DOCUMENT RESUME

ED 195 934 CS 005 742

AUTHOR Pepinsky, Harold B.; DeStefano, Johanna S.
TITLE Toward a Differentiation of Texts to be Comprehended.
INSTITUTION Ohio State Univ., Columbus. Mershon Center.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Apr 80
GRANT NIE-G-79-0032
NOTE 15p.: Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cognitive Processes; *Computational Linguistics; *Discourse Analysis; *Reading Comprehension; *Reading Processes; *Structural Analysis (Linguistics)
IDENTIFIERS *Schemata

ABSTRACT

To conceptualize a reader's comprehension of text as a semantic and interpretive processing of information, it is necessary to take note of interactions among persons and texts and conditions under which the texts are to be comprehended. A Computer-Assisted Language Analysis System (CALAS) was constructed which focuses on the text as any interpretable record of the employment of a language. In making its interpretations, CALAS, utilizes a model of the English language, which imputes to its texts the properties of a syntactic and semantic grammar. Data analysis is accomplished in three stages: (1) analyzing the text to identify each word in sequence in terms of its grammatical equivalent; (2) gathering the individual words into phrases, which are again identified in terms of their grammatical equivalents; and (3) gathering phrases into clauses, with the component phrases displayed within each clause and identifying the phrases within a clause in terms of the roles each plays within the clause. This macro/micro analysis of text invites identification of and comparison among texts in terms of their structural properties. (HOD)

Toward a Differentiation of Texts to be Comprehended

Harold B. Pepinsky.

Co-Director (with Johanna S. DeStefano)

Program in Language and Social Policy

Mershon Center

The Ohio State University

Part of a Symposium on

"Application of Semantic Grammars to Individual Difference Research"

Victor M. Rentel, Chair

Presented at the Annual Convention of the

American Educational Research Association,

Boston, Massachusetts, 7-11 April, 1980

Running Head: A Differentiation of Texts

Research on which this paper is based is supported in part by Grant No. G-79-0032 from the National Institute of Education and by the Mershon Center for Research and Education in Leadership and Public Policy, and is under the auspices of the Mershon Center's Program in Language and Social Policy

Computer-simulation of cognitive processes in humans offers a startling vision of "artificial intelligence" in machines. The work has captured public attention with claims that are often exciting, unsettling, and overstated (Restak, 1980). A contemporary vogue in research on reading draws upon this effort in attempts to portray the complex processes by which people select, interpret, and subsequently retrieve the information conveyed to them by texts. The underlying idea is that readers comprehend and remember texts by recourse to hierarchically-ordered schemata, developed out of prior experience (Adams & Collins, 1979). Much of what is reported, understates the case for individual differences among persons in their development and use of any such schemata.

The research of colleagues like Bonnie Meyer (1980) and Bruce Dunn (1980) confronts us with the necessity of accounting for the fact that people do differ in their comprehension of texts. Meyer's (Note 1) and Dunns's (Note 2) research points up the added relevance of accounting for differences among texts to be comprehended and among conditions under which any such comprehension of text is invited to occur (cf. DeStefano, 1978, and Scribner, 1979, on sociocultural conditions to be accounted for). However tempting it may be for us to conceptualize a reader's comprehension of text as a semantic and interpretive processing of information (Adams & Collins, 1979), it is equally necessary to take note of interactions among persons and texts and conditions under which the texts are to be comprehended. At the very least, the investigation of "reading comprehension" calls for a strategy of inquiry that is multi-faceted.

"Multi-faceted" in this sense implies that what goes on in people's heads as they read or what is in the texts themselves to be comprehended are necessary but not sufficient conditions of reading comprehension and that we are warranted in taking account of individual differences in each as well as in their interactions. Having belabored that point, which I think we need to be reminded of, I want to focus the remainder of my remarks on the important problem of differentiating among texts to be comprehended as "multi-faceted" in its own right. Again, Bonnie Meyer (Note1) gives us a clue as to dimensions of this problem in describing her construction of a system in which texts can be analyzed in terms of their structural properties at a top, middle, and bottom level of analysis. Fortunately, she is with us to speak more eloquently for herself on this symposium.

What strikes me is that there are critical respects in which our views of the problem are consistent with each other. A first is the assumption of levels at which texts can be analyzed, ranging from a more global, macroanalytic perspective to one that is more narrowly focused, hence more microanalytic. A second is the presumption that structure can be imputed to texts at any of the levels into which they may be composed. At the risk of putting words into Dr. Meyer's mouth, which she may not wish to utter, let me suggest that for both of us the concept of "structure" postulates the existence of named classes of phenomena and their relations to each other. In the domain of analyzing texts where I am most at home, namely that of clauses and larger blocks of main and subordinate clauses, essential structural ingredients are postulated to exist in the form of noun phrases as named things and of verb phrases as things that define relations between noun phrases.

In the remainder of this paper, I shall describe briefly the system for analyzing texts from a microanalytic perspective, at once the primary source and object of my remarks about text on this occasion. Elsewhere, I

have provided the foundation for a theory of meaning in the context of which the system has been constructed (Pepinsky, Note 3), so I'll but touch upon it here. The idea is that when people employ language in communication with each other, they are both (a) governed by implicit linguistic rules--which make possible a common sensing and understanding of things--and (b) are prone to innovate and/or modify their linguistic rules--which enable them to enhance their sense of common understanding. Hence, in their communication, people are inferred to be both structured and structuring. Their language itself is thus identified as a system of formulations which enable people to make evident to each other their access to information in a mode of communication (Pepinsky & Patton, 1971). Based on these principles, A Computer-Assisted Language Analysis System (CALAS) has been constructed. CALAS centers our attention on text as any interpretable record of the employment of a language. The texts of events like counseling interviews or talk in a classroom afford familiar and pertinent examples of such records. Given present technology, printed transcripts, as the text of spoken English, may be fed as inputs directly into a computer, which, when properly instructed, promptly reads and interprets their contents.

In making its interpretations, CALAS utilizes a model of the English language, which imputes to its texts the properties of a syntactic and a semantic grammar. The syntactic grammar presupposes the text to occur as a string of words, such that each word or cluster of words may be assigned a grammatical label in terms of the slot it occupies and the purpose it serves in an ordered sequence. In contrast, a semantic grammar attributes relationships to component parts of the sequence, apart from the order in which they appear. These linear and non-linear structures imposed upon texts by the model comprise a metalanguage, which incorporates the rationale

and methodological features of a computational psycholinguistics (Pepinsky, Note 3). Today, I may but outline briefly how this is done (for a fuller description, along with a partial list of earlier research studies in which it has been employed, see the CALAS Manual: Pepinsky, Baker, Matalon, May, Staubus, Note 4).

For its computer operations, CALAS relies upon a series of four programs of language analysis which, along with their implementing rules, make use of two programming languages: SPITBOL and PL/I. These programs, designed to be run on an IBM System 370/Model 168 Computer, have been adapted for use on certain other computers such as our present Amdahl system. But that is only part of the story. By design, CALAS also includes human editors who, according to instructions, assist the computer and its human programmer in the processing of data. As its raw material, CALAS ingests "machine-readable" text, which has been key-punched onto cards or tapes from original text, or transcripts of speech.

With these resources, data analysis is accomplished in three stages. It is in Stage 1, called EYEBALL, that CALAS makes the syntactic analysis of text, identifying each word in sequence in terms of its grammatical equivalent, e.g., noun, verb, adjective, adverb, preposition. CALAS does this by reference to a small dictionary (of approximately 600, mainly "function (Fries, 1952) words) and, as importantly, a set of rules for identifying other words in terms of where in sequence each appears and what role it is thus intended to play. Where alternative roles are plausible for a word, the computer is programmed to list these in the order of their most likely occurrence for that word in that slot, e.g., as adjective, verb, noun. At this point, one or more humans (we recommend at least two persons) edit the text according to instructions, rapidly correcting the relatively few evident "errors" made by the computer. That editing is an important addition to the

process of analysis because, as will become apparent, earlier errors become compounded in later stages of analysis.

The edited output of EYEBALL becomes input for a second stage of analysis called PHRASER. Here, guided by a second set of programs, the computer aggregates the individual words into phrases, which are again identified in terms of their grammatical equivalents, e.g., as noun phrases, verb phrases, adverbial phrases, prepositional phrases; also, as conjunctions and subordinators (i.e., terms that introduce main and subordinate or partial clauses). PHRASER, like EYEBALL, is then edited, and the system is ready for its third and--at present--final stage of analytic display.

Stage 3 in the analysis of data processed by CALAS is called CLAUSE/CASE. At this stage, the computer is instructed to do three things, and by reference to a third set of programs. First, phrases are aggregated into clauses, with the component phrases displayed within each clause. Second, the phrases within a clause are identified in terms of the roles each plays within the clause. Because by definition a clause contains one and only one predicate--notably, a verb phrase--the verb phrase itself becomes an essential feature of the clause, with other phrases as optional ones. Verb phrases, then, are identified as particular types, basically as verbs of state, action, process, or action-process; and secondarily as compounds of experiential or benefactive states or actions. Noun phrases that accompany the verb phrase are thus identified in terms of their case roles within a clause, as objects, agents, experiencers, or beneficiaries of a state or activity. Finally, the clauses themselves are exhibited to display a main or independent clause, along with clauses subordinate to it and in the order of their embedding within the block of clauses.

In thus applying our linguistic system to the analysis of language used in interactive talk, which we've been doing since 1974 (e.g., Bieber, Patton, & Fuhriman, 1977; Hurndon, Pepinsky, & Meara, 1979; Meara, Shannon, & Pepinsky, 1979; Patton, Fuhriman, & Bieber, 1977), two major kinds of information have been quantified to date. The first includes content measures: prominently, of the relative frequencies with which the different types of verb phrases are used. The second includes measures of structural or stylistic complexity: again, prominently, the ratio of the total number of clauses to main clauses and a measure of the average embeddedness of clauses within blocks. It seems strange to be telling you here that this microsystem of analysis, like the macrosystems we are using (DeStefano, Pepinsky, & Sanders, Note 5; Pepinsky, DeStefano, & Sanders, Note 6; cf. Halliday & Hasan, 1976; Sinclair & Coulthard, 1975), is as easily applied to texts already in written form. That is because impetus for the technical development of CALAS came from the need for efficient, effective methods of indexing and abstracting scientific and technical or other written documents (Rush, Pepinsky, Meara, Landry, Strong, Valley, & Young, 1973; Strong, Note7). After suffering for intervening years the slings and arrows of outrageous conversations that we have been attempting to analyze, it's a relief even to contemplate dealing with the expository or narrative prose of texts originally designed to be read and comprehended in that form.

A noteworthy feature of CALAS is that it invites identification of and comparisons among texts in terms of their structural properties. To repeat myself, I mean by "structure" the specification of things named and the designation of how they are related to each other. From the relatively microscopic perspective of CALAS, again, the things named are grammatical surrogates for words called noun phrases; the relations between them are

specified by another class of grammatical surrogates called verb phrases; more peripheral relationships may be found in terms of such things as adverbial or prepositional phrases. These and other features of a case grammar (ours is patterned after that evolved by Cook, 1979, who synthesized the earlier work of Fillmore, 1968, and Chafe, 1970), have been drawn upon for the purpose of interpreting texts and of differentiating among them as structural phenomena. One of my former colleagues, Sue Strong (1974), nicely extended my proposal (Pepinsky, 1974) for thus viewing and comparing texts at empirical, analytic, and formal levels of display. She then proceeded to outline a series of steps for translating texts thus analyzed into two- and three-dimensional graphic forms. Accordingly, named things could be represented as nodes and relations among them, as connecting lines of various shapes and slopes. She then demonstrated how the idea of names and relations embodied in informational blocks of clauses ("sentences") could be extended to include those embodied in still larger segments of text.

I regret that a change of jobs, though it benefited Sue Strong, also made it necessary for her to abandon her promising research. Stimulated by the work of Bonnie Meyer (Note1) and others on micro-/macroanalytic schemata for interpreting texts, however, I have returned with enthusiasm to Strong's (1974) long-neglected proposal for integrating research on texts at various levels of analysis. The concepts of names and relations form a cornerstone for inquiry along these lines.

In my opening remarks, I suggested that Bonnie Meyer's (Note1) rationale and methods for analyzing text similarly presumes the existence of named phenomena and relations between them, at successively more global levels of analysis. There is a bonus to be had for viewing texts in this manner. Namely, it becomes possible to postulate for all interpretable texts the existence of named phenomena and relations between them that render the

texts isomorphic to each other by virtue of structural properties that they possess in common. Moreover, it becomes possible to specify for any given text peculiar attributes of structure and content that set it apart from any other text.

There is a methodological problem lurking in all of this, which in conclusion, I'd like to call to your attention. In my experience, it has become a truism that the most richly meaningful harvest yielded by analyses and differentiations among texts also demands the most highly skilled, knowledgeable, and otherwise thoroughly indoctrinated of human raters. The same principle holds for the most globally inclusive purviews, i.e., the most encompassing of entire texts, and of all that is implied--linguistically and paralinguistically--when people are understood to communicate with each other by means of natural language. Conversely, the most rigorously specified, reliable, evidential, and replicable kinds of analyses are also the most trivial and the dullest, and the least related to events that are "environmentally probable" (Brunswik, 1956) in everyday life. The trick is to learn the most about texts and with the least amount of self-deception, in describing and differentiating among them. I should hope that increased attention to the structural properties of texts, via a specification of their constituent features as named classes of phenomena and relations between them, would make our differentiations among texts more amenable to sensible talk about what readers are being exposed to. Above all, I should hope that multi-faceted inquiry would be encouraged so as to keep things both interesting and explicable.

My remarks in this paper presuppose a major requirement for students of reading to be the persistent one of differentiating among texts to be read and digested--treating these, if you will, as stimulus conditions whose systematic manipulation can afford us a clearer picture of what it is that readers are being invited to comprehended. I have proposed for this purpose the prior task of identifying and categorizing texts in terms of their structural properties, essential features of which are postulated to exist as names that can be imputed to things in the text and as relations among those named things. I have proposed further that structural elements of this kind can be identified concurrently at microscopic and macroscopic levels of analysis, rendering the varieties of analysis as much as the varieties of text to be analyzed--isomorphic to one another by virtue of their common structural properties. Examples are the Computer-Assisted Language Analysis System (CALAS), a microanalytic system described briefly in this paper, and the macroanalytic system which Meyer (Note 1) and Dunn (Note 2) will now proceed to introduce and discuss on this symposium.

My concluding remarks about the gains and opportunity costs to be realized in choosing one over another mode of analyzing text, can be extended to encompass the larger problem of determining what and how people comprehend when they read. Given the state of the art, this is no time for restricting our purview. What I have elected to press for here is a systematizing of knowledge about the phenomenon of text itself as a congeries of stimulus materials that people are exposed to when they read. What people do with these materials and the conditions under which any such exposure takes place are inescapably important components of whatever we may choose to identify as reading comprehension. My argument is that the identification and comparative analysis of texts in terms of their structural properties is as inescapably important to us if we are to make better sense out of their readers.

Reference Notes

1. Meyer, B. J. F. Text structure and its use in the study of reading comprehension across the adult lifespan. Part of a symposium on "Application of semantic grammars to individual difference research," V. M. Rentel, Chair. Presented at the annual convention of the American Educational Research Association, Boston, Mass., 7-11 April 1980.

2. Dunn, B. R. Individual differences in semantic recall from texts. Part of a symposium on "Application of semantic grammars to individual differences research," V. M. Rentel, Chair. Presented at the annual convention of the American Educational Research Association, Boston, Mass., 7-11 April 1980.

3. Pepinsky, H. B. A metalanguage of text. In V. M. Rentel (Ed.) Psychophysiologic aspects of reading. Elmsford, New York: Pergamon, Manuscript in production.

4. Pepinsky, H. B., Baker, W. M., Matalon, R., May, G. D., & Staubus, A. User's manual for the computer-assisted language analysis system, Second Edition. Columbus, Ohio: Program in Language and Social Policy, Mershon Center, Ohio State University, 1977.

5. DeStefano, J. S., Pepinsky, H.B., & Sanders, T. S. Making policy: a preliminary note on the language of cultures-in-contact in the educational domain. Presented at the International Conference on Language and Power, Bellagio, Italy. 4-8 April 1980.

6. Pepinsky, H. B., DeStefano, J. S., & Sanders, T. S. Discourse rules taught to and learned from culturally diverse children during literacy instruction. Part of a Symposium on "Discourse Processes

6. (cont'd.) in School Settings, " J. S. DeStefano, Chair. Presented at the annual convention of the American Educational Research Association, Boston, Mass, 7-11 April 1980.

7. Rush, J. E., Pepinsky, H. B., Meara, N. M., Landry, B. C., Strong, S. M., Valley, J. A., & Young, C. E. A computer-assisted language analysis system. Columbus, Ohio: Computer and Information Science Research Center, OSU-CISRC-TR-73-9, Ohio State University, 1974.

References

Adams, M. J., & Collins, A. A schema-theoretic view of reading. In R. O. Freedle (Ed.) New directions in discourse processing. Norwood, New Jersey: Ablex, 1979. Pp. 1-22.

Bieber, M. R., Patton, M. J., & Fuhriman, A. J. A metalanguage analysis of counselor and client verb usage in counseling. Journal of Counseling Psychology, 1977, 24, 264-271.

Brunswik, E. Perception and the representative design of psychological experiments. Berkeley, California: University of California Press, 1959.

Chafe, W. L. Meaning and the structure of language. Chicago: University of Chicago, 1970.

Cook, W. A. Case grammar development of the matrix model (1970-1978). Washington, D. C.: Georgetown University Press, 1979.

DeStefano, J. S. Language, the learner and the school. New York: Wiley, 1978.

Fillmore, C. J. The case for case. In E. Bach & R. Harms (Eds.) Universals in linguistic theory. New York: Holt, Rinehart & Winston, 1968. Pp. 1-88.

Fries, C. C. The structure of English. New York: Harcourt, Brace, 1952.

Halliday, M. A. K., & Hasan, R. Cohesion in English. London: Longman Group, 1976.

Hurndon, C. J., Pepinsky, H. B., & Meara, N. M. Conceptual level and structural complexity in language. Journal of Counseling Psychology, 1979, 26, 190-197.

Meara, N. M., Shannon, J. W., & Pepinsky, H. B. Comparison of the stylistic complexity of the language of counselor and client across three theoret cal orientations. Journal of Counseling Psychology, 1979, 26, 181-189.

Patton, M. J., Fuhriman, A. J., & Bieber, M. R. A Model and a metalanguage for research on psychological counseling. Journal of Counseling Psychology, 1977, 24, 25-34.

Pepinsky, H. B. A metalanguage for systematic research on human communication via natural language. Journal of the American Society for Information Science, 1974, 25 (1), 59-69.

Pepinsky, H. B., & Patton, M. J. Informative display and the psychological experiment. In H. B. Pepinsky & M. Patton (Eds.) The psychological experiment: a practical accomplishment. Elmsford, New York: Pergamon, 1971. Pp. 1-30.

Restak, R. M. Smart machines learn to see, talk, listen, even 'think' for us. Smithsonian, 1980, 10, 48-57.

Scribner, S. Modes of thinking and ways of speaking: culture and logic reconsidered. In R. O. Freedle (Ed.) New directions in discourse processing. Norwood, New Jersey: Ablex, 1979. Pp. 223-243.

Sinclair, J. C. H., & Coulthard, R. M. Towards an analysis of discourse. London: Oxford University, 1975.

Strong, S. M. An algorithm for generating structural surrogates of English language text. Journal of the American Association for Information Science, 1974, 25 (1), 10-24.