Claus Huitfeldt
University of Bergen

Multi-Dimensional Texts in a One-Dimensional Medium

The paper discusses one of the tools which may be used for representing texts in machine-readable form, i.e. encoding systems or markup languages. This discussion is at the same time a report on current tendencies in the field. An attempt is made at reconstructing some of the main conceptions of text lying behind these tendencies. It is argued that although the conceptions of texts and text structures inherent in these tendencies seem to be misguided, nevertheless text encoding is a fruitful approach to the study of texts. Finally, some conclusions are drawn concerning the relevance of this discussion to themes in text linguistics.

— * —

My aim in this paper is to show one of the ways in which information technology opens the door to an entirely new approach to text studies. That computing provides us with powerful tools for manipulating and analyzing texts is a well documented fact. My claim, however, is that the use of information technology in textual studies may also help us explicate traditional concepts of text by way of stimulating a new kind of text analysis.

If possible, I would have liked to start out by suggesting possible answers to questions like: What is a text? What is the ontological status of the text? What is the epistemological status of texts? However, I have come to think that these questions do not represent a fruitful first approach to our theme at all.

The answer to the question what a text is depends on the context, methods and purpose of our investigations.

Texts have been studied by many and diverse disciplines – in so-called analytical bibliography [Kraft p 77-79] or codicology texts are studied as physical objects with physical properties. In classical, medieval, and biblical philology and text criticism the physical objects containing texts are called text witnesses, the text being an abstract entity. In linguistics texts are sometimes regarded as discourse events, sometimes as strings of sentences [de Beaugrande and Dressler, Halliday and Hasan].

– * –

Our concept of text has to a large extent been shaped by the limits and possibilities of the media which have traditionally carried texts. In this perspective, the computer is a new medium which will create new kinds of texts, i.e. change the subject matter of our study, thus changing also our concepts of and ways of dealing with texts.

My main concern here is that adapting our traditional concepts of texts to the use of the new medium may also help us explicate our traditional concepts of texts and give us a better understanding of existing ways of dealing with and relating to texts.

This effect becomes particularly clear in attempts to transfer texts from traditional media to the new medium, a process which may be seen as an attempt to represent multi-dimensional texts in a one-dimensional medium.

– * –

Why I call texts multi-dimensional will hopefully become clear from the discussion further below. But what is the reason for calling the computer a one-dimensional medium? After all, computers display text on screens and in paper printout, which may look exactly like traditional printed text.

Internally, a computer represents a text as a long string of characters, which in turn will be represented by a series of numbers, which in turn will be represented by a series of binary digits, which in turn will be represented by variations in the physical properties of the data carrier.

For the present purpose, we may regard the conventional data text file format as essentially a one-dimensional string of characters. This format is significantly different from the traditional book or printed paper. It is even more different from *handwritten* material. Handwritten material is structurally more informal then printed texts. Variations which in print will be discrete and easily identifiable may in manuscripts be gradual and hardly discernible.

If the means of representation placed at our disposal is essentially only a long string of characters, how can all the information contained in a written manuscript page be mapped on to this one-dimensional format? The answer is simple enough:

We insert special, reserved character combinations, so-called codes, into our long character string. These codes indicate features such as line endings, page brakes, start of underlining, end of underlining, and so on.

Still, there is only one way of reproducing *all* the information, and that is by the production of an exact duplicate. First of all, we have

to ask: Which parts of the information conveyed by a document are to be retained? How do we distinguish between form and content, between the ("relevant") information contained in the manuscript and the more accidental traits of its actual appearance on paper?

The very fact that questions like these are asked, and that attempts are made at both asking and answering them in a systematic manner, is perhaps one of the most promising and fruitful outcomes of recent discussions on text encoding.

There are several different kinds and uses of text encoding. The purpose of *descriptive* text encoding is not to prepare for some specific mode of presentation or analysis, but to represent as accurately as possible the textual information, the logical structure of the text, and the internal relationships between different text segments. In this way, modes of presentation and analysis may be decided on afterwards, independently of the initial preparation of the text.

In order to facilitate exchange of computerized texts and text software, there is an urgent need for standardization, not only of hardware and internal representation formats, but also of markup languages. (In the following, I will use the terms 'markup language' and 'markup' interchangeably with 'encoding systems' and 'text encoding'.)

In 1986, Standard Generalized Markup Language (SGML) was established as an international standard by the International Standards Organization [ISO 8879-1986].

SGML is, strictly spoken, not itself a markup language, but a formal grammar for the design and specification of markup languages.

In SGML, a text is associated with a Document Type Definition (DTD). The DTD defines a document type, declaring which basic constituents a document may have, how they should be marked up, and how these marked-up elements may be combined.

An SGML-encoded text is a hierarchy of serially ordered text elements, the structure of which adheres to the declarations given in the associated DTD.

Since the DTD is specified in a highly structured formal language it is possible to design computer programs to check whether any given text adheres to the specifications and definitions given in a DTD. This has several advantages and increases control over composition, analysis, and manipulation of texts.

The "father" of SGML, ... Charles Goldfarb of IBM Corporation, suggests that from now on, "...the techniques available for processing rigorously-defined objects like programs and data bases can be used for processing documents as well." [Goldfarb, p 60]

SGML was not only launched with a great deal of optimism, but also received with enthusiasm, particularly in bureaucratic and administrative milieus. SGML has a strong prescriptive power which makes it well suited for exerting control over the structure of documents. E.g., SGML has already been adopted by the US Defence Department and the EEC's administration.

This kind of optimism and enthusiasm persists also to day. I recently received an invitation to an international workshop on document processing, which says: "...document processing can be fairly assessed as being in a state similar to that obtaining for programming languages prior to the development of syntax- and

semantics-directed compilation techniques, and that obtaining for databases prior to the development of relational and deductive data models. It is time to exploit current techniques and ideas from computer science to raise principles and models of document processing to the same intellectual level as principles and models for programming languages and databases." [From an announcement of "The First International Workshop on Principles of Document Processing" on TEI public discussion list, mid-May 1992]

Surprisingly perhaps, SGML was met with the same kind of enthusiasm in humanistic research disciplines. Already in 1988, the Text Encoding Initiative (TEI) was launched. TEI is a large cooperation project aiming at the establishment of standards for text encoding in the humanities within the framework of SGML.

TEI includes several dozens of text scholars within nearly all humanities disciplines, – ranging from linguistics over philosophy, literature, and history, to classicists and biblicists.

TEI as such does not commit itself to any particular theory of texts, neither are the views expressed by TEI necessarily shared by all members of the project. TEI has actively encouraged the expression of conflicting views, and has been an extremely stimulating forum for the discussion of text theory.

Nevertheless, there are some basic conceptions of text laying behind this project which are in my eyes rather dubious. Though they still seem to persist, these tendencies were particularly clear in the early phases of the project. As the project has proceeded these views have been modified and diversified.

For the purposes of discussion, I will allow myself to construe something which might be called an "early prototypical" TEI view. This view can be expressed in the following theses:

1. To mark up texts descriptively means first and foremost to mark up the logical structure of the texts. In printed or written texts, established conventions of typography or paleography convey the logical structure. Therefore, we should not encode the typography, but the underlying feature.

2. Since traditional typography is inaccurate and unstable, we may also mark up structural elements which are only implied by the text and a result of our subjective interpretation or analysis.

3. In this manner, we will be able to maintain a sharp and clear distinction between the text itself and the encoding. Markup is not itself part of the text but tells us something about it. Markup makes the structural organization and our interpretation and analysis of the text explicit.

4. Although details of text structure differ from genre to genre and from text to text, all texts are hierarchies of linearly ordered objects. In this respect, SGML is well suited for the encoding of texts. Admittedly, some texts contain elements which overlap. In such cases, however, the overlapping elements belong to different hierarchies, and since SGML allows for the coexistence of several hierarchies in one and the same text this poses no technical problem.

(It should be kept in mind, then, that these views are not necessarily representative neither of the TEI (at least not any more), nor of any individual member of TEI. However, see Coombs et al (esp. p 934 and 942-945), DeRose et al, TEI P1, and also to some extent Sperberg-

McQueen 1991. All of those come very close to several of the theses above.)

In opposition to the "early prototypical" TEI view expressed in the above four statements, I will claim:

1. What is regarded as the structure and what as the content of a text depends on the purpose of analysis. Any text may be said to have many kinds of structure (physical, compositional, narrative, grammatical). It is not clear which of these is to be counted as the 'logical' structure. Thus the definition of 'descriptive markup' says nothing, unless we also say what it is that we are describing.
The identification of the "underlying feature" of typographical convention is interpretational. Besides, the relationship between (outer) appearance and (inner) structure is sometimes very close, e.g. in realistic poetry.

2. There are no facts about a text which are objective in the sense of not being interpretational. However, being interpretational does not mean being entirely subjective – there are some things about which all competent readers agree, at least for all practical purposes.
A simple example: We normally regard capitalization and full stop as typographical evidence of a sentence, which is the underlying, 'structural' feature. However, when encoding manuscripts, we often have to decide whether we regard a letter as capitalized or not and a mark as a comma or a full stop, partly on the basis of the visual evidence, partly on our interpretation of the text.

3. Unless further qualified, the notion of making 'the structure' explicit in the codes is of little help, because (a) all structure cannot be made explicit at the same time (there are endlessly many structures), and (b) as soon as something has been made explicit it

has become part of the text, which has thereby changed, and acquired a new structure. 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. It is quite symptomatic that the "text itself" on the TEI view seems to consist roughly of the encoded elements, i.e. that part of a text which occurs between tags. Taking punctuation as an example once again — although a full stop is mostly represented not by a tag but rather as part of a tagged element, it would be highly appropriate to regard it as an indication of an underlying structural feature — the sentence.

4. It is a serious limitation that SGML enforces a prescriptive, top-down approach to text analysis and presupposes hierarchical structures.

Any formal language is bound to have its limitations and to favor certain biases — "...devising a representational system that does not impose but only maps linguistic structures" [Coulmas p 270] is impossible.

If not in theory, then at least in practice, any use of SGML, with its DTDs, invites us to prescribe or predict the structural order of the elements encoded in a text. Since SGML presupposes that the entire text is somehow marked up, this enforces a top-down approach to document design.

Furthermore, SGML presupposes that the design is hierarchical, or alternatively that the text is represented as consisting of a number of concurring hierarchies.

In many cases, a prescriptive, top-down approach presupposing (one or several) hierarchical structure(s) may be well suited to the goals of analysis and composition. In a large number of cases, however, these features of SGML will be detrimental to the purposes of investigation and analysis.

There is a notable tendency in TEI to distinguish between, on the one hand, the information and, on the other, its actual representation on a physical medium. What we seem to be searching for, then, is the key, the "mode of representation", or the specific rules governing the representation of each different kind and feature of textual information on a specific physical medium.

As soon as we have identified these rules, it would seem like an easy task to specify their corollaries for the representation of the same kinds of information on another kind of medium.

However, the kinds and features of information contained in printed texts are probably shaped just as much by the means of expression at our disposal, as vice versa. Our concept of a text is partly a product of the historically mediated knowledge of limits and possibilities of expression posed by the medium carrying texts.

This exemplifies a general point concerning the cultural impact of innovation throughout the history of information technology. Sinding-Larsen makes a similar observation in his studies of the development of musical notation in the medieval ages: "An improvement of the tools for *description* of a certain domain will, in general, also be the starting point for new design and *prescription* which will change the domain originally to be described." [Sinding-Larsen 1988b, p 111.]

When TEI started, I was working on an improvement of the Norwegian Wittgenstein Project's encoding scheme for Wittgenstein's manuscripts [cf Huitfeldt & Rossvær 1989]. The fourth point above, i.e. the prescriptive, top-down approach and hierarchical structure of SGML, convinced me that I had to design a quite different encoding scheme for the Wittgenstein Archives. This lead to the development

of what I have called a Multi-Element Code System (MECS) [Huitfeldt 1992].

MECS is in many respects similar to SGML. As in SGML, codes may be declared in a separate "document definition". The syntax of this document definition lacks much of the expressive power of SGML's DTD. I have therefore found another name for it: code definition table (CDT). Alternatively, codes may be declared simply by using them in the text (in-line declaration). MECS allows for the reconstruction of CDTs (what I have called "minimal CDTs) from encoded texts. MECS does not presuppose any hierarchical structure – any element may overlap with any other element. Finally, MECS includes syntactical means for the representation of structures which in SGML have to be treated in a more roundabout way.

One might say that in SGML everything is forbidden unless it is explicitly permitted or mandatory; while in MECS everything is permitted unless it is explicitly forbidden or mandatory.

Paradoxically, perhaps (since SGML is advocated by so many adherents of so-called "descriptive" markup), SGML is excellent for prescriptive purposes, where the aim is to exert strict control over the structure and content of documents which are still to be created.

MECS, however, is better suited for descriptive purposes. When our aim is to describe already existing documents, we cannot expect to know all about their structure and content in advance. A code system which forces us to prescribe an order in advance may easily lead us to prescribe an order which is perhaps not there in the document at all.

To sum up, I allow myself to characterize MECS, in contrast to SGML, as a code system which encourages a descriptive, bottom-up approach to text analysis, not presupposing a hierarchical structure of texts. This has lead to some in my eyes rather illuminating discussions with other members of TEI.

For example, on the top-down vs. bottom-up approach to text structures and the prescriptive vs. descriptive attitude, my problem with SGML is the following: Designing a registration standard of a project aiming at a machine-readable version of Wittgenstein's manuscripts, we do not want to superimpose a structure on these texts which is not in accordance with a sound interpretation of them. This is precisely the risk we run by predesigning a DTD to which all documents have to conform.

The reaction from other members of TEI has been that I suffer from an illusion that theory-independent gathering of data should be possible. SGML enforces you to make your hypothesis about texts explicit. This does not mean that you may not revise your DTD if you find that your hypothesis was wrong.

However, in our project we are not particularly interested in testing any specific theory about the structure of Wittgenstein's manuscripts in terms of possible structural relationships between text elements encoded in certain specific ways. What we want, is a representation based on a sound interpretation. Therefore, we want the transcriber, who is a highly competent reader, to interpret the text and to mark it in accordance with his interpretation. The transcriber's interpretation is not theory-independent, but it is not couched in terms of markup structures either. An exhaustive description of structural relationships between differently marked-up text elements

may be an interesting by-product, but cannot be the starting point of our work.

My problem with (concurrent) hierarchies is similar: Even if SGML allows me to have several concurrent hierarchies in a text, I am not convinced that Wittgenstein's manuscripts are basically hierarchical structures. Potentially, for all that I know, any feature may overlap with any other feature. Besides, I do not even know what the hierarchies should consist of, or whether the identification of such hierarchies would be particularly illuminating.

Other members of TEI have recently suggested [Renear et al 1992] a very interesting answer to this objection: That two text elements overlap is in itself a criterion that they belong to different conceptual frameworks, theoretical perspectives, or modes of analysis, such as the compositional, the metrical, the physical, the narrative etc.

This view is difficult to assess, since not much specific has so far been said about what a theoretical perspective or conceptual framework is. I have three comments:

1. It is still unclear why such a conceptual framework should demand that the features/elements recognized in a text must be hierarchically ordered. Is this an empirical or an a priori observation? Is the possibility of overlap the only criterion that two features belong to different frameworks? If it *did* turn out that all analyses based on different frameworks do yield different hierarchies, this might be an extremely interesting empirical discovery. But what if it turns out that the hierarchical order is an a priori truth. Would this then be a discovery about our concept of a text?

2. Admittedly, our experience at the Wittgenstein Archives seems to confirm the observation to a very large extent, – in most cases when we find that two features overlap we also find that they are very different kinds of features, e.g. the one belonging to the physical organization of the text (pages, lines), the other e.g. to what we might call the semantic macrostructure (paragraphs, sentences). But this is far from always the case. Some examples will illustrate this:

If a deletion overlaps with an underlining, there is no problem recognizing these features as belonging to very different "perspectives". However, what if two tokens of the same type, e.g. two underlinings, overlap? How could we possibly justify that they belong to different "perspectives"?
Wittgenstein, like many others, used one kind of underlining to indicate emphasis and another kind of underlining to indicate uncertainty or dissatisfaction. These features often overlap. Does that necessarily mean that they belong to different "perspectives"?

Chapters, sections and sentences are normally regarded as features belonging to the same "perspective" (the compositional?). Normally, they form nice hierarchies. But what if we find a chapter break in the middle of a sentence? Should we conclude that contrary to what we believed, sentences and chapters belong to different "perspectives", or should we conclude that what we believed to be one complete sentence divided by a chapter break is really two (perhaps incomplete) sentences, or not sentences at all?

3. Finally, and most importantly, I am struck by the lack of imagination in this approach: Why on earth should texts by all means be hierarchies? No doubt, there are many hierarchical structures, and no doubt this is important, but there are countless other relations between text elements which are worth while finding

and investigating – overlap, substitution, discontinuity, parallel texts, cross-references, etc.

– * –

I will not pursue this discussion any further here. Irrespective of which of the parties are judged to be on the right track, I believe the discussion serves to establish my main point: The use of modern information technology in textual studies may help us reach a better understanding of traditional concepts of and ways of dealing with texts.

This is one of the relationships between our understanding of linguistic phenomena and the development of a new technology. It has been suggested that our language in general has to a large extent been shaped by the technology of writing [Ong, Goody]. It has also been suggested that linguistics draws many of its most basic concepts from features peculiar to written language. [Harris, Coulmas] This is a bit surprising, since at least in the early stages of modern linguistics speech was regarded as the primary form of language.

Linguists have traditionally concentrated on microstructures of language on or well below sentence-level. It is therefore interesting that the recent call (during the last one or two decades) for a concern with larger chunks of language has taken the form of an urge for linguists to concern themselves with texts, and typical that some linguists immediately started to talk about texts as "discourse events". [cf. de Beugrande]. The primacy of the spoken seems to persists, even though most non-linguists would probably regard the written and not the spoken as the primacy of texts.

One of the alternative approaches is also typical of modern linguistics – typical, that is, of its concern with sentences as primary units: The attempt to study texts as strings of sentences displaying a certain degree of cohesion and coherence. Texts must consist of sentences, since they must be grammatically well-formed. And since not any arbitrary collection of sentences constitutes what we would like to call texts, there must be some connection between them – that of cohesion and coherence.

Linguists claim that writing is a secondary form of language – writing represents speech, and does so only more or less successfully. However, Florian Coulmas suggests that the prominence of such objects as phonemes, words and sentences as basic units of linguistic analysis is a reflection not so much of their prominence in speech, but rather of their prominence in writing. While the earliest alphabetic writing systems were scripta continua, and thus had no way of representing word and sentence boundaries, the later invention of punctuation and spacing made writing a more precise tool for the description of these crucial elements of speech. Coulmas points out that those features of speech which are typically relegated to appendixes and play subordinate roles in linguistic text books are precisely those features which have not found any expression in writing, so-called suprasegmental or prosodic features like melody, rhyme, rhythm, and intonation. [Coulmas, p 39-40 and 270]

Linguistics has concentrated on features which have already found their expression in writing, and at least to some extent tended to disregard features which have not. It looks as if linguistics, while claiming speech to be the primary form of language, gets some of its basic concepts of analysis from writing: 'The units of linguistic

analysis are derivative of the units of written language" [Coulmas, p 270]

REFERENCES

Barnard, Hayter, Karababa, Logan, and McFadden:
"SGML-Based Markup for Literary Texts: Two Problems and Some Solutions", in Computers and the Humanities vol 22 (1988), 265-276

de Beaugrande and Dressler:
"Introduction to Text Linguistics", Longman, London 1981

Coombs, Renear, and DeRose:
"Markup Systems and the Future of Scholarly Text Processing" in Communications of the ACM, november 1987, vol 30 (11), p 933-947

Coulmas, Florian:
"The Writing Systems of the World", Basil Blackwell, Oxford 1989, 1991

DeRose, Durand, Mylonas, and Renear:
"What is Text, Really?" in Journal of Computing in Higher Education, Winter 1990, Vol I (2), p 3-26.

Goldfarb, Charles F.:

"Introduction to Generalized Markup", Annex A to ISO 8879-1986. Adapted from "A Generalized Approach to Document Markup", SIGPLAN Notices, June 1981. The annex does not form an integral part of the International Standard.

Goody, Jack:

"The Interface Between the Written and the Oral", Cambridge UP 1987

Halliday, M.A.K., and Hasan, Ruqaiya:

"Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective", Oxford UP 1985, 1990.

Halliday, M.A.K.:

"Spoken and Written Language", Oxford UP 1985, 1990.

Huitfeldt, Claus:

1989 with Viggo Rossvær: "The Norwegian Wittgenstein Project Report 1988", Norwegian Computing Centre for the Humanities, Report Series no 44, Bergen 1989.

1990 "Towards a Machine-Readable Version of Wittgenstein's Nachlaß", unpublished working paper (draft), 23. February 1990.

1991 "Das Wittgenstein-Archiv der Universität Bergen. Hintergrund und erster Arbeitsbericht" in Mitteilungen aus der Brenner-Archiv no 10/1991, p 93-106, Innsbruck 1991.

1992 "MECS – A Multi-Element Code System", forthcoming in Working Papers from the Wittgenstein Archives at the University of Bergen, no 3, 1992

Ide, Nancy, and Sperberg-McQueen, Michael:
"Encoding Guidelines Meeting", ACH Newsletter, 9, 4(1987), 2, 4, 6.

ISO 8879-1986:
"Information Processing – Text and Office Systems Standard Generalized Markup Language (SGML)", International Organization for Standardization, ISO 8879-1986

Kraft, Herbert:
"Editionsphilologie", Wissentschaftliche Buchgesellschaft, Darmstadt 1990

Ong, Walter J.:
"Writing is a Technology that Restructures Thought", in Baumann, Gerd (ed.): "The Written Word", Clarendon Oxford 1986

Renear, Durand, and Mylonas:
"Overlapping Hierarchies of Text Objects: Refining our Notion of What Text Really is" in 'Conference Abstracts and Programme' from the 1992 ALLC-ACH Conference, OUP April 1992.

Sinding-Larsen, Henrik:
1988a "Information Technology and the Externalization of Knowledge", in Sinding-Larsen (ed) "Artificial Intelligence and Language", Tano, Oslo 1988, pp 77-89
1988b "Notation and Music: The History of a Tool of Description and its Domain to be Described", in Sinding-Larsen "Artificial Intelligence and Language", Tano, Oslo 1988, pp 90-114.

Sperberg-McQueen, Michael:

"Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts" in Literary and Linguistic Computing, vol 6 no 1, 1991, Oxford UP

TEI EDP1:

"Design Principles for Text Encoding Guidelines", Text Encoding Initiative, document number TEI EDP1, 18.07.89

TEI P1:

C.M. Sperberg-McQueen and Lou Burnard (eds.): "Guidelines for the Encoding and Interchange of Machine-Readable Texts", Draft Version 1.1, Chicago and Oxford November 1990

Wittgenstein, Ludwig:

"Tractatus Logico-Philosophicus", Routlege and Kegan Paul, London 1922, 1966