4-8-1975

Literary Computing: Some Methodological Pitfalls

Steven Sondrup

LITERARY COMPUTING: SOME METHODOLOGICAL PITFALLS

Steven Sondrup

Languages and Linguistics Symposium

April 7-8, 1975

Brigham Young University

LITERARY COMPUTING: SOME METHODOLOGICAL PITFALLS

Steven Sondrup

Since literary critics and historians have long fancied that they dealt exclusively with matters of the spirit and of human intuition, there was considerable resistence on the part of many when computers were first introduced as a tool for literary analysis. Some regarded the computer as an unwelcome interloper that had escaped from the technological, if not the military-industrial complex. The simple fact of the matter, though, is that quantitative judgments, which computers facilitate, have long been part and parcel of literary criticism and analysis. In historical but more particularly in stylistic studies, quantitative judgments have been used to explain many aspects of style and stylistic change. Observations like "The mood of this novel is created by the preponderance of adjective." or "The dynamism of this poem results from the concentration of verbs." are frequent and typical. For such evaluations to be accurate and statistically significant, the relevent stylistic element must be counted and compared to some kind of standard or norm. Typically, though, the judgment is subjective and approximate. Although the intuition of the researcher or critic may in the end be absolutely correct, we deserve and should expect more specific and concrete evidence: we should expect that such subjective observations will be supported by facts.

In the past the expense and tedium of actually counting any stylistic feature has precluded furnishing this kind of specific data, and the public has generally been rather indulgent and understanding. Computers, however, can quite literally in seconds make these counts and comparisons with unrivaled accuracy. It should be noted too that this is essentially all that computers can do: they cannot interpret the data, they cannot explain the significance of the data. In short they cannot replace the mind and soul of the sensitive critic. Computers can only make the perceptive and understanding critic's work easier, more accurate, and more penetrating.

Although there are many things that computers cannot do, what they can do, they do very well. The great precision and exactitude that computers bring to literary analysis, while a welcome antidote against unsupported approximations and intuitive guessing, are awesome. Many traditional stylistic and grammatical categories and definitions are too inexact or ambiguous to be used in conjunction with the great accuracy afforded by computer-aided tabulation. The critic, is, therefore, necessarily faced with the need to rethink and reconsider some of the most basic categories and concepts of his discipline, lest the precision of the computing techniques be diluted by the ambiguity of poorly defined o ill-considered categories.

One of the first steps in any computer-aided study is the generation of a word-frequency chart. The procedure is simple from a computing point of view; it involves counting how many times each word in the corpus is used and then computing what percentage of the total each word represents. Although this is a simple, straightforward procedure from a technological standpoint, it requires the literary critic to ask himself some very probing questions. What is, for example, meant by the term "word" in this context and how does one "word" differ from every other "word"? Although Chomsky and Halle provided a penetrating and precise definition within the framework of transformational grammar[1], the literary critic must nonetheless consider how, within the context of his particular study, this crucial term is to be defined. If an unedited text is submitted to the computer, and the computer is programmed to count how many times each word in the corpus appears, the results supplied will be based solely on orthography. All items that have exactly the same spelling will be grouped together and counted together, and conversely variant spellings, inflected forms, and abbreviated forms of what is usually considered the same lexical item will be grouped separately and counted separately. For a study concerned with a particular author's orthographic habits, such a count would be useful, but for studies concerned with more subtle elements of style, such statistics would be only marginally applicable, if not completely misleading. The first step in going beneath the orthographic surface should involve coding the most obvious homographs, so that the computer will be able to distinguish between them, list them separately, and count them separately. Thus book in the sense of a printed volume should manually be distinguished from book in the sense of making a reservation. Or in German der Arm-- the arm-- should be distinguished from the adjective arm--poor-- in some way.

Closely related to this segregation of the most obvious homographs, but perhaps slightly more subtle, is the question of words that in common parlance seem to have a wide range of overlapping meanings but syntactically and perhaps grammatically behave very differently. In many cases it is a matter of two very different deep structures emerging in the same surface structure. Consider for example the German word auf meaning roughly "on". It can function as an ordinary transitive preposition--that is a preposition that takes an object-- as in the sentence: Der Hund springt auf den Stuhl. (The dog jumps onto the chair.) The same lexical item -- the word auf--can also function as a verbal partical as in the sentence: Die Musik hört sofort auf. (The music stopped immediately.) Now the inclination of some might be simply to regard the word auf as a single item that has two or more different meanings, and in some situations such a procedure might be perfectly satisfactory. But the matter is more complex than it appears on the surface. The basic question is not one of conceptual meaning in the ordinary sense of the word, but rather a matter of syntactical behavior of the

1 Noam Chomsky and Morris Halle, The Sound Pattern of English (New York: Harper and Row, 1968), pp. 12-14.

two items. If the computer-assisted analysis of any literary text is going to be useful in making judgments about the grammatical patterns or syntax of a particular poet, such syntactic subtleties must be reflected in the preparation of the text. In this case auf as a preposition and auf as a verbal partical would necessarily have to be distinguished in such a way that the computer would be able to tabulate them separately.

Perhaps an even more telling and extreme example can be found in the English word have. It is such a common word that the complexities beneath the surface are often overlooked. Have can be used as the main verb of a sentence: I have a house in the forest. In such sentences have is the only verb in the sentence and takes an object as normal transitive verbs do. Have though can also be used as an auxillary verb to form the past tenses of other verbs: They have already come. Although the transitive form of have and the auxillary form of have in many respects seem to be the same word, syntactically and grammatically they are two very different words: their spelling and some aspects of the functioning appear to be similar, but they derive from very different deep structures and just happen to look alike at first glance. The very different nature of these words is clearly reflected in other languages. Italian, for example, in some stylistic respects illustrates aspects of this difference, but in Spanish the distinction is clear and obligatory. Have as a main verb requires the use of the tener: Tengo dos cabillos. (I have two horses.) Tiene una casa en Mexico. (He has a house in Mexico.) As an auxillary verb Spanish requires the use of haber: Ha hablado. (He has spoken.) or Ya ha comido. (He has already eaten.) Although the difference between the two words is not as apparent in English, it is nonetheless as real and as critical. In English moreover it is not only a matter of accurately representing the syntactic structure of any particular literary text: there are a number of phonetic implications that play a role in any metrical analysis of the text in question. Have and its conjungated forms --has and had-- when used as main verbs in clause are stressed, but as auxillary verbs, they are not stressed. Computers have recently proven extremely useful in scanning large quantities of poetry and suggesting the favored metrical patterns of different poets, but in order to accomplish this task effectively, the distinction between the stressed and unstressed forms of have necessarily must be indicated.

Although in English there are relatively few other words that function in the same way, in other languages the verb to be can function both as a main verb and as an auxillary verb. Consider the French Il est ici.(He is here) and Il est venu aujourd'hui. (He came today.) A French text would thus have to be examined for occurrences of etre both as a main verb and as an auxillary. The matter is even more complicated in the case of German: in addition to the corresponding forms of have and be--haben and sein--werden, the verb corresponding roughly to become, can be both an auxillary and a main verb. Consider for example, Er wird bald König . (He soon becomes king.) and Er wird bald hier sein. (He will soon be here.) Just as have must be examined in English and avoir and etre in French, in German this procedure of segregation would necessarily include haben,. sein, and werden. This list is by no means exhaustive: depending on the language and the text in question many other similar matters would have to be considered. This list, though, should suggest at least the nature and, to a degree, the extent of the problem.

Just as the accuracy and precision of computer-aided tabulating techniques require precision in distinguishing one word or lexical item from another, great care must be exercised in classifying words. It would be of considerable interest to know whether a given poet, for example, used more adjectives in his verse than is typical in prose. It would also be very enlightening to know whether nouns rhymed with nouns or with verbs or with adjectives that modify them. It would be useful to know what class of word most frequently appeared in stressed positions and what class appeared most frequently in unstressed positions. But before this kind of information can be provided a working definition of the various parts of speech must be established. Too many studies have been based on the traditional largely Latin-based eight parts of speech. Though these eight parts of speech may be a useful point of departure, they are generally too poorly defined to provide significant information. In endeavoring to overcome this difficulty, some researchers have carefully defined the parts of speech in terms of the structure of the relevant language, in one case producing twenty-four different parts of speech. In another case the emphasis was placed on simplicity: only four parts of speech plus a large class of undefined or unspecified words were established. Both the approach of defining categories with great precision and thus multiplying their number as well as that of simplifying and thus working with a relatively small number of classes have advantages and disadvantages. The crucial factor is simply establishing a system that will yield the kind of information sought, whether it be very general or highly specific.

Although it would be very difficult if not entirely impossible to establish criteria that would apply in all cases, general guidelines can be suggested. Perhaps the first important consideration to be borne in mind is that categories should be established that are useful in analyzing the language in question. Although this may seem obvious, critics with a distinctly traditional, literary rather than linguistic background have in the past applied categories that make eminently good sense in Latin to language where they do not fit well at all. If, for example, the language in question does not distinguish between adjectives and adverbs, it makes little sense to set up the categories; a general category modifier would probably make more sense. Secondly, the categories should be defined specifically in terms of the kind of data that is being sought. If only questions of a very general nature are being asked, then quite obviously general categories will suffice. If, however, more specific questions are to be asked, more precise categories would be required. Ideally, though, more than surface structure should be reflected in both cases. Nominalized adjectives, for example, in many respects are nouns, but in terms of the deep structure they modify a substantive

that does not emerge in the surface structure. How then should such adjectives be counted? Ultimately the answer will depend on the kind of information being sought, but the linguistic facts of the matter should at least be kept in mind.

While the general goal of computer-aided literary analysis is to bring a degree of precision and exactness to a field where these qualities have been notably lacking for many years, it must be remembered that language and especially literary language is not mathematically exact. Ironically therefore the final suggestion for setting up a framework within which the usefulness of the computer can be optimized will in many respects necessarily relativize and perhaps temper the foregoing pleas for exactitude. Since language and literature are full of ambiguities--matters that ultimately must be left open to individual judgment and interpretation--any critical framework that is extablished to analyze literary language must leave room for these ambiguities. There is much poetry that is based on the tentative and at times inexact nature of language. To resolve ambiguities in one direction or another, even if this is done consistently and with great care, introduces a most unwelcome element of arbitrary judgment and destroys some of the fundamental meaning of the passage, especially if the poet's intent is particularly ambiguous. Examples of poetic ambiguity are legion, but by way of example consider the opening lines of Gerard Manley Hopkins' "Spring":

> The glassy peartree leaves and blooms,
> they brush
> The descending blue. . . .

What are the words leaves and blooms? They can be regarded as verbs whose subject is peartree, but they are also the antecedents of they and therefore necessarily nouns. Hopkins quite intentionally introduces this element of ambiguity, and in resolving it one way or the other, an important element of poetry would be destroyed.

Thus in an ironic way, the precision that computer-aided techniques provide leads to an awareness of the ambiguous and approximate nature of poetic diction. In spite of the irony involved, it is certainly preferable to persue new standards of accuracy and precision with an awareness of the irresolvable factors that will be encountered than to work in the dark, unaware of this critical aspect of the nature of language and especially poetic diction.