1980

Comprehensive Text Processing and the Papers of Henry Laurens, Part 2

David R. Chesnutt
University of South Carolina

Comprehensive Text Processing and the Papers of Henry Laurens *Part 2*

DAVID R. CHESNUTT*

[The system which I described in the opening part of my paper at Princeton was one in which the central computer is used primarily for the storage and manipulation of files. We now have a National Endowment for the Humanities grant that supports the development of a new series of programs which will allow us to automatically encode our computer files for typesetting. We have done further testing of this concept, and a report on those tests appears below.]

Automated coding is based on the fact that each part of our printed volume has a standardized format. Beginning with the table of contents and going through the introduction, list of abbreviations, the documents themselves, the source notes, the footnotes, and the index—a careful examination will reveal that each of these sections is printed in a certain way. An examination of the corresponding typescript reveals a rough correlation with the printed page. As an example of this rough correspondence, I have selected the index for volume 7 of the *Laurens Papers* (HL7).

In the typescript (figure 1), each main entry begins flush left and is typed within a seventy-two-space line. If the entry is more than one line in length, subsequent lines are indented five spaces from the left margin. Each line of an entry is separated by one line of space; each entry is separated from other entries by two lines of space. On the printed page (figure 2), the line length is less than half that of the typescript; the hanging indention is smaller; the amount of space between lines of an entry is smaller; and the space between entries is smaller. All of which is simply to say that while the typescript looks somewhat like the printed page, the correlation is general not specific.

The point I want to make is that both the typescript and the printed page have regular structures, even though the structures are not identical. Each line is within a certain length; the hanging indentions are of so much space; the space between lines is such and such; the space between entries is such and such. This regular structure of the typescript—not its general correlation to the printed page—is the factor which makes automated coding for

*David R. Chesnutt is the editor of the Laurens papers at the University of South Carolina. Part 1 of this paper, which was presented at the Association's 1979 meeting in Princeton, New Jersey, was published in the May issue of the *Newsletter*.

typesetting possible. What is most important about the typescript is that you can tell at a glance:

1) where an entry begins
2) where an entry ends;
3) where a shift from roman to italic occurs; and
4) where a shift from italic to roman occurs.

In coding a typescript to get the printed output on a computer typesetter, these are the only four points at which a typesetting code must be inserted.

You can clearly identify each of these code points in the typescript—and if the typescript is placed in a machine-readable file the computer can also identify each of the code points. The regularization of our typed format makes it possible for us to define for the computer each code point. After we have defined the code points, the computer can then insert the appropriate typesetting codes.

I have used the HL7 index as my example for a special reason. In the spring of 1979, we used this index to test the concept of automated coding. We chose the index because of the minimal coding required and because our indexes are already in machine-readable form. The Social and Behavioral Sciences Lab at the University of South Carolina wrote a special program modification of CINDEX to identify the code points and to insert the appropriate typesetting codes. Every code required for typesetting was then automatically inserted. An encoded tape was shipped to Graphics Composition in Athens, Georgia, and a sample galley (figure 3) was produced from the tape.

The coding scheme we used was very simple. We devised a series of mnemonic format codes which were simple combinations of letters and numbers. We actually made up the codes and Graphics Composition defined them according to our specifications. The definition of the codes by the printing contractor has several advantages. The most obvious advantage is that the coding system does not restrict us in the choice of contractors. We can use any printing firm with mid-range sophistication in its computer typesetting equipment. Another advantage is that the codes can be defined according to whatever set of design specifications are required. This means that anyone who uses our computer indexing system—CINDEX—can produce a machine-readable file with typesetting codes.

To further test the validity of using format codes which could be defined typographically in a variety of ways, we turned to a typesetting contractor here in Columbia, South Carolina—the State Printing Company. We gave State a copy of the computer tape which contained the encoded HL7 index, and we asked State to change the definition of our format codes. State redefined all of the variables: the type font, the type size, the line size, the hanging indention, even the shift from roman to italic. In the two State Printing Company examples printed here (figures 4 and 5), the changes are readily apparent. The new tests demonstrated two points: the flexibility of using format codes and the fact that many companies now have the

Figure 1

```
LAUFENS    PAGE:    1

Academies, 213n. See also Philadelphia Academy

Account books, 156, 409

Active, a schooner: case of, 12-13, 72-74, 75, 77, 78, 82, 117;
     owners of, 90

Acts, 232; of Ga., 504

Acts, of Parliament, 13, 23, 26, 61, 65, 69, 70, 72, 73, 74, 75, 78,
     81, 82, 86, 117, 168, 172, 189-190, 226, 266, 293, 298, 307,
     332, 348n, 504, 524; repeal of, 345, 355, 360. See also Mutiny
     Act; Stamp Act; Tea Act; Townshend Duties
```

Figure 2

Academies, 213n. *See also* Philadelphia Academy
Account books, 156, 409
Active, a schooner: case of, 12–13, 72–74, 75, 77, 78, 82, 117; owners of, 90
Acts, 232; of Ga., 504
Acts, of Parliament, 13, 23, 26, 61, 65, 69, 70, 72, 73, 74, 75, 78, 81, 82, 86, 117, 168, 172, 189–190, 226, 266, 293, 298, 307, 332, 348n, 504, 524; repeal of, 345, 355, 360. *See also* Mutiny Act; Stamp Act; Tea Act; Townshend Duties
Acts, of S.C., 39, 153; disallowance of, 408, 432, 433n; incorporating Fellowship Society, 276; on powder magazines, 271; to establish parishes, 432n; to promote tobacco and flour, 180n. *See also* Circuit Court Act; Currency Act (1765); Currency Act (1769); Negro Act (1714); Negro Act (1740)
Adams, John, 295n
Addington, Joseph, 569
Addison, John (V, 312n), 37
Additional Instruction (April 14, 1770), 336n, 338, 408, 432n, 434, 447n, 462; memorial against, 413; report on, 478
Administrator of an estate, 156, 173, 195, 235, 303, 304n, 306, 307, 322n, 510n, 578n. *See also* Letters of administration
Admirals, 449n
Admiralty, 46, 292, 459. *See also* Commissioners of the admiralty; Vice-admiralty courts
Adultery, 376
Advertisements, 141, 161, 162, 243, 245, 256, 323, 326, 422n, 500n, 526n, 573, 576n; by Egerton Leigh, 13n; by HL, 2–3, 200–201, 253–254, 258, 311, 405, 420, 421, 448, 518. *See also* Handbills
Advocate general, 154n. *See also* Dering, John; Troup, John
Affidavits, 25n, 34, 66, 123, 390, 451
Africa, 344n, 356n. *See also* Bance Island; Deloes; Gambia; Guinea; Sierra Leone

Figure 3

Academies, 213n. *See also* Philadelphia Academy
Account books, 156, 409
Active, a schooner: case of, 12–13, 72–74, 75, 77, 78, 82, 117; owners of, 90
Acts, 232; of Ga., 504
Acts, of Parliament, 13, 23, 26, 61, 65, 69, 70, 72, 73, 74, 75, 78, 81, 82, 86, 117, 168, 172, 189–190, 226, 266, 293, 298, 307, 332, 348n, 504, 524; repeal of, 345, 355, 360. *See also* Mutiny Act; Stamp Act; Tea Act; Townshend Duties
Acts, of S.C., 39, 153; disallowance of, 408, 432, 433n; incorporating Fellowship Society, 276; on powder magazines, 271; to establish parishes, 432n; to promote tobacco and flour, 180n. *See also* Circuit Court Act; Currency Act (1765); Currency Act (1769); Negro Act (1714); Negro Act (1740)
Adams, John, 295n
Addington, Joseph, 569
Addison, John (V, 312n), 37

Figure 4

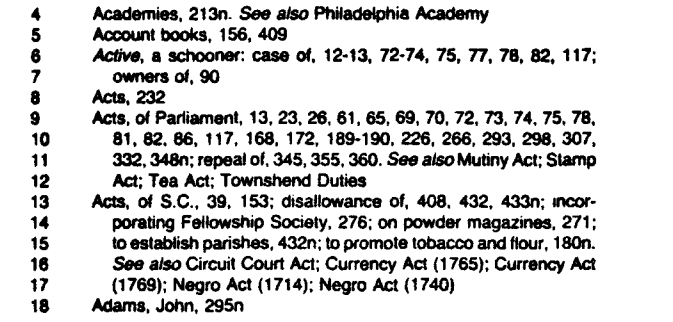
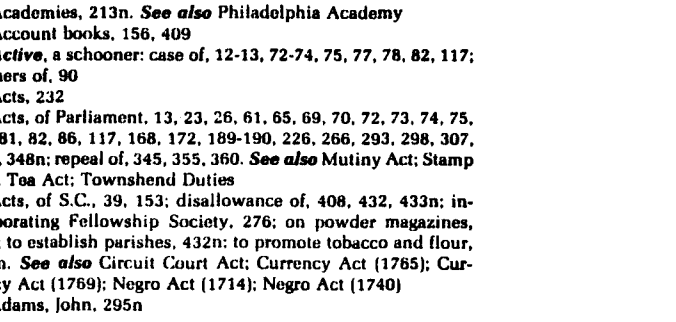


Figure 5



equipment to process machine-readable files into type. We recently began a nationwide survey of typesetting contractors and we expect to publish those results at a later time. Preliminary responses to the survey have already confirmed our belief that many typesetting firms were upgrading their equipment in order to be able to process the machine-readable files generated by their customers.

But let me return to the larger question of automatically encoding the rest of the manuscript. With our success in coding indexes, we have been encouraged to develop a series of programs which will handle the other parts of our printed volumes. We have assembled a panel of local editors from a variety of disciplines at the University of South Carolina to help us design the programs that will be needed for Laurens and for other projects. Our first priority is Laurens, but that priority is almost a "first among equals" because we are committed to develop programs which can serve the needs of others.

Our experience has made us aware that we cannot expect to do 100 percent of our coding automatically. In most cases, however, less than 15 percent of the codes will have to be inserted by an operator. Looking at it another way, this will mean that an operator has to learn no more than 15 percent of the codes and that the chance of an error in coding will be 15 percent or less.

One of my major concerns in developing computer applications is to make the computer work for us, not the other way around. If we are careful, we can use the computer to help us eliminate many of the repetitive steps now required in publishing a volume. And we can do so with a minimal knowledge of computers. The computer is rather like a car. As long as we don't overload it with luxury options of marginal value, we can get where we want to go with a minimum of fuss and a high degree of reliability.