Digital Text Representation
Expression and Content

Dino Buzzetti
University of Bologna

Processable data and text representation

What is a digital text representation essentially helpful for? In computer science, digital data, that is to say information represented in whatsoever binary format, are essentially functional to specific forms of processing, and it can be assumed that the same obtains in computer applications to humanities. Text in digital form is essentially data to be processed. In my opinion, as I have written elswhere, '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.'[1] But how?

First of all, a clarification on the notion of text in a digital environment is here in order. From a computer scientist's point of view, the text is conceived of as 'information coded as characters or sequences of characters' and not as 'literary material as originally written by an author.'[2] The text is then defined simply as a data type for the manipulation of alphanumeric symbols arranged in a linear order. This notion of the text is obviously discrepant from a literary critic's understanding of the text. What cannot be overlooked, from this point of view, is the semiotic nature of the text, that consists not only in a sequence of graphic symbols, but also in what these symbols mean for a writer and a reader. Accordingly, it has been maintained that 'the text is not a physical reality at all but a concept-limit [*Grenzbegriff*]'; and, if 'the nature of the text is not material,' we may say that 'the text is, therefore, only' and 'always an image.'[3]

There is something remarkable about the nature of the text to be gathered here. If the text cannot but be an image of something, which is not of a physical nature, the text is made of two components. According to the stucturalist linguist Louis Hjelmslev, 'the sign,' or for that matter what we refer here to as an image of the text, 'is an entity generated by the

1 D. Buzzetti, *Digital Editions and Text Processing*, in M. Deegan and K. Sutherland (eds.), *Text Editing, Print, and the Digital World*, Aldershot, Ashgate, 2009, pp. 45-62, p. 46.

2 A. C. Day, *Text Processing*, Cambridge, Cambridge University Press, 1984, p. 1.

3 C. Segre, *Introduction to the Analysis of the Literary Text* [1999], Engl. transl. by John Meddemmen, Bloomington, Ind., Indiana University Press, 1988, pp. 301, 315.

connection between an expression and a content.'[4] *Expression* and *content*, then, are the two basic components of the text. Broadly speaking, 'the expression plane refers to the material aspect of the linguistic sign, the content plane to the semantic aspect, there not necessarily being a one-to-one correspondence between both aspects of the linguistic sign.'[5] Apparently, this is due to the fact that the material aspect, or the image of the text, is always an approximation to the semantic aspect, whose complete characterization cannot be exhausted by any of its possible images. It is an intrinsic property of natural language, that there are many ways of expressing the same content, just as there are many ways of assigning a content to the same expression. The relationship between expression and content, therefore, is not a one-to-one relationship, being rather a one-to-many compensation relationship, obeying a kind of indetermination principle: if you fix the expression, its content remains undetermined, and vice versa, if you fix the content, its expression results equally indeterminate. Accordingly, *synonymy* and *polysemy*, that 'are relations between [linguistic] form and meaning,' or in other words relations between expression and content, can be defined as follows: 'synonymy: more than one form having the same meaning,' and 'polysemy: the same form having more than one meaning.'[6] The relation between expression and content is then, essentially, an indetermination relationship between the two basic components of the text.

All this has a bearing on the adequacy of the kind of processing we may apply to a digital representation of the text. On the one hand, processing text in the literary or ordinary sense cannot consist in processing mere strings of characters, for one would process only the image, or the expression, and not the content of the text. On the other hand, the approach proposed by the practitioners of the so-called hard artificial intelligence, grounded as it is on formalization, seems to overlook the basic indeterminacy of the relationship between expression and content in ordinary language. Formalization amounts to ensure a direct correspondence between syntax and semantics. According to Donald Davidson, to formalize or to 'give the logical form' to a sentence is 'to describe it in terms that bring it within the scope of a semantic theory.'[7] And more explicitly, John Haugeland advocates the following 'Formalist's Motto: "You take care of the syntax and the semantics will take care of

---

[4] L. Hjelmslev, *Prolegomena to a Theory of Language* [1943], Engl. transl. by Francis J. Whitfield, Madison, Wis., University of Wisconsin Press, 1963, p. 47.

[5] H. Bussmann, *Routledge Dictionary of Language and Linguistics,* translated and edited by Gregory Trauth and Kerstin Kazzazi, London, Routledge, 1996, p. 245.

[6] G. N. Leech, *Semantics: The Study of Meaning*, Harmondsworth, Penguin Books, 1974, pp. 101-102.

[7] D. Davidson, Essays on Actions and Events, Oxford, Oxford University Press, 2001, p. 144.

itself".'[8] This assumption is based on the adoption of the Physical Symbol System Hypothesis (PSSH), first formulated by Newell and Simon in their famous Turing Award paper, which states that 'a physical symbol system has the necessary and sufficient means for intelligent action.'[9] This hypothesis implies that physical symbol systems, such as digital computers, 'when we provide them with the appropriate symbol-processing programs, will be capable of intelligent action,' such as the assignment of semantic interpretations.[10] Because the operations of a physical symbol system are formal, and because its symbols directly designate content, the formalist's motto supports the belief that formal symbol manipulations preserve meaning. But again such an approach endorses the fallacy that text processing can focus on the formal or syntactic properties of symbol systems, that is to say on the expression of the text, and not worry about processing its content, through the assignment of suitable semantic interpretations.

The presence of an intrinsic indetermination relationship between the two fundamental constituents of language and text has significant consequences. If there is no one-to-one correspondence between syntax and semantics, or in other words between each and every element in the structure of the expression and each and every element in the structure of its content, the relationship between the two components can only obtain between their respective structures taken as a whole. As a result, a variation in either structure, affects the other in its entirety and produces a gestaltic shift. Variation, in a text, is thus essentially governed by an holistic principle. Textual holism impinges on the relation between language and what it means, be it purely conceptual – its *sense* – and so confined to sole meanings, or be it actually existing – its *reference* – and so concerned with observational evidence. Holism and indetermination tie in with each other: Quine's well-known theses of translational indeterminacy,[11] epistemological holism,[12] and ontological relativity,[13] all have to do with the relation between language and what it means, and all

8 J. Haugeland, *Artificial Intelligence: The very idea* [1985], Cambridge, Mass., MIT Press, 1989, p. 118.

9 A. Newell and H. A. Simon, 'Computer Science as Empirical Inquiry: Symbols and Search,' *Communications of the ACM*, 19:3 (1976), 113-126, p. 116.

10 N. J. Nilsson, 'The Physical Symbol System Hypothesis: Status and prospects,' in Max Lungarella, Fumiya Iida, Josh Bongard (eds.), *50 Years of Artificial Intelligence: Essays dedicated to the 50th Anniversary of Artificial Intelligence* (Lecture Notes in Computer Science, 4850), Berlin, Springer, 2007, 9-17, p. 9.

11 Cf. W. V. Quine, *Word and Object*, [Cambridge, Mass.], Technology Press of the Massachusetts Institute of Technology, [1960]. Quine sometimes refers to this kind of indeterminacy as *holophrastic indeterminacy*, and the claim, which 'involves the whole language,' is that 'there is more than one correct method of translating sentences where the two translations differ not merely in the meanings attributed to the sub-sentential parts of speech but also in the net import of the whole sentence' (P. Hylton, 'Willard Van Orman Quine,' in *Stanford Encyclopedia of Philosophy*, URL = http://plato.stanford.edu/entries/quine, accessed 26.02.2011).

confirm the point. Textual indeterminacy, then, implies that language and text cannot be restrained to a unique assignment of syntax and semantics. Language and text, therefore, are to be conceived of as essentially mobile and dynamical systems.

A textual and literary critic, Jerome McGann, describes this constitutive feature of the text in a paradoxical way: 'Text is not self-identical.'[14] Textual ambiguity comes about precisely because 'text is dynamic and mobile and textual structures are essentially indeterminate,' so that 'neither the *expression* nor the *content* of a text are given once and for all.'[15] According to McGann, a text is endowed with 'perceptual features,'[16] that relate to its expression, and a *Gestalt* shift in the perception of formal textual patterns opens 'doors of perception' towards 'new' interpretational 'opportunities and points of view.'(288) So, if it can be said, that 'the invariant rule of the textual condition' is 'variation,'[17] this is clearly due to textual indetermination and holism. The two basic textual components, the expression and the content, behave very much in the same way as do organic wholes in biology. The biologist William Morton Wheeler states that in biological organisms 'the whole is not merely a sum, or resultant, but also an emergent novelty, or creative synthesis,' and describes the 'unique qualitative character of organic wholes' as 'due to the peculiar non-additive relations or interactions among their parts.'[18] In a living system,

---

[12] Cf. Id., 'Two Dogmas of Empiricism,' in *The Philosophical Review*, 60:1 (1951), 20-4. 'We could sum up this thesis – sometimes called epistemological holism – in two points: (1) no knowledge is a priori and immune to empirical refutation, and no knowledge is completely theory-independent; (2) in cases of conflict between theory and observations we cannot summon certain statements in isolation; the whole system of beliefs, or large parts thereof, must stand to trial' (S. Bem and H. Looren de Jong, *Theoretical Issues in Psychology : An introduction*, London, SAGE, 2006², p. 68).

[13] Cf. Id., *Ontological Relativity and Other Essays*, New York and London, Columbia University Press, 1969. According to Quine, 'while it is possible to verify or falsify whole theories, it is not possible to verify or falsify individual statements' (Entry 'W.V. Quine,' in *Wikipedia*, URL = http://en.wikipedia.org/wiki/ W._V._Quine, accessed 26.02.2011), for 'there is more than one way of translating sentences' and 'the various versions differ in the reference that they attribute to parts of the sentence but not in the overall net import that they attribute to the sentence as a whole' (P. Hylton, 'Willard Van Orman Quine,' cit.).

[14] D. Buzzetti and J. McGann, 'Critical Editing in a Digital Horizon,' in L. Burnard, K. O'Brien O'Keeffe, and J. Unsworth (eds.), *Electronic Textual Editing*, New York, The Modern Language Association of America, 2006, 53-73, p. 64. For a more thorough discussion of this assertion, see J. McGann, *Radiant Textuality: Literature after the World Wide Web*, New York, Palgrave, 2001, especially chapter 5 and the Appendix to chapter 6.

[15] Ibid.

[16] J. McGann, 'Visible and Invisible Books: Hermetic Images in N-Dimensional Space,' in *New Literary History*, 32:2 (2001), 283-300, p. 297.

[17] I d., *The Textual Condition*, Princeton University Press, 1991, p. 185.

[18] W. M. Wheeler, 'Emergent Evolution and the Social,' in *Science*, 64:1662 (1926), 433-440, p. 443.

the 'relations' that 'determine the dynamics of interactions and transfonnations it may undergo' is what another biologist, Francisco Varela, calls its 'organization.'[19] Organization acts as an informational holistic principle of 'mutual interconnection' (102) and as Norbert Wiener would say, organization has to be though of as ' "information," not matter or energy.'[20] Moreover, a living system, can be described as 'autopoietic,' in as much as it 'generates and specifies its own organization,'[21] and since in a living system 'what makes it a unity with identity and individuality' is its own 'invariant organization,'(26) that is, its intrinsic, self-definig and self-regulating information content, autopoietic systems 'are unities because, and only because, of their specific autopoietic organization.'(15) Also these features of a biological system can be observed in a text. The relations the organization consits of are described by Varela as 'codependent,'(xv) and in McGann's opinion, textual artefacts – that is to say, 'print and manuscript encoding systems' and 'technologies' – are 'organized under a horizon of co-dependent relations.'[22] So we can say that 'like biological forms and all living systems, not least of all language itself, textuality is a condition that codes (or simulates) what are known as autopoietic systems,'[23] and we can think of literary texts precisely as 'paradigms of those interactive and feedback mechanisms that Humberto Maturana and Francisco Varela have studied as, and called, autopoiesis.'[24] Of his book, *The Textual Conditition*, McGann writes that it 'attempts to sketch a materialist hermeneutics,' in as much as it 'considers text as autopoietic mechanisms operating as self-generating feedback systems that cannot be separated from those who manipulate and use them.' To be more precise, textual autopoiesis 'functions through a pair of interrelated textual embodiments,' namely a content and its expression, that 'we can study as systems of linguistic and bibliographic codings.'(15) Finding a suitable computational model for complex textual phenomena of this kind is thus the obvious challenge faced by scholars aiming at a digital representation of the text up to befitting standards of critical enquiry.

Available technologies

A digital representation of the two basic components of the text, the expression and the content, has already been implemented, but through

[19] F. J. Varela, *Principles of Biological Autonomy*, New York, North Holland, 1979, p. 9.

[20] N. Wiener, *Cybernetics*, 2nd ed., Cambridge, Mass., MIT Press 1961, p. 132.

[21] Varela, *Principles*, p. 13.

[22] J. McGann, 'Marking Texts ofMany Dimensions,' in S. Schreibman, R. Siemens and J. Unsworth (eds.), *A Companion to Digital Humanities*, Maiden, Mass., Blackwell Publishing, 2004, 198-217, p. 200.

[23] Id., 'Texts in N-Dimensions and Interpretation in a New Key,' in *Text Technology*, 12:2 (2003), pp. 1-18, p. 7.

[24] Id., *The Textual Condition*, p. 11.

different techniques for each component and with a different approach. The primary base for the processing of textual data was laid by the introduction, back in the 1960s, of character codes, such as ASCII and EBCDIC, proposed and developed by the U.S. computer manufacturing companies. Character codes were soon extended to handle special characters for other European languages and non-Latin alphabets. But the mere sequence of encoded characters is not enough to represent all the information contained in a manuscript or in a printed page. This need led to the development and use of *markup languages*, first introduced to provide instructions as to how a printed document should look.[25] The use of markup languages was then extended to provide not only display instructions, but also meaning or semantics to words or phrases, or to provide processing instructions. What made that possible, was the introduction of 'generic' markup languages, such as Generalized Markup Language (GML) developed at IBM, which later became an ISO standard as Standard Generalized Markup Language (SGML).[26] This last so-called 'generic' or 'descriptive' language is the parent of current languages now generally used, such as HyperText Markup Language (HTML) and eXtensible Markup Language (XML). HTML is a markup language designed to allow links between documents and enriched, after the introduction of graphical interfaces, to serve the visualization of documents in a browser. XML is a more powerful markup language to provide structure and meaning within documents, and to enable the exchange of data. As a matter of fact, SGML and XML are not properly markup languages, but they have been described as metalanguages, because they provide the rules to define particular markup languages or set of tags applicable to a given purpose. HTML, or, to give another example, the *Guidelines for Electronic Text Encoding and Interchange* of the Text Encoding Initiative (TEI),[27] a language introduced for the

---

[25] *Markup* consists of codes or tags attached to given strings of characters to describe their properties.

[26] 'Historically, electronic manuscripts contained control codes or macros that caused the document to be formatted in a particular way ("specific coding"). In contrast, generic coding, which began in the late 1960s, uses descriptive tags (for example, "heading", rather than "format-17"). Many credit the start of the generic coding movement to a presentation made by William Tunnicliffe,' of the Graphic Communications Association (GCA), to a meeting of the Canadian Government Printing Office in September 1967, entitled 'The separation of information content of documents from their format' (Ch. F. Goldfarb, *The SGML Handbook*, New York, Oxford University Press, 1990, 'Appendix A: A brief history of the development of SGML, p. 567). The 'content of documents,' made, as it is, of coded characters, is not to be confused with the *content* of the *text*, for a *document* is an 'image' or the *expression* of the text, that is made of graphic signs or, for that matter, of their digital representations.

[27] 'The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics' (URL = http://www.tei-c.org/, accessed 05.03.2011). The earlier releases of the Guidelines, P1, P2 and P3, were

processing of literary texts, are properly called languages or SGML/XML applications. SGML is a generalized data representation language and all descriptive markup languages are suited to the representation of data structures such as a marked up string of characters. They are applicable, therefore, to the representation of the *expression* of the text.

To represent the information *content* of a text, either data modelling languages or semantic annotation languages can be used. A data modelling language is 'a mathematical formalism with a notation for describing [(a)] data structures and [(b)] a set of operations used to manipulate and validate that data.'[28] The two elements, (a) and (b), constitute a data model, and its definition requires the specification of both a syntax and a semantics: 'the syntax, or notation, may be given formally in a grammar,' and the semantics is needed 'to refer to the properties of objects within a data model [...] and in particular to the effect (behaviour or abstract meaning) of operations on those objects.'[29] Data modelling languages, then, provide a semantics for structured data.[30] Notable examples of data modelling languages are the so-called entity-relationship-attribute (ERA) diagrams and the Unified Modeling Language (UML). Semantic annotation languages, on the other hand, have been developed, in particular, in the framework of the Semantic Web initiative to describe the information content of Web documents and resources. They are used to provide a semantics for semistructured data.[31] Notable examples are the Resource Description Framework (RDF) and the Web Ontology Language (OWL), a family of knowledge representation languages for authoring ontologies, characterised by formal semantics and RDF descriptions – or, in a more technical jargon, serialisations.

How far can we get by means of these languages in the construction of a dynamic model for text representation? Let us examine some examples. In a medieval manuscript, the following proposition might be found:

> Animal currere si homo currit est necessarium.[32]

But how is it to be interpreted? We may understand it *in sensu composito* and, by adding a comma, we can rewrite it in this form:

---

issued as SGML applications, whereas the last two, P4 and P5, have been issued as XML applications.

[28] 'Data model,' in D. Howe, *FOLDOC: Free On-Line Dictionary of Computing*, URL = http://foldoc.org/ data+model, accessed 05.03.2011.

[29] M. L. Brodie, 'Axiomatic Definitions for Data Model Semantics,' in I*nformation Systems,* 7:2 (1982), 183-197, p. 184.

[30] Data are usually described as *structured*, when their structure and meaning is formally defined by a data model.

[31] Although so-called *semistructured* data may have some structure, they lack a formal data model.

[32] Cf. Ricardus Sophista, *Abstractiones*, URL = http://www.hs-augsburg.de/~harsch/Chronologia/Lspost13/ RicardusSophista/ric_abst.html, accessed 05.03.2011.

Animal currere si homo currit**,** est necessarium.

Or else we may construe it *in senso diviso*, and then we would rewrite it in this way:

Animal currere**,** si homo currit**,** est necessarium.

In the first case, the proposition would have to be taken as true, whereas, in the second case, it would have to be taken as false. But what does a modern editor actually do by adding punctuation in his or her transcription? Does she make explicit what is implicitly there, and assume that the sense of the proposition is already given and objectively recognizable, or does she add punctuation marks to convey an external and truly metalinguistic description of the structure of the sentence, that she takes for objective as it stands, devoid of punctuation marks? As it stands, the sentence is open to different interpretations, that we may signal with diverse punctuation marks. Textual variants are the result of the indeterminacy of interpretation. On the contrary, variant interpretations are the result of textual indetermination. Any new act of writing, any variation in the expression of the text, just as a new punctuation mark, requires a variation in the content or in the interpretation of the text. In the same way, any new act of reading, any variation in the information content of the text, just as a new interpretation of a given phrase, requires a variation in the expression of the text. Again, as it has been convincingly argued by Jerome McGann, 'we may usefully regard all criticism and interpretation as deformance.'[33]

Another example, taken from a recent bestseller on punctuation, 'a marvellous punctuation-fan joke about a panda who "eats, shoots and leaves",' allows us to examine these textual phenomena in more detail.[34] Here again, the presence of a comma is decisive in determining the meaning of the entire sentence. By removing the comma, the whole meaning of the sentence undergoes a restructuring shift: the two verbs (V) 'shoots' and 'leaves' are now seen as two nouns (N) that express the object of the predicate 'eats'. It is not only the whole comprehensive meaning of the compound sentence, as in the former example, that gets changed, but here the meaning and grammatical status of individual words is recast just as well.

Now, as it as been maintained, 'punctuation is not simply part of our writing system,' for it is also 'a type of document markup that may vary and be replaced by other types of markup.'[35] So, for the sake of simplicity,

---

[33] J. McGann and L. Samuels, 'Deformance and Interpretation,' in *New Literary History*, 30:1 (1999), 25-56, p. 46.

[34] L. Truss, *Eats Shoots & Leaves: The zero tolerance approach to punctuation*, London, Profile Books, 2003, p. 2.

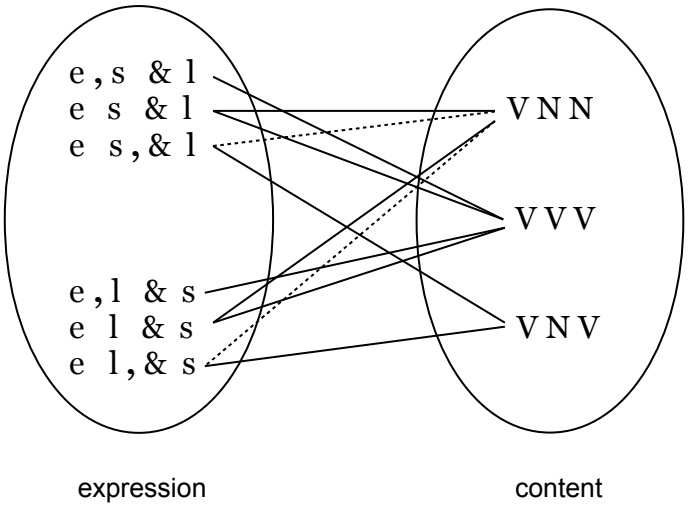
[35] . H. Coombs, A. H. Renear and S. J. DeRose, 'Markup systems and the future of scholarly text processing,' in *Communications of the ACM*, 30:11 (1987), 933-47, p. 940.

we may reason on punctuation instead of markup and use this example to consider the relationship between the structure of the document – or the expression of the text – and the structure of its information content. It is the markup, or punctuation, that assigns a structure to the document. Markup is introduced precisely for the sake of 'rendering "digital stuff" as structured information,' so as to transform unstructured string of characters – mere 'digital stuff ' – into stuctural units 'accessible as information within intelligent computing environments.'[36] Punctuation then, or markup for that matter, signals possible alternatives between different structural arrangements, as it can be seen from our example:

	Eats , shoots and leaves
Semantic structure	V V V
Markup	Eats **<add>** , **</add>** shoots and leaves

	Eats shoots and leaves
Semantic structure	V N N
Markup	Eats **~~** , **~~** shoots and leaves

By way of illustration, we can also outline a (partial) mapping between two sets of possible structural configurations of both the expression and the content of the text (Fig. 1). Here again we can see that punctuation, or markup, acts as a special textual marker for the semantic structure of the text.



— Fig. 1 —

A dynamic model of the interactions between the two kinds of structure, the structure of the expression and the structure of the content, should be

[36] R. Cover, N. Duncan, and D. T. Barnard, 'The Progress of SGML (Standard Generalized Markup Language): Extracts from a Comprehensive Bibliography,' in *Literary and Linguistic Computing*, 6:3 (1991), 197-209, pp. 197-98.

capable of accounting for the procedures of both textual and literary criticism. An editor tries to reduce to a coherent unity the several documents and variant representations through which the text was handed down to her, whereas an interpreter proceeds from a single, coherent form of text representation to several interpretative structures, all compatible with it. In both cases, our dynamic model should take into account a one-to-many mapping: from several textual variants to their unique 'logical sum,'[37] or, vice versa, from a single 'logical sum' of compatible interpretations to their several and distinct representations. In the first case, we have to consider an ambiguous content as an indeterminate whole compatible with different forms of expression; in the other, we have to consider a number of different interpretations, or assignments of information content, all compatible with an ambiguous expression taken as an indeterminate whole.

From these observations, we may draw some further conclusions about markup and see that we may take advantage precisely of its intrinsic features to contrive a conceptual model of textual dynamics. In the first place, we can see that through punctuation, or markup, a variant interpretation can be transformed into a textual variant and vice versa.
We can also elicit that the same indetermination relationship that occurs between the expression and the content of the text is to be found between the the set of possible markup structures and the set of their possible semantic representations. As it has been clearly pointed out, 'the same markup can convey different meanings in different contexts,' just as 'markup can communicate the same meaning in different ways using very different syntax.'[38] This means that the outcome of applying 'mapping rules,' from 'syntactic relations' between markup elements into semantic relations between the elements of an 'object level domain,' (6) would amount either to 'the re-tagging of documents with richer markup,' or to a new semantic description or serialisation 'in the form of RDF or a topic map' (8).[39]

But there is another feature of the markup that is crucial for the contrivance of a dynamic model of text representation in digital form. Markup exhibits a distinctive duality, that it shares again with punctuation and, more generally, with all diacritical marks. For, as it has been observed, the markup is 'simultaneously embedded and separable' from

[37] Cf. M. Thaller, 'Historical Information Science: Is There Such a Thing? New Comments on an Old Idea,' in T. Orlandi (ed.), *Discipline umanistiche e informatica*, Roma, Accademia Nazionale dei Lincei, 1993, 51–86, p. 64.

[38] D. Dubin and D.J. Birnbaum, 'Interpretation beyond markup,' presented at *Extreme Markup Languages 2004*, p. 1.

[39] Topic Maps 'is an ISO standard for describing knowledge structures and associating them with information resources' (V. Lombardi, *Metadata Glossary*, URL = http://www.noisebetweenstations.com/personal/essays/metadata_glossary/metadata_glossary.html, accessed 10.03.2011).

the text, and we can say that it 'is part of the text and yet it is distinct' from it.[40] Accordingly, the markup has been described either as an external 'technique for representing structure,' (3) or as that very 'structure' (4) embedded in the text. So the 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.[41] Markup shares this property with diacritical signs. Markup tags are then, in fact, diacritics and like all diacritical marks can be considered as an expression of the 'reflexive metalinguistic nature'[42] of natural language, the capability that all natural languages – and texts – possess of saying something about themselves. A diacritical mark or phrase, a punctuation mark or a markup construct, is an expression of the object language, as to its expression, just as it is an expression of a metalanguage, as to its content. Diacritics can be viewed as part of the text and as separate from the text, and markup has actually been described in both ways. On the one side, it has been maintained that 'the markup is not part of the text or content of the expression, but tells us something about it,'[43] or in other words that the markup belongs to a metalanguage. On the other side, it has been acknowledged that the markup is 'constitutive of the text it characterizes,' albeit with a reservation, for this 'recognition' raises 'new puzzles 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.'[44] But it can indeed be both, without posing any problem, for diacritics and markup are essentially ambiguous.

And it is precisely this kind of *diacritical ambiguity* possessed by markup that can be exploited to devise a dynamic model of the text. Such a model can be expounded through a diagram, a kind of multidimensional matrix, whose elements are connected by a series of operations. The resulting process is a kind of loop. But let us examine it in detail.

---

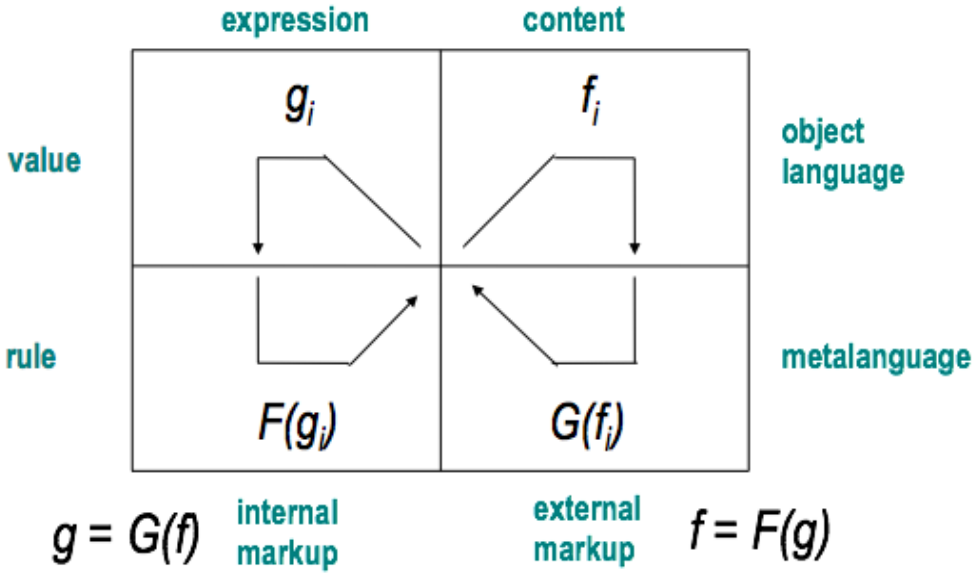
40 D. R. Raymond, F. W. Tompa and D. Wood, 'Markup Reconsidered,' paper presented at the First International Workshop on Principles of Document Processing, Washington DC, 22-23 October 1992, URL = http://softbase.uwaterloo.ca/~drraymon/papers/markup.ps, accessed 10.03.2011, p. 3.

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

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

43 Coombs et al., 'Markup systems,' p. 934.

44 A. Renear, 'The descriptive/procedural distinction is flawed,' in *Markup Languages: Theory & Practice*, 2:4 (2001), 411-420, p. 419.



– Fig. 2 –

The structural elements of the *expression* of the text, represented by embedded or *internal markup*, are here arranged in the first column ot the table. In the second column we find structural elements of the *content* of the text, as described by data modelling or semantic description languages, that can also be regarded as a form of *markup*, albeit *external*. Now, one and the same internal markup construct can be seen either as belonging to the *object language* of the text, in as much as it is a structural element of its expression, or else as a representation of that very element, separate from the text and belonging to a *metalanguage*. These two aspects of a markup construct can be severed, and the operation that converts the one into the other is a logical move, that rests on the assumption that the 'meaning of the markup' is 'the set of inferences about the document that are licensed by the markup.'[45] Accordingly, this move posits a markup construct as an inference-licence. If so, we can place it in the lower part of the first column and regard it, to recall Gilbert Ryle's famous description, as an 'inference-ticket,' or a *rule*-statement 'to move from asserting factual statements to asserting other factual statements'[46] – in our case, to infer from a statement about an observed textual property, to a statement about a property of its content. That content property, in its turn, expressed in a semantic annotation language, can be placed in the upper compartment of the second column as the *value* of the operation prompted by the instruction found in the lower compartment of the first column. All this

[45] C. M. Sperberg-McQueen, C. Huitfeldt and A. Renear, 'Meaning and Interpretation of Markup,' in *Markup Languages: Theory & Practice*, 2:3 (2000), 215-34, p. 231.

[46] G. Ryle, *The Concept of Mind*, London, Hutchinson's University Library, 1949 , p. 121.

means that markup can have both 'descriptive' and 'performative' force,[47] and what has just been said about markup constructs, or the structural elements of the expression of the text, applies also to semantic annotation constructs, or the stuctural elements of its content. We can therefore posit a semantic description as a rule, place it in the lower part of the second column, and move from it to the value of the operation it commands, ending up again with a property of the expression, in the upper part of the first column. And so the cycle is complete.

A plausible logical explanation can actually be provided for the whole series of operations represented by this conceptual model. As Ryle reminds us, inference-licences 'belong to a different and more sophisticated level of discourse from that [...] to which belong the statements of the facts that satisfy them.'[48] In other words, rule-statements are to be seen as 'second-order object-language statements,' or statements based on a second-order form of predication, that 'are equivalent to first-order metalinguistic statements,' or statements based on an ordinary form of predication.[49] So the logical import of markup expressions understood as rules, or inference-licences, is different from the logical import of markup expressions construed as factual statements about observed textual properties. And this ambivalence of markup expressions is paralleled by the 'double sense' acquired by 'the only explicit symbol' of Spencer Brown's calculus of indications, the mark ⌝ of distinction, since 'on the one hand it represents' an operation, namely 'the act of distinction,' and 'on the other hand it is a value,' namely 'the content of a distinction,' or the result of an operation.[50] The mark, then, can be a mark of ambiguity, for in a sense 'there can be no' real 'separation between distinctions and acts of distinguishing.'[51] The separation can be only formal and this kind of ambiguity may also be seen as an instance of the notorious *distinctio formalis a parte rei*, Duns Scotus has more than often been blamed for.[52] Be it as it may, the analogy shall actually be of further use.

Can we now derive from this model, merely conceptual, a formal mathematical model of textual dynamics? This is the challenge that a digital tepresentation of the text apparently has to face. But before turning to this question, a clarification on another technological issue is here in

47 Renear, 'The descriptive/procedural distinction,' p. 419.

48 Ryle, *The Concept of Mind*, p. 121.

49 Buzzetti, *Digital Editions and Text Processing*, p. 57.

50 Varela, *Principles*, p. 111. Cf. G. Spencer Brown, *Laws of Form*, London, Allen & Unwin, 1969.

51 L. H. Kauffman and F. J. Varela, 'Form dynamics,' in *Journal of Social and Biological Systems,* 3:2 (1980), 171-206, p. 205.

52 For a formal reconstruction of Scotus' formal distinction, see D. P. Henry, *Medieval Logic and Metaphysics: A modern introduction*, London, Hutchinson University Library, [1972], pp. 88-95.

order. The adoption of SGML, or XML, as 'a standard for encoding textual data,'[53] has had the doubtful consequence of suggesting an 'OHCO structure,' i.e. 'an ordered hierarchy of content objects,' as the 'basic model of the text.' (6) A 'content object' is a portion of a 'document' (5) that contains or is contained within other content objects, or portions of the document, and that forms with them a 'hierarchy' of containment relations, the smallest elements of which are 'ordered' in succession, in the sequence of characters that form the document. The shortcomings of the OHCO model are quite obvious. By equating the structure of the text with the structure of its expression, it engenders a confusion between the document and the text. Moreover, it does not take into account the 'limitations'[54] of current 'SGML-based markup systems, that cannot handle "overlapping hierarchies",'[55] that is to say features that do not nest within other features. Structuring a poem by verse would not allow for a concurrent structure by grammatical units spanning over different verses.

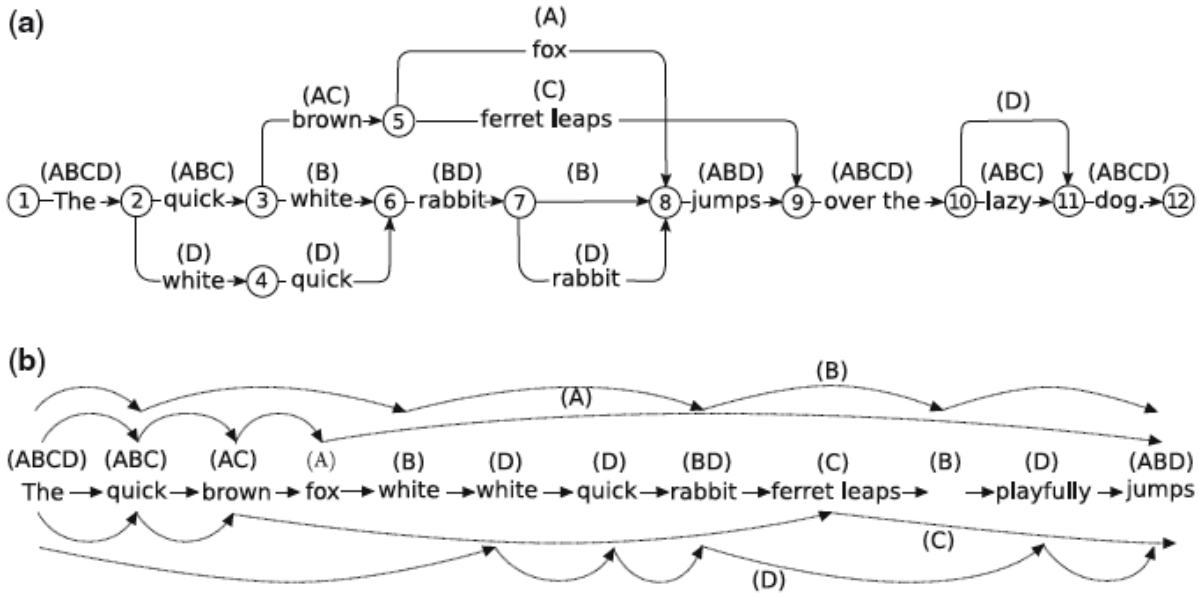
A solution to this problem has been proposed through the introduction of the so-called Multi-Version Document (MVD) model. According to Desmond Schmidt, who has devised it, 'overlap is a serious problem in the encoding of cultural heritage texts,' and a problem that cannot be solved through markup, for it resides in 'the technical limitations of embedded markup itself.'[56] An MVD model represents 'all the versions of a work' as 'a directed graph, with one start node and one end-node,' (350) as shown in Fig. 3 (a). This

[53] S. J. DeRose, D. G. Durand, E. Mylonas, and A. H. Renear, "What Is Text, Really?," in Journal of Computing in Higher Education, 1:2 (1990), 3-26, p. 18.

[54] W. Piez, 'Form and Format: Towards a Semiotics of Digital Text Encoding,' in *Digital Humanities 2007: Conference Abstracts*, 153-57, p. 156. According to the author, current markup systems 'can gracefully handle only a single organizational hierarchy at a time' (Ibid.). But see note 53 below for a partial amendment of this view.

[55] A. Witt, 'Multiple Hierarchies: New aspects of an old solution,' *Extreme Markup Languages 2004*, p. 4. The author points out, though, that 'different structures often can be expressed within one hierarchy,' when 'none of [the] elements belonging to [...] different tag sets overlap': so, 'in practice,' it is possible to find also 'multi-hierarchically' structured texts compatible with 'SGML-based markup systems,' although 'from a formal point of view' they can only 'allow for the representation of exactly one hierarchy' (p. 1). The 'problem of overlapping hierarchies' was first discussed in a paper by D. Barnard, R. Hayter, M. Karababa, G. Logan, and J. McFadden, 'SGML-based Markup for Literary Texts: Two problems and some solutions,' in *Computers and the Humanities*, 22:4 (1988), 265-276.

[56] D. Schmidt, 'The Inadequacy of Embedded Markup for Cultural Heritage Texts,' in *Literary and Linguistic Computing*, 25:3 (2010), 337-356, p. 348.



– Fig. 3 –

'variant graph,' as it has also been called, comprises 'a direct analogue' of 'each type of editing operation: deletion, insertion, replacement and transposition.'[57] Alternatively, an MVD model 'can be serialized as a list of paired values, each consisting of a fragment of text and a set of versions to which that fragment belongs,'[58] as shown in Fig. 3 (b). An MVD graph actually works as the 'logical sum' of all the versions of a text and since 'XML documents are simply text files'[59] all different semantic descriptions of a document produced by an XML semantic annotation language can be represented through an MVD graph. A mapping between markup structural elements and semantic description structural elements becomes thus a mapping between elements of two isomorphic graphs. It remains to enquire whether this circumstance can help in finding a mathematical function to represent the relations between the textual and the interpretational variants of a text.

A mathematical model ?

A formal model for dynamic text representation should account for two essential aspects of textual mobility, namely self-reflexivity and indetermination. In what follows, only a few hints shall be surmised on how these two basic issues could possibly be approached.

[57] D. Schmidt, R. Colomb, A data structure for representing multi-version texts online, in International Journal of Human-Computer Studies, 67:6 (2009), 497-514, p. 503.

[58] Schmidt, 'The Inadequacy of Embedded Markup,' p. 350.

[59] K. Williams [et al.], *Professional XML databases*, Wrox Press, Birmingham, 2000, p. 2.

Textual self-reflexivity originates from the interconnexion between the expression and the content of the text. The relation of mutual dependence between the expression and the content is undetermined and conveys mobility and dynamism to textual structures, 'which may become stable as definite structural forms either of the expression or of the content, and may, reciprocally, determine the instability either of the corresponding content or of the corresponding expression.'[60] To one and the same arrangement of the expression many possible content assignments may be related, and vice versa. The text, then, is not identical to itself, because of its self-reflexive instability. Jerome McGann derives this very 'law of non-identity'

$$A = A \Leftrightarrow A \neq A \qquad [1]$$

from the distinction produced by the primary partition of the text into expression and content, as expounded formally by means of the 'form of distinction' introduced by George Spencer Brown.[61] Through self-reflexivity the text determines its own internal organization and can be analysed by means of the logic of autopoietic systems, as proposed by Lou Kauffman's and Francisco Varela's extension of Spencer Brown's calculus of indications.[62]

Because of the indeterminacy of the relationship between expression and content, the set of internal relations between the constituent parts of the text remains mostly implicit, and the structure of the text may be defined as the 'set of latent relations'[63] among its structural elements. The text can then be assumed as an indetrminate whole, and applied to it, Spencer Brown's primary partition can be provided by the distinction of its primary subunits, expression and content.

---

60 D. Buzzetti, 'Diacritical Ambiguity and Markup,' in Id., G. Pancaldi, and H. Short (eds.), *Augmenting Comprehension: Digital tools and the history of ideas*, London-Oxford, Office for Humanities Communication, 2004, 175-188, p. 180. In this article the following considerations on textual self-reflexivity are developed in greater detail.

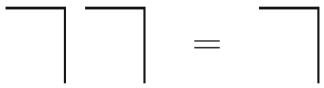
61 'Texts and their field spaces,' such as the expression and the content of the text, 'are autopoietic scenes of co-dependent emergence. As such, their primal state is dynamic and has been best characterized by G. Spencer Brown's *Laws of Form* (1969), where "the form of distinction" – the act of making indications by drawing a distinction – is taken as "given" and primal (1). This means that the elementary law is not the law of identity but the law of non-identity (so that we must say that "*a* equals *a* if and only if *a* does not equal *a*"). Identities emerge as distinctions are drawn and redrawn, and the acts of drawing out distinctions emerge as co-dependent responses to the field identities that the form of distinction calls to attention' (J. McGann, 'Marking Texts of Many Dimensions,' in S. Schreibman, R. G. Siemens, J. M. Unsworth (eds.), *A companion to digital humanities*, Malden, Mass., Blackwell Publishing, 2004, 198-217, p. 212).

62 See Kauffman and Varela, 'Form dynamics,' cit.

63 C. Segre, *Introduction to the Analysis of the Literary Text* (1985), transl. by J. Meddemmen, Bloomington Ind., Indiana University Press, 1988, p. 44.

The structural instability of any image or representation of the text is the immediate outcome of this distinction. Expression and content constitute two subsystems of the whole textual system; but once considered as separate units, they constitute two new distinct and related wholes. By analysing the expression, we determine its structure in relation to its integral whole. However, for any given and self-identical structure of the expression we can have several ways of analysing its content. By equating the identity of the text with the partial unit that consists in its mere expression, the other partial unit that consists in its content remains undetermined. A symmetrical and similar phenomenon occurs by equating the identity of the text with the other partial unit that consists in its content. The phenomena of instability and indeterminacy of the structure of the text come about when we reduce the integral identity of the text to the identity of one of its two partial subunits.

The 'indication' of the expression, or its representation in Spencer Brown's calculus, and the 'indication' of the content presuppose their distinction, produced by the primary partition operating upon the whole of the text. The 'indication' of the expression makes it a subunit of the text identical to itself and determines its structure. The determination and the identity of the expression with itself is expressed formally by the law of idempotence of the expression with respect to its representation, as specified by Spencer Brown's first axiom, or 'law of calling,'[64] and 'form of condensation'(5):



The 'indication' of the content makes it a subunit of the text identical to itself, and determines its structure. In the same way, the determination and the identity of the content with itself is expressed formally by the law of idempotence of the content with respect to its representation. But how can the identity of the text with itself, expressed as the idempotence of its partial subunits with respect to their representation, both depend on and at the same time be cancelled by the primary distinction that defines them?

It can be shown that the 'law of non-identity' [1] – or law of compensation between determination and indetermination of the expression and the content of the text, as the case may be – presupposes and implies an endomorphism (f) between the structural constituents of the text:[65]

---

64 Spencer Brown, *Laws of Form*, p. 1.

65 See D. Buzzetti, 'Digital Representation and the Text Model,' in *New Literary History*, 33:1 (2002), 61-88, pp. 82-84.

$$(A = A \Leftrightarrow A \neq A) \Leftrightarrow A \xrightarrow{f} A \qquad [2]$$

This endomorphism maps elements of the content into elements of the expression, or conversely, elements of the expression into elements of the content. The compensation between the reciprocal determination and indetermination of the expression and the content of the text is represented by the inversion of the domain and co-domain of the endomorphism and shows itself explicitly in the ambiguity of markup constructs and the oscillation between their dual logical function, declarative and performative as it may be.

The relationship between the structure – or the logical form – of the expression of the text and the structure – or the semantic model – of its content can in turn be considered an example of Spencer Brown's second axiom, or 'law of crossing,'[66] and 'form of cancellation' (5):

For, the reference of the structural articulation of one of the two textual subunits to the structural totality of the other cancels its identity with itself and brings about its indeterminacy. In conclusion, then, the identity of the text with itself is posited by the primary partition between expression and content, and is cancelled by the crossing from one subunit to the other, which revokes the separate identity of each determinate partial unit and reintegrates the indeterminate totality of the text. The text can be considered and described, in brief, only as a holistic unit.

As the extended calculus of indications may account for the self-reflexive features of the text, another mathematical approach may account for its indeterminacy.[67] But the convergence of the two possible points of view should not be altogether ruled out.

As it has been pointed out, markup constructs are a kind of diacritical expressions that oscillate between their dual logical function, and can be seen both as a metalinguistic representation of a textual structure, and as an objectivised object-language textual structure themselves. From this

[66] Spencer Brown, *Laws of Form*, p. 2.

[67] The following considerations have already been exposed, in the essence, in D. Buzzetti, *Text, Science, and Technology: Construing text as a system*, in G. Castellani, V. Fortunati, E. Lamberti and C. Franceschi (eds.), *Biocomplexity: At the cutting edge of physics, systems biology and humanities*, Bologna, Bononia University Press, 2008 (Quaderni del Centro Interdipartimentale "L. Galvani," 1), pp. 295-320.

point of view, a 'principle of representation-theoretical self-duality' can apply to markup conceived of both as the structure, and as a representation of the structure of the text. For it is precisely the 'identification [...] beween structures and the collection of all representations of the structure' that is 'expressed in the principle of self-duality' as introduced by Shahn Majid.[68]

According to Majid, 'an evaluation *f(x)* can also be read *x(f)*, where *f* is an element of a dual structure.' This kind of self-duality holds in a pure mathematical language, where 'such an "observer-observed" reverse interpretation of the mathematical structure can always be forced,' but 'will the dual interpretation also describe physics?'[69] or for that matter a text, viewed as an information carrier and physical device? In physics, as Majid has shown, Hopf algebras, one of the simplest self-dual 'categories,' or types of mathematical structures, can provide 'models in which quantum mechanics and gravity are unified into one mathematical structure.'[70] Likewise, in a text, a diacritical sign or markup element of the expression can be seen as a representation of the structure of the content, just as a structural unit of the content can be seen as a representation of the structure of the expression. A restructuring operation from an expression unity to compatible content assignments can be easily reversed, and markup elements, either internal or external, can be seen both as signs, or values, as well as instructions, or operations. In physics, self-duality implies that a theory 'should admit a "polarisation" into two halves each of which is the set of representations of the other,' so that we 'should be able to reverse interpretations.'[71] And that is precisely how we can construe the polarisation between the expression and the content of the text.

The analogy with self-dual physical systems can be assumed as a starting point for a formal description of textual phenomena and the construction of what McGann has called 'quantum poetics.'[72] In quantum mechanics, observables or 'coordinates like *x*, *p*,' that describe the position and momentum of particles, 'become operators **x**, **p**' that 'do not commute,' so that '**xp** no longer equals **px**.' The non-commutativity of position and momentum coordinates

---

[68] S. Majid, 'Principle of Representation-theoretic Self-duality,' in *Physics Essays*, 4:3 (1991), 395-405, p. 396. On Majid's philosophy of physics, see M. Heller, 'Algebraic Self-Duality as the "Ultimate Explanation",' in *Foundations of Science*, 9:4 (2004), 369-385.

[69] I d., *Foundations of Quantum Group Theory*, Cambridge, Cambridge University Press, p. 293.

[70] Id., 'Principle of Representation-theoretic Self-duality,' p. 402.

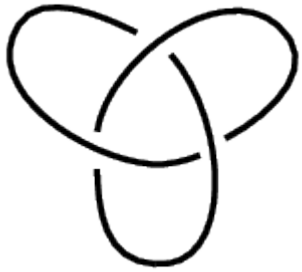
[71] Id., 'Quantum Groups and Noncommutative Geometry,' in *Journal of Mathematical Physics*, 41:6 (2000), 3892-3942. Also at URL = http://arxiv.org/PS_cache/hep-th/pdf/0006/0006167v1.pdf, accessed 19.03.2011, p. 61.

[72] Cf. McGann, 'Visible and Invisible Books,' p. 297.

has the interpretation that it matters which you measure first, **x** or **p**, and this in turn is related to the famous Heisenberg uncertainty principle, that you cannot measure both of them accurately at the same time.[73]

Likewise, in a text, markup elements behave on the one side as the observable representations of structural units of its expression or its content and, on the other, as operators that rearrange textual units and produce a restructuring of the expression or the content of the text. In a text, 'a structural shift changes the entire network of internal relations and affects the whole range of textual units.' This kind of '*Gestalt* leap,' is then a 'discrete' and 'discontinuous' phenomenon.[74] Similarly, in physics, 'non-commutativity leads to a kind of "finite difference" or discretization,' which is a 'general feature' of physical self-dual structures.[75] The shift from an object-languag to a metalinguistic interpretation of a diacritical mark can be seen as a shift from a classical to a quantum interpretation of the textual condition.

A structural shift introduces a temporal dimension. To take that into account, the braided structure of Fig. 2 should be extended: it should comprise a third dimension besides expression and content to represent perceptual restructuring operations.[76] The result would be a trefoil knot structure (Fig. 4), whose 'invariant,' or defining characteristic, can be described in terms of a non-commutative geometrical structure such as a quantum group.[77]



– Fig. 4 –

According to Majid, the time dimension could be introduced as shown in Fig. 5 (b), (99) where the vertical axis is interpreted as time and the knot

---

[73] S. Majid, 'Non-commutative Geometry and Quantum Groups,' in *Philosophical Transactions of the Royal Society of London*, Series A, 358:1765 (2000), 89-109, p. 90.

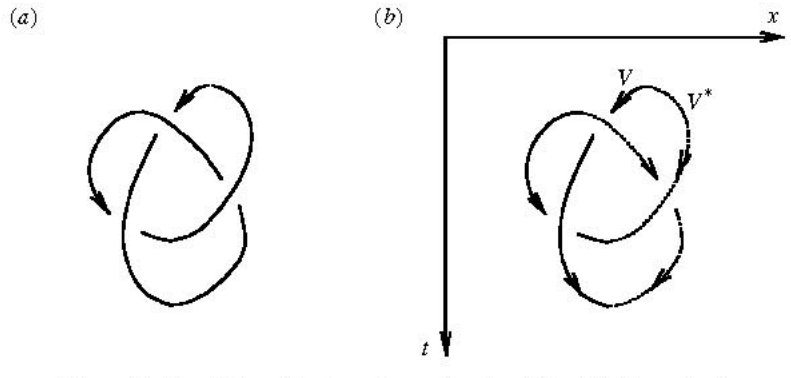
[74] Buzzetti, *Text, Science, and Technology*, p. 312.

[75] Majid, 'Non-commutative Geometry and Quantum Groups,' p. 91.

[76] According to Charles Sanders Peirce, a sign always involves 'thirdness.' As he writes, 'thirdness is the triadic relation existing between a sign, its object, and the interpreting thought, itself a sign, considered as constituting the mode of being of a sign.' (*A Letter to Lady Welby*, CP 8.331-332, 1904).

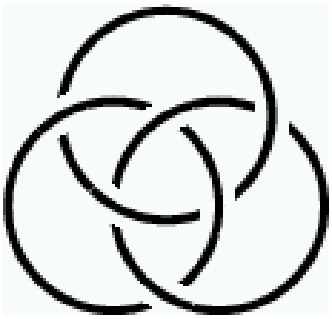
[77] See Majid, 'Non-commutative Geometry and Quantum Groups,' pp. 98ff.

as describing the trajectories of self-dual elements *V* and *V** flowing down the page.



– Fig. 5 –

These last considerations are purely tentative and are meant only to suggest an evocative line of research. In this respect, it may be interesting to note that psychoanalysts find it necessary to expose the kind of discourse that constitutes their analytic practice expressly through an interlacing of the Real, the Symbolic and the Imaginary[78] as represented by a structure akin to the trefoil knot, known as the Borromean link – or rings, or knot (Fig. 6):



– Fig. 6 –

'The Borromean knot, is defined as the way in which we *imagine* the *real* effect of the *symbolic*.'[79] Could we say that a textual structure lives its life

[78] Cf. J. Lacan, 'Au-delà du "principe de réalité",'(1936) in *Écrits*, Paris, Seuil, 1966, pp. 73-92.

[79] Ph. Julien, *Pour Lire Jacques Lacan*, 2e éd., Paris, E.P.E.L., 1990, p. 221.

precisely in its enacting an analytic practice? which, after all, is as language based as a text is? But perhaps more in line with the thrust of our argument is to recognise that structures like the trefoil knot or the Borromean rings, as shown in Fig. 4 to 6, 'are topological diagrams, not geometrical representations.'[80] As such, they provide, to use Maturana and Varela's terminology, the *organisation* of an information carrying physical *structure* of sorts – be it a biological organism, a semiotic system, or a molecule, as in the use of 'DNA components' to forge molecular Borromean rings.[81]

So far we have only ventured to provide some data to build a mathematical model for a dynamic text representation, but we have not even attempted to solve the problem. Whether that would be feasable along the lines of reasoning that we have tried to suggest is left to further, more mature considerations.

[80] Chengde Mao, Weiqiong Sun, and N. C. Seeman, 'Assembly of Borromean Rings from DNA,' in *Nature*, 386 (1997), 137-138, p. 137.
[81] Ibid.

Contexts

Proceedings of ANPA 31

Arleta D. Ford, *Editor*

Proceedings of the 31st Annual International Meeting of the
Alternative Natural Philosophy Association

Wesley House, Jesus Lane, Cambridge
August 2010