-day tasks.

REFERENCES


