Representation schemes for language data: the Text Encoding Initiative and its potential impact for encoding African languages

Nancy Ide

Department of Computer Science
Vassar College
Poughkeepsie, New York 12601 (U.S.A.)
e-mail: ide@cs.vassar.edu

and

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

Keywords. text encoding, electronic texts, language resources, Text Encoding Initiative.

Abstract. The Text Encoding Initiative (TEI)*Guidelines for the Encoding and Interchange of Machine-Readable Texts* provide standardized encoding conventions for a large range of text types and features relevant for a broad range of applications. Given the potential challenges of encoding texts in the African languages, it will be important to establish collaboration between the TEI and projects encoding language resources in these languages.

Résumé. Les *Recommandations pour l'Encodage et l'Echange des Textes Informatisés* de la Text Encoding Initiative (TEI) fournissent des conventions d'encodage pour un large éventail de types de textes et d'éléments textuels utiles pour de nombreuses applications. L'encodage des textes africains pose des difficultés particulières, et il est important d'établir des collaborations entre la TEI et les projets qui encodent des ressources linguistiques et textuelles pour les langues africaines.

1. Introduction

The explosion in the amount and different types of texts and other language data that are now being rendered in electronic form and, in many cases, being made available in the Internet and via the World Wide Web, has led to concerted efforts to study and develop adequate and appropriate representation schemes for language data. Above all, the need for encoding standards for machine readable texts and language data has been recognized. Without encoding standards, we risk the proliferation of language data in multiple formats which cannot be processed with common software, and which are incompatible for use in common applications.

The development of substantial electronic language data in the various African languages is likely to explode with equal force in the near future. These data will be used for a variety of purposes, including:

- the development of natural language applications for African languages, such as machine translation;
- preservation and access to national literatures and linguistic resources;
- development of tools such as spell checkers, grammar checkers, etc.;
- development of computer-assisted language learning software;
- creation of term banks, lexicons, dictionaries, text bases;
- book publication, and electronic publishing;
- development of document libraries for access via the Internet and the World Wide Web.

Unlike many of the electronic text resources in western European languages, which were rendered into machine readable form over a decade ago in a variety of formats which have proved very difficult to re-use or translate into more usable representation schemes, there is no pre-existing large body of electronic texts in African languages. Therefore, the vast majority of electronic texts in these languages will be created over the next few years. This presents a unique opportunity to create these resources using a standard encoding format that can ensure maximum reusability and flexibility.

A standard encoding format adequate to represent a broad range of potentially complex textual data must be (1) capable of representing the different kinds of information across the spectrum of text types and languages, 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, grammatical information, notation, editorial comment, etc. as well as links to other documents and media); (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 many potential types of processing. The development of such a flexible and comprehensive encoding system is a substantial intellectual task, demanding first of all 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.

In 1994, the Text Encoding Initiative (TEI) published a set of Guidelines for text encoding explicitly intended to meet these criteria, in order to provide a standard representation scheme with the power and flexibility to cover the full range of text types and applications. The TEI was and is a collaborative, on-going project; there remains technical work to be done in simplifying and rationalizing some of the exisiting features, and further tag set development is required in some areas. In particular, the scheme must now be extended to accomodate languages other than Western European languages and deal with problems posed by different text types and encoding requirements. Therefore, the TEI is actively seeking collaboration with projects creating electronic resources in these other languages. The potential for collaboration with African projects is especially interesting, since these languages demand consideration of previously unaddressed encoding problems. This paper provides an overview of the TEI and its encoding scheme, and discusses its application to encoding language data in African languages, in order to open the possibility of such collaboration.

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 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;
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: character sets, language corpora, general linguistics, dictionaries, terminological data, spoken texts, hypermedia, literary prose, verse, drama, historical source materials, text critical apparatus

3. Basic architecture of the TEI scheme

3.1. General architecture

From the beginning of the project it was clear that the specification produced by the TEI would be very large. It was also clear that most documents would only make use of some parts of the specification. It was thus necessary to structure the specifications in such a way that users could select those parts that apply to a specific document. The mechanism for making this selection should be easy to understandand use, and it should not require any formalism or software beyond what is provided in SGML.

The TEI Guidelines are built on the assumption that there is a common core of textual features shared by virtually all texts. Beyond this, the Guidelines provide sets of features in several categories, each of which is encoded using a set of SGML tags. Therefore, the Guidelines provide an extensible framework containing a common core of features, a choice of frameworks or bases,

---

[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).

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. In general, the full tagset for a given document is put together in such a way that sets of tags can be included or excluded from it, and thus the tags are allowed in a document or prohibited, respectively.

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.

A critical need for encoding many text is to provide for encoding multiple *views* of the 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.

Here is a simple document encoded using the TEI scheme:

```
 <!DOCTYPE tei.2 system 'tei2.dtd' [
   <!ENTITY % TEI.prose 'INCLUDE'>
   <!ENTITY english.wsd system 'teien.wsd' SUBDOC>
]>
<tei.2>
<teiHeader>
  <fileDesc>
    <titleStmt><title>Short document.</title>
    <publicationStmt><p>Unpublished.</p></publicationStmt>
    <sourceDesc><p>Electronic form is original.</p></sourceDesc>
  </fileDesc>
  <profileDesc>
    <langUsage><language id=eng wsd=english.wsd></langUsage>
  </profileDesc>
</teiHeader>
<text>
  <body>
    <p>A very short TEI document.</p>
  </body>
</text>
</tei.2>
```

The first line of this document identifies the document type by pointing to an external definition. The next two lines (contained within the square brackets on lines 1 and 4) contain definitions that override those in the external definition; their purpose here is to include the definitions of the tag set for prose documents, and to indicate that the document being encoded is written in English. From line 5 through to the end of the document, there are two parts. The *teiHeader* contains information describing the electronic document; this is akin to an extended bibliographic description. The *text* contains the encoded text itself; in this case, there is a single line of text.

Realistic documents have these same constituent parts: a reference to the TEI definitions, some local slections of parts of those definitions, a header, and the text itself.

3.2. The TEI base tagsets

There is a *base tag set* for each major document category that can be encoded. These are:

- prose (contains only the core features)
- verse (structure within lines, rhyme, metrical structure)
- drama (cast lists, performances, stage directions)
- transcribed speech (utterances, pauses, temporal information)
- dictionaries (structure of entries, grammatical and typographical information)
- terminological databases (terms and definitions)

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. The TEI bases defined are similar in their basic structure, though 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. Some documents are more complex; they contain material that is from more than one of these bases. For example, a critical essay that included lengthy quotations from poems under discussion would require both prose and verse structures. There are mechanisms for combining base tag sets in a single document; the details can be found in the Guidelines.

To accomodate the variety of text features that might be encoded, the constituents of all divisions of a TEI text element are not defined explicitly, but in terms of SGML *parameter entities*: 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. Modular construction is a familiar concept. Procedural, functional and object oriented programming languages all have mechanisms for combining parts into a larger whole in a disciplined manner. In particular, object oriented notions of instantiation and inheritance support such constructions in a natural way. However, SGML was designed with very limited abstraction mechanisms; there is a SUBDOC construction allowing the inclusion of portions of documents conforming to a separate DTD, but it does not serve the purpose here.

To implement the modular structure that was required, the TEI DTD uses marked sections. There is a marked section for each of the tag sets. The section has a guard that is a parameter entity--i.e., a string-valued variable that is defined in the DTD. When the value of the guarding entity is set to IGNORE, the marked section is not included as the DTD is parsed; when the value of the guarding entity is set to INCLUDE, the marked section is used. By default, all of the entities are set to IGNORE. The user explicitly resets the value of an entity to INCLUDE to select a tag set. Here is an example:

```
<!DOCTYPE TEI.2 system 'tei2.dtd' [
<!--  base tag set                -->
<!ENTITY % TEI.prose       'INCLUDE'>
<!--  additional tag sets         -->
<!ENTITY % TEI.names.dates 'INCLUDE'>
<!ENTITY % TEI.figures     'INCLUDE'>
]>
```

This use of parameter entities, together with careful structuring of the DTD, does achieve a usable modularization. Bases and additional tag sets can be chosen as required.

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:

- Paragraphs
- Segmentation, for example into orthographic sentences.
- Lists of various kinds, including glossaries and indexes
- Typographically highlighted phrases, whether unqualified or used to mark linguistic emphasis, foreign words, titles etc.
- Quoted phrases, distinguishing direct speech, quotation, terms and glosses, cited phrases etc.
- Names, numbers and measures, dates and times, and similar data-like phrases.
- Basic editorial changes (e.g. correction of apparent errors; regularization and normalization; additions, deletions and omissions)
- Simple links and cross references, providing basic hypertextual features.
- Pre-existing or generated annotation and indexing
- Passages of verse or drama, distinguishing for example speakers, stage directions, verse lines, stanzaic units, etc.
- Bibliographic citations, adequate for most commonly used bibliographic packages, in either a free or a tightly structured format
- 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[2]-compatible bibliographic description for the electronic text, covering all of the following:

- the electronic document itself
- sources from which the document was derived
- encoding system
- 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

---

[2] American Association of Catalogue Records.

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 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.[3]

4. Relevance of the TEI for texts in African languages

A major advantage for the encoding of texts in African languages, many of which will be created in the near future, is that the TEI Guidelines and related standards (SGML, HyTime, etc.) are well in place. At the same time, SGML handling software is increasingly available. Unlike limited encoding schemes such as HTML, which are intended mainly to provide for text presentation, the TEI is a common encoding scheme for complex texts that is flexible and extensible to support research and to facilitate the reuse of texts, and should support most needs for encoding this new data.

We can envision several encoding needs for the development of electronic texts for the African languages. These texts are potentially very complex, for at least the following reasons:

- the use of non-standard character sets and transcription schemes, and the existence of transcription variants;
- the need to accomodate the transcription of speech, which presents significant encoding challenges, as well as to link this material to other, possibly parallel data.

The character set problem is not inconsequential for representing African languages, in large part because many of these languages have not been considered in the development of existing standards. There are many properties of African languages, such as accentual and tonal properties, which demand special diacritics, and transcription of certain African languages relies on the use of the International Phonetic Alphabet. Existing standards, in particular the ISO 8859 series of character sets (including standard codes for Latin-based languages, Cyrilic, Greek, Hebrew, and Arabic), do not accomodate many aspects of the transcription of African languages, and even the emerging Universal Chracter Set (UCS) does not take into account all African languages.

---

[3] 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.

The TEI provides a mechanism for declaring a *Writing System Declaration* (WSD), in which a user-defined transcription scheme is declared in which special characters are represented by SGML *entities.* The use of entities to represent special characters is reliable and uninversal, and avoids problems often associated with transmission across networks. The specification of sets of standard entities using the TEI WSD system, ensures that all parties can appropriately decode the entities since it provides the mapping between entity and character. Therefore, an important task to be undertaken for the African languages is the development of pre-defined TEI WSDs for these languages.

Another aspect of African language texts that must be considered is the existence, in many cases, of variant transcription schemes. In many cases it will be desirable that texts rendered in different orthographies should have the potential to be displayed in any one of them. The TEI has developed sophisticated mechanisms for the encoding of variants, including not only means to specify alternatives, but means to associate discontinuous portions of text (which may be the case if one orthography demands the interposition of some character, syllable, etc. while the other does not). The TEI also provides mechanisms for mixing languages and character sets in the same document, which is likely to arise when texts reference or utilize transcription variants.

Much language data in the African nations, especially historical, literary, and linguistic materials, exists in oral form due to the long history of oral transmission of history, stories, even laws and norms, in many areas. As a result, we can anticipate the development of large bodies of materials comprising transcribed speech. Speech transcription presents particular encoding problems (Johansson, 1995: 149-58). Transcription practice can vary widely, both with respect to the types of features represented and the way particular features are rendered. In non-linguistic applications the transcription is usually heavily edited, so that the text differs minimally from an ordinary written text. In linguistic applications there may be a careful record of speech- specific features, such as prosodic and paralinguistic features, speaker overlap, pauses, hesitations, repetitions, interruptions, and other disfluencies.

In addition, we can anticipate that it will be desirable, especially for historical, linguistic, and literary studies and for the preservation of the national literatures, to link together transcription, recording, video, and, depending upon the application, possibly also parallel linguistic information (phonetic transcription, prosodic annotation, syntactic analysis, etc.), or commentary and other contextual information.

To handle many of the problems described above, sophisticated linkage mechanisms, often beyond what is provided by SGML alone, are necessary--for instance, for the following cases:

- One document makes reference to an element in another document--possibly, a document not encoded in SGML, possibmy consisting of acoustic signal, graphics, video, etc.
- The target of a pointer might be a point or a span of text, not necessarily co-extensive with some element or elements.
- Alignment of parallel texts.
- Alignment of non-parallel correspondences. For example, if one wants to represent two or more different possible syntactic analyses of a natural language sentence, the text will be the same but the structures will be different.
- Indication of temporal correspondence , for example, in a spoken text.
- Elements aggregated from various pieces of a document occurring elsewhere. For example, the linking of variant orthographies, described above.
- An entire document might be represented with more than one structure, and the strucutres are not heirarchical In essence, this means that more than one parse tree spans the same string (has the same frontier). The SGML CONCUR feature appears to provide for this, but it is verbose, requires modification of existing tags if new structures are added to the encoding (to specify the DTDs to which an element pertains), and is not implemented by most SGML software systems.[4]

---

[4] See (Murata, to appear) and (Barnard, et al., 1988: 265-76).

The TEI uses a number of different mechanisms to overcome the problems addressed above. The basis for these is the use of the SGML ID and IDREF mechanism to link and align elements in various ways and with different structures; however, not all the places to which pointers are to be attached can be assumed to have SGML IDs. Therefore, a special system of *extended pointers* was devised for the TEI in order to handle this case, which provides for pointing within or between documents. 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; however, the TEI system is simpler and easier to implement. The TEI extended pointer mechanism incorporates ideas from the SGML ID mechanism; HyTime; HyQ, the HyTime query language (so that pointers may need to be dynamically evaluated).

The pointer mechanism can be used to solve various linking and alignment problems in texts such as those outlined above, 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.

Appenix I provides a few examples of the TEI encoding of spoken transcription, linkage mechaisms, etc.

5. Conclusion

Many of the documents that are likely to be encoded in the African languages in the near future are potentially complex , some containing multiple features that must be marked. Many of these documents contain features that are inherently non-hierarchical, and there exists the potential need to link complex structures. This means that a dynamic environment is needed for the specification of document encoding--an environment in which it is simple to encode simple structures, but in which more complicated structures can be encoded. In addition, it is important to have specifications that can be gracefully extended as new features of interest are identified.

The Text Encoding Initiative has produced a complex DTD that is modular and extensible. The DTD supports the encoding of a rich set of features in texts. This should provide a sound basis for encoding African language texts in the near future. However, the development of the TEI Guidelines is an on-going project; it is necessary to apply the existing mechanisms to more and different texts, in different languages, and presenting different encoding problems. There remains technical work to be done in simplifying and rationalizing some of the existing features, and further tag set development is required in some areas. Most importantly, the Guidelines need to be validated as they are used by projects throughout the world.

The TEI is developing its collaborative relationships with such projects to facilitate the validation and development of the Guidelines. Given the potential challenges of encoding texts in the African languages, it will be important to seek and establish collaborative efforts. At the same time, the development of significant electronic language resources in the African languages can benefit from the encoding groundwork already laid by the TEI and ensure the potential for maximum usability and reusability of these resources.

6. Information about the TEI

The TEI Guidelines for Electronic Text Encoding and Interchange are available 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

They are also available lectronically via the World Wide Web at **<http://www-tei.uic.edu/orgs/tei>** and via anonymous ftp from **<ftp-tei.uic.edu>** (in **pub/tei** and its subdirectories).

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")

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 various TEI participants.

References

Barnard, D. Hayter, R., Karababa, M., Logan, G. and McFadden, J. (1988) "SGML-Based Markup for Literary Texts: ‚Two Problems and Some Solutions". *Computers and the Humanities* 22, 4, 265-76.

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.

DeRose, S.J., Durand, D.G. 1994. *Making HyperMedia Work: A Users's Guide to HyTime.* Kluwer Academic Publishers, Boston.

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". *The Text Encoding Initiative: Background and Context.* Dordrecht: Kluwer Academic Publishers, 5-15.

Ide, N., Véronis, J. (Eds.) 1995. *The Text Encoding Initiative: Background and Context.* Dordrecht: Kluwer Academic Publishers.

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.

Johansson, S. 1995. "The Encoding of Spoken Texts." *The Text Encoding Initiative: Background and Context.* Dordrecht: Kluwer Academic Publishers, 149-58.

Makoto Murata (to appear). "File Format for Documents Containing Both Logical Structures and Layout Structures". *Electronic Publishing*, 1.

Sperberg-McQueen, C.M., Burnard, L. 1995. "The Design of the TEI Encoding Scheme". *The Text Encoding Initiative: Background and Context.* Dordrecht: Kluwer Academic Publishers, 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 I: Examples of TEI encoding

The following simple examples are intended to provide the flavor of TEI encoding, so that the application to more complex cases can be assumed. The examples comprise excerpts of various kinds of encoding from the fully encoded text, which would contain a header, etc.

```
<s type=sentence>
  <w full=usingekuja>
     <m type="2ps personal pronoun>u</m>
     <m type=negation>si</m>
     <m type=conditional>nge</m>
     <m type="verbal marked infinitive">kuja</m>
  </w>
  <w>leo</w>
  <w full=nisingepika>
     <m type="1ps personal pronoun">ni</m>
     <m type=negation>si</m>
     <m type=conditional>nge</m>
     <m type="radical verbal, no mark for infinitive">pika</m>
  </w>
  <w type=singular>keki</w>

<s type=sentence>
  <w full=amerudishiwa>
     <m type="3ps personal pronoun>a</m>
     <m type="3ps personal pronoun" function=object>me</m>
     <m type="radical verbal" baseform=kurudi>rud</m>
```

```
      <m type="verbal derivative of causative">ish</m>
      <m type=verbal derivative of passive">iw</m>
      <m type=terminating>a</m>
    </w>
    <w full=vitabu>
      <m type="class 8">vi</m>
      <m type="nominal, theme">tabu</m>
    </w>
  </s>
```

Figure 1. A possible linguistic encoding for the sentences "usingekuja leo nisingepika keki" and "amerudishiwa vitabu", including markup for orthographic words and morphemes The specification of linguistic information is open and can be as simple or complex as required for the application.

```
<u who=P1 id=U1>Can I have ten oranges and a kilo of
    bananas please?</u>
<u who=P2 id=U2>Yes, anything else?</u>
<u who=P1 id=U3>No thanks.</u>
<u who=P2 id=U4>That'll be dollar forty.</u>
<u who=P1 id=U5>Two dollars</u>
<u who=P1 id=U6>Sixty, eighty, two dollars. Thank you.</u>

<spanGrp type=transactions>
  <span from=U1 value='sale request'>
  <span from=U2 to=U3 value='sale compliance'>
  <span from=U4 value='sale'>
  <span from=U5 value='purchase'>
  <span from=U6 value='purchase closure'>
  </spanGrp>
```

Figure 2. Encoding of a simple spoken transcription, with identification of sub-parts using ID references.

```
<div1 id=E lang=EN>
<seg id=E1>
  <s>According to our survey, 1988 sales of mineral water and soft
  drinks were much higher than in 1987, reflecting the growing popularity
   of these products.</s>
  <s>Cola drink manufacturers in particular achieved above-average
   growth rates.</s>
</seg>

<!-- ... -->

<div1 id=F lang=FR>
<seg id=F1>
  <s id=fs1>Quant aux eaux min&eacute;rales et aux limonades, elles
   rencontrent toujours plus d'adeptes.</s>
  <s id=fs2>En effet, notre sondage fait ressortir des ventes nettement
   sup&eacute;rieures &agrave; celles de 1987, pour les boissons &agrave;
   base de cola notamment. </s>
</seg>
<linkGrp type=alignment
         domains='E F'
  <link targets='E1 F1'>
  ...
</linkGrp>
```

Figure 3. Encoding of two parallel translations, with links showing their alignment. The French text includes SGML entities (e.g., "é") for special characters.