Encoding Standards for Linguistic Corpora

Nancy Ide

Department of Computer Science
Vassar College
Poughkeepsie, New York 12601 (U.S.A.)

and

Laboratoire Parole et Langage
CNRS/Université de Provence
29, Avenue Robert Schuman
13621 Aix-en-Provence Cedex 1 (France)

e-mail: ide@cs.vassar.edu/ide@univ-aix.fr

Abstract. The Text Encoding Initiative (TEI) is an international project established in 1988 to develop guidelines for the preparation and interchange of electronic texts for research, and to satisfy a broad range of uses by the language industries more generally. The need for standardized encoding practices has become inxreasingly critical as the need to use and, most importantly, reuse vast amounts of electronic text has dramatically increased for both research and industry, in particular for natural language processing. In May 1994, the TEI issued its *Guidelines for the Encoding and Interchange of Machine-Readable Texts*, which provide standardized encoding conventions for a large range of text types and features relevant for a broad range of applications.

Keywords. Encoding, markup, large text resources, corpora, SGML.

1. Introduction

The past few years have seen a burst of activity in the development of statistical methods which, applied to massive text data, have in turn enabled the development of increasingly comprehensive and robust models of language structure and use. Such models are increasingly recognized as an invaluable resource for natural language processing (NLP) tasks, including machine translation.

The upsurge of interest in empricial methods for language modelling has led inevitably to a need for massive collections of texts of all kinds, including text collections which span genre, register, spoken and written data, etc., as well as domain- or application-specific collections, and, especially, multi-lingual collections with parallel translations. In the latter half of the 1980's, very few appropriate or adequately large text collections existed for use in computational linguistics research, especially for languages other than English. Consequently, several efforts to collect and disseminate large mono- and multi-lingual text collections have been recently established, including the *ACL Data Collection Initiative (ACL/DCI)*, the *European Corpus Initiative (ECI),* the U.S. *Linguistic Data Consortium (LDC), MULTEXT* in Europe, etc. It is widely recognized that such efforts constitute only a beginning for the necessary data collection and dissemination efforts, and that considerable work to develop adequately large and appropriately constituted textual resources still remains.

The demand for extensive reusability of large text collections in turn requires the development of standardized encoding formats for this data. It is no longer realistic to distribute data in *ad hoc* formats, since the effort and resources required to clean up and reformat the data for local use is at best costly, and in many cases prohibitive. Because much existing and potentially available data was originally formatted for the purposes of printing, the information explicitly represented in the encoding concerns a particular physical realization of a text rather than its logical structure (which is of greater interest for most NLP applications), and the correspondence between the two is often difficult or impossible to establish without substantial work. Further, as data become more and more available and the use of large text collections become more central to NLP research, general and publicly available software to manipulate the texts is being developed which, to be itself reusable, also requires the existence of a standard encoding format.

A standard encoding format adequate for representing textual data for NLP research must be (1) capable of representing the different kinds of information across the spectrum of text types and languages potentially of interest to the NLP research community, including

prose, technical documents, newspapers, verse, drama, letters, dictionaries, lexicons, etc.; (2) capable of representing different levels of information, including not only physical characterstics and logical structure (as well as other more complex phenomena such as intra- and inter-textual references, alignment of parallel elements, etc.), but also interpretive or analytic annotation which may be added to the data (for example, markup for part of speech, syntactic structure, etc.); (3) application independent, that is, it must provide the required flexibility and generality to enable, possibly simultaneously, the explicit encoding of potentially disparate types of information within the same text, as well as accomodate all potential types of processing. The development of such a suitably flexible and comprehensive encoding system is a substantial intellectual task, demanding (just to start) the development of suitably complex models for the various text types as well as an overall model of text and an architecture for the encoding scheme that is to embody it.

2. The Text Encoding Initiative

In 1988, the Text Encoding Initiative (TEI) was established as an international co-operative research project to develop a general and flexible set of guidelines for the preparation and interchange of electronic texts. The TEI is jointly sponsored by the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. The project has had major support from the U.S. National Endowment for the Humanities (NEH), Directorate XIII of the Commission of the European Communities (CEC/DG-XIII), the Andrew W. Mellon Foundation, and the Social Science and Humanities Research Council of Canada.

In May 1994, the TEI issued its *Guidelines for the Encoding and Interchange of Machine-Readable Texts*, which provide standardized encoding conventions for a large range of text types and features relevant for a broad range of applications, including natural language processing, information retrieval, hypertext, electronic publishing, various forms of literary and historical analysis, lexicography, etc. The Guidelines are intended to apply to texts, written or spoken, in any natural language, of any date, in any genre or text type, without restriction on form or content. They treat both continuous materials (*running text*) and discontinuous materials such as dictionaries and linguistic corpora. As such, the TEI Guidelines answer the fundamental needs of a wide range of users: researchers in computational linguistics, the humanities, sciences, and social sciences; publishers; librarians and those concerned generally with document retrieval and storage; as well as the growing language technology community, which is amassing

substantial multi-lingual, multi-modal corpora of spoken and written texts and lexicons in order to advance research in human language understanding, production, and translation.

The rules and recommendations made in the TEI Guidelines conform to the ISO 8879, which defines the Standard Generalized Markup Language, and ISO 646, which defines a standard seven-bit character set in terms of which the recommendations on character-level interchange are formulated.[1] SGML is an increasingly widely recognized international markup standard which has been adopted by the US Department of Defense, the Commission of European Communities, and numerous publishers and holders of large public databases.

2.1. Overview

Prior to the establishment of the TEI, most projects involving the capture and electronic representation of texts and other linguistic data developed their own encoding schemes, which usually could only be used for the data for which they were designed. In many cases, there had been no prior analysis of the required categories and features and the relations among them for a given text type, in the light of real and potential processing and analytic needs. The TEI has motivated and accomplished the substantial intellectual task of completing this analysis for a large number of text types, and provides encoding conventions based upon it for describing the physical and logical structure of many classes of texts, as well as features particular to a given text type or not conventionally represented in typography. The TEI Guidelines also cover common text encoding problems, including intra- and inter-textual cross reference, demarcation of arbitrary text segments, alignment of parallel elements, overlapping hierarchies, etc. In addition, they provide conventions for linking texts to acoustic and visual data.

The TEI's specific achievements include:

1. a determination that the Standard Generalized Markup Language (SGML) is the framework for development of the Guidelines;

---

[1] For more extensive discussion of the project's history, rationale, and design principles see TEI internal documents EDP1 and EDP2 (available from the TEI) and (Ide and Sperberg-McQueen 1995: 5-15) and (Sperberg-McQueen and Burnard 1995: 17-39), both published in a special triple issue on the TEI in *Computers and the Humanities.*

2. the specification of restrictions on and recommendations for SGML use that best serves the needs of interchange, as well as enables maximal generality and flexibility in order to serve the widest possible range of research, development, and application needs;

3. analysis and identification of categories and features for encoding textual data, at many levels of detail;

4. specification of a set of general text structure definitions that is effective, flexible, and extensible;

5. specification of a method for in-file documentation of electronic texts compatible with library cataloging conventions, which can be used to trace the history of the texts and thus assist in authenticating their provenance and the modifications they have undergone;

6. specification of encoding conventions for special kinds of texts or text features, including:

    a. character sets
    b. language corpora
    c. general linguistics
    d. dictionaries
    e. terminological data
    f. spoken texts
    g. hypermedia
    h. literary prose
    i. verse
    j. drama
    k. historical source materials
    l. text critical apparatus

3. Basic architecture of the TEI scheme

3.1. General architecture

The TEI Guidelines are built on the assumption that there is a common core of textual features shared by virtually all texts, beyond which many different elements can be

encoded. Therefore, the Guidelines provide an extensible framework containing a common core of features, a choice of frameworks or bases, and a wide variety of optional additions for specific applications or text types. The encoding process is seen as incremental, so that additional markup may be easily inserted in the text.

Because the TEI is an SGML application, a TEI conformant document must be described by a *document type definition (DTD),* which defines tags and provides a BNF grammar description of the allowed structural relationships among them. A TEI DTD is composed of the *core tagsets,* a single *base tagset,* and any number of user selected *additional tagsets,* built up according to a set of rules documented in the TEI Guidelines.

At the highest level, all TEI documents conform to a common model. The basic unit is a *text,* that is, any single document or stretch of natural language regarded as a self-contained unit for processing purposes. The association of such a unit with a *header* describing it as a bibliographic entity is regarded as a single TEI element. Two variations on this basic structure are defined: a collection of TEI elements, or a variety of composite texts. The first is appropriate for large disparate collections of independent texts, for example in language corpora, or collections of unrelated papers in an archive; the second applies to cases such as the complete works of a given author, which might be regarded simultaneously as a single text in its own right and as a series of independent texts.

Often, it is necessary to encode more than one view of a text--for example, the physical and the linguistic or the formal and the rhetorical. One of the essential features of the TEI Guidelines is that they offer the possibility to encode many different views of a text, simultaneously if necessary. A disadvantage of SGML is that it uses a document model consisting of a single hierarchical structure; often, different views of a text define multiple, possibly overlapping hierarchies (for example, the physical view of a print version of a text, consisting of pages sub-divided into physical lines, and the logical view consisting of, say, paragraphs sub-divided into sentences) which are not readily accomodated by SGML's document model. The TEI has identified several possible solutions to this problem in addition to SGML's concurrent structures mechanism, which, because of the processing complexity it involves, is not a thoroughly satisfactory alternative.

The TEI Guidelines provide sophisticated mechanisms for linking and alignment of elements, both within a given text and between texts, as well as links to data not in the form of ASCII text such as sound and images. Much of the TEI work on linkage was accomplished in collaboration with those working on the Hypermedia/Time-based

Document Structuring Language (HyTime), recently adopted as an SGML-based international standard for hypermedia structures.

3.2. The TEI base tagsets

Eight distinct TEI base tagsets are proposed:

1. prose
2. verse
3. drama
4. transcribed speech
5. letters or memos
6. dictionary entries
7. terminological entries
8. language corpora and collections

The first seven are intended for documents which are predominantly composed of one type of text; the last is provided for use with texts which combine these basic tagsets. Additional base tag sets will be provided in the future.

Each TEI base tagset determines the basic structure of all the documents with which it is to be used. More exactly, it defines the components of text elements, combined as described above. In practice, so far, almost all the TEI bases defined are similar in their basic structure, though they can vary if necessary. However, they differ in their components: for example, the kind of sub-elements likely to appear within the divisions of a dictionary will be entirely different from those likely to appear within the divisions of a letter or a novel. To accomodate this variety, the constituents of all divisions of a TEI text element are not defined explicitly, but in terms of SGML *parameter entities,* which behave similar to a variable declaration in a programming language: the effect of using them here is that each base tag set can provide its own specific definition for the constituents of texts, which can, moreover, be modified by the user.

3.3. The core tagsets

Two core tagsets are available to all TEI documents unless explicitly disabled. The first defines a large number of elements which may appear in any kind of document--coinciding more or less with that set of discipline-independent textual features concerning

which consensus has been reached. The second defines the header, providing something analogous to an electronic title page for the electronic text.

The core tagset common to all TEI bases provides means of encoding with a reasonable degree of sophistication the following list of textual features:

1. Paragraphs
2. Segmentation, for example into orthographic sentences.
3. Lists of various kinds, including glossaries and indexes
4. Typographically highlighted phrases, whether unqualified or used to mark linguistic emphasis, foreign words, titles etc.
5. Quoted phrases, distinguishing direct speech, quotation, terms and glosses, cited phrases etc.
6. Names, numbers and measures, dates and times, and similar data-like phrases.
7. Basic editorial changes (e.g. correction of apparent errors; regularization and normalization; additions, deletions and omissions)
8. Simple links and cross references, providing basic hypertextual features.
9. Pre-existing or generated annotation and indexing
10. Passages of verse or drama, distinguishing for example speakers, stage directions, verse lines, stanzaic units, etc.
11. Bibliographic citations, adequate for most commonly used bibliographic packages, in either a free or a tightly structured format
12. Simple or complex referencing systems, not necessarily dependent on the existing SGML structure.

There are few documents which do not exhibit some of these features, and none of these features is particularly restricted to any one kind of document. In most cases, additional more specialized tagsets are provided to encode aspects of these features in more detail, but the elements defined in this core should be adequate for most applications most of the time.

Features are categorized within the TEI scheme based on shared attributes. The TEI encoding scheme also uses a classification system based upon structural properties of the elements, that is, their position within the SGML document structure. Elements which can appear at the same position within a document are regarded as forming a *model class*: for example, the class *phrase* includes all elements which can appear within paragraphs but not spanning them, the class *chunk* includes all elements which cannot appear within

paragraphs (e.g., paragraphs), etc. A class *inter* is also defined for elements such as lists, which can appear either within or between chunk elements.

Classes may have super- and sub-classes, and properties (notably, associated attributes) may be inherited. For example, reflecting the needs of many TEI users to treat texts both as documents and as input to databases, a sub-class of phrase called *data* is defined to include data-like features such as names of persons, places or organizations, numbers and dates, abbreviations and measures. The formal definition of classes in the SGML syntax used to express the TEI scheme makes it possible for users of the scheme to extend it in a simple and controlled way: new elements may be added into existing classes, and existing elements renamed or undefined, without any need for extensive revision of the TEI document type definitions.

3.4. The TEI header

The TEI header is believed to be the first systematic attempt to provide in-file documentation of electronic texts. The TEI header allows for the definition of a full AACR2-compatible bibliographic description for the electronic text, covering all of the following:

1. the electronic document itself
2. sources from which the document was derived
3. encoding system
4. revision history

The TEI header allows for a large amount of structured or unstructured information under the above headings, including both traditional bibliographic material which can be directly translated into an equivalent MARC catalogue record, as well as descriptive information such as the languages it uses and the situation within which it was produced, expansions or formal definitions for any codebooks used in analyzing the text, the setting and identity of participants within it, etc. The amount of encoding in a header depends both on the nature and the intended use of the text. At one extreme, an encoder may provide only a bibliographic identification of the text. At the other, encoders wishing to ensure that their texts can be used for the widest range of applications can provide a level of detailed documentation approximating to the kind most often supplied in the form of a manual.

A collection of TEI headers can also be regarded as a distinct document, and an auxiliary DTD is provided to support interchange of headers alone, for example, between libraries or archives.

3.5. Additional tagsets

A number of optional additional tagsets are defined by the Guidelines, inlcuding tagsets for special application areas such as alignment and linkage of text segments to form hypertexts; a wide range of other analytic elements and attributes; a tagset for detailed manuscript transcription and another for the recording of an electronic variorum modelled on the traditional critical apparatus; tagsets for the detailed encoding of names and dates; abstractions such as networks, graphs or trees; mathematical formulae and tables etc.

In addition to these application-specific specialized tagsets, a general purpose tagset based on feature structure notation is proposed for the encoding of entirely abstract interpretations t, either in parallel or embedded within it. Using this mechanism, encoders can define arbitrarily complex bundles or sets of features identified in a text. The syntax defined by the Guidelines formalizes the way in which such features are encoded and provides for a detailed specification of legal feature value/pair combinations and rules (a *feature system declaration*) determining, for example, the implication of under-specified or defaulted features. A related set of additional elements is also provided for the encoding of degrees of uncertainty or ambiguity in the encoding of a text.

A user of the TEI scheme may combine as many or as few additional tagsets as suit his or her needs. The existence of tagsets for particular application areas in the Guidelines reflects, to some extent, accidents of history: no claim to systematic or encyclopedic coverage is implied. It is expected that new tagsets will be defined as a part of the continued work of the TEI and in related projects.[2]

---

[2] For example, the European project MULTEXT, in collaboration with EAGLES, is developing a specialized Corpus Encoding Standard for NLP applications based on the TEI Guidelines.

4. Information about the TEI

The TEI Guidelines for Electronic Text Encoding and Interchange are available as follows:

- in paper (1300 pp., 2 volumes), at a cost of $75 US or 50 pounds sterling, sent to:

TEI Orders
Oxford University Computing Services
13 Banbury Road
Oxford OX2 6NN

- electronically via the World Wide Web at the following sites:

http://www-tei.uic.edu/orgs/tei
http://etext.virginia.edu/TEI.html

- electronically via anonymous ftp from any of the following:

ftp-tei.uic.edu (in pub/tei and its subdirectories)
sgml1.ex.ac.uk (in tei/p3 and its subdirectories)
ftp.ifi.uio.no (in pub/SGML/TEI)

- electronically in formatted ASCII-only via Listserv, by sending electronic mail to listserv@uicvm.uic.edu containing the following line:

get p3ascii package

The TEI also maintains a publicly-accessible ListServ list, TEI-L, housed at the University of Illinois at Chicago. To subscribe, send electronic mail to **listserv@uicvm.uic.edu** containing the text **Subscribe TEI-L J. Q. Public** (substitute your name for "J. Q. Public")

Additional information can be obtained by contacting one of the TEI editors:

C. M. Sperberg-McQueen
University of Illinois at Chicago (M/C 135)
Computer Center
1940 W. Taylor St.
Chicago, Illinois 60612-7352 US
email: u35395@uicvm.uic.edu
tel: +1 (312) 413-0317
fax: +1 (312) 996-6834

Lou Burnard
Oxford University Computing Service
13 Banbury Road
Oxford OX26NN
United Kingdom
email: lou@vax.ox.ac.uk
tel: +44 (865) 273200
fax: +44 (865) 273275

Acknowledgments -- The TEI has been funded by the U.S. National Endowment for the Humanities (NEH), Directorate XIII of the Commission of the European Communities (CEC/DG-XIII), the Andrew W. Mellon Foundation, and the Social Science and Humanities Research Council of Canada. Some material in this paper has been adapted from other TEI documents written by the TEI editors Michael Sperberg-McQueen and Lou Burnard, and chairs and members of various TEI committees.

References

Bryan, M. 1988. *SGML: An Author's Guide,* New York: Addison-Wesley.

Coombs, J.H., Renear, A.H., and DeRose, S.J. 1987. "Markup systems and the future of scholarly text processing". *Communications of the ACM* 30(11): 933-947.

Goldfarb, C.F. 1990.*The SGML Handbook.* Oxford: Clarendon Press.

Ide, N., Sperberg-McQueen, C.M. 1995. "The Text Encoding Initiative: Its History, Goals, and Future Development". *Computers and the Humanitites* (Special Issue on the Text Encoding Initiative) 29(1): 5-15.

Ide, N., Véronis, J. (eds.) 1995. *Computers and the Humanitites* (Special Issue on the Text Encoding Initiative) 29(1-3).

International Organization for Standards 1986. ISO 8879: Information Processing--Text and Office Systems--Standard Generalized Markup Language (SGML). Geneva: ISO.

International Organization for Standards 1992. ISO/IEC DIS 10744: Hypermedia/Time-based Document Structuring Language (Hytime). Geneva: ISO.

Sperberg-McQueen, C.M., Burnard, L. 1995. "The Design of the TEI Encoding Scheme". *Computers and the Humanitites* (Special Issue on the Text Encoding Initiative) 29(1): 17-39.

Sperberg-McQueen, C.M., Burnard, L. 1994. *Guidelines for Electronic Text Encoding and Interchange.* Chicago and Oxford: Text Encoding Initiative.

van Herwijnen, E. 1991. *Practical SGML*. Dordrecht: Kluwer Academic Publishers.

Appendix : TEI Guidelines Table of Contents