Documents in Language Engineering : Application of the Text Encoding Initiative

Nancy Ide

Laboratoire Parole et Langage
CNRS/Université de Provence
29, Avenue Robert Schuman
13621 Aix-en-Provence Cedex 1 (France)
ide@univ-aix.fr

Abstract. Treatment of texts in the language industries, which is fast becoming a primary activity due to the increased availability of large quantities of texts in electronic form, places stringent demands on the requirements for data representation, such as the ability to maintain and provide access simultaneously to information about a document's physical, logical, linguistic, etc. structures. The aim of this presentation is to outline the document representation needs of language engineering and demonstrate the applicability of the Text Encoding Initiative's (TEI) Guidelines for encoding electronic documents answer these needs.

Keywords. Standardization, reusability, document structures, TEI, SGML

1. Introduction

Language engineering has specific demands for natural language processing (NLP) tasks, often very different from those applied in more theoretically-based academic work in the area. A dominant example is the recent development and application 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 NLP tasks such as machine translation, terminology extraction, etc. They are of particular interest for industry because they are robust, and can most often be implemented using relatively simple engineering techniques, which is an important factor in the industrial environment. The upsurge in the application of empirical methods has led necessarily 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) and for standardized methods for representing these texts and multiple kinds of information within them, in order to enable their widespread reusability.

Industrial applications typically deal with actual texts which may be very complex and highly structured. The encoding of such texts often requires the representation of a variety of different types of information, including its original *physical* format (font renderings, page breaks, line breaks, etc.) and its *logical* structure (sections, subsections, titles, figures, notes, etc.), as well as other more complex phenomena such as intra- and inter-textual references, alignment of parallel elements, etc. Most industrial applications also enhance the information in the text to include interpretive or analytic annotation which may be added to the data (for example, markup for part of speech, syntactic structure, etc.). These different *views* of a text as a physical, logical, linguistic, etc. object are independent: in general; i.e., the elements of one view do not coincide with or depend on the elements of another (e.g., physical page breaks are not coincident with or dependent upon paragraph breaks, etc.).

Unlike academic work in computational linguistics, industrial applications typically demand that multiple views of a document are accessible to the application, or at least, are preserved throughout the course of treatment of the document. For example, in automatic translation, it is obviously the goal to translate the *linguistic objects* in the text; but in many applications it is necessary at the same time to retain the physical and logical features of the base text--including page breaks, font renderings, section numbering, etc.-- throughout the translation process, in order to duplicate them in the translated version of the document. The need to retain simultaneous access to different views of a document is evident in the process of morphosyntactic annotation, in order to enable intelligent retrieval taking into account, at the same time, linguistic information (for example, sequences of "N1 prep N2" for terminological extraction) and the logical structure of the document (for example, such sequences found in titles might be given a higher probability of constituting actual terms). It is also possible to imagine utilizing information about the physical and logical views of a document for NLP itself, possibly to activate different tools or processes for different text components (e.g., titles do not follow the same syntactic rules as the content of paragraphs), or to use such information for processing (for example, physical rendering in italics can signal a foreign word). The examples can be multiplied: the need to take into account document updates, the association of NLP with other kinds of processing (e.g., information retrieval, hypermedia), etc.

In short, language engineering demands from an encoding scheme the ability to represent potentially very complex documents. In addition, a representation scheme adequate for language engineering applications be reusable, and therefore it must be standardized as well as easily extendible. It must also be capable of representing the different kinds of information across the spectrum of text types and languages potentially of interest to language engineering applications, including prose, technical documents, newspapers, verse, drama, letters, dictionaries, lexicons, etc. Finally, it must be 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 accommodate all potential types of processing.

These document representation needs of language engineering are met by the scheme developed by the Text Encoding Initiative (TEI), which is rapidly becoming a standard for representing electronic texts for use in research and industry. The following sections describe the TEI and provide a brief overview of the encoding scheme it has developed, pointing out its features of greatest importance for language engineering applications.

2. The TEI

In 1988, the 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 January 1994, the TEI issued its *Guidelines for the Encoding and Interchange of Machine-Readable Texts* (Sperberg-McQueen and Burnard, 1994), 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. Initially aimed at the needs of scholarly and industrial research, the Guidelines apply to a broad range of uses by the 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 TEI Guidelines formally specify a detailed and extensible encoding system based on SGML, the Standard Generalized Markup Language (ISO 8879), 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.

3. Achievements of the TEI

The TEI began by identifying the encoding needs for interchange and for the varied processing and analysis requirements of the academic and research communities. It then established encoding principles appropriate to a general purpose model and specified key text types and features for which encoding conventions needed to be developed. The TEI's specific achievements include:

- determination that the Standard Generalized Markup Language (SGML) is the best available framework for development of a general-purpose encoding scheme, and thus for the *Guidelines*;
- the definition of specific restrictions on and recommendations concerning the use of SGML which best serves the needs of interchange, while guaranteeing maximal generality and flexibility in order to serve the widest possible range of research and application needs;
- analysis and identification of categories and features for encoding textual data, at many levels of detail;
- specifications for ways of representing arbitrary interpretive and analytic structures within textual resources without needing to extend basic SGML syntax
- specifications for the integration of non-textual objects such as figures, graphs, tables, and multimedia resources within complex documents
- specification of a set of general text structure definitions that is effective, flexible, and extensible;
- detailed specification of a method for in-file documentation of electronic resources that is compatible with existing library cataloging practices
- an integrated set of conventions for the encoding of meta-textual information such as version control, topic indexing, code books etc.

---

[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 (1994) and Burnard and Sperberg-McQueen (1994), both forthcoming in a special triple issue on the TEI in *Computers and the Humanities.*

- specification of encoding conventions for special kinds of texts or text features, in addition to the material usually handled by existing document production systems, (for example, quotations, footnotes, bibliographic references, editorial corrections). These include:

    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

4. Overview of the TEI scheme

4.1. 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. The TEI differs from other general schemes which might be considered for use in language engineering applications (e.g. ISO 12083 or HTML) in providing much more extensive facilities for customizing the DTD, whether by restricting or extending it, or by translating the element identifiers into other languages than English, and in defining specialized modules for areas not addressed at all by other existing general text markup languages. 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.

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, for example, 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.

4.2. The TEI base tagsets

Eight distinct TEI base tagsets are provided:

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. Specifically, it defines the components of text elements, combined as described above. Almost all the TEI bases defined are similar in their basic structure (although 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 be modified by the user if desired.

4.3. 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, and which coincide more or less with the 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.

4.4. 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.

4.5. 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; 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 included for the encoding of entirely abstract interpretations of a text, 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. 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.

5. Conclusion

The TEI scheme places at the disposal of designers and implementers of language engineering applications a complete and customizable system for representing and managing both textual and other resources in electronic form. In summary, the TEI scheme is *flexible*, in that it

- comprises a set of guidelines adaptable to the needs of different application areas and projects, rather than a single set of rigid requirements
- can be used for texts in any natural language
- is not biased towards any one type of processing or application, but can be used for any type of processing desired
- allows text, images, and other types of material to be combined into complex information objects for multimedia or other presentation
- allows the application to record as much or as little information about the text as is desired: handles simple applications simply, but can also support realistically complex applications
- can be used equally well be used for the creation of new electronic texts and the transcription of existing texts into electronic form
- allows for user-defined extensions to the encoding scheme, to simplify the customization of the scheme and its use in specialized application areas

Furthermore, the TEI scheme is "future-proof", by virtue of being both application- and system-independent. It

- provides a rigorous formal definition of textual structures
- does not depend on a single hardware or software platform
- can be used in interchanging documents among unlike machines, software, or projects

Finally, the TEI is unique in the sophistication of the facilities it includes for resource-discovery and bibliographic control. The TEI scheme defines methods compatible with current library practice, for documenting electronic materials of any kind in a "header" which can be used :

- for purposes of bibliographic control
- to assist in the management of text, images, time-based or other electronic resources in a controlled way
- to provide data for network information discovery and retrieval purposes
- as a prototype for a general-purpose metadata information structure

In the long run, the needs of language engineering can only be met with a markup language rich enough to handle many different kinds of text and many different kinds of application; it must be well documented, so as to make it easier for different applications to inter-operate on the same data; it must be flexible enough to handle simple tasks simply, but must also scale up and provide the structure needed to handle complex tasks reliably. Such goals require the use of a common markup language where possible, to ensure that data can be used by many different applications. The TEI is the best candidate for such a general purpose mark-up language yet to have appeared.

References

Ide, N. (1994) Encoding standards for large text resources: The Text Encoding Initiative. *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94,* Kyoto, Japan, August 1994.

Ide, N., Sperberg-McQueen, C.M. (1995) The Text Encoding Initiative: History and Background. Special Issue on the Text Encoding Initiative, *Computers and the Humanitites,* 29, 1-3 , (in press).

Ide, N., Veronis, J. (Eds.) (1995) Special Issue on the Text Encoding Initiative, *Computers and the Humanities,* 29, 1-3 (in press).

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

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

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