ALLC Bulletin

# ENCODING LITERARY TEXTS: SOME CONSIDERATIONS

J. B. Smith *(Pennsylvania State University, USA)*

* * * *

*Abstract*

This article on encoding conventions differs from most earlier treatments in that it does not propose a prescriptive set of conventions. Rather, it attempts to locate a range of problems that a researcher is likely to face when encoding a text, and to consider the ramifications of various decisions. By examining the 'trade-offs' resulting from various sets of protocols, one can establish principles that are applicable or modifiable for virtually any text. Topics considered include segmentation, punctuation, multiple fonts, accents, deletions, and additions; special emphasis is given to distinguishing between linguistic and physical aspects of a text. The article is cast in terms of an illustrative example taken from James Joyce's *Ulysses*.

* *

*Introduction*

Hopefully, within the next few years a discussion of encoding procedures for natural language texts will seem as antiquated and irrelevant as a discussion of problems involved in turning spokes for buggy wheels seems today. The 'ancient' promise of optical scansion of conventional, printed texts - not texts re-typed in some special font - has been realized in the recently published concordance of Pope. Publishers may even come to view themselves as formatters and purveyors of information instead of producers of a product and make new materials directly available in a machine-readable form. But today, 9 January 1976, most of us, because of economy or access to equipment, still must encode language material by typing it into the computer through a keyboard terminal or using a keypunch.

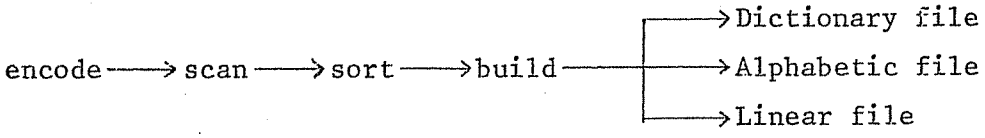
Over the past ten years or so there have appeared a number of articles, both implicit as well as explicit in nature, that have examined encoding conventions. One of the earliest and most complete was Martin Kay's Rand Corporation memorandum, 'Standards for Encoding Linguistic Data', in which he presented a set of specific standards and conventions applicable to a wide variety of encoding problems. More recently F. de Tollenaere has explored some of the problems inherent in encoding Dutch texts;[1] many of the solutions he reached are applicable to English as well. Implicit sets of standards are present in the prescribed conventions required of language analysis packages. See, for example, the conventions required for Sally Y. Sedelow's VIA System[2] or George Borden's CLAS program.[3] In these and all other encoding documents with which I am familiar, the discussion rests with specific conventions for specific textual phenomena that one should or must follow. Little if any attention is given to the rationale behind the conventions or to general principles that might be applied to problems not covered by the particular system.

In considering a set of encoding conventions, the first thing the user must realize is the impossibility of his task. A text is a physical object and, as such, has associated with it a potentially infinite amount of information. To encode all of the information in the text would include not only the inked characters on the page, but in many instances the spacing of those characters (left justified, offset, centred, etc.), their size and shape (bold face, italics, etc.); it is not

inconceivable that with some texts, particularly older ones, the paper and ink composition may be important. Since it is impossible to include all the potential information associated with a text, the user should approach the set of decisions he must make from the standpoint of weighted alternatives: there is no absolutely right or wrong way to proceed, only ways that produce certain results at a particular cost.

Secondly, the set of conventions should be viewed from a systematic perspective. They should be well-defined and as unambiguous as possible. Since encoding is necessarily a many-one phenomenon, the user should be sure that ambiguities are intentional and information omitted will not be needed. For example, in denoting line spacing, one set of conventions might be to note: (a) left justification, (b) paragraph indentation, (c) centering, and (d) 'all other spacing'. Of the above, (d) is ambiguous, but few applications will probably require further definition; of course, if such information is important, it should be encoded. In establishing a set of well-defined and unambiguous (or, at least, tolerably unambiguous) conventions, care should be taken in the complexity of the resulting conventions. Elaborate conventions require more expensive labour and results in more errors. One should strive for the simplest set that gets the job done.

Another aspect of regarding encoding conventions as a system is that the encoding stage is one step in a series of computer-related steps. It is very likely that encoding will be followed by scanning where individual words will be isolated and numbered in some relation to context. In establishing conventions, the user should try to anticipate later processing stages not only to select conventions that will simplify processing but also to take full advantage of textual modification or re-arrangement to do more easily a later task that may be difficult or laborious at the encoding stage. Below, I shall consider encoding problems within the context of the Random Accessible Text System, a set of programs that provides a highly flexible representation of a text.[4] In that system, the encoding stage is succeeded by a scan step, a sort, and a third processing step that builds a type dictionary of the text and two token files of index information.




Thus, factors that may be defined for all occurrences of a word type that must be encoded manually can be delayed until the type file is created in the BUILD step.

Before considering some actual textual problems a few remarks might be appropriate concerning professional attitude. Tremendous savings in time, labour, and money could be obtained if researchers would take into account that future researchers may wish to use their text or that their own future needs may change. Often, decisions that seem arbitrary or unimportant may drastically affect the general usefulness of a text. For example, some researchers omit punctuation because their particular study does not need it; this makes the text useless for many other purposes. If one's needs are limited, no one could ask for inclusion of information that would result in a 40% or 50% increase in labour; but if a 10% increase will make considerable difference in usefulness, one's commitment to one's profession should warrant this. There are at least six different encodings of Milton's *Paradise Lost*, each different with inherent limitations. It would seem that a single carefully encoded version could suffice. Finally, when dealing with literary texts, it is so tempting when one does not absolutely need definitive textual accuracy to hand a paperback version to the typist. Surely, it is worth the

effort to identify and locate the standard or best edition!

*A Specific Problem*

Below is a reproduction of two pages from the text of James Joyce's *Ulysses*, a novel that presents several interesting encoding problems. I have circled for reference some of the textual features that must be considered. (See Figure 1.)

Segmentation. Unless we wish to consider, computationally, the text as one long string of characters, each time some search is made or statistical measure computed, we must identify textual segments. There are at least two kinds of segments we may wish to consider: those that are linguistic (words, sentences, paragraphs, etc.) and hence, inherent in the author's work and, secondly, those that are physical (pages, lines, etc.) and most likely to be determined by the printer. While the first is certainly the more important, physical segmentation should not be ignored: it is much easier for the human reader to locate a word on, say, page 116, line 23, than to find word 50,578! The opposite, of course, is true for the computer; hence one should probably indicate both physical as well as linguistic segmentation.

Decisions concerning physical segmentation for most texts include denotation of physical volume, physical page within volumes, physical line within page, and, finally, position within line. To facilitate proof-reading, I prefer to have typists encode the text one textual line per terminal/keypunch line; if a volume line is too long, continuation can be indicated by some convention, such as a symbol in a specific column or, preferably, some special symbol (I use an '@' character) anywhere to the right of the last word on the line. Since line segmentation is usually arbitrarily decided by the printer, I instruct typists to complete words that are hyphenated at the end of the line. Thus, the radi- and the ance, three quarters of the way down page 133, would be joined without the hyphen to read radiance.

Physical spacing within lines involves a set of 'trade-off' decisions; to mark absolute position can cause considerable trouble. Consequently, the researcher may wish to adopt specific conventions to mark frequent spacings and use absolute spacing for others. For example, rather than trying to count spaces for centred lines, such as the headlines or the centred lines of poetry on page 132, a symbol placed in a fixed column or a reserved symbol can indicate that only the words on that line are centred. (The convention I use is a c in a specified column, but this is, of course, arbitrary.) For more complex spacing problems - the poetry of E. E. Cummings, for example - the researcher may wish to indicate the absolute position of the line by encoding in a fixed field the column or position where the line begins or by flagging this numeric value with some reserved symbol (%23, for example) before or after the line.

Since paragraphs are both physical and linguistic, I mark them in two ways. To facilitate proof-reading, typists indent paragraphs on the left three spaces. Because of the difficulty both during encoding and proof-reading of accurately determining an exact number of blanks, however, I prefer to attach no significance to particular numbers of blanks: whatever one blank means to the scan program, three blanks or twenty-three blanks mean the same thing. To indicate a new paragraph to the computer, the typist doubles the final terminal mark of punctuation. Thus, the full stop after job on line 3, page 133, would be encoded as ..; the scan program, after noting the end of paragraph and taking appropriate action, converts the double punctuation back to a single character.

Pages can conveniently be indicated by placing a number in some fixed field or by using a reserved character in relation to the numerals. While the page for each line can be indicated, the computer can, of course, generate such numbers if only the first line of a new page is indicated. To facilitate proof-reading, I prefer to set off the page number, indicated only for the first line on that page, in the right margin of the line or card, using a fixed field. Physical volume and chapter, which are also linguistic segments, will be considered below.

Linguistic segments, for prose, include volume, chapter, paragraph, sentence, and word; analogous units exist for other genres. I shall consider linguistic segmentation within the context of prose texts with the understanding that similar decisions can be made for other genres. A word might be defined as any string of non-blank characters bounded by a blank or syntactic marks of punctuation; however, to apply this definition is not quite as obvious as it might appear. For example, in the first sentence on page 133, how can we distinguish the full stop after J. from the full stop that ends the sentence? (The first is part of the word itself, while the second, not part of paperweight, should probably be considered as a separate unit or word.) The obvious choices - a list of abbreviations or a frequently employed convention - are both unattractive. My own preference is for the latter since it is unambiguous and is not data dependent. The specific convention I use is to separate all character strings that are to be considered words - including syntactic punctuation - by blanks. Thus, the first sentence on page 133 would be encoded as follows:

--- The moot point is did he forget it ? J. J. O'Molloy said
quietly , turning a horseshoe paperweight .[5]

Semantic punctuation would be encoded as it is: J. J. O'Molloy, etc.

While this convention requires a continuous modification of the text by the typist, I find after a few hours that it becomes habitual and requires no significant effort.

A sentence, using the convention described above, is marked by the occurrence of one of several terminal punctuation marks that is bounded by blanks. Similarly, a paragraph is marked by double-terminals separated by blanks, as described in the section on physical segmentation. Since chapter and volume are rather infrequent encoding tasks, some arbitrary symbol, not expected in the text, can be used to mark them. The conventions I use are $$$ for chapter and $$$$ for volume. These symbols may appear anywhere on a line but are usually placed at the left margin on a line by themselves. So indicated, these divisions become readily apparent for proof-reading.

Type Font. A second set of decisions that must be made concerns encoding of information indicating type fonts. Again, the researcher should balance his immediate personal needs with some reasonable set of conventions that will extend the usefulness of the text. Certainly, upper and lower case should be indicated, but the bold face 'headlines' on pages 132 and 133 might simply be encoded as 'caps'. This convention might not seriously affect someone's reading of the text, but it could limit or make more difficult certain kinds of analyses. In this case, critics have suggested that the headlines reflect a pejorative development in journalism over a period of time; to examine or extend this observation to other stylistic attributes, the critic may wish to locate, computationally, the vocabulary and context of all such headlines. This capability can be built into the encoding conventions by including some marking of bold face type.

The most common form of indicator of a shift in type font has been the escape

character. Under this system, a shift in font is indicated by the appearance of a reserved character (the hash symbol '#' has often been used to signify upper case); the scan program would then mark the following characters as of different font, mark all succeeding characters until the next shift character is encountered, or follow some other prescribed procedure. When correctly employed, this system is fine; but proof-reading is difficult and a single error can cause such catastrophic results that the entire scan operation will have to be repeated.

A similar convention uses an escape character attached to each word or character string that is from a different font. The only decision to be made under this system is whether to place the shift character before or after the word or character string of different font. If placed before the word, all designated words, if the text is ever sorted, will sort out into a clump somewhere in the vocabulary sequence, usually at the top. Placing the character immediately after the word or character string has the advantage that tokens of different font will be grouped with the standard font tokens of the same word type. For this reason, I follow the practice of using shift characters on an individual word basis and append it to the right of the character string.

Looking at pages 132 and 133, there are some five type fonts that require encoding decisions. First and foremost, one should adopt a convention to indicate upper and lower case in the standard font. To indicate upper case by a shift character is a horrendous job; consequently, in my judgement, the keypunch is simply not a viable encoding device for large texts. There are terminals with standard typewriter shifts for upper and lower case priced competitively with keypunches. Some of these can even be used off-line. With this hardware, upper and lower case fonts would be encoded as with a standard typewriter.

Other fonts that are apparent are the bold face headlines and the liberally used italics. Both of these can be encoded by using shift characters: say, a dollar sign '$' for bold face and a hash sign '#' for italics. Note, further, that upper and lower case italics can be encoded with standard upper and lower case strokes when a shift character convention is used. The fifth type font present is the upper case digraph, seven lines from the bottom of page 133. Here, one is faced with the decision of adopting a new escape character or resorting to some other convention. I, personally, would consider this one of those areas of marginal importance that approaches the lower bound of 'reasonableness' and would probably have it encoded as simply AE. For early English texts and for texts in some foreign languages, another decision would be mandatory.

Encoding texts that are represented primarily in some other standard font (Greek, Cyrillic, etc.) requires special consideration. The more common and least attractive solution is to set up a correspondence between the set of characters in the text and the standard keyboard characters; the typist must then mentally translate each character encountered. The opportunities for error, both during encoding as well as proof-reading, are obvious. If one can gain access to one of the several terminals that use the IBM Selectric mechanism, there is a more attractive solution. One simply obtains a relatively inexpensive type ball with the appropriate characters and a set of characters to stick onto the keys. The text may then be typed as it is. It is a simple matter to adjust the internal binary representation of the character for the 'proper' collating sequence during the scan phase.

Other Common Problems. Accents are a common problem, particularly with texts in Romance languages. Since there are character strings that mean two different things with and without accents in many languages, one should encode this information. If numerous accents are present one may wish to take the approach outlined

above: gain access to a Selectric mechanism terminal, obtain a type element with the necessary accents, and type the accented character as it is followed by the accent. Most systems, however, use a shift character immediately after the accented character. Some try to pick a character that looks something like the intended accent. I prefer to use characters that are distinctly different so that they stand out during proof-reading. The specific convention I use for common accents is the following:

1 - ´ 2 - ` 3 - ^ 4 - " 5 - ¸

Thus été would be encoded as e1te1.

Other areas where decisions must be made involve additions and deletions to the text. For the latter, I urge that one encode as complete a text as possible. Delete only what is absolutely unreasonable to include or what is 'accidental' to the text. For example, hyphens at the ends of lines are 'accidents' determined by the printer, not the author; one could delete the hyphen and complete the word without seriously altering one's reading of the text. Additions, such as part of speech or semantic category, are less serious in implication if reasonable care is exercised. So long as additions are marked in a systematic way (so that they could be eliminated algorithmically) they do not alter the integrity of the text; additions that cannot be distinguished from text are a different matter. A second consideration is efficiency. If the text is going to be processed (with a system such as RATS) and sorted at some time, it may be far easier to make additions during some later stage rather than when encoding. If, for example, one wishes to mark functors, one may do so by encoding only several hundred words and letting the computer mark all occurrences of these words. For a text of 100,000 words this is likely to be 200 words v. 50,000 words. Similar savings can be obtained in determining words that have the same root and differ by affix, in dealing with contractions, in delineating among homonyms, and numerous other problem areas. One should, thus, look closely at the entire computing process and adopt encoding additions on a token basis only as something of a last resort.

*Summary*

Above I have tried to look at some of the decisions that must be made in encoding a text and to examine some of the implications that these choices contain. The discussion is not complete, partially because it cannot be complete: there will always be unique features of a text or unique features of the research design that require unique, *ad hoc* solutions. But by going through the process of making a number of basic encoding decisions, the reader may be better able to confront these problems when they arise. As a further step in that direction I shall summarize below the major principles that underlie these decisions and the major areas where decisions must be made.

1. Integrity of the Text

The encoded text is the axiomatic basis on which all further analysis rests. It should obviously be the most authoritative text available. It should be encoded so that at some later time it can be recreated computationally in a form such that the reader's experience of it would not be materially different from his experience of reading the original.

2. Realization of the Impossibility of the Task

It is literally impossible to encode all of the information present in the physical text. Encoding is a series of compromises and 'trade-offs'; however, the encoder, guided by the principle of recreation stated above,

should strive for the fullest reasonable encoding to meet unanticipated future needs.

3. Systematic Perspective

The encoding conventions should be viewed as a system: simple, well-defined, and as unambiguous as possible. They should also be viewed as part of a larger computational system that will most probably modify and re-arrange the text. Care should be taken to add information where it is easiest to do so. Finally, one cannot get out of the far end of the system what one does not put into the system.

4. Major Areas of Decision

(a) Segmentation: linguistic and physical.

(b) Punctuation: syntactic and semantic.

(c) Fonts: base fonts (upper and lower), other fonts (italic, bold face, etc.), other symbol systems (Greek, Cyrillic, etc.).

(d) Accents.

(e) Deletions (as few as practical) and additions (included where easiest; marked so that they could be removed algorithmically).

I have not tried to promulgate any specific set of conventions; however, throughout the article I have indicated the set of choices that I have adopted in my own work. Should the reader decide that they meet his needs, they are summarized in Figure 2 in the form of the actual directions I give a typist for encoding.

*Notes*

1. F. de Tollenaere, 'Encoding Techniques in Dutch Historical Lexicography', *Computers and the Humanities*, 6 (1972), 147-52.

2. Sally Y. Sedelow, *Automated Analysis of Language Style and Structure: 1969-70* (Chapel Hill, 1970).

3. George A. Borden and James J. Watts, 'A Computerized Language Analysis System', *Computers and the Humanities*, 5 (1971), 129-42.

4. John B. Smith, 'RATS: A Middle Level Text Utility System', *Computers and the Humanities*, 6 (1972), 277-83.

5. This sentence is actually more complicated than it may appear. The quotation, which Joyce indicated by the initial dash, is a complete sentence embedded within the sentence. If the space is omitted before the question mark, the character string it? would be recognized as a word; if the space is left in, the sentence will be recognized as two sentences by the scan program. Either a complex convention must be established for this rather rare event or the scanned data set 'patched up' later. I would suggest the latter, where context indication can be altered rather easily on an *ad hoc* basis.

*Editor's Note*

The '#' symbol is referred to as 'hash' in Great Britain but 'pound' in America. It is therefore referred to as 'hash' in the main body of the article, but appears as 'pound' in Figure 2 since this is a copy of the actual instructions given to the typist in America. (The 'pound' sign in Great Britain is '£'.)

? ? ?

Lenehan said to all:

—Silence! What opera resembles a railway line? Reflect, ponder, excogitate, reply.

Stephen handed over the typed sheets, pointing to the title and signature.

—Who? the editor asked.

Bit torn off.

—Mr Garrett Deasy, Stephen said.

—That old pelters, the editor said. Who tore it? Was he (130) short taken.

*On swift sail flaming*
*From storm and south*
*He comes, pale vampire,*
*Mouth to my mouth.*

—Good day, Stephen, the professor said, coming to peer over their shoulders. Foot and mouth? Are you turned . . . ?

Bullockbefriending bard.

## SHINDY IN WELLKNOWN RESTAURANT

—Good day, sir, Stephen answered, blushing. The letter is not mine. Mr Garrett Deasy asked me to . . .

—O, I know him, Myles Crawford said, and knew his wife too. The bloodiest old tartar God ever made. By Jesus, she had the foot and mouth disease and no mistake! The night she threw the soup in the waiter's face in the Star and Garter. Ohol.

A woman brought sin into the world. For Helen, the runaway wife of Menelaus, ten years the Greeks. O'Rourke, prince of Breffni.

—Is he a widower? Stephen asked.

—Ay, a grass one, Myles Crawford said, his eye running down the typescript. Emperor's horses. Habsburg. An Irishman saved his life on the ramparts of Vienna. Don't you forget! Maximilian Karl O'Donnell, graf von Tirconnel in Ireland. Sent his heir over to make the king an Austrian fieldmarshal now. Going to be trouble there one day. Wild geese. O yes, every time. Don't you forget that!



—The moot point is did he forget it? J. J. O'Molloy said quietly, turning a horseshoe paperweight. Saving princes is a thank you job.

Professor MacHugh turned on him.

—And if not? he said.

—I'll tell you how it was, Myles Crawford began. Hungarian it was one day . . .

## LOST CAUSES
## NOBLE MARQUESS MENTIONED

—We were always loyal to lost causes, the professor said. Success for us is the death of the intellect and of the imagination. We were never loyal to the successful. We serve them. I (131) teach the blatant Latin language. I speak the tongue of a race the acme of whose mentality is the maxim: time is money. Material domination. *Dominus!* Lord! Where is the spirituality? Lord Jesus! Lord Salisbury. A sofa in a westend club. But the Greek!

Figure 1

## KYRIE ELEISON!

A smile of light brightened his darkrimmed eyes, lengthened his long lips.

—The Greek! he said again. *Kyrios!* Shining word! The vowels the Semite and the Saxon know not. *Kyrie!* The radiance of the intellect. I ought to profess Greek, the language of the mind. *Kyrie eleison!* The closetmaker and the cloacamaker will never be lords of our spirit. We are liege subjects of the catholic chivalry of Europe that foundered at Trafalgar and of the empire of the spirit, not an *imperium,* that went under with the Athenian fleets at Aegospotami. Yes, yes. They went under. Pyrrhus, misled by an oracle, made a last attempt to retrieve the fortunes of Greece. Loyal to a lost cause.

He strode away from them towards the window.

—They went forth to battle, Mr O'Madden Burke said greyly, but they always fell.

—Boohoo! Lenehan wept with a little noise. Owing to a

Figure 2

Standard Encoding Conventions:

1. Type one text line per R.J.E. line
complete hyphenated words at end of line

2. Blank delimit:
separate syntactic punctuation by blanks
do not separate semantic punctuation by blanks
ex:
... end of sentence . (syntactic)
Mr. Jones (semantic)

3. Mark end of paragraph by doubling the final, terminal punctuation.
ex:
... end of paragraph ..
for quotations, double the terminal punctuation within the quotation marks but not the quotation mark itself.
ex:
... end of paragraph and quotation .. "

4. Mark beginning of new chapter by
$$$ on separate line. Any chapter heading comes on line with $$$ preceded by a blank.

5. Mark end of volume by
$$$$ on separate line.

6. Place page number for first line of new page in columns 75-77. Right justify or precede by zero's. Leave the columns blank for other lines on the same page.

7. Italics typed in upper and lower case, as is, but with pound sign (#) attached to end of word.
ex:
Italics#

8. Lines centered on page, such as poetry, etc., type with visual offset (indent a few more spaces than for a paragraph) and put a c in col. 74.

9. Accents character, if any, should be followed by a numerical digit indicating the accent.
1 -- ´ as in été: e1te1
2 -- `
3 -- ^
4 -- ¨
5 -- ¸

10. For any problems not covered, jot down the page number and ask about them.

11. Other Special Conventions: