

# communications

July 1972  
Volume 15  
Number 7

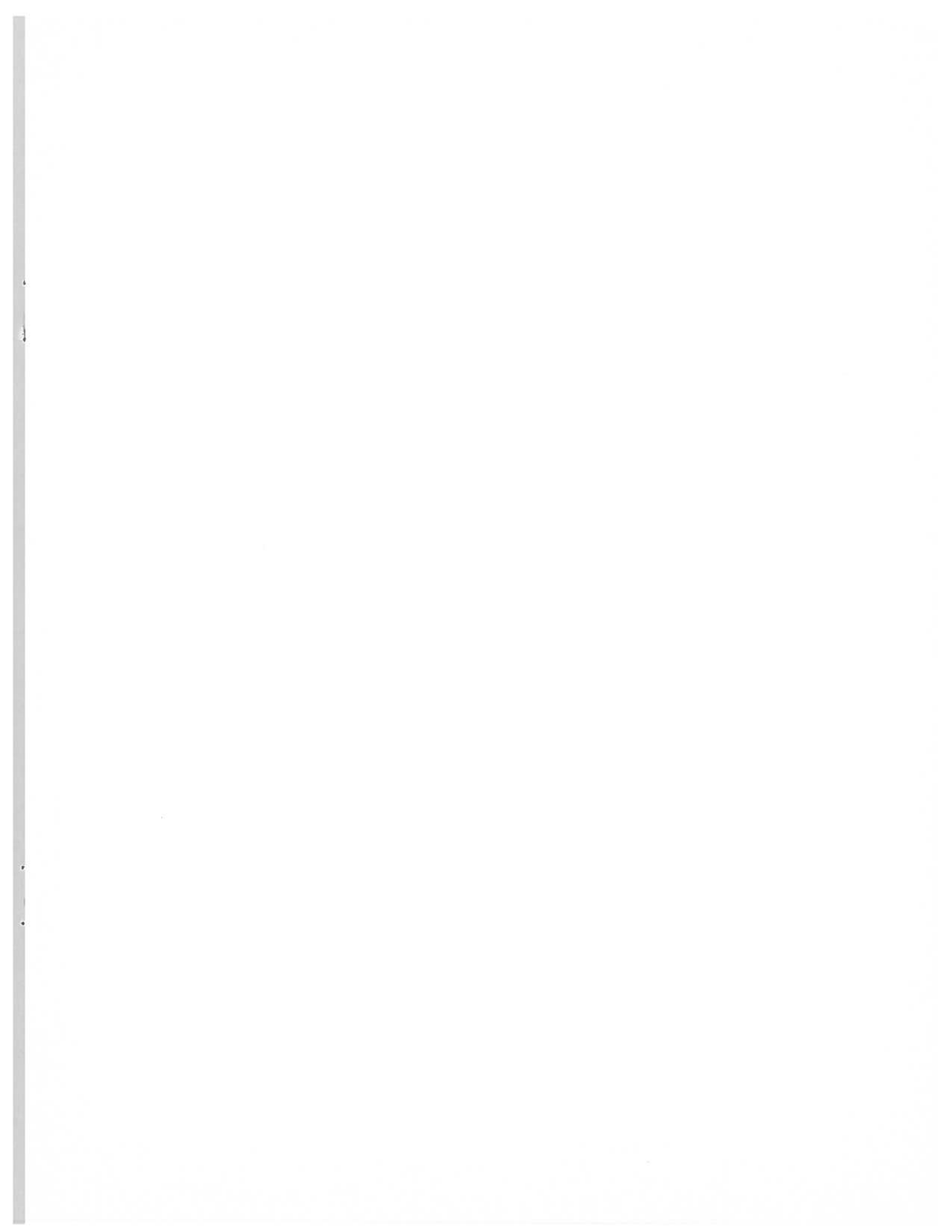
of the

# acm



acm  
1947-1972





---

# Language Analysis in the Humanities

Sally Yeates Sedelow  
University of Kansas

---

**The use of the computer in the language-oriented humanities for exhaustive listing of detail (as in indices and concordances) is widespread and accepted as desirable. The implications of the computer for a "science" of the humanities—a science entailing gathering data for the construction and testing of models—are neither widely recognized nor accepted. This paper argues that the computer's major role as to language analysis in the humanities will be the establishing of such a science. Thus, for those areas of the humanities for which rigor and precision are necessary (e.g. analyzing literature or teaching a student to write a composition) the computer can be a critically important facilitator.**

**Key Words and Phrases:** language analysis, humanities, science of the humanities, pattern recognition, pattern generation, interdisciplinary cooperation

**CR Categories:** 1.3, 2.19, 3.42, 3.43

One could safely assume that the use of the computer for language analysis in the humanities has been prompted by a broad range of motivations and needs. But surely one major thrust has resulted from a desire for rigor: both for the rigor, always acceptable in the humanities, implied by the exhaustive listing of word occurrences in indices and concordances, and for a rigor, hitherto generally unattained in the humanities, in the description of language use and behavior—a rigor which would be so precise and exact that the teaching of composition, literature, and languages would be greatly facilitated [19].

Today the use of the computer for both exhaustive and exhausting humanistic drudge work is both accepted and expected by many humanists. The manual production of concordances (a concordance reproduces each occurrence of a textual word along with a specified quantity of context) is now largely a matter of history. The use of the computer for the compilation of bibliographies also appeals to many humanists, although the utilization of computers for bibliographical research in the humanities is still relatively modest. The Modern Language Association's annual bibliography now uses the computer, and the annual Shakespeare bibliography received a strong impetus in that direction during the course of the World Shakespeare Congress in Vancouver, B.C., in August 1971. Thus far the major obstacle to such "clerkly" efforts has not been a reluctance on the part of the humanist, but rather a paucity of viable computer-based approaches to the automatic establishment of subject heading groupings and other systems of categorization to which humanists have become accustomed. In summary, insofar as rigor connotes an exhaustive listing of detail, the computer is seen by a large set of humanists as an indispensable tool, and its much more extensive use for that type of task is simply facilitation through funding and computer accessibility.

---

Copyright © 1972, Association for Computing Machinery, Inc. General permission to republish, but not for profit, all or part of this material is granted, provided that reference is made to this publication, to its date of issue, and to the fact that reprinting privileges were granted by permission of the Association for Computing Machinery.

Author's address: University of Kansas, Lawrence, KS 66044.

When the definition of rigor is shifted toward the notion of constructing and testing models for the analysis of language, both the concept of rigor and any uses of the computer it entails seem much less congenial to many humanists. The problem is basically one of understanding. Faced with this latter approach to rigor, the humanist often defensively takes "refuge" in a cloud of such terms as "creativity" and "individual sensitivity," perhaps in part because he suspects the outsider doesn't really know what these concepts mean. Unfortunately, the humanist himself doesn't know, insofar as he tries to be analytical about such terms—and as a teacher and scholar he really should be analytical in his consideration [22]. To a degree, such essentially mindless resistance to the use of the computer has given way, as some humanists have used the machine to make apparent certain verbal relationships or relationships among categories—still leaving ample latitude for final interpretation by the scholar. The scope of this brief paper does not permit a survey of such efforts, but for both explication and leads to relevant research, the interested reader may refer to references [4, 5, 17]. For example, individual scholars have worked at problems of author identification, at syntactic patterns in the writing of a particular author, at an author's word choice, and many other aspects of literature.

A sign of the increasing acceptability of such efforts was the establishment of an investigative committee—one of the five such research committees—on the computer and Shakespearean studies for the World Shakespeare Congress of August 1971. As an illustrative indication that scholars from other disciplines also are interested in the use of the computer for language analysis in the humanities, it might be noted that one of the members of this Shakespeare Congress committee was a mathematician from the University of Toronto. It is well known that statisticians have quarried linguistic data in testing statistical techniques [2, 10, 13]. In short, for diverse reasons a number of types of scholars and scientists have used the computer to look at aspects of language.

In addition to the work directed toward specific problems, some effort has been directed toward developing systems of programs which will perform a range of language-analytic tasks. The research with which I have been associated—having as its goal the comprehensive analysis of style—is one such effort [18, 20, 21]. Dilligan and Ule at the University of Southern California are interested in compiling extensive literary statistical data with the help of the computer [8]; and, as the "grammar" of linguistics becomes ever more encompassing (semantics, in addition to the more conventionally recognized syntactics, is now accepted as a part of "grammar"; as, indeed, are aspects of phonology), efforts to implement "grammars" on the computer become more and more comprehensive.

These latter efforts, supplemented and strengthened by many ad hoc projects, represent a powerful move

toward what I would describe as a science of the humanities. It is in the establishing of such a science that the computer will play its strongest role as to language analysis in the humanities. I would argue that, though as yet not integrated, there currently exist strong forces which could help to bring about the crystallization of such a science. As an example of a social force, albeit at work indirectly (in contrast to the more direct and obvious scientific forces indicated in this article), there are efforts at colleges and universities throughout the nation to do away with course requirements in literature, in foreign languages, and in composition. As a teacher of English, I wearied both of hearing colleagues (not in the English Department) say that their students were not learning to write and of hearing my students say that they had no very clear idea about what writing a good theme meant. I wearied of it chiefly because at present there are no very satisfactory responses to these laments. Once a student has been taught to use a dictionary to clean up his spelling, and taught to manage agreement of tense, number, and mood, and to cope with other syntactic rules, there remain much less clearly defined problems alluded to by such terms as coherence, or unity, or diction (e.g. "avoid purple prose") [19]. Exactly the same problems face the teacher of literature as he tries to explain why Milton's "Il Penseroso" is a *better* poem than "On the Death of a Fair Infant Dying of a Cough." At one time the teacher could usually keep ahead of the student in this game (until the student encountered another teacher who held opposite views); but students are now less willing to accept a teacher's literary judgments, and they have not necessarily been able to understand why, for example, the title of a given composition strikes the teacher as *good* and that of another as *bad*. Thus, insofar as the humanist wants to teach students to write in such fashion as to communicate an idea or perception, and to respond to literature in such fashion as to look for relationships, to search for patterns, and in general to develop analytical capacities which depend on explicit evidence, then the humanist must move toward being precise, toward measuring and specifying the components of any given linguistic artifact. Then, if the student and teacher and society at large approve of the writing in that particular artifact, the student would have a good idea as to how to produce something analogous. Insofar as the humanities are moving away from even as much rigor as was implied by the "New Criticism" (typified by the "close reading" advocated by Cleanth Brooks and Robert Penn Warren in, for example, *Understanding Poetry* [3]) they are moving toward sometimes congenial, sometimes uncongenial "rap" sessions on literature and writing. In a sense, such courses are in the realm of "fine arts," and certainly should not be *required* of every student in the university. On the other hand, if it is acknowledged that the world in which we live is increasingly a world of words, as well as of actions, then it is exceedingly important for every student to be able



to deal with words easily, skillfully, and thoroughly. It is also important for him to develop his analytical powers at least as well for languages as, for example, with respect to biological specimens or physical phenomena.

Assuming for the sake of discussion that a science of the humanities is desirable, how will the computer help attain it? As I have stated elsewhere [17], the use of computers in the humanities can be categorized in terms of pattern recognition and pattern generation. The computer's facility for dealing with masses of data has been and will be even more extensively used to pick out (recognize) patterns of character or letter occurrence, of word occurrence, of word type occurrence, of sentence type, paragraph type, and of any other linguistic element which can be specified—so that such patterns can be incorporated into hypotheses or models about the structure or behavior of language which, in turn, can be tested through the use of the model for pattern generation. These efforts to recognize regularities in language and then to attempt to reproduce them will involve not humanists only but with them mathematicians, statisticians, psychologists, computer scientists, sociologists, and scholars/research scientists from any discipline concerned with man's symbolic behavior.

The computer has already facilitated such interdisciplinary cooperation. Before computers, humanists ordinarily did not have sufficient quantities of statistical data in usable form to interest a statistician. Now statisticians find verbal data a rich storehouse of events which behave in a sufficiently aberrant fashion (relative to normal curves and even to transforms which produce a semblance of normality) so as sometimes to challenge the forefront of statistical knowledge, as well as of computer implementations of that knowledge. I predict that the effort to use statistics in pattern recognition in natural language will result in further extensive development of nonparametric statistical methods, as well as of computer packages which will cope with such problems as extremely large but extremely sparse matrices. My own research group is working toward such a package as part of a general language analysis system. I am sure that the work of John Carroll at the Educational Testing Service [6, 7], of Dilligan and Ule at the University of Southern California [8], of Barron Brainerd at the University of Toronto [2], and many other scholars and scientists in this country and abroad will contribute to the development of statistics for pattern recognition in language.

Insofar as the linguistic aspect of the humanities becomes more scientific, there is much greater convergence between the traditional humanistic disciplines and linguistics. Thus, linguists such as Robert Longacre are attempting to construct models of extended discourse analysis which in fact encompass in new ways many of the approaches which humanists have used—often in an isolated, noncoherent fashion—for the analysis of literature. Longacre is attempting to pull together such

approaches and tie them to at least some aspects of a transformational grammar [12]. His effort is representative of the many which apply linguistic models to language and literature. Some of these models are themselves being implemented and tested on the computer [11], which is effecting much greater rigor in the models. One can predict that the use of these models for language analysis in the humanities will become more extensive and, insofar as possible, more completely computer-based.

As linguistic models of grammars have become more comprehensive, the emphasis upon semantics as well as upon syntactics and phonology has strikingly increased. This interest meshes with that of psychologists interested in conceptual networks, of lexicographers interested in dictionaries, of sociologists interested in social discourse, of computer scientists who want to communicate with computers and robots in natural language, of political scientists interested in content analysis, of computer scientists interested in question-answering systems, of computer scientists and information scientists interested in information retrieval, of linguists interested in machine translation, of information specialists interested in automatic abstracting, and so on. One aspect of the work on semantics has been the construction of very small thesauri or semantically related dictionaries put together for demonstration purposes [15, 16, 24]. Another component has been research on the nature of dictionaries and thesauri themselves [9, 14, 23]. I anticipate massive efforts, all computer-based, to characterize thoroughly existing thesauri and dictionaries with a view toward modifying them—again through the use of the computer—so that they can be used as input for generating analogues to the semantic nets used by Winograd [24] or Quillian [15, 16]. Clearly, it is the tedium of hand-constructing semantic relationships for programs such as Winograd's and Quillian's which limits the generality of the system. Thus, it seems eminently desirable to develop computer-based techniques for constructing semantic relationships. Dictionaries and, most especially, thesauri represent a source of such relationships as developed and culturally accepted through time. It seems extremely likely that a significant breakthrough in computer-based semantics will imply major efforts in the analysis of dictionaries and thesauri and the semantic nets implicit in them.

In summary, both scientific research from a broad range of disciplines (including the humanities) and social pressures may be expected to effect a much more extensive and intensive use of the computer for language analysis in the humanities than has been true thus far. The use of the computer—for pattern recognition to provide inputs for hypotheses or models to be tested through computer-based pattern generation—will be central to the rigorous characterization of language. That will indeed contribute, when aesthetically and socially desired, to science in the humanities.

## References

1. Bailey, Richard W., and Dolezel, Lubomir. *An Annotated Bibliography of Statistical Stylistics*. Dept. of Slavic Languages and Literatures, U. of Michigan, Ann Arbor, Mich., 1968.
2. Brainerd, Barron. *Introduction to the Mathematics of Language Study*. American Elsevier, New York, 1971.
3. Brooks, Cleanth, and Warren, Robert Penn. *Understanding Poetry*. Holt, Rinehart and Winston, New York, 1938.
4. *Computers and the Humanities*, (bimonthly), Queens College, Flushing, New York.
5. *Computer Studies in the Humanities and Verbal Behavior*, (quarterly), Mouton; Editorial Office, U. of Kansas, Lawrence, Kansas.
6. Carroll, John B. An alternative to Juillard's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (SFI). *Computer Studies in the Humanities and Verbal Behavior* 3, 2 (Aug. 1970) 61-65.
7. Carroll, John B. Vectors of prose style. In *Statistics and Style*, Lubomir Dolezel and Richard W. Bailey (Eds.). American Elsevier, New York, 1969.
8. Dilligan, Robert, and Ule, Louis. The Mathematics of Style. Letter in *London Times Literary Supplement*. Oct. 22, 1971, p. 1336.
9. Dillon, Martin, and Wagner, David J. Models of thesauri and their applications. In *Automated Analysis of Language Style and Structure in Technical and Other Documents*. S.Y. Sedelow, et al. U. of Kansas, Tech. Rept. No. 1, Sept. 1971, pp. 11-47.
10. Dolezel, Lubomir, and Bailey, Richard W. *Statistics and Style*. American Elsevier, New York, 1969.
11. Friedman, Joyce. *A Computer Model of Transformational Grammar*. American Elsevier, New York, 1971.
12. Longacre, Robert. Extended Discourse Analysis. Invited Lecture, Midwestern Linguistics Conf., Columbia, Mo., (Nov. 1971). To appear in Proceedings.
13. Mosteller, F., and Wallace, D.L. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, Mass., 1964.
14. Olney, John, Revard, Carter, and Ziff, Paul. *Toward the Development of Computational Aids for Obtaining a Formal Semantic Description of English*. System Development Corp., Santa Monica, Cal., SP-2766/001/00, Oct. 1968.
15. Quillian, M. Ross. The teachable language comprehender; a simulation program and theory of language. *Comm. ACM* 12, 8 (Aug. 1969), 459-476.
16. Quillian, M. Ross. Word concepts: a theory and simulation of some basic semantic capabilities. *Behavioral Science* 12, 5 (May, 1967), 410-430.
17. Sedelow, Sally Yeates. The computer in the humanities and fine arts. *Computing Surveys* 2, 2 (June 1970), 89-110.
18. Sedelow, Sally Yeates, et al. *Automated Analysis of Language Style and Structure*. U. of North Carolina, Sept. 1970.
19. Sedelow, Sally Yeates. Computers and Language. *Iowa Alumni Review* (June-July 1971), 6-7; 18-19.
20. Sedelow, Sally Yeates, et al. *Automated Analysis of Language Style and Structure in Technical and Other Documents*. U. of Kansas, Tech. Report # 1, Sept. 1971.
21. Sedelow, Sally Yeates, and Sedelow, Walter A. Jr. Categories and procedures for content analysis in the humanities. In *The Analysis of Communication Content*. Gerbner, et al. (Eds.). Wiley, New York, 1969, pp. 487-499.
22. Sedelow, Sally Yeates, and Sedelow, Walter A. Jr. Models, computing, and stylistics. In *Current Trends in Stylistics*. B. Kachru and H. F. W. Stahlke (Eds.). Linguistics Research, Inc., Edmonton, Alberta, Canada, 1972.
23. Wagner, David J. Thesaurus research. In *Automated Analysis of Language Style and Structure*. S.Y. Sedelow, et al. U. of North Carolina, Sept. 1970, pp. 15-18.
24. Winograd, Terry. *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*. MIT, MAC-84, Feb. 1971.



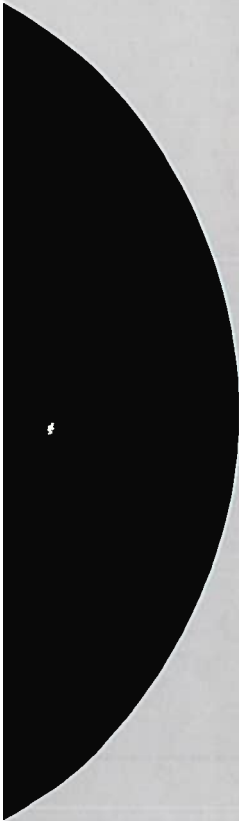




# Twenty-fifth year



**This special issue of Communications commemorates the twenty-fifth year of the Association. It complements the Silver Anniversary activities of the 1972 ACM Annual Conference.**



**Containing articles which focus on the future of the computing scene while using the past as a backdrop, this commemorative issue is a unique tribute to the role ACM has played as a professional society during the formative and rapid growth years of the field. The new president and past presidents of the Association comment on the future of ACM and its relationship to the information processing milieu as a whole —remarks which are counterpointed by perspectives provided by a number of industrial leaders, who join the Association in looking ahead to its next twenty-five years.**