BRUCE A. ROSENBERG
and JOHN B. SMITH

The Computer and the Finnish Historical-Geographical Method

THE FINNISH, OR HISTORICAL-GEOGRAPHICAL, METHOD for the comparative study of folktales needs no introduction to folklorists. The methodology has endured for nearly a century, and the surest signs of its soundness have been the improvements fashioned upon it by its most devastating critics. Recently, it has been attacked again by the current generation of structuralists as being "the study of folklore for its own sake," as though that were a fault.[1] Its endurance, as Stith Thompson's classic study of the "Star Husband Tale" demonstrates, establishes it as one of the procedural standards in folklore methodology.[2]

The demands made upon the researcher, however, are forbidding to all but the most vigorous and dedicated, mainly because of the Herculean housekeeping tasks that this approach requires. When a popular tale is studied—and the popular tales tend to be almost by definition the most "important"—a thousand or more versions may have to be analyzed and compared. Even with the symbol coding commonly used to simplify the description of each tale, the folklorist's tasks of analysis are staggering. For this reason alone Thompson can say, wistfully yet with painful truthfulness, that not many scholars are able to produce more than one or two thorough studies of folktales in their lifetime; most of the books that present the research and results of the Finnish method are the life's work of their author.[3] Obviously the situation cries out for some labor- and time-saving process: the folklorist's burden would be lessened, and he would not only be able to produce more work, but the quality of it would also improve. Alan Dundes remarks, rather cavalierly, in his headnotes to Thompson's model study of the "Star Husband Tale," that "no doubt some enterprising folklorist will one day use an electronic

[1] Alan Dundes, *The Study of Folklore* (Englewood Cliffs, N.J., 1965), 415.
[2] Collected in Dundes, 416–474.
[3] Stith Thompson, *The Folktale* (New York, 1946), 442.

computer to help him separate groups of tales from an unwieldy corpus of texts."[4]

That day has arrived. We have developed a computer procedure for the sorting and classification of symbol-coded folktales. The encoding process required by the computer is nothing more than is presently done in analyzing folktales for the Finnish method. Much of the comparative analyses can be performed by the computer, so programmed, removing a great deal of the mechanical work from these folktale studies. In the following pages, we will describe the system's functions in terms that will be useful to the folklorist, rather than to the computer technician; technical procedures (the algorithm itself) will be kept on file at the Computer Center of the Pennsylvania State University and will be available on an unrestricted basis to all requests made to the authors (who are in the Department of English). We have designated this algorithm the "Folktale Analysis, Retrieval, and Tabulating System" (FARTS), and have filed it under this acronym.

The initial stage of computerization requires that the folklorist present the basic motifs (Thompson's "traits") symbolically, and that he encode each narrative as a sequence of these symbols. Although the Finnish method was designed to study folktales, the computer can analyze any narrative or non-narrative if the item can be coded. This shorthand is nothing more than the coding now performed, and in fact Thompson's coding system of the "Star Husband Tale" can be used. These symbols may be computed even though they have various categorical and hierarchical relations among themselves, although this variation is not necessary. For instance, the trait "A" may exist in three forms, represented as A1, A2, and A3 (the computer does not acknowledge subscriptures), while the trait "B" may exist in only one form: in this example the composite symbols A1, A2, and A3 are considered as a subcategory of "A" and, hence, are "hierarchically" lower than "A." This distinction, which will be discussed more fully below, is essential in comparing the variations of traits—again as Thompson's study indicated. Our system has been designed to accept symbols with as many as ten hierarchical levels; in the "Star Husband Tale" the traits never had more than three levels (A2a, for example). To use the system, the folklorist translates each narrative into symbols of this form, just as he would ordinarily, and then punches this representation onto IBM cards or types the information directly into the computer using some sort of remote terminal.

In addition to the symbolic representation of motifs, the folklorist may also include additional information, such as the tale's language, date of collection, and geographic occurrence. This information enables the researcher to compare at a glance all the tales collected in a particular area, or those told in a specific language or within a certain time limit, if such comparisons are valuable to the research. Coded narratives may also be read into the memory bank and "forgotten" by the

[4] Dundes, 415. In the past few years an impressive number of computer aids have been contemplated for folklore research, as well as for the entire range of humanistic study. Just one indication of this trend is the scheduling of a section at the 1973 American Folklore Society meeting on "Computers and Folklore." Three papers were read: "The Indexing and Analysis of Motifs in a Collection of American Tall Tales" by Russell Reaver; "The General Inquirer System as an Aid to the Cross-Cultural Interpretation of Folktales" by Gary Alan Fine; and "The Use of an Open-Ended Verbal Model for the Processing of Folkloric Information: A Suggestion" by Richard S. Thill. These papers discussed the computer's theoretical ability to process great masses of material quickly and accurately, as well as the use of computer-generated graphs and other visual aids. None of them, however, proposed a specific algorithm.

folklorist until he has sufficient data to begin analysis, when he will have instant retrieval capability by each category.

Once the complete data for a set of tales have been entered into the system and have undergone some preliminary "housekeeping" manipulation (which causes the folklorist no extra labors), the researcher can begin his analysis. A number of analytic options are available, and requests to the machine for information may be made as the folklorist's research progresses. Before describing these various options, however, we must emphasize the essentially passive role of the computer. Its programs provide an extensive array of research aids, including some statistical analytic tools, but in no sense does it produce interpretations; that responsibility lies wholly with the researcher. The computer is his ready slave, following his requests for information quickly, thoroughly, and explicitly, sometimes embarrassingly so. Working with the materials provided by the computer system, he may check and follow his intuitions as his research progresses; but the computer must always be regarded only as a tool that can multiply, not replace, his own mental powers.

The electronic aids are most profitably thought of as those that are "lexical" and those that take cognizance of sequence. By the former we mean the set of symbols used to encode the motifs or traits found in a tale or corpus of tales. Lexical aids consider only the symbols or motifs included in or absent from the tales. Sequential aids, however, take into consideration the order in which the symbols appear: the abstracted narrative itself. We will outline below the basic options available; specific instructions for the actual use of the system are, as mentioned before, included in a separate user's manual available at Penn State.

Lexical Measures (analyzing traits or motifs)

1. Basic lexical counts. Each symbol found in the corpus of tales can be listed in the print-out along with the following statistics: frequency of occurrence, relative frequency (frequency divided by collective frequency for all symbols), the number of tales in which the symbol occurs, and the proportion of tales in which it occurs. These counts are computed for each unique symbol (A2b, for example), but they may also be accumulated at various hierarchical levels (all the A's, all the A2's) at the folklorist's discretion. These statistics are normally printed out in alphabetical order by symbol, but they may also be printed in rank order; by the latter method the statistics for the most frequent symbol are listed first, and the least frequent last. These counts can be used in a variety of ways but are particularly useful in determining archetypes.

2. Sequences lexically similar. This option indicates those sequences that contain the same set of symbols regardless of the order in which the symbols occur; that is, it will identify narratives with the same motifs even though one or more of them may be in some variant order. In addition to sequences that have identical sets of traits, sequences that differ by at most *n* symbols (for some value of *n* specified by the folklorist) may be selected. Thus, identical sequences can be selected by specifying a value of *ø* for *n*, and all sequences will be selected for an *n* equal to the length of the longest sequence. That is, variants of a tale with, for example, fifteen traits can be analyzed for up to fifteen variations. Selection may be based on the entire symbol (such as A1b or B2a), or it may be based on any hierarchical

level; at level 2 the symbols "A2b" and "A2c" would be considered equivalent.

3. Lexical distribution over sequence positions. This option computes the frequency of each symbol for each position in the sequence. That is, it determines the number of times a symbol appears as the first symbol in a sequence, as the second symbol, as the third, and so on. Again, these statistics may be accumulated for any and all hierarchical levels.

4. Lexical cluster analysis. Using a spatial representation of lexical items, this option provides a statistical technique for determining groups of sequences that are "close" to one another, such as would be true of sub-types and oicotypes, in terms of a spatial model. Again, this option may be applied to any hierarchical level.

Sequential Measures (analyzing the entire narrative)

All the options described above are based on the occurrence (or non-occurrence) of symbols or categories of symbols; the following options compare and analyze the sequential order of the symbols as well.

5. Similar sequence measures. This option compares all pairs of sequences and determines those that are identical or that differ at most by *n* symbols, for some specified *n*. All tales (symbol sequences) are compared and classified according to similarity of trait sequence, and similar tales (sequences) are listed together in the print-out. This option can be applied at any hierarchical level.

6. Concordance. A concordance may be produced that will list the sequences in which each symbol appears, grouped by symbol. It would be particularly useful in isolating oicotypes or in listing all the tales in which a certain sequential pattern occurs. A second feature computes the frequency for all pairs of symbols that are within *m* positions of one another, for some specified *m*. It might be used to analyze the tendency of certain traits to occur in proximity to certain others or to analyze their position within the sequence. This data may be printed for reference or passed along to the following option. Both features may be applied at any hierarchical level.

7. Sequential cluster analysis. This option, again using a spatial model, develops clusters of symbols that occur "close" to one another. It uses option 6 to produce the data to which the model is applied. Since option 6 may be applied at various hierarchical levels, the clusters will be in terms of the corresponding level of symbols.

The capability of these procedures will be clarified if we can see how the computer would have assisted Stith Thompson's "Star Husband Tale" research. To begin with, as each version was collected it could have been read into the computer so that the corpus of tales at any given moment in the research would be ready for almost instant tabulation. Preliminary studies or trial runs could have been performed at any time almost effortlessly, possibly allowing Thompson to follow through on hunches and trends that might have altered or short-cut his subsequent work.

One of Thompson's first analytical tasks was to determine the "basic tale," the archetype. It was established by tabulating each of the hierarchical variations of each trait and designating as "basic" the trait that occurred most frequently. We do not know how long this process took, although a few days seems likely, even for so limited a corpus of texts as he used for his demonstration: eighty-six versions. Assuming that all the tales had been encoded and read into the machine, using our system this task would have taken several minutes, including turn-around time at many installations. Several hundred variants would have increased Thompson's time proportionately, but they would have resulted in no noticeable increase in computer time. Option 1 would accomplish this job and a few others besides, as we have indicated.

Thompson had then to compare his archetype with all the known variants to see whether any of the actual tales corresponded with his theoretical archetype. Again, depending on how rapidly he was able to work, this stage might have consumed several hours or even a day or two; options 2 and 5 of the Folktale Analysis, Retrieval, and Tabulation System could have been used with no appreciable expense in terms of time. The folklorist could spend the time "saved" on other work.

Thompson then determined which forms of the tale appeared with only a single difference in detail, which enabled him to isolate the subtypes and oicotypes. Our estimate is that more time would have been needed to make this determination than in discovering structural analogues. Again, this information can be produced almost instantaneously by the computer. But, in addition, had Thompson wished to know which tales varied by two traits (or by three or four), the computer could have informed him of this as well. The print-out would also have shown him at which position in the sequence each variant trait had occurred. One of the major variants of the "Star Husband Tale" involved the simple addition of a single item; we do not know whether such a variation would have been readily noticeable to the eye, but the computer would have made it immediately clear.

Once Thompson determined the principal redaction (the "Porcupine Redaction," so named because the heroines of the tales in this subtype are led to the upper world by following a porcupine), he then tabulated an archetype or basic tale type for it. Option 5 would have identified this subtype for him; and a rerun of option 1 for just these subtype tales would have given him this second subtype abstract. Intermediate versions even more complex and time-consuming to determine would have been readily identified by options 2 and 5.

By listing the tales according to language and geography, the computer-assisted researcher can determine in a very short time what influences language frontiers and linguistic families have played in the tale's transmission and form. The print-out can show us which tales occur where, and, conversely, which tales occur within a specified linguistic or geographical area. Further, option 6 can also show the clustering of certain traits, which would indicate—again as Thompson remarks—the second kind of modification of the archetype, those in which a chain of variants has been caused by a single change of detail. Again, this option can go beyond Thompson's intentions by indicating those traits that appear with significant frequency in proximity to certain others.

The hand worker has some difficulty in isolating narratives containing the same traits if one or more of them are out of the "usual" order; the computer can do this work effortlessly. The researcher's tasks are multiplied when tales vary by several features, but again the computer can make such an analysis in a second and compute the frequency with which a particular trait is found "displaced." Finally, options 4 and 7 present some of their findings graphically so that similarities (and divergences)—and not only of analogues—of both traits and tales can be seen plotted on one plane. To search manually for all these variations is to compare over and over the encoded sequences in all their bewildering and befuddling combinations. With eighty-six separate sequences (a rather modest instance) the task is formidable; with several hundred, it is impractical.

A thorough analysis of several hundred tales may still take several years, but not the decade of effort now common in this field. We do not pretend that the computer will make the job of the historical-geographical researcher easy; only easier. The researcher must still read each narrative (folktales, let us repeat, are not the only narratives that can be coded for analysis by this algorithm) and encode each one, much as he now does. Each encoded tale must also be accurately keypunched (or otherwise "read" into the computer, often by a keypunch operator, enabling the folklorist to trade off money for his time saved). Thus, while much of the mechanical effort is eliminated, some new tasks are introduced.

The analysis, itself, must still be done by the folklorist, and the conclusions drawn by him. The computer, for instance, cannot by itself make the necessary adjustments in evaluating the influence of literary versions on tales. This matter is still one to be determined by the researcher's skill, good sense, and knowledge; the computer "merely" assists him in assembling that knowledge. Similarly, the cultural factors that cause the variations in narratives cannot be deduced by electronic means; they too must be determined by human intelligence. Yet, the work that the computer will perform enables the researcher to make his conclusion more accurately and certainly much more rapidly. We hope that by possibly saving the folklorist years of mechanical work he will not only be more accurate in his conclusions but may also even attempt further study.

The Pennsylvania State University
University Park, Pennsylvania

BASCOM LA
others enjoy
the North C
of such a w
Lunsford in
dance, and r
Beginnin
simple (anc
He visited—
son's Summ
had been a
prosperous
lector and r
A generc
ings), he is
the Library
over 300 fc
items in all
night" in m
In 1928
lina, the fc
ginnings in
Festival. H
His hosp
the Ritchie
Bascom La

University
Lexington,