

Stephen N. Matsuba  
Department of English  
Stong College  
York University  
4700 Keele Street  
North York, Ontario  
CANADA M3J 1P3  
(416) 736-5166  
MATSUBA@WRITER.YORKU.CA (Bitnet)

The "Cunning Pattern of Excelling Nature":

Literary Computing and Shakespeare's Sonnets

\*\*NOTE: The original version of this paper included a number of diagraphs illustrating some of its material. Because of the format used to storing these papers, I am not able to include them here. I have, however, kept the references to them in this version. Anyone interested in obtaining a copy of the diagraphs and Table 1 can request them from me at the above address.

--SNM

In his summary of the proceedings of the Literary Data Processing Conference held in 1964, S. M. Parrish declares:

The thing we may not understand, though we ought to soon enough, is that in a revolution of this sort there is no holding back, and no turning back. The movement of events becomes compelling--inevitable. The successful completion of a computer concordance makes the making of concordances by hand old-fashioned, expensive . . . and obsolete. The making of dictionaries or of large bibliographies by hand will soon enough in the same way become obsolete. And not only the making but the using of them is involved. As Professor [Alan] Markman has observed, when all the libraries or at least all pertinent bibliographical references are readily available on tape or in core memory, there will be no excuse for ignorance. But the real force of the revolution has not even yet been intimated. More ominous, some of us may think, but surely just as inevitable, the perfection of attribution study or source study or influence study by computer techniques will make obsolete the studies that rely on the judgment and the memory of one poor fallible human scholar.  
(5)

Parrish's prediction has not only become true, but has been surpassed. The application of the computer in literary research has gone beyond the types of studies that he outlines into areas involving critical inquiry and theory.

Computer-generated studies of Shakespeare's work are not new. In 1973, Dolores M. Burton published a study of grammatical style in \*Richard II\* and \*Anthony and Cleopatra\*, and Walter A.

Sedelow edited a series of papers on the application of the computer in Shakespeare Studies in *Computer Studies in the Humanities*. Stanley Wells and Gary Taylor's *William Shakespeare: A Textual Companion* (1987) provide tables of stylometric data to support claims about authorship, Shakespeare's style, and the chronology of the plays. The sonnets have also been the subject of this kind of research. M. G. Tarlinskaja and L. K. Coachman pursued a statistically-based study correlating text and theme in seven sonnets by Shakespeare. Employing an algorithm to set an "objective" parsing of semantic elements, they identified "thematically relevant semantic components of the content" (339) while outlining a method of comparing texts containing similar themes. The computer-assisted study of Shakespeare's sonnets outlined in this paper was first conceived in 1988 as a test of the application of DiscAn, and the preliminary results were presented with Ira Nadel at the Dynamic Text conference held in Toronto in 1989.

DiscAn is an IBM-PC compatible version of a mainframe computer package for content and discourse analysis designed by Pierre Maranda, Professor of Anthropology at Laval University. The program can process a single machine-readable text or an entire canon, with the only limitation being the space available on one's disk drive. Originally designed to assist in the analysis of myths and folktales, following the work of Propp and Levi-Strauss, it has two main components. The content analysis section includes word-frequency generators, contingency searchers, and programs that assist the user in creating a library of codes and tagging a database. These codes, or tags, can be whatever elements interest the user: rhetorical devices, sound patterns, imagery, et cetera. Once the text has been processed, it can be run through the frequency generators to determine the paradigmatic weights of each tag. The discourse analysis section calculates the probabilities of incidence linking the various tags to each other using Markovian analysis. The output can then be converted into diagraphs, thereby providing a visual presentation, or map, of not only the patterns of co-relatives within the corpus being analysed, but also the strength of the connection between them.

Our study was limited to a rudimentary stylistic analysis of 15 sonnets chosen at random: 4, 26, 34, 57, 68, 82, 116, 122, 137, and 149. Shakespeare's sonnets provided the ideal test: they are known by most people, but are reasonably unencumbered by a vast amount of criticism. Moreover, they allowed us to look at a relatively short body of work with diverse themes and structures written by a single author. The tags we designed followed, to some degree, the syntactic units described by John Porter Houston in *Shakespearean Sentences* (see table 1). After processing the tags, we used DiscAn's content analysis component to determine their frequencies. The tags with the highest occurrence were transitive verbs and conjunctions (each making up 8.18 percent of the total). Pronouns involving the speaker as the subject and the direct object made up 2.97 percent and 0.19 percent respectively. Pronouns involving the addressee as the subject occurred only 0.37 percent of the time.

The discourse analysis output allowed us to examine common patterns within Shakespeare's syntax (figure 1). DiscAn lists each tag in alphabetical order, indicates the tags that precede

and follow it, and notes the probabilities of moving from one tag to another. As well, measures of each tag's frequency and dynamics are given at the end of the output. The most interesting patterns that emerged involved pronouns denoting the speaker and the addressee (PV and PW). The strength of the connections between auxiliary verbs (VA) and these pronouns is the same (24.14 percent). This pattern is the only one in which the latter appear. Following the diagraph, the most likely syntactic order is: subject (SU), followed by a \*wq\*-question word (WQ), then an auxiliary verb, a pronoun involving either the speaker or the addressee as subjects, and finally a transitive verb. We discovered that this syntactic pattern denoted a stylistic pattern in Shakespeare's sonnets in which a noun (or noun phrase) is placed at the beginning of the clause to act as a description of either the speaker or the addressee, as is the case in Sonnet 4: "Unthrifty loveliness, why dost thou spend/ Upon thyself thy beauty's legacy?"

When we compared this result to an analysis of five sonnets from John Donne's *\*Divine Poems\** (see figure 3), we noted a significant difference between the two corpora. The reflexive element that appears in Shakespeare's sonnets do not appear in Donne's. In fact, \*wq\*-question words play no significant role in the latter. And while the Shakespeare-corpus showed a significant co-relation between the verb *\*to be\** and negation (NG), the Donne-corpus was much more prone to one between *\*to be\** and adjectives (AJ).

We also examined the effect that eliminating specific syntactical tags would have on the structure of the sonnets (see figure 2). Removing conjunctions from Shakespeare's sonnets affected only two tags: intransitive verbs (VI) and coordinate nominal *\*that\**-clauses (CT). Conjunctions, therefore, have a primarily coordinating role in this corpus. But in Donne's sonnets, a large number of tags were affected when conjunctions were removed, indicating a subordinating role.

The larger conclusions drawn from the original study focussed on the application of DiscAn in traditional methods of literary study. Professor Nadel and I noted the program's value in traditional literary studies, and felt that it could be used as a supplement to the critical positions outlined by critics like Houston to verify, or refute, their hypotheses and conclusions. Is there proof, as he asserts, that Shakespeare's sentence development advances, and then remains stable for the rest of his career? Can one, as Joel Fineman does, assert a linguistic disjunction in the treatment of imagery in Shakespeare's sonnets which, in turn, creates a new poetics of subjectivity? DiscAn can easily generate evidence that could support or refute such claims.

The program could also be used in matters concerning disputed authorship. Rather than relying on the paradigmatic weights within a corpus, DiscAn allows the further step of using syntagmatic dynamics. Profiles of the patterns of language and style in the known plays could be statistically compared to the same type of profile in, say, *\*The Two Noble Kinsmen\**. Significance tests such as Chi2 or the Mann-Whitney test, as well as factor and cluster analyses, would point to whether variations between the two corpora indicated a real difference, and allow

the critic to determine which passages were written by Shakespeare and which by Fletcher with a greater degree of surety.

The present study includes all 154 of Shakespeare's sonnets, and incorporates the tagging of not only syntax, but also semantic elements and imagery. In this kind of study, a more thorough process of conceptualization is necessary in which a "coding manual"--consisting of a structured list of null words, a structured list of tags, and the operational principles to define both these lists--will emerge. This "filter" rests on reduction formulas that the analyst must make explicit as well as ensure replicability. I am currently experimenting with a more complex set of syntactic tags based on the three-digit York Syntactic Code developed by Robert Cluett (see \*Prose Style and Critical Reading\*, 20-21), and developing a set denoting meaning. My semantic tagging system focusses on the relation of meaning to the speaker and addressee in the poem. Thus I would mark the first three lines of Sonnet 142:

Love is my sin, and thy dear virtue hate,  
Hate of my sin, grounded on sinful loving.  
O but with mine compare thou thine own state,

as follows:

love/be/speaker/transgression/addressee/purity/hate  
hate/speaker/transgression/build/transgression/love  
speaker/compare/addressee/addressee/addressee/condition

This set of tags is still in the developmental stage. Note that some words, like conjunctions, were not included in this group of codes. However, I now feel that they should not only be left in this coding scheme, but that there should also be a differentiation made between coordinating and subordinating conjunctions.

A test run of the Markovian analyser on these semantic tags revealed some interesting clusters. For example, the tag indicating "culpability" is used only in relation to the tag indicating "hate". Moreover in those words associated with "hate", one finds a higher incidence of clustering involving the speaker than is the case with the addressee. I plan to analyse individual sonnets to see if some have patterns that are statistically close enough to say that they can be grouped together. I also plan to compare the corpus as a whole to the sonnets of other writers, and to those that appear in Shakespeare's plays (most notably \*Romeo and Juliet\*).

Some have questioned the validity of statistical analysis in literary study. Houston criticizes its use in stylistic studies:

Some statistics are almost inevitable in a stylistic study. I do not consider them admirable in themselves; nor do I like tables of them, since it is easy to miss the really important figure buried among the trivia. It is quite possible, despite my rechecking, that some of the figures I give here and there are not absolutely accurate. However, I do not regard the difference between, say, nineteen and twenty-two occurrences of a

stylistic device to be significant, and I base no argument on such slight variations. (ix)

Houston supports his declarative statements by a traditional method: with referential evidence. But his claim that no argument will be based on "slight variations" may actually bury "the really important figure." A thorough study of the stylistic patterns within a body of work cannot present \*all\* the relevant passages that support a claim. Something must be left out. However, statistical studies condense the same material into a form that can reveal both large patterns and minute variants within a text or an entire canon. In this way, one can more effectively support claims involving large-scale studies involving areas like character-types, the literature of specific periods, and even entire genres. And the computer is the most logical tool for this kind of inquiry.

But when the computer is used in analyses involving meaning and connotation, the researcher using this tool must consider matters involving critical theory. Creating an encoded text invites a deconstructive critique. The need to produce replicable results raises issues concerning reader response and intentionality. In fact as computer-assisted research becomes more complex, the raising of theoretical questions becomes inevitable. So far, many computer applications in literary study have not dealt with the implications of intertextuality, semiotics, or deconstruction. But the ignoring of these issues will become more and more problematic as computer-aided studies advance.

By way of example, I point to my projected doctoral thesis. The larger study of Shakespeare's sonnets is a part of an inquiry into the development of expert systems for literary study. My interest focusses on the nature of allusions, particularly those identified in Shakespeare's works by various critics. Despite their importance in literary criticism, and in Shakespearian criticism in particular, little in the way of a detailed, systematic study has been done to determine how allusions work, or even to define what they are. Carmela Perri notes that "allusion remains a notion inadequately defined as 'indirect or tacit reference', and is used with no further agreement concerning its characteristics and theoretical status" (289).

My thesis will examine allusions in three main stages, each involving a computer-analysis of the texts. The first will examine the positioning of Shakespeare's allusions to determine if they are clustered together, taking into account different character types, acts and scenes, and genres. In the second stage, I will analyse the syntactic, semantic, and connotative patterns of Shakespeare's allusions. These results will be compared to critical material on the subject, and, based on this analysis, I will design a computer program that searches for allusions in different texts. The third section of my thesis will compare the allusions identified by the computer to those identified by different scholars. And while I am limiting my main research to Shakespeare's works, I hope to draw some conclusions about a more general theory of allusion.

But in defining allusions and their structure, one runs into the question of whether or not a single system exists. Reader

response, intertextuality, and all the theoretical implications they entail further complicate matters. For example, can one measure how a Victorian critic identifies an allusion? Will a single parser be able to determine the relation between signifiers and signifieds over different periods and genres, or must a separate one be created for each? I will have to deal with all of these issues in my dissertation, and other researchers will have to do the same if they wish to pursue similar work.

But computer-assisted research will not stop here. Fuzzy logic and chaos theory presents interesting connections between discourse analysis, theories of the brain, and artificial intelligence. Parallel processing brings the computer closer to the structure of the human brain, and the development of optical computer processors may eventually eliminate existing limitations on processing times. There will be a day, not very long in the future, when the computer will be as much a part of literary study as the book.

This is not to say that the student or scholar will be able to rely on the computer for critical output. Parrish noted that the computer will not replace us as critics, but will make us better ones (7-8). The computer can only act as a tool, albeit an extremely powerful one, that manipulates and condenses masses of data for us. The coding of texts requires an understanding of those texts, and the determination of what may or may not be significant will always be our own. It is as Shakespeare's supposed double, Christopher Marlowe, observed:

If all the pens that ever poets held  
Had fed the feeling of their master's thoughts  
And every sweetness that inspired their hearts,  
Their minds and muses on admired themes;  
If all the heavenly quintessence they still  
From their immortal flowers of poesy,  
Wherein as in a mirror we perceive  
The highest reaches of human wit--  
If these had made one poem's period  
And all combined in beauty's worthiness,  
Yet should there hover in their restless heads  
One thought, one grace, one wonder at the least,  
Which into words no virtue can digest.

(\*Tamburlaine\* 5.1.163-73)

#### Works Cited

- Bessinger, Jess B., Jr., Stephen M. Parrish, and Harry F. Arader, eds. *Literary Data Processing Conference Proceedings*. New York: MLA, 1964.
- Burton, Dolores M. *Shakespeare's Grammatical Style: A Computer-Assisted Analysis of Richard II and Anthony and Cleopatra*. Austin: U of Texas P, 1973.
- Cluett, Robert. *Prose Style and Critical Reading*. Pref. John Stedmond. New York: Teachers College, 1976.
- Fineman, Joel. *Shakespeare's Perjured Eye: The Invention of*

Subjectivity in the Sonnets\*. Berkeley: U of California P, 1986.

Houston, John Porter. \*Shakespearean Sentences: A Study in Style and Syntax\*. Baton Rouge: Louisiana State UP, 1988.

Nadel, Ira B., and Stephen N. Matsuba. "Literary Applications of DISCAN: A Content and Discourse Analysis Program." Literary Computing sess. 1. 9th International Conference on Computers and the Humanities and 16th International Association for Literary and Linguistic Computing--"The Dynamic Text," Toronto, 6 Jun. 1989.

Perri, Carmela. "On Alluding." \*Poetics\* 7 (1978): 289-307.

Tarlinskaja, M. G., and L. K. Coachman. "Text-Theme-Text: Semantic Correlation Between Thematically Linked Poems (seven Sonnets by Shakespeare)." \*Language and Style\* 19.4 (Fall 1986): 338-67.

Wells, Stanley, and Gary Taylor. \*William Shakespeare: A Textual Companion\*. Oxford: Clarendon, 1987.