Text Editing, Print and the Digital World

Edited by
MARILYN DEEGAN
King's College London, UK
KATHRYN SUTHERLAND
University of Oxford, UK

ASHGATE

Published by
Ashgate Publishing Limited
Wey Court East
Union Road
Farnham
Surrey, GU9 7PT
England

Ashgate Publishing Company
Suite 420
101 Cherry Street
Burlington
VT 05401-4405
USA

www.ashgate.com

British Library Cataloguing in Publication Data
Text editing, print and the digital world. - (Digital research in the arts and humanities)
1. Criticism, Textual 2. Electronic books
I. Deegan, Marilyn II. Sutherland, Kathryn
801.9'59

Library of Congress Cataloging-in-Publication Data
Text editing, print, and the digital world / [edited] by Marilyn Deegan and Kathryn Sutherland.
p. cm. -- (Digital research in the arts and humanities)
Includes bibliographical references and index.
ISBN 978-0-7546-7307-1
1. Editing. 2. Criticism, Textual. 3. Transmission of texts. 4. Electronic publishing. I. Deegan, Marilyn. II. Sutherland, Kathryn.

PN162.T434 2008
808'.027--dc22

2008024692

ISBN 978 0 7546 7307 1

Mixed Sources
Product group from well-managed forests and other controlled sources
www.fsc.org Cert no. SA-COC-1565
© 1996 Forest Stewardship Council

Printed and bound in Great Britain by
MPG Books Ltd, Bodmin, Cornwall.

Chapter 3
Digital Editions and Text Processing

Dino Buzzetti

In a number of recent articles Peter Robinson[1] suggests that much of the continuing resistance of the scholarly community, both editors and readers, to supporting the adoption of digital editions can be attributed to the lack of user-friendly tools and manageable and easily available applications. The argument implies that many scholars are reluctant to devote too much time to becoming conversant with the sophisticated technologies required for the design and production of a complex and expedient digital object. This is certainly true, but in my opinion it is only one part of the truth. Another part of the truth can be elicited from the answer to this predicament that has been given, for instance, by Tito Orlandi. According to Orlandi, the main purpose of a digital edition is transferring the competence of the reader to the machine. But that goal is far from being achieved either in single editions or in general solutions. One might be tempted to say that present-day digital editions, for all their merits, are not yet fully digital, since they do not fully exploit the distinctive features of the digital form of textual representation to obtain better critical and analytical results. For a digital edition is in the first place a representation, a particular form of representation of textual information.

Recognizing this fact has two important consequences. A digital edition is, we may say, an 'image' of the text. Needless to say, the term 'image' is taken here in a figurative not a literal sense. I use the term deliberately, because it is tantamount to acknowledging the semiotic nature of the text. In fact, I take the notion of the 'image' of the text from Cesare Segre's *Introduction to the Analysis of the Literary Text*, together with a series of other basic assumptions, namely that 'the text does not have a material nature'; that 'the text is only' and 'always an image'; and that any attempt to identify it with a material witness whatsoever, even an autograph original, is 'an attempt to conceal [its] unavoidable problematic nature'.[2] For, indeed, 'the notion of an original', or of an 'autograph' for that

1 Cf. P. Robinson, 'Where We Are with Electronic Scholarly Editions, and Where We Want to Be', *Jahrbuch für Computerphilologie, 5 (2003)*: 125–46; P. Robinson, 'Current Issues in Making Digital Editions of Medieval Texts – or, do Electronic Scholarly Editions have a Future?', *Digital Medievalist*, 1/1 (2005). Available online at <http://www.digitalmedievalist.org/article.cfm?RecID=6>, accessed 19 June 2006.

2 C. Segre, *Avviamento all'analisi del testo letterario* (Torino: Einaudi, 1985); Engl. edn, *Introduction to the Analysis of the Literary Text*, trans. J. Meddemmen (Bloomington, IN: Indiana University Press, 1988), p. 378 and p. 376.

matter, taken 'in the sense of an authentic text, that expresses the author's will', is, in D'Arco Silvio Avalle's words, 'one of the most elusive and ambiguous notions of textual criticism'.[3] Any particular witness or any particular edition, even the most authoritative one, is only and always a representation of the text. It is in this sense, in a semiotic and not in a material sense, that a digital edition may be described as an 'image', or a manifestation of the text, that is, as that particular kind of text representation which is produced in digital form. It is primarily digital editions as representations, digital editions from a semiotic point of view, that I shall deal with in this chapter.

On the other hand, a digital representation is data, and data is processable. Data is the representation of information in a form that can be processed by a machine. And this is a point worth insisting upon; for the means of rendering a text – spoken, written, printed, digital – affords a different and distinctive approach to seizing it. In this respect, an 'image,' or representation of the text in digital form, can considerably enhance our opportunities of penetrating deeply into its discourse. But in fact, in present-day digital editions, all the foreseeable potentialities that a digital form of representation can afford are not yet fully exploited. In my opinion, the true rationale of a genuine digital edition consists precisely in taking advantage of the digital form of representation to improve our critical engagement with the text through effective computational processing. But how?

Let us start by considering some of the reasons why this goal does not seem to have been fully achieved as yet. The first reason lies with the already-mentioned reluctance of the humanities scholar to devote more than cursory attention to informatics and computer science. It is a real and in many respects justified resistance, but on the other hand, in a recent paper on markup systems that appeared in the *Jahrbuch für Computerphilologie*, Claus Huitfeldt regrets 'that many humanities scholars still regard markup as a product of computing technology and thus a purely technical tool of no interest or concern to humanities scholarship'. He makes a good point in recommending that 'textual scholars should not relate to markup technology as passive recipients of products from the computing industry, but rather should be actively involved in its development and in setting the agenda'. He aptly observes that 'the experience and expertise of textual scholars may turn out to be essential [...] as they possess insight which is essential to a successful shaping of digital text technology'.[4] This seems to be very much the case in the domain of digital editions, so that a decisive contribution to the advancement of their methods and design can be expected only from humanities scholars.

Another reason may be found in the persistence of conventional habits and practices in the production of digital editions. The edition continues to be seen chiefly as something for a human to read and only to a very limited extent for a

3 D'A.S. Avalle, *Principi di critica testuale* (Padova: Antenore, 1972), p. 33.

4 C. Huitfeldt, 'Scholarly Text Processing and Future Markup Systems', *Jahrbuch für Computerphilologie*, 5 (2003): 219–36. Available online at <http://www.computerphilologie.uni-muenchen.de/jg03/huitfeldt.html>, accessed 20 March 2006.

computer to process. Furthermore, transcription practices still seem to be carried out in a conventional frame of mind. In most cases descriptive markup practices consist of singling out detailed information and simply banking it for an otherwise undisclosed eventual conversion or some other remotely possible future use.

An edition in digital form, then, is considered here, alongside its conventional use, as textual data to be processed. From this point of view, our primary concern becomes that of extracting processable semantics from character data, by assigning to them a functional structure and a suitable formalism. For a humanities scholar, a reliable edition essentially serves the purpose of allowing interpretation. In this respect, visualization and string processing can be invaluable. Stylometric analysis and authorship attribution are based on processing of this kind, but apart from such special cases, the overall limitations of string manipulation are evident enough. Not much can be expected beyond what has already been obtained. The sort of results string processing can afford and their conceptual quality cannot substantially improve. As anyone can see, the Web provides an extraordinary visualization contrivance, but it does not meet the demands attended to by its content-aware counterpart, the Semantic Web. The chief concerns of text processing for interpretational purposes remain information retrieval, content management, or knowledge representation and extraction.

An example may be of use in clarifying the point. One might be tempted to equate a string of characters to the notion of a 'chain,' as defined by Louis Hjelmslev:

> The signs form *chains*, and the elements within each sign likewise form chains. We shall use the term *relation* for the function (dependence, relationship) between signs or between elements within one and the same chain: the signs, or the elements, are *related* to one another in the chain.[5]

But processing a string is not processing linguistic information, for a string of characters does not contain all the information that qualifies a chain as a linguistic unit. To that effect, representing the signs is not enough and all the functional relations between the elements of the chain should also be expressed explicitly. It is precisely that kind of information which makes up a linguistic unit out of a sequence of signs and if that information is not comprised in our textual representation, we cannot process texts as linguistic materials. Transferring the competence of the reader to the computer means processing textual data as linguistic units and implies representing all the relevant information. But the basic form of text representation as a simple string of characters does not explicitly embody it. This is a fairly obvious fact and the introduction of markup is the straight and direct answer to it. But there is a more subtle kind of confusion we should be wary of. A string of characters, or any other kind of data for that matter, is not information but an information carrier. To mistake a string for information actually means carrying

5 L. Hjelmslev, *Language: An Introduction* (Madison, University of Wisconsin Press, 1970), p. 32.

over into a digital environment the positivistic notion that identifies the text with a material object: the usual confusion between a witness and the text becomes the confusion between an information carrier and the information it conveys. A text has to be conceived, equally in a digital environment, as a semiotic system. A digital representation is a digital type of 'image' of the text, and a proper recognition of a digital edition as a digital representation of the text implies the recognition that text processing should be primarily concerned with the treatment of its information content and should not be mistaken for the processing of its information carrier. But, again, how might that aim be achieved?

In that respect, the problem of text representation becomes of paramount importance in order to qualify a digital edition as a properly digital one, namely as a kind of edition that makes allowance for effective processing of its information content. Now, as far as text representation is concerned, its standard form is provided by a marked-up string of characters. The introduction of 'structural markup', enabling as it does the move from simple, flat and unstructured text files to structured or semi-structured textual data, has been cheered as a decisive step 'to distinguish information and knowledge from mere "digital stuff"'.[6] One of the main motives for introducing descriptive markup has actually been the idea of processing textual data as structured information. But the mere processing of marked-up character data still falls short of adequate treatment.

Markup has been described as a 'technique for representing structure',[7] for it consists in 'the denotation of specific positions'[8] within a linear sequence of characters. But, again, we should not mistake the structure of the string for the structure of the text. Quite consistently, the ordinary notion of a text, that in a book on text processing is considered as 'literary material as originally written by an author', is clearly kept apart from the computer scientist's notion of the text, defined as 'information coded as characters or sequences of characters'.[9] For, as Michael Sperberg-McQueen has contended, 'claiming that the only essential part of a text is its sequence of graphemes' is indeed 'a misguided and inadequate theory of texts'.[10] But when it comes to its digital representation, we still find considerable uncertainty hovering around the notion of text.

---

6 R. Cover, N. Duncan and D.T. Barnard, 'The Progress of SGML (Standard Generalized Markup Language): Extracts from a Comprehensive Bibliography', *Literary and Linguistic Computing*, 6 (1981): 197–209, pp.197–8.

7 D.R. Raymond, F.W. Tompa and D. Wood, 'Markup Reconsidered', paper presented at the First International Workshop on Principles of Document Processing, Washington DC, 22–23 October 1992, Abstract available online at <http://softbase.uwaterloo.ca/~drraymon/papers/markup.ps>, accessed 14 June 2006.

8 Ibid., p. 4.

9 A.C. Day, *Text Processing* (Cambridge: Cambridge University Press, 1984), pp. 1–2.

10 C.M. Sperberg-McQueen, 'Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts', *Literary and Linguistic Computing*, 6 (1991): 34–46, p. 35.

A lack of clarity on the semiotic nature of the digital representation of the text can be found in the very definition of markup[11] proposed by the editors of the Text Encoding Initiative (TEI),[12] the ground-breaking scholarly enterprise whose aim is 'to provide a format for data interchange in humanities research' and 'to suggest principles for the encoding of texts in the same format'.[13] If markup is defined, as it has been by the editors of the TEI *Guidelines*, as 'all the information contained in a computer *file* other than the text itself', how can it be maintained at the same time that '*any* aspect of the text of importance to a researcher' could 'be signalled by markup'?[14] For either markup is thought to be information that '*is not* part of the text'[15] and is *different* from text – and in that case the text is identified with the string of characters representing it – or markup is understood as expressing certain aspects of that information which '*is* part of the text, and is *the same as* text'[16] – and in that case the text is identified with the information content expressed by that string of characters. To overlook that difference is to overlook Hjelmslev's distinction between the 'expression' and the 'content' of a text,[17] and to ignore that 'the representation of any information content is not the information content that is represented by that representation'.[18] Clearly the structure of the representation is not the structure of what is represented. By singling out definite positions in a stream of characters, the markup assigns a structure to the expression of the text that does not necessarily coincide with the structure of its content.

---

11 L. Burnard and C.M. Sperberg-McQueen, *Living with the Guidelines: An Introduction to TEI Tagging*, Text Encoding Initiative, Document Number: TEI EDW18, 13 March 1991, p. 2.

12 Cf. D. Buzzetti and M. Rehbein, 'Textual Fluidity and Digital Editions,' in M. Dobreva (ed.), *Text Variety in the Witnesses of Medieval Texts*, Proceedings of the International Workshop (Sofia, 21–23 September 1997) (Sofia: Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences, 1998), pp. 14–39; and D. Buzzetti, 'Digital Representation and the Text Model', *New Literary History*, 33/1 (2002): 61–87.

13 L. Burnard, 'An Introduction to the Text Encoding Initiative', in D. Greenstein (ed.), *Modelling Historical Data* (St. Katharinen: Max-Plank-Institut für Geschichte i.K.b. Scripta Mercaturae Verlag, 1991), p. 83.

14 Burnard and Sperberg-McQueen, *Living with the Guidelines*, p. 2 (my italics).

15 J.H. Coombs, A.H. Renear and S.J. DeRose, 'Markup Systems and the Future of Scholarly Text Processing', *Communications of the ACM*, 30 (1987): 933–47, p. 934 (my italics).

16 Buzzetti and Rehbein, 'Textual Fluidity', p. 31.

17 Cf. L. Hjelmslev, *Prolegomena to a Theory of Language*, trans. F.J. Whitfield (Madison: University of Wisconsin Press, 1961), pp. 47–70.

18 D. Buzzetti, 'Text Representation and Textual Models', in *ACH-ALLC '99 Conference Proceedings* (Charlottesville VA: University of Virginia, 1999). Available online at <http://www.iath.virginia.edu/ach-allc.99/proceedings/buzzetti.html>, accessed 14 June 2006.

In a lucid and incisive paper, John Unsworth describes humanities computing essentially as 'a practice of representation',[19] and digital textual editing is certainly no exception. So in order to establish 'the first elements of a theory of the digital text', we actually need 'to outline a semiotics of its machine-readable representation'.[20] If we consider the text as a semiotic system, there is a fundamental feature that cannot be disregarded. From a semiotic point of view the text is intrinsically and primarily an indeterminate system. To put it briefly, there are many ways of expressing the same content just as there are many ways of assigning content to the same expression. Synonymy and polysemy are two well-known and mutually related linguistic phenomena. This brings us to the conclusion, to express it in Jerome McGann's words, that 'no text is self-identical'.[21] And actually, as he observes, 'variation, in other words, is the invariant rule of the textual condition'.[22] The 'material and the conceptual "text"' are then characterized by a 'radical instability,' for 'the freedom of the reader' leads to 'interpretive differentials' and 'the ordering of the words in every text is *in fact*, at the factive level, unstable'.[23]

But such a circumstance, which is essential to the textual condition, does not compromise 'the possibility of a reliable knowledge of text'.[24] For text can also be described as an 'immutable mobile' form of representation, just as the maps and records so designed by Bruno Latour.[25] In a representation of that kind, either of its two components, the material expression or the conceptual content, may be assumed as an invariant and a mark of identity, while the other varies in an indefinite number of ways. Indetermination is thus a structural feature of textual semiotics and its basic components are related as two correlative variables in a quantum physical environment: if you determine and fix either of them with observational precision, the other remains by necessity uncertain and indeterminate. The identity of the text can then be traced in the specific form of the mutual relationship between its variation and invariance, and established in the co-dependent series of its transient and unstable manifestations.

---

19 J. Unsworth, 'What is Humanities Computing and What is Not?', in G. Braungart, K. Eibl and F. Jannidis (eds), *Jahrbuch für Computerphilologie, 4 (2002)*. Available online at <http://computerphilologie.uni-muenchen.de/jg02/unsworth.html>, accessed 14 June 2006.

20 D. Buzzetti, 'Diacritical Ambiguity and Markup,' in D. Buzzetti, G. Pancaldi and H. Short (eds), *Augmenting Comprehension: Digital Tools and the History of Ideas* (London: Office for Humanities Communication, 2004), p. 178.

21 J.J. McGann, 'Rethinking Textuality', <http://www.iath.virginia.edu/~jjm2f/old/jj2000aweb.html>, accessed 15 June 2006. For a more thorough discussion, see J.J. McGann., *Radiant Textuality: Literature after the World Wide Web* (New York: Palgrave, 2001), especially ch. 5 and the Appendix to ch. 6.

22 J.J. McGann, *The Textual Condition* (Princeton, NJ: Princeton University Press, 1991), p. 185.

23 Ibid.

24 Ibid.

25 Cf. B. Latour, 'Visualization and Cognition: Thinking with Eyes and Hands', in *Knowledge and Society: Studies in the Sociology of Culture Past and Present*, 6 (1986): 1–40.

It is important, therefore, not only to recognize the presence of the two textual components, expression and content, and not to mistake the one for the other, but also to consider the way they are mutually related. This is the core of the problem. The dominant paradigm has been that of formalization, that is translating natural language expressions into symbolic language and formulas. This was to ensure a thorough correspondence and isomorphism between the syntactic and the semantic form of linguistic expressions. Grammatical form had to be reduced to strict logical form. As Donald Davidson put it, 'to give the logical form of a sentence is [...] to describe it in terms that bring it within the scope of a semantic theory that meets clear requirements'.[26] The whole programme of artificial intelligence was committed to that ideal: according to John Haugeland's reassuring recommendation, 'If you take care of the syntax, *the semantics will take care of itself*.'[27]

But the coincidence between the syntactic and the semantic structure of textual data is by no means a condition for their automatic processing. Their disparity is, on the contrary, what essentially characterizes the textual condition. A symbolic formula is indeed a textual expression, but it constitutes only a very special case. The burden of being absolutely univocal would deprive textuality of the richness and flexibility which are continually produced by its indetermination: a symbolic calculus generates textuality only in a very impoverished and diminutive sense. In any case, textual data can be processed despite discrepancy between the syntactical structure assigned to a string of characters and the semantic model assigned to its information content. The problem is precisely how to relate the two kinds of structure in a consistent and productive way.

The discrepancy between the 'form of the expression' and the 'form of the content,' to appeal again to Hjelmslev's distinctions, is an intrinsic feature of the alphabetic form of textual representation. The ancient Stoics had already noticed that the word 'dog' does not bite, and obviously we cannot assume that the word 'ball' be round, for even in a *calligramme*, or an ideogram for that matter, it would not be three-dimensional. Incidentally, it may be observed that in an alphabetic text we read the words, and that in an ideographic text we read the meanings. More precisely, in an alphabetic text we read the meanings through the words, and in an ideographic text we read the words through the meanings. We cannot completely separate these two related aspects, for they actually concur in the concrete working of spoken and written discourse to shape the relationship between words and meanings, expression and content, the syntactic and semantic structure of the textual materials. Again, that relationship has to be properly recognized and the problem of a digital representation of the text is precisely how to implement its dynamics in a properly exhaustive and functional way.

---

26 D. Davidson, *Essays on Actions and Events* (Oxford: Oxford University Press, 1980), p. 144.

27 J. Haugeland, 'Semantic Engines: An Introduction to Mind Design', in J. Haugeland (ed.), *Mind Design: Philosophy, Psychology, Artificial Intelligence* (Montgomery VT: Bradford Books, 1981), p. 23.

It has to be observed that it is the tacit assumption of isomorphism between the syntactic and the semantic structure of the text that lies behind the assertion of the controversial OHCO model, which presumes to define 'what text really is' as 'an "ordered hierarchy of content objects" or "OHCO"'.[28] This definition relies on 'the use of SGML [Standard Generalized Markup Language] as a basic text description language'[29] and assumes that the structure assigned by markup to a string of characters is *tout court* the structure of the text. In fact, 'SGML defines a document in terms of its OHCO structure',[30] for indeed:

> one fundamental premise of SGML is that texts are composed of discrete content objects, and that supplying meaningful names for these delimited textual objects, their attributes and their hierarchical relationships independent of possible appearances is one of the most powerful means of transforming text into information units that may be addressed sensibly by knowledgeable software.[31]

One should note that here 'text' is equated throughout with 'document' or 'string of characters', and that the OHCO definition consequently identifies the 'content structure'[32] of a document with the 'hierarchical structure'[33] assigned by the SGML scheme to its representation as a sequence of characters. However, it is only on the assumption of a complete isomorphism between the syntactic and the semantic form that we may identify the structure of the expression with the structure of the content of a text.

Markup languages such as SGML and XML (eXtensible Markup Language) are data representation languages: they serve to specify data structures, but they do not provide a data model or formalism to process their information content. 'SGML explicitly declaims semantics'[34] and XML, 'just like its parent metalanguage (SGML)', 'formally governs syntax only – not semantics.'[35] It is no surprise, then, that 'the current approach for "fixing the meaning" of a data exchange/archival format', which 'is to provide an XML DTD', should be found unsatisfactory, for a DTD (Document Type Definition) is nothing but a formal

---

28 S.J. DeRose, D.D. Durand, E. Mylonas and A.H. Renear, 'What is Text, Really?', *Journal of Computing in Higher Education*, 1/2 (1990): 3–26, p. 6.

29 Ibid., p. 13.

30 Ibid., p. 12.

31 Cover et al., 'The Progress of SGML', p. 198.

32 'What is Text, Really?', p. 23.

33 Ibid., p. 22.

34 D. Raymond, F. Tompa and D. Wood, 'From Data Representation to Data Model: Meta-semantic Issues in the Evolution of SGML', *Computer Standards and Interfaces*, 10 (1995): 25–36, <http://softbase.uwaterloo.ca/~drraymon/papers/sgml.ps>, p. 2, accessed 16 June 2006; cf. C.F. Goldfarb, *The SGML Handbook* (Oxford, Oxford University Press, 1990).

35 R. Cover, 'XML and Semantic Transparency' <http://www.oasis-open.org/cover/xmlAndSemantics.html>, accessed 16 June 2006.

specification of the hierarchical structure assigned by SGML and XML to a stream of characters. 'Many communities and organizations define their own standard "community language" via DTDs', and the endorsement of the TEI DTD by the humanities computing community is no exception. But as the NARA (National Archives and Records Administration) project for the long-term preservation of digital data developed by the San Diego Supercomputer Center (SDSC) openly acknowledges, the 'shortcomings of DTDs for data modeling and validation have been widely recognized'.[36] So, the realization that 'although SGML/XML-based markup languages provide explicit rules for syntactic well-formedness and validity, they provide nothing analogous for semantic correctness',[37] or that an XML DTD cannot provide a suitable semantic model to process the information content of textual data, has led 'to a flood of extensions'[38] and attempts, such as 'XML Schema, RDF, the Semantic Web',[39] to develop a 'semantics of document markup'.[40]

The most consistent and systematic endeavour of this kind has been 'the BECHAMEL Markup Semantics project, which is developing a formal framework for the interpretation of markup, and a multi-layer knowledge representation and inferencing environment (in Prolog) with which to express theories of markup semantics'.[41] The BECHAMEL system comprises a syntactic layer to describe 'SGML/XML syntactic relations' between 'string representations';[42] an object modelling layer to describe 'the distinctive semantic relationships that they represent';[43] and a mapping layer, which 'concerns mechanisms and rules for mapping syntactic markup structures to instances of objects, properties, and

---

36 B. Ludäscher, R. Marciano and R. Moore, 'Preservation of Digital Data with Self-validating, Self-instantiating Knowledge-based Archives', *ACM SIGMOD Record*, 30/3 (2001): 54–63, p. 59. Available online at <http://users.sdsc.edu/~ludaesch/Paper/kba.pdf>, accessed 17 June 2006.

37 C.M. Sperberg-McQueen, D. Dubin, C. Huitfeldt and A. Renear, 'Drawing Inferences on the Basis of Markup', in B.T. Usdin and S.R. Newcomb (eds), *Proceedings of the Extreme Markup Languages 2002 Conference* (Montreal, Quebec, 2002), p. 1. Available online at <http://www.mulberrytech.com/Extreme/Proceedings/xslfo-pdf/2002/CMSMcQ01/EML2002CMSMcQ01.pdf>, accessed 17 June 2006.

38 Ludäscher et al., 'Preservation of Digital Data', p. 59.

39 A. Renear, D. Dubin, C.M. Sperberg-McQueen and C. Huitfeldt, 'Towards a Semantics for XML Markup', in R. Furuta, J.I. Maletic and E. Munson (eds), *Proceedings of the 2002 ACM Symposium on Document Engineering*, McLean VA, November 2002 (New York: ACM Press, 2002), p. 119.

40 Ibid., p. 123.

41 D. Dubin and D. Birnbaum, 'Interpretation Beyond Markup', in B.T. Usdin (ed.), *Proceedings of the Extreme Markup Languages 2004 Conference* (Montreal, Quebec, 2004), p. 2. Available online at <http://www.mulberrytech.com/Extreme/Proceedings/xslfo-pdf/2004/Dubin01/EML2004Dubin01.pdf>, accessed 17 June 2006.

42 Ibid., p. 6.

43 Ibid., p. 9.

relations'.[44] The identification of relevant syntactic and semantic textual structures and the invention of a bridging mechanism between them can be conceived as the basic and essential tasks of digital textual editing and processing. But it is even more important to recognize that they cannot be univocally envisaged and designed lest we should disregard text indetermination and mobility.

In a paper related to 'an ongoing project to develop a system for extracting rhyme schemes from Russian verse',[45] which illustrates 'how BECHAMEL can be used in a synthesis of different semantic models'[46] across different semantic domains, David Dubin and David Birnbaum clearly acknowledge the indeterminate relationship between the syntactic expression and the semantic content of the text, for they assume 'that the same markup can convey different meanings in different contexts' and 'that markup can communicate the same meaning in different ways using very different syntax'.[47] The recognition of this fundamental principle entails substantial consequences about markup and the role it plays in the digital representation of text:

> We don't intend to suggest that markup presents any more of an inherent problem than other methods of encoding and representing information. All the distinctions that we're able to explicate using BECHAMEL could, in principle, guide the re-tagging of documents with richer markup that would eliminate the need for an inferential step in each case. Or BECHAMEL's network of properties and relations could be serialized in the form of RDF or a topic map.[48]

In other words, the role performed by the BECHAMEL 'mapping rules'[49] could be performed either by appropriate markup at a syntactic level or by a suitable data model at a semantic one. And this leads us to our central and concluding point, for markup within a text can really be construed as a rule. To expound this assertion properly, some considerations 'on the status of markup in relation to the bibliographically coded text'[50] are given here.

---

44 D. Dubin, 'Object Mapping for Markup Semantics', in B.T. Usdin (ed.), *Proceedings of the Extreme Markup Languages 2003 Conference* (Montreal, Quebec, 2003), p. 2. Available online at <http://www.mulberrytech.com/Extreme/Proceedings/xslfo-pdf/2003/Dubin01/EML2003Dubin01.pdf>, accessed 17 June 2006.

45 Dubin and Birnbaum, 'Interpretation Beyond Markup,' p. 4.

46 Ibid., p. 2.

47 Ibid., p. 1.

48 Ibid., p. 8.

49 Ibid., p. 6.

50 D. Buzzetti and J. McGann, 'Electronic Textual Editing: Critical Editing in a Digital Horizon', in L. Burnard. K. O'Brien O'Keeffe and J. Unsworth (eds), *Electronic Textual Editing* (New York: Modern Language Association of America, 2006), p. 59.

The 'function of markup' has actually been described as that of expressing 'inference rules' about the properties it labels in a text.[51] Accordingly, as it has also been convincingly maintained, 'to describe the meaning of the markup in a document, it suffices to generate the set of inferences about the document which are licensed by the markup', and 'in some ways, we can regard the meaning of the markup as being constituted, not only described, by that set of inferences'.[52] A markup expression, then, can be regarded as an inference-licence or an 'inference-ticket', to use Gilbert Ryle's famous description, a statement which licenses us 'to move from asserting factual statements to asserting other factual statements'[53] – in our case – about 'passages in the marked-up material'.[54] But, as Ryle reminds us, inference licences and rule-statements 'belong to a different and more sophisticated level of discourse from that [...] to which belong the statements of the facts that satisfy them'.[55] So the linguistic status of markup expressions understood as rules, or inference licences, is different from the linguistic status of markup expressions construed as factual statements 'about the structure and properties of the text'.[56]

This ambivalence of markup expressions should not worry us, for as we shall see, it constitutes an essential and functional property of markup. What we may already observe, though, is that markup expressions can be construed either as the mark of a rule, or as the mark of an observed textual property – or to use a more formal language, either as the sign of an operation, or as the sign of an operation value or result. This means that markup can have both 'descriptive' and 'performative' force – or mood, to choose Searle's instead of Austin's phraseology – and as Allen Renear discerningly points out 'the recognition that markup has modality and that some of it is performative, constitutive of the text it characterizes' raises fundamental problems 'about just what markup really is, and in particular, when it is *about* a text and when it is *part* of a text ... and when, and how, it may sometimes be both'.[57] Thus, from the ambiguity between values and rules we are led to another kind of ambivalence of markup expressions, that between

51 C.M. Sperberg-McQueen, C. Huitfeldt and A. Renear, 'Meaning and Interpretation of Markup', *Markup Languages*, 2/3 (2000): 215–34, pp. 215 and 218.

52 Ibid., p. 231.

53 G. Ryle, *The Concept of Mind* ([1949]; 2nd edn, Harmondsworth: Penguin Books, 1963), p. 117.

54 Sperberg-McQueen et al., 'Meaning and Interpretation of Markup', p. 215.

55 Ryle, *The Concept of Mind*, p. 116.

56 Sperberg-McQueen et al., 'Meaning and Interpretation of Markup', p. 216.

57 A. Renear, 'The Descriptive/Procedural Distinction is Flawed', *Markup Languages*, 2/4 (2001): 411–20, p. 419. Cf. J.L. Austin, *How to Do Things with Words*, ed. J.O. Urmson, (Oxford: Clarendon Press, 1962); J.R. Searle, *Speech Acts: An Essay in the Philosophy of Language* (Cambridge: Cambridge University Press, 1969); and J.R. Searle, 'A Taxonomy of Illocutionary Acts', in K. Gunderson (ed.), *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, 7 (Minneapolis: University of Minnesota Press, 1975), pp. 344–69, repr. in J.R. Searle, *Experience and Meaning: Studies in the Theory of Speech Acts* (Cambridge: Cambridge University Press, 1975), pp. 1–29.

expressions that are part of the text and expressions that are not – or in more formal terms, between object-language and metalinguistic expressions.

What that means in fact, is that markup may be considered either as belonging to text, 'as an extension of the expressive resources of the [...] very language constituting the text', or as an external description of the structure and properties of the text, i.e. 'as a form of metalinguistic notation'.[58] For, on the one hand, markup makes explicit some implicit feature of the text and 'as soon as something has been made explicit it has become part of the text, which has thereby changed, and acquired a new structure'. As Claus Huitfeldt aptly remarks, 'there is a similarity here to Wittgenstein's distinction in the *Tractatus* between showing and saying – the structure of the text shows itself in the text'.[59] As Wittgenstein puts it, 'that which mirrors itself in language, language cannot represent';[60] structure and logical form can only show themselves within the object-language and 'the only way to *represent* a logical form is to describe it by means of a metalanguage'. But that is also, on the other hand, what markup can actually do. So markup can both exhibit and describe a structural feature of the text, and 'it can perform both functions only by changing its logical status' and commuting between object-language and metalanguage.[61] Here is another case of markup ambiguity and we should not find it particularly disconcerting. As a matter of fact, it has been observed that markup 'is part of the text and yet it is distinct' from it,[62] and, to the same effect, markup has been described both as an external 'technique for representing structure'[63] and as that very 'structure'[64] within the text. We have to acknowledge, again, that this kind of ambiguity is another essential and functional property of markup.

All these seemingly inconsistent aspects of markup are grounded in a fundamental property of markup expressions that can be described as *diacritical*. Diacritics are special notation signs that are 'used to distinguish different sounds or values of the same letter or character',[65] and they may be considered as an expression of the 'reflexive metalinguistic nature'[66] of natural language, the capability that all natural languages possess of saying something about themselves. Whenever the notational form is ambiguous, they explicitly flag out a discriminating mark, that exhibits self-reflexively what can only be shown and cannot be said in the object-

58 Buzzetti, 'Diacritical Ambiguity and Markup', p. 178.

59 C. Huitfeldt, 'Multi-dimensional Texts in One-dimensional Medium', *Computers and the Humanities*, 28 (1995): 235–41, pp. 237–8.

60 L. Wittgenstein, *Tractatus Logico-philosophicus*, 4.121; see also 4.1212: 'What can be shown cannot be said.'

61 Buzzetti and McGann, 'Critical Editing in a Digital Horizon', p. 63.

62 Raymond et al., 'Markup Reconsidered', p. 3.

63 Ibid., p. 3.

64 Ibid., p. 4.

65 *Oxford English Dictionary*, s.v.

66 Cf. T. De Mauro, *Minisemantica dei linguaggi non verbali e delle lingue* (Bari: Laterza, 1982), pp. 93–4, and T. De Mauro, *Prima lezione sul linguaggio* (Bari: Laterza, 2002), pp. 89 and 91–93.

language itself. But they have themselves a 'reflexive metalinguistic nature', or a self-describing metalinguistic force, and can be converted into equivalent expressions of a genuinely metalinguistic kind. In formal terms, it can be shown that second-order object-language statements – statements based on a second-order form of predication – are equivalent to first-order metalinguistic statements – statements based on an ordinary form of predication. The notational ambiguity of markup, incapable as it is of distinguishing between the two different forms, does not prevent it from being duly construed in either of these two absolutely acceptable ways.

All this means that we can explain McGann's assertion that 'no text is self-identical'[67] through the fact that text is endowed with self-reflexive functions that allow us to discriminate and switch between different values assigned to notationally identical characters and character strings. No text is self-identical just because it is self-reflexive. To put it formally, we may say that the non-identity of the text with itself is logically equivalent to an endomorphism, or a mapping, of the text onto itself – in formula:

$$(A = A \text{ if and only if } A \neq A) \leftrightarrow A \xrightarrow{\textit{markup}} A^{68}$$

The markup can then be construed as the expression of a self-reflexive function that relates the semiotic elements of the text to other semiotic elements of the same text.

We can try to illustrate all these phenomena by means of a diagram (Figure 3.1). As we have seen, structural distinctions can refer either to the *expression* or to the *content* of the text. A markup expression can also be understood both as a *rule*, or an operation, and as a *value*, or the result of an operation. Moreover, the markup can be considered as belonging to the *object-language* of the text, or to a *metalanguage* describing it. And all these dimensions can be displayed in the diagram. To round it off, we can recall that markup structures can be 'embedded' and 'present in the data',[69] or apart as 'non-embedded structure'; accordingly, we can distinguish between *internal* structure or markup, and *external* structure or markup – 'so-called *out-of-line* markup'.[70] These two dimensions are also displayed in our diagram to complete its multi-dimensional space.

We may now try to expound the use of the diagram to represent textual mobility and build a dynamic model of textual instability. Our purpose is to show that the fundamental ambiguity of markup, due to its self-reflexive and diacritical nature, far from being 'an obstacle to an automatic processing of textual information' can actually serve that very purpose, by providing a 'formal representation of textual

67 Cf. note 21 above.

68 Buzzetti, 'Digital Representation and the Text Model', p. 84.

69 Raymond et al., 'Markup Reconsidered', p. 3.

70 Ibid., p. 4.

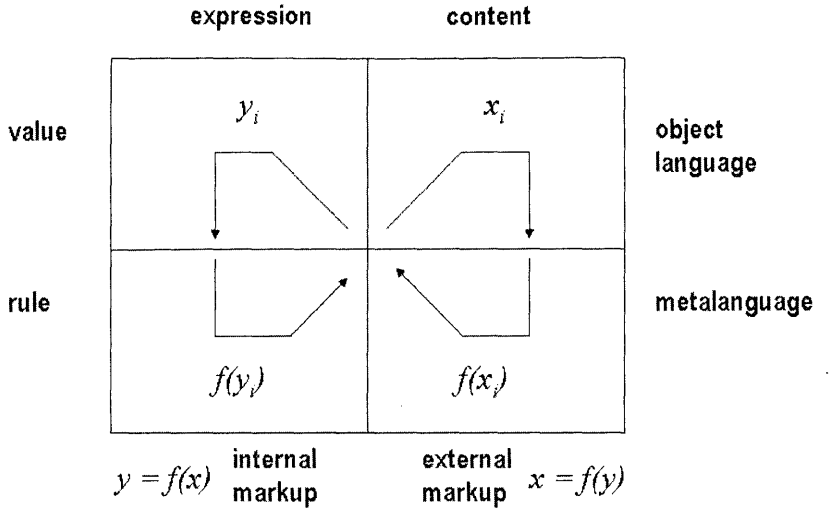

Figure 3.1 The multiple dimensions of markup

dynamics'.[71] We can refer, by way of example, to that traditional form of markup, or diacritical notation, that is represented by punctuation. For 'punctuation is not simply part of our writing system; it is a type of document markup that may vary and be replaced by other types of markup';[72] and therefore what can be said about punctuation can also be said about markup.

What a modern scholar does, for instance, by adding punctuation to a medieval text, is to mark it up. Let y_i be the result of a marking operation of this kind, which consists in inserting a punctuation mark in a certain position within the string of characters that represents the text. Our mark y_i is the *value* of the marking operation and a piece of *internal markup*. But as soon as it is drawn, it becomes part of the *expression* of the text, and a construct of its *object-language*. So we can place it in the upper left sector of our diagram. As a new component of the expression of the text it constitutes a textual variant.

But our mark can also be construed as a *rule* for mapping the syntactical phrase it belongs to onto the object domain of what it signifies. The mapping operation is a function of y_i and we may indicate it by $f(y_i)$. Considered as a function, our mark changes its linguistic status and becomes a higher-order statement in the object-language. Such a statement has metalinguistic import and can be converted into its logically equivalent statement expressed in a proper *metalanguage*. We can thus locate it in the lower left quadrant of our diagram.

71 Buzzetti and McGann, 'Critical Editing in a Digital Horizon', p. 64.

72 Coombs et al., 'Markup Systems and the Future of Scholarly Text Processing', p. 935.

The mapping operation $f(y_i)$ imposed by the markup expression understood as a rule assigns a structure to the content of the text. Let $x_i = f(y_i)$ be the result of that mapping. As a *value* of the mapping operation, x_i designates a structural element of the object domain signified by the text and constitutes a component of the external structure described by the text. Consequently it does not belong any more to the expression or to the internal structure of the text and it is not a part of it. It is rather to be seen as a piece of stand-off or *external markup*. As a structure assignment onto the object domain signified by the text it constitutes an interpretational variant of its *content*. It can also be denoted by an expression of a data modelling language (e.g. ERA, RDF, UML, etc.), which assigns a suitable formal model to the content of the text and which provides an *object-language* for the description of its components. Our external structure component x_i is then to be positioned in the upper right quadrant of the diagram.

The structural construct x_i that belongs to the *content* of the text can in its turn be understood as a *rule*, or as an instruction for a structuring or mapping operation onto the expression of the text. We may denote this function as $y_i = f(x_i)$. Again, understood as a function, the expression of the external structural component x_i changes its linguistic status and becomes a second-order statement in the external markup object-language, or in the data modelling language used to describe the object domain signified by the text. Such an expression, assuming as it does x_i as its argument, has a metalinguistic force and can be converted into its equivalent construct expressed in a suitable *metalanguage* referring to the text. Accordingly, we can place the functional expression $y_i = f(x_i)$ in the lower left section of the diagram. The functional expression $f(x_i)$ has y_i as its value and so the cycle is complete. External markup constructs x_i that depend on the mapping of internal markup constructs y_i can be mapped back onto internal markup constructs y_i.

The *cycle* we have described can be seen as an unfolding of textual instability and indetermination. We have described the indetermination of the text as a non-identity relation of the text with itself and this relation has been analysed into the mapping of components of the text onto other components. The non-identity relation of the text with itself is a constrained one. We can define the structure of the text as 'the set of latent relations'[73] between all its virtual parts and components, and if we conceive the structure of the text as 'the whole of all its possible variant readings and interpretations', we can also think of the text as 'a virtual unity identical with itself'.[74] Within this virtual unity we can draw a primary distinction between the expression and the content of the text and 'the variation of either component is dependent upon the invariance of its related counterpart'.[75] So the relation of the text with itself is a one-to-many-relationship and the mapping can have more than one value.

73 Segre, *Avviamento all'analisi del testo letterario*, p. 44; Engl. trans., p. 34.

74 Buzzetti and McGann, 'Critical Editing in a Digital Horizon,' p. 64.

75 Ibid., p. 64.

The mapping of the text onto itself can be performed by markup that gives explicit expression to implicit structural features of the text. But markup is ambiguous and can be both internal and external to the text. So markup can transform textual variants into interpretative variants and vice versa: as it happens, it can map structural components of the expression onto structural components of the content, or it can map structural constructs of the content onto structural constructs of the expression. Since the linguistic status of markup can commute between logically equivalent forms, the inversion of the mapping can be explicitly expressed and markup can provide an explicit representation of compensation constraints. Textual inner mobility can then be described as 'a permanent cycle of compensating actions between determination and indetermination of the expression and the content of the text'.[76]

What I have tried to do so far is to sketch a structural scheme for a formal representation of the semiotic working of a digital representation of a text. There may undoubtedly be better ways of doing it, but my purpose was simply to give an idea of the complexities and extent of the phenomena that must be analysed. For it is unlikely that we may ensure a suitable and efficient processing of textual information without relying on a functional and exhaustive representation of all its semiotic components and constraints. An edition can be said to be properly digital only if it affords that kind of processing. In other words, the major task seems to be that of improving the adequacy of the basic form of text representation.

The fact that texts are currently represented in computer systems as 'linear strings of atomic characters' is 'usually accepted as an immutable fact of technology';[77] but as we have seen there are reasons to think that this need not be so. Both the basic form of text representation and the current markup schemes do not afford suitable means to process a string of characters as a proper linguistic unit and so cannot ensure that the linguistic competence of the reader can be transferred to a computer. But alternative means have been proposed.

The notion of an enhanced or 'extended string' as a new data type for processing textual information was introduced some years ago by Manfred Thaller[78] and it stems from the realization that a string of characters 'is a representation of an underlying meaning with a specific information density' and that it is usually able to 'transfer only part of the meaning originally available'.[79] The extension of the data type was based on the principle that a string made up of 'information carrying

76 Ibid., p. 66.

77 M. Thaller, 'Strings, Texts and Meaning', in *Digital Humanities 2006*, 1st AHDO International Conference Abstracts (Paris: Université Paris-Sorbonne – Centre de Recherche Cultures Anglophones et Technologies de l'Information, 2006), p. 212.

78 Cf. M. Thaller, 'Text as a Data Type', in *ALLC-ACH'96: Conference Abstracts* (Bergen: University of Bergen, 1996), pp. 252–54. Available online at <http://gandalf.aksis.uib.no/allc/thaller.pdf>, accessed 7 August 2007.

79 Thaller, 'Strings, Texts and Meaning', p. 213.

tokens' can 'be understood to exist in an *n*-dimensional conceptual universe'[80] and that 'low level programming tools' currently used to develop text-handling applications can directly 'tackle the implications of this model'[81] – for example, by allowing the application to toggle between different ways of processing a string perhaps by sensitivity or insensitivity to a given textual feature.

The application of the 'extended string' data type 'to text critical problems' has 'proved to be a substantial step towards reaching satisfactory solutions', and 'its application to problems of analysis and interpretation looks just as promising on the same grounds'.[82] Moreover, the extension of the basic string representation from a one-dimensional to a many-dimensional case 'in turn leads to the notion, that not only the handling of information carrying tokens can be generalized', but that 'the properties of markup languages can as well'.[83] Thus, in principle, all the injunctive and performative force of markup could be transferred to effective processing and be of use in enhancing the functionality of text representation and digital editions.

The examination and testing of these new possibilities opens up a new, promising direction for research, in the conviction that only an improved form of low-level text representation can allow semantic and content-based text processing and afford an effective transfer of linguistic competence from the human reader to the machine. Without this decisive step, I fear that digital editions are doomed to fail in gaining generalized support or a discerning preference among textual scholars, and they will probably continue to lag behind in comparison or confrontation with the conventional form of text representation and transmission.

80 Ibid., p. 212.

81 Ibid., p. 213.

82 D. Buzzetti, 'Digital Editions: Variant Readings and Interpretations', in *ALLC-ACH'96: Conference Abstracts* (Bergen, University of Bergen, 1996), p. 256. Available online at <http://gandalf.aksis.uib.no/allc/thaller.pdf>, accessed 7 August 2007.

83 Thaller, 'Strings, Texts and Meaning', p. 213.