The Association for Computers and the Humanities (ACH)

The Association for Computational Linguistics (ACL)

The Association for Literary and Linguistic Computing (ALLC)

Guidelines for Electronic Text Encoding and Interchange

Edited by

C. M. Sperberg-McQueen and Lou Burnard

TEI P2, Chapter 26
Text Encoding Initiative
Chicago, Oxford

March 10, 1993

Draft Version 2, March 10, 1993

Note

This is the current draft (as of March 10, 1993) of one section of the *Guidelines for Electronic Text Encoding and Interchange*, version 2 (also known as TEI P2), produced by the Text Encoding Initiative. This fascicle contains:

- a table of contents for TEI P2, indicating the status of each section as of the release date of this fascicle
- a draft for one section of TEI P2
- entries derived from this section for the reference list of tags of TEI P2 (if any)
- a User Response and Comment Form

Please feel free to send us your comments on the contents of this draft and their applicability to your needs. Comments may be sent to the TEI-L list server for general discussion, or directly to either of the TEI editors (adresses on the User Response and Comment form). Bear in mind in reading, however, that TEI P2 is a reference manual, not a tutorial in the use of electronic text, and unavoidably assumes either a certain amount of prior knowledge or patience with unfamiliar technical material. Note also that some matters which may appear obscure in one section may well be treated in more detail elsewhere in the full text of TEI P2.

TEI P2, like P1, is a working draft. The observant will find several typographic and presentational flaws in the current text: for example, cross-references are not properly resolved in the LaTeX versions and some parts of the reference material are not yet complete. We hope to rectify as many as possible of these flaws in the coming months, but have decided not to allow them to delay any further the release of the draft. We hope that those interested in the Guidelines will find the current state of the text readable and usable enough for the purpose of substantive comment. As far as possible, all comments received on this section will be taken into account in the preparation of the next version of the Guidelines, which is due by the end of 1992.

Updates on the status of the draft as a whole will be distributed automatically to all subscribers of the Listserv list TEI-L. To subscribe to TEI-L, send electronic mail to the address LISTSERV@UICVM (or Listserv@uicvm.uic.edu) containing the single line

```
subscribe tei-l J. Smith
```

(substituting your name for “J. Smith”). TEI-L is also the appropriate place to pose questions or offer public comments on the TEI guidelines and other relevant issues. The TEI-L file server contains all the sections of TEI P2 thus far released, as well as other TEI materials. Its contents are shadowed at various other sites around the world; for more information consult one of the following documents:

- TEI ED J8, “Obtaining the Second Version of the TEI Guidelines (TEI P2)” (describes how to retrieve electronic copies of TEI P2 and the various formats they are available in)

- TEI ED J9, “Obtaining Paper Copies of the Second Version of the TEI Guidelines (TEI P2)” (describes how to request paper copies of TEI P2, for those without electronic mail access)

These documents are available from the TEI-L file server or through the editors at the addresses given on the User Response and Comment Form.

This document is made possible in part by financial support from the U.S. National Endowment for the Humanities, an independent federal agency; Directorate General XIII of the Commission of the European Communities; the Andrew W. Mellon Foundation; and the Social Science and Humanities Research Council of Canada. Direct and indirect support has also been received from the University of Illinois at Chicago, the Oxford University Computing Service, the University of Arizona, the University of Oslo, Queen's University (Kingston, Ont.), and Ohio State University. We acknowledge assistance with software acquisition from the following firms: SEMA Group and Yard Software (for Mark-It and Write-It), Software Exoterica (for CheckMark and Xtran), SoftQuad, Inc., (for Author/Editor and RulesBuilder), Electronic Book Technology (for demonstration copies of DynaText), Xerox Corporation (for Ventura Publisher).

Overall Table of Contents for TEI P2

The following list shows the expected contents of TEI P2 as a whole and indicates the current status of each chapter.

Contents

Part 4

Additional Tag Sets

Chapter 26

Additional Tag Set for Language Corpora

The term *language corpus* is used to mean a number of rather different things. It may refer simply to any collection of linguistic data (written, spoken, or a mixture of the two), although many practitioners prefer to reserve it for collections which have been organized or collected with a particular end in view, generally to characterize a particular state or variety of one or more languages. Because opinions as to the best method of achieving this goal differ, various subcategories of corpora have also been identified. For our purposes however, the distinguishing characteristic of a corpus is that its components have been selected or structured according to some conscious set of design criteria.

These design criteria may be very simple and undemanding, or very sophisticated. A corpus may be intended to represent (in the statistical sense) a particular linguistic variety or sublanguage, or it may be intended to represent all aspects of some assumed "core" language. A corpus may be made of whole texts or of fragments or text samples. It may be a "closed" corpus, or an "open" or "monitor" corpus, the composition of which may change over time. However, since an open corpus is of necessity finite at any particular point in time, the only likely effect of its expansibility from the encoding point of view may be some increased difficulty in maintaining consistent encoding practices (see further section 26.5, "Recommendations for the Encoding of Large Corpora," on page 31). For simplicity, therefore, our discussion largely concerns ways of encoding closed corpora, regarded as single but composite texts.

Language corpora are regarded by these Guidelines as *composite texts* rather than *unitary texts* (on this distinction, see further 7, "Default Text Structure," [in separate fascicle]). This is because although each discrete sample of language in a corpus clearly has a claim to be considered as a text in its own right, it is also regarded as a subdivision of some larger object, if only for convenience of analysis. Corpora share a number of characteristics with other types of composite texts, including anthologies and collections. Most notably, different components of composite texts may exhibit different structural properties (for example, some may be composed of verse, and others of prose), thus potentially requiring elements from different TEI bases. Composite texts are thus especially likely to require the techniques for combining base tag sets described in chapter 14, "Composite Texts and Combining Bases," [in separate fascicle].

Aside from these high-level structural differences, and possibly differences of scale, the encoding of language corpora and the encoding of individual texts present identical sets of problems. Any of the encoding techniques and elements presented in other chapters of these Guidelines may therefore prove relevant to some aspect of corpus encoding and may be used in corpora. However, we do not repeat here the discusssion of such fundamental matters as the representation of multiple character sets (see chapter 4, "Characters and Character Sets," [in separate fascicle]); nor attempt to summarize the variety of elements provided for encoding basic structural features such as quoted or highlighted phrases, cross references, lists, notes, editorial changes and reference systems (see chapter 6, "Tags Available in All TEI DTDs," [in separate fascicle]). In addition to these general purpose elements, these Guidelines offer a range of more specialized sets of tags which may be of use in certain specialized corpora, for example those consisting primarily of verse (chapter 8, "Base Tag Set for Verse," [in separate fascicle]), drama (chapter 9, "Base Tag Set for Drama," [in separate fascicle]), transcriptions of spoken text (chapter 10, "Base Tag Set for Transcriptions of Spoken Texts," [in separate fascicle]), letters and memoranda (chapter 11, "Base Tag Set for Letters and Memoranda," [in separate fascicle]) etc. Chapter 3, "Structure of the TEI Document Type Declarations," [in separate fascicle] should be reviewed for details of how these and other components of the Guidelines should be tailored to create a document type definition appropriate to a given application. In sum, it should not be asssumed that only the matters specifically addressed in this chapter are of importance for corpus creators.

Though entitled "Additional Tag Set for Language Corpora", this chapter also includes some other material relevant to corpora and corpus-building, for which no other location appeared suitable. It begins with a review of the distinction between unitary and composite texts, and of the different methods provided by these Guidelines for representing composite texts of different kinds (section 26.1, "Varieties of Composite Text,"). Section 26.2, "Contextual Information," on page 7 describes a set of additional header elements provided for the documentation of contextual information, of importance largely though not exclusively to language corpora. This is the additional tag set for language corpora proper. Section 26.3, "Associating Contextual Information with a Text," on page 20 discusses a mechanism by which individual parts of the TEI Header may be associated with different parts of a TEI-conformant text. Section 26.4, "Linguistic Annotation of Corpora," on page 25 reviews various methods of providing linguistic annotation in corpora, with some specific examples of relevance to current practice in corpus linguistics. Finally, section 26.5, "Recommendations for the Encoding of Large Corpora," on page 31 provides some general recommendations about the use of these Guidelines in the building of large corpora.[1]

26.1 Varieties of Composite Text

Both unitary and composite texts may be encoded using these Guidelines; composite texts, including corpora, will typically make use of the following tags for their top-level organization.

<TEI.corpus.2> : contains the whole of a TEI encoded corpus, comprising a single corpus header and one or more TEI.2 elements, each containing a single text header and a text.

<TEI.2> : contains a single TEI-conformant document, comprising a TEI header and a text, either in isolation or as part of a Teicorpus.

<TeiHeader> : supplies the descriptive and declarative information making up an "electronic title page" prefixed to every TEI-conformant text. Attributes include:

[1] Parts of this chapter are derived from working papers of the Work Group on Corpora, chaired by Douglas Biber, and the Work Group on Linguistic Description, chaired by D. Terence Langendoen. The membership of these work groups is listed in the preface.

type : specifies the kind of document to which the header is attached. Legal values are:
text : the header is attached to a single text.
corpus : the header is attached to a corpus.
status : indicates whether the header is new or has been substantially revised. Sample values include:
new : the header is a new header.
update : the header is an update (has been revised).
creator : identifies the creator of the TEI Header.
date.created : indicates when the first version of the header was created.
date.updated : indicates when the current version of the header was created.
<text> : contains a single text of any kind, whether unitary or composite, for example a poem or drama, a collection of essays, a novel, or a corpus sample.
<group> : contains the body of a composite text, grouping together a sequence of distinct texts (or groups of such texts) which are regarded as a unit for some purpose, for example the collected works of an author, a sequence of prose essays, etc.

Full descriptions of these may be found in chapter 3, “Structure of the TEI Document Type Declarations,” [in separate fascicle] (for **<TEI.corpus.2>** and **<TEI.2>**), chapter 5, “The TEI Header,” [in separate fascicle] (for **<TeiHeader>**), and chapter 7, “Default Text Structure,” [in separate fascicle] (for **<text>** and **<group>**); this section discusses their application to composite texts in particular.

In these Guidelines, the word *text* refers to any stretch of discourse, whether complete or incomplete, unitary or composite, which the encoder chooses (perhaps merely for purposes of analytic convenience) to regard as a unit. The term *composite text* refers to texts within which other texts appear; the following common cases may be distinguished:

- language corpora
- collections or anthologies
- poem cycles and epistolary works (novels or essays written in the form of collections or series of letters)
- otherwise unitary texts, within which one or more subordinate texts are embedded

The tags listed above may be combined to encode each of these varieties of composite text in different ways.

In corpora, the component samples are clearly distinct texts, but the systematic collection, standardized preparation, and common markup of the corpus often make it useful to treat the entire corpus as a unit, too. Some corpora may become so well established as to be regarded as texts in their own right; the Brown and LOB corpora are now close to achieving this status.

The **<TEI.corpus.2>** element is intended for the encoding of language corpora, though it may also be useful in encoding newspapers, electronic anthologies, and other disparate collections of material. The individual samples in the corpus are encoded as separate **<TEI.2>** elements, and the entire corpus is enclosed in a **<TEI.corpus.2>** element. Each sample has the usual structure for a **<TEI.2>** document, comprising a **<TeiHeader>** followed by a **<text>** element. The corpus, too, has a corpus-level **<TeiHeader>** element, in which the corpus as a whole, and encoding practices common to multiple samples may be described. The overall structure of a TEI-conformant corpus is thus:

```
<TEI.corpus.2>
   <TeiHeader type=corpus> ...
   <!-- TEI header for corpus-level information -->
   </TeiHeader>

   <TEI.2 id=T1>
        <TeiHeader type=text> ... </TeiHeader>
        <text> ... </text>
   </TEI.2>
   <TEI.2 id=T2>
        <TeiHeader type=text> ... </TeiHeader>
        <text> ... </text>
   </TEI.2>
   <!-- ... etc. -->
</TEI.corpus.2>
```

Header information which relates to the whole corpus rather than to individual components of it should be factored out and included in the **<TeiHeader>** element prefixed to the whole. This two-level structure allows for contextual information to be specified at the corpus level, at the individual text level, or at both. Discussion of the kinds of information which may thus be specified is provided below, in section 26.2, “Contextual Information,” on page 7, as well as in chapter 5, “The TEI Header,” [in separate fascicle]. Information of this type should in general be specified only once: a variety of methods are provided for associating it with individual components of a corpus, as further described in section 26.3, “Associating Contextual Information with a Text,” on page 20.

In some cases, the design of a corpus is reflected in its internal structure. For example, a corpus of newspaper extracts might be arranged to combine all stories of one type (reportage, editorial, reviews, etc.) into some higher-level grouping, possibly with sub-groups for date, region, etc. The **<TEI.corpus.2>** element provides no direct support for reflecting such internal corpus structure in the markup: it treats the corpus as an undifferentiated series of components, each tagged **<TEI.2>**.

If it is essential to reflect a single permanent organization of a corpus into sub- and sub-sub-corpora, then the corpus or the high-level subcorpora may be encoded as composite texts, using the **<group>** element described in section 7, “[Groups and Composite Texts],” [in separate fascicle]. The mechanisms for corpus characterization described in this chapter, however, are designed to reduce the need to do this. Useful groupings of components may easily be expressed using the text classification and identification elements described in section 26.2.1, “The Text Description,” on page 9, and those for associating declarations with corpus components described in section 26.3, “Associating Contextual Information with a Text,” on page 20. These methods also allow several different methods of text grouping to co-exist, each to be used as needed at different times. This helps minimize the danger of cross-classification and misclassification of samples, and helps improve the flexibility with which parts of a corpus may be characterized for different applications.

Anthologies and collections are often treated as texts in their own right, if only for historical reasons. In conventional publishing, at least, anthologies are published as units, with single editorial responsibility and common front and back matter which may need to be included in their electronic encodings. The texts collected in the anthology, of course, may also need to identifiable as distinct individual objects for study.

Poem cycles, epistolary novels, and epistolary essays differ from anthologies in that they are often written as single works, by single authors, for single occasions -- nevertheless, it can be useful to treat their constituent parts as individual texts, as well as the cycle itself. Structurally, therefore, they may be treated in the same way as anthologies: both are texts whose body is composed largely of other texts.

The **<group>** element is provided to simplify the encoding of collections, anthologies, and cyclic works; as noted above, the **<group>** element can also be used to record the potentially complex internal structure of language corpora. For full description, see chapter 7, “Default Text Structure,” [in separate fascicle].

Some composite texts, finally, are neither corpora, nor anthologies, nor cyclic works: they are otherwise unitary texts within which other texts are embedded. In general, they may be treated in the same way as unitary texts, using the normal **<TEI.2>** and **<body>** elements. The embedded text itself may be encoded using the **<text>** element, which may occur within quotations or between paragraphs or other chunk-level elements inside the sections of a larger text. For further discussion, see chapter 7, “Default Text Structure,” [in separate fascicle].

All composite texts share the characteristic that their different component texts may be of structurally similar or dissimilar types. If the components may all be encoded using the same base tag set, then the composite text should use that same base. If the components require different base tag sets, however, the encoding should use the *general base* tag set or the *mixed base* tag set described in chapter 14, “Composite Texts and Combining Bases,” [in separate fascicle].

26.2 Contextual Information

Contextual information is of particular importance for collections or corpora composed of samples from a variety of different kinds of text. Examples of such contextual information include: the age, sex and geographical origins of participants in a language interaction, or their socio-economic status; the cost and publication date of a newspaper story; the topic, register or factuality of an extract from a textbook. Such information may be of the first importance, whether as an organizing principle in creating a corpus (for example, to ensure that the range of values in such a parameter is evenly represented throughout the corpus, or represented proportionately to the population being sampled), or as a selection criterion in analysing the corpus (for example, to investigate the language usage of some particular vector of social characteristics).

Such contextual information is potentially of equal importance for unitary texts, and these Guidelines accordingly make no particular distinction between the kinds of information which should be gathered for unitary and for composite texts. In either case, the information should be recorded in the appropriate section of a TEI Header, as described in chapter 5, "The TEI Header," [in separate fascicle]. In the case of language corpora, such information may be gathered together in the overall corpus header, or split across all the component texts of a corpus, in their individual headers, or divided between the two. The association between an individual corpus text and the contextual information applicable to it may be made in a number of ways, as further discussed in section 26.3, "Associating Contextual Information with a Text," on page 20 below.

Chapter 5, "The TEI Header," [in separate fascicle], which should be read in conjunction with the present section, describes in full the range of elements available for the encoding of information relating to the electronic file itself, for example its bibliographic description and those of the source or sources from which it was derived (see section 5.2, "The File Description," [in separate fascicle]); information about the encoding practices followed with the corpus, for example its design principles, editorial practices, reference system etc. (see section 5.3, "The Encoding Description," [in separate fascicle]); more detailed descriptive information about the corpus' creation and content, such as the languages used within it and any descriptive classification system used (see section 5.4, "The Profile Description," [in separate fascicle]); and version information documenting any changes made in the electronic text (see section 5.5, "The Revision Description," [in separate fascicle]).

In addition to the elements defined by chapter 5, "The TEI Header," [in separate fascicle], several other elements can be used in the TEI header if the additional tag set defined by this chapter is invoked. These additional tags make it possible to characterize the social or other situation within which a language interaction takes place or is experienced, the physical setting of a language interaction, and the participants in it. Though this information may be relevant to, and provided for, unitary texts as well as for collections or corpora, it is more often recorded for the components of systematically developed corpora than for isolated texts, and thus the additional tag set is referred to as being "for language corpora". Included in this tag set are the following elements:

- **<textDesc>** : provides a description of a text in terms of its *situational parameters.*
- **<particDesc>** : describes the identifiable speakers, voices or other participants in a linguistic interaction.
- **<settingDesc>** : describes the setting or settings within which a language interaction takes place, either as a prose description or as a series of **<setting>** elements.

These elements form an optional extension to the **<profileDesc>**, defined in section 5.4, "The Profile Description," [in separate fascicle] and are further described in the remainder of this section. They are formally defined as follows:

```
<!-- 26.2:  Header extensions for Corpus Texts               -->
<!ELEMENT textDesc      - o  (channel, constitution, derivation,
                               domain, factuality, interaction,
                              preparedness, purpose+)          >
<!ATTLIST textDesc           %a.global;
                              %a.declarable;                  >
<!ELEMENT particDesc    - o  (p+ | ( (participant | particGrp)+,
                               particLinks?) )                 >
<!ATTLIST particDesc         %a.global;                        >
<!ELEMENT settingDesc   - o  (p+ | setting+)                   >
<!ATTLIST settingDesc        %a.global;                        >
<!-- (continued in sec. 26.2.1, 26.2.2, 26.2.3)              -->
```

The additional tag set for language corpora will be invoked, thus enabling the use of these elements, if a parameter entity called *TEI.corpus* is declared with the value “INCLUDE”, somewhere within the DTD subset. If the document is structured as a TEI corpus (that is, using the **<TEI.corpus.2>** element), its document type declaration will resemble this:

```
<!DOCTYPE TEI.corpus.2 system 'TEI2.dtd' [
    <!ENTITY % TEI.corpus 'INCLUDE' >
]>
```

26.2.1 The Text Description

The **<textDesc>** element provides a full description of the situation within which a text was produced or experienced, and thus characterizes it in a way relatively independent of any *a priori* theory of text-types. It is provided as an alternative or a supplement to the common use of descriptive taxonomies used to categorize texts, which is fully described in section 5.4.3, “The Text Classification,” [in separate fascicle], and section 5.3.5, “The Classification Declaration,” [in separate fascicle]. The description is organized as a set of values and optional prose descriptions for eight *situational parameters*, each represented by one of the following eight elements:

<channel> : describes the medium or channel by which a text is delivered or experienced. For a written text, this might be print, manuscript, e-mail, etc.; for a spoken one, radio, telephone, face-to-face, etc. Attributes include:
- *mode* : specifies the mode of this channel with respect to speech and writing. Legal values are:
  - *s* : spoken
  - *w* : written
  - *sw* : spoken to be written (e.g. dictation)
  - *ws* : written to be spoken (e.g. a script)
  - *m* : mixed modes
  - *x* : unknown or inapplicable

<constitution> : describes the internal composition of a text or text sample, for example as fragmentary, complete, etc. Attributes include:
- *type* : specifies how the text was constituted. Legal values are:
  - *single* : a single complete text

composite : a text made by combining several smaller items, each individually complete
frags : a text made by combining several smaller, not necessarily complete, items
unknown : composition unknown or unspecified

<derivation> : describes the nature and extent of indebtedness or derivativeness of this text with respect to others. Attributes include:
type : categorizes the derivation of the text. Sample values include:
original : text is original
revision : text is a revision of some other text
translation : text is a translation of some other text
abridgment : text is an abridged version of some other text
plagiarism : text is plagiarized from some other text
traditional : text has no obvious source but is one of a number derived from some common ancestor

<domain> : describes the most important social context in which the text was realized or for which it is intended, for example private vs. public, education, religion, etc. Attributes include:
type : categorizes the domain of use. Sample values include:
art : art and entertainment
domestic : domestic and private
religious : religious and ceremonial
business : business and work place
education : education
govt : government and law
public : other forms of public context

<factuality> : describes the extent to which the text may be regarded as imaginative or non-imaginative, that is, as describing a fictional or a nonfictional world. Attributes include:
type : categorizes the factuality of the text. Legal values are:
fiction : the text is to be regarded as entirely imaginative
fact : the text is to be regarded as entirely informative or factual
mixed : the text contains a mixture of fact and fiction
inapplicable : the fiction/fact distinction is not regarded as helpful or appropriate to this text

<interaction> : describes the extent, cardinality and nature of any interaction among those producing and experiencing the text, for example in the form of response or interjection, commentary etc. Attributes include:
type : specifies whether or not there is any interaction between active and passive participants in the text. Legal values are:
none : no interaction of any kind, e.g. a monologue
partial : some degree of interaction, e.g. a monologue with set responses
complete : complete interaction, e.g. a face to face conversation
inapplicable : this parameter is inappropriate or inapplicable in this case
active : specifies the number of active participants (or *addressors*) producing parts of the text. Legal values are:
singular : a single addressor
plural : many addressors
corporate : a corporate addressor
unknown : number of addressors unknown or unspecifiable

passive : specifies the number of passive participants (or *addressees*) to whom a text is directed or in whose presence it is created or performed. Suggested values include:
- *self* : text is addressed to the originator e.g. a diary
- *single* : text is addressed to one other person e.g. a personal letter
- *many* : text is addressed to a countable number of others e.g. a conversation in which all participants are identified
- *group* : text is addressed to an undefined but fixed number of participants e.g. a lecture
- *world* : text is addressed to an undefined and indeterminately large number e.g. a published book

<preparedness> : describes the extent to which a text may be regarded as prepared or spontaneous. Attributes include:

type : a keyword characterizing the type of preparedness. Sample values include:
- *none* : spontaneous or unprepared
- *scripted* : follows a script
- *formulaic* : follows a predefined set of conventions
- *revised* : polished or revised before presentation

<purpose> : characterizes a single purpose or communicative function of the text. Attributes include:

type : specifies a particular kind of purpose. Suggested values include:
- *persuade* : didactic, advertising, propaganda etc
- *express* : self expression, confessional etc
- *inform* : convey information, educate etc
- *entertain* : amuse, entertain etc

degree : specifies the extent to which this purpose predominates. Legal values are:
- *high* : this purpose is predominant
- *medium* : this purpose is intermediate
- *low* : this purpose is weak
- *unknown* : extent unknown

To make up a TEI-conformant text description, each of the above elements must be supplied in the order specified, though its value may indicate that it is not applicable. Except for the **<purpose>** element, which may be repeated to indicate multiple purposes, no element may appear more than once within a single text description. Each element may be empty, or may contain a brief qualification or more detailed description of the value expressed by its attributes. It should be noted that some texts, in particular literary ones, may resist unambiguous classification in some of these dimensions; in such cases, the situational parameter in question should be given the content “not applicable” or the equivalent.

Texts may be described along many dimensions, according to many different taxonomies. No generally accepted consensus as to how such taxonomies should be defined has yet emerged, despite the best efforts of many corpus linguists, text linguists, sociolinguists, rhetoricians, and literary theorists over the years. Rather than attempting the task of proposing a single taxonomy of “text types” (or the equally impossible one of enumerating all those which have been proposed previously), the closed set of *situational parameters* described above can be used in combination to supply useful distinguishing descriptive features of individual texts, without insisting on a system of discrete high-level “text-types”. Such text-types may however be used in combination with the parameters proposed here, with the advantage that the internal structure of each such text-type can be specified in terms of the parameters proposed. This approach has the following analytical advantages:[2]

- it enables a relatively continuous characterization of texts (in contrast to discrete categories based on type or topic)
- it enables meaningful comparisons across corpora
- it allows analysts to build and compare their own text-types based on the particular parameters of interest to them
- it is equally applicable to spoken and written texts

Two alternative approaches to the use of these parameters are supported by these Guidelines. One is to use pre-existing taxonomies such as those used in subject classification or other types of text categorization. Such taxonomies may also be appropriate for the description of the topics addressed by particular texts. Elements for this purpose are described in section 5.4.3, “The Text Classification,” [in separate fascicle], and elements for defining or declaring such classification schemes in section 5.3.5, “The Classification Declaration,” [in separate fascicle]. A second approach is to develop an application specific set of *feature structures* and an associated *feature system declaration,* as described in chapters 18, “Feature Structure Analysis,” [in separate fascicle] and 29, “Feature System Declaration,” [in separate fascicle].

Where the organizing principles of a corpus or collection so permit, it may be convenient to regard a particular set of values for the situational parameters listed in this section as forming a *text-type* in its own right; this may also be useful where the same set of values applies to several texts within a corpus. In such a case, the set of text-types so defined should be regarded as a *taxonomy.* The mechanisms described in section 5.3.5, “The Classification Declaration,” [in separate fascicle] may be used to define hierarchic taxonomies of such text-types, provided that the **<catDesc>** component of the **<category>** element contains a **<textDesc>** element rather than a prose description. Particular texts may then be associated with such definitions using the mechanisms described in sections 5.4.3, “The Text Classification,” [in separate fascicle].

Using these parameters, an informal domestic conversation might be characterized as follows:

```
<textDesc id=t1 n='Informal domestic conversation'>
  <channel mode=s>informal face-to-face conversation</channel>
  <constitution type=single>each text represents a continuously
        recorded interaction among the specified participants
        </constitution>
```

---

[2] Schemes similar to that proposed here were developed in the 1960s and 1970s by researchers such as Hymes, Halliday, and Crystal and Davy, but have rarely been implemented; one notable exception being the pioneering work on the Helsinki Diachronic Corpus of English, on which see M. Kytö and M. Rissanen, “The Helsinki Corpus of English Texts,” in *Corpus Linguistics: hard and soft*, ed. M. Kytö, O. Ihalainen, and M. Rissanen (Amsterdam: Rodopi, 1988).

```
  <derivation type=original>
  <domain type=domestic>plans for coming week, local affairs</domain>
  <factuality type=mixed>mostly factual, some jokes</factuality>
  <interaction type=complete active=plural passive=many>
  <preparedness type='spontaneous'>
  <purpose type=entertain degree=high>
  <purpose type=inform degree=medium>
 </textDesc>
```

The following example demonstrates how the same situational parameters might be used to characterize a novel:

```
 <textDesc n='novel'>
   <channel mode=w>print; part issues</channel>
   <constitution type=single>
   <derivation type=original>
   <domain type=art>
   <factuality type=fiction>
   <interaction type=none>
   <preparedness type=prepared>
   <purpose type=entertain degree=high>
   <purpose type=inform degree=medium>
</textDesc>
```

The formal declarations for these elements are given below:

```
<!-- 26.2.1:  Header extensions for Corpus Texts (cont'd)     -->
<!-- (continuation of sec. 26.2)                              -->
<!ELEMENT channel       - o  (%phrase.seq)                      >
<!ATTLIST channel            %a.global;
          mode               (s | w | ws | sw | m | x)
                                                 x              >
<!ELEMENT constitution  - o  (%phrase.seq)                      >
<!ATTLIST constitution       %a.global;
          type               (single | composite | frags |
                             unknown)            single         >
<!ELEMENT derivation    - o  (%phrase.seq)                      >
<!ATTLIST derivation         %a.global;
          type               CDATA               #IMPLIED       >
<!ELEMENT domain        - o  (%phrase.seq)                      >
<!ATTLIST domain             %a.global;
          type               CDATA               #IMPLIED       >
<!ELEMENT factuality    - o  (%phrase.seq)                      >
<!ATTLIST factuality         %a.global;
          type               (fiction|fact|mixed|inapplicable)
                                                 #IMPLIED       >
<!ELEMENT interaction   - o  (%phrase.seq;)                     >
<!ATTLIST interaction        %a.global;
          type               (none|partial|complete|
                             inapplicable)       #IMPLIED
          active             CDATA               #IMPLIED
          passive            CDATA               #IMPLIED       >
<!ELEMENT preparedness  - o  (%phrase.seq)                      >
```

```
<!ATTLIST preparedness       %a.global;
          type               CDATA              #IMPLIED      >
<!ELEMENT purpose       - o  (%phrase.seq)                    >
<!ATTLIST purpose            %a.global;
          type               CDATA              #IMPLIED
          degree             (high | medium | low | unknown)
                                                #IMPLIED      >
```

26.2.2 The Participants Description

The **<particDesc>** element in the **<profileDesc>** element provides additional information about the participants in a spoken text or, where this is judged appropriate, the persons named or depicted in a written text. Individual speakers or groups of speakers may be named or identified by a code which can then be used elsewhere within the encoded text, for example as the value of a *who* attribute. Demographic and descriptive information may be supplied about their individual characteristics and the relationships between them.

It should be noted that although the terms *speaker* or *participant* are used throughout this section, it is intended that the same mechanisms may be used to characterize fictional personae or “voices” within a written text, except where otherwise stated. For the purposes of analysis of language usage, the information specified here should be equally applicable to written and spoken texts.

The element **<particDesc>** contains one or more **<participant>** or **<participantGrp>** elements, followed by an optional **<particLinks>** element, as described below:

- **<participant>** : describes a single participant in a language interaction. Attributes include:
  - *role* : specifies the role of this participant in the group.
  - *sex* : specifies the sex of the participant. Sample values include:
    - *m* : male
    - *f* : female
    - *u* : unknown or inapplicable
  - *age* : specifies the age group to which the participant belongs.
- **<particGrp>** : describes a group of individuals treated as a single participant. Attributes include:
  - *role* : specifies the role of this group of participants in the interaction.
  - *sex* : specifies the sex of the participant group. Sample values include:
    - *m* : male
    - *f* : female
    - *u* : unknown
    - *x* : mixed
  - *age* : specifies the age group of the participants.
  - *size* : specifies the size or approximate size of the group.
- **<particLinks>** : describes the relationships or social links existing between participants in a linguistic interaction.

Both participants and participant groups have the same substructure. This may be a prose description, or, more formally, a series of specialized subelements providing more specific details. Such details will vary enormously for different kinds of analysis; the set of demographic characteristics presented here as sub-elements should therefore be regarded as providing only an indication of the kinds of descriptive information which have been found to be generally useful, for example in socio-linguistics. Users of these Guidelines are free to extend or modify this set of demographic characteristics, by redefining the parameter entity *m.demographics*, associated with the class *demographics*, as further described in chapter 32, “Modifying TEI DTDs,” [in separate fascicle]. Where well-known classification schemes exist, e.g. for socio-economic class or occupation, these should be used and may be documented in the same way as for text classification (see section 5.3.5, “The Classification Declaration,” [in separate fascicle])

The following elements are the default members of the class *demographics*:

<person> : contains a name of a person, or a reference to a person.
<birth> : contains information about a person's birth, such as its date and place. Attributes include:
  date : specifies the date of birth in a ISO standard form (yyyy-mm-dd).
<firstLang> : specifies the first language of a participant.
<langKnown> : contains an informal description of a person's competence in different languages, dialects, etc.
<residence> : describes a person's present or past places of residence.
<education> : contains a brief prose description of the educational background of a participant.
<affiliation> : contains an informal description of a person's present or past affiliation with some organization, for example an employer or sponsor.
<occupation> : contains an informal description of a person's trade, profession or occupation. Attributes include:
  scheme : identifies the classification system or taxonomy in use by supplying the identifier of a **<taxonomy>** element elsewhere in the header.
  code : identifies an occupation code defined within the classification system or taxonomy defined by the *source* attribute.
<socecStatus> : contains an informal description of a person's perceived social or economic status. Attributes include:
  scheme : identifies the classification system or taxonomy in use.
  code : identifies a status code defined within the classification system or taxonomy defined by the *source* attribute.

For example, an individual might be described informally by the following **<participant>** element:

```
<participant id=P1 sex=F age='mid'>
  <p>Female informant, well-educated, born in Shropshire
   UK, 12 Jan 1950, of unknown occupation.
   Speaks French fluently. Socio-Economic status B2
   in the PEP classification scheme.
</participant>
```

Provided that the “PEP classification scheme” has been defined elsewhere in the heading (as a **<taxonomy>** element within the **<textClass>** element; see 5.3.5, “The Classification Declaration,” [in separate fascicle]), the same individual might more formally be described as follows:

```
<participant id=P1 sex=F age='mid'>
  <birth date='1950-01-12'>
     <date>12 Jan 1950</date>
     <place>Shropshire, UK</place>
  </birth>
  <firstLang>English</firstLang>
  <langKnown>French</langKnown>
  <residence>Long term resident of Hull</residence>
  <education>University postgraduate</education>
  <occupation>Unknown</occupation>
  <socecstatus source=PEP code=B2>
</participant>
```

An identified character in a drama or a novel might be defined using a subset of the same tags as follows:[3]

```
<participant id=EM01 sex=F age=young>
  <person>Emma Woodhouse</person>
</participant>
```

As noted above, the **<particLinks>** element is used to document personal or social relationships between individual participants, where this is felt to be of importance in the analysis. This may be done either as an informal prose description, or more formally using the special purpose **<relation>** element described in this section.

<relation> : describes any kind of relationship or linkage amongst a specified group of participants. Attributes include:
- *type* : categorizes the relationship in some respect, e.g. as social, personal or other. Suggested values include:
  - *social* : relationship concerned with social roles
  - *personal* : relationship concerned with personal roles, e.g. kinship, marriage, etc.
  - *other* : other kinds of relationship
- *desc* : briefly describes the relationship.
- *active* : identifies the “active”participants in a non-mutual relationship, or all the participants in a mutual one.
- *passive* : identifies the “passive” participants in a non-mutual relationship.
- *mutual* : indicates whether the relationship holds equally amongst all the participants. Legal values are:
  - *Y* : the relationship is mutual
  - *N* : the relationship is directed

A *relationship*, as defined here, may be any kind of describable link between specified participants, for example a social relationship (such as employer/employee), a personal relationship (such as sibling, spouse, etc.) or something
&@fn@sep

[3] It is particularly useful to define participants in a dramatic text in this way, since it enables the *who* attribute to be used to link **<sp>** elements to a definition for their speaker; see further section 9, “[Base Tag Set for Drama],” [in separate fascicle]

less precise such as “possessing shared knowledge”. A relationship may be *mutual*, in that all the participants engage in it on an equal footing (for example the “sibling” relationship); or it may not be if participants are not identical with respect to their role in the relationship (for example, the “employer” relationship). In the case of non-mutual relationships, only two kinds of role are supported by the present proposals, somewhat arbitrarily named “active” and “passive”. These names are chosen to reflect the fact that many non-mutual relations are *directed*, in the sense that they are most readily described by a transitive verb, or a verb phrase of the form “[is] X [of]” or “[is] X [to]”. The subject of the verb is classed as *active*, the direct object of the verb, or the object of the concluding preposition, as *passive*. Thus parents are “active” and children “passive” in the relationship “parent” (interpreted as “[is] parent [of]”); the employer is “active”, the employee “passive”, in the relationship “employs”. These relationships can be inverted: parents are “passive” and children “active” in the relationship “[is] child [of]”; similarly “works for” inverts the active/passive roles of “employs”.

For example:

```
<particLinks>
  <relation desc='parent' active="P1 P2"
         passive="P3 P4" mutual=N>
  <relation desc='spouse' active="P1 P2">
  <relation class='social' desc='employer'
          active=P1 passive='P3 P5 P6 P7' mutual=N>
</particLinks>
```

This example defines the following three relationships among participants P1 through P7:

> P1 and P2 are parents of P3 and P4.
> P1 and P2 are linked in a mutual relationship called “spouse” — i.e. P2 is the spouse of P1, and P1 is the spouse of P2.
> P1 has the social relationship “employer” with respect to P3, P5, P6, and P7.

The elements discussed in this section are formally defined as follows:

```
<!-- 26.2.2:  Header extensions for Corpus Texts (cont'd)     -->
<!-- (continuation of sec. 26.2)                              -->
<!ENTITY % x.demographic '' >
<!ENTITY % m.demographic '%x.demographic birth |
education | firstLang | occupation | person |
place | residence | socecStatus' >
<!ELEMENT participant   - o  (p+ | (%m.demog)* )              >
<!ATTLIST participant        %a.global;
          role               CDATA              #IMPLIED
          sex                (m | f | u )       #IMPLIED
          age                CDATA              #IMPLIED       >
<!ELEMENT particGrp     - o  (p+ | (%m.demog)* )              >
<!ATTLIST particGrp          %a.global;
          role               CDATA              #IMPLIED
          sex                (m | f | u | x)    #IMPLIED
          age                CDATA              #IMPLIED
```

```
          size              CDATA              #IMPLIED      >
<!-- PERSON is defined in the core.                         -->

<!ELEMENT birth         - -  (%phrase.seq)                   >
<!ATTLIST birth              %a.global;
          date               %ISO-date          #IMPLIED      >
<!ELEMENT firstLang     - o  (%phrase.seq)                   >
<!ATTLIST firstLang          %a.global;                      >
<!ELEMENT langKnown     - -  (%phrase.seq)                   >
<!ATTLIST langKnown          %a.global;                      >
<!ELEMENT residence     - o  (%phrase.seq)                   >
<!ATTLIST residence          %a.global;                      >
<!ELEMENT education     - o  (%phrase.seq)                   >
<!ATTLIST education          %a.global;                      >
<!ELEMENT affiliation   - -  (%phrase.seq)                   >
<!ATTLIST affiliation        %a.global;                      >
<!ELEMENT occupation    - -  (%phrase.seq)                   >
<!ATTLIST occupation         %a.global;
          scheme             IDREF              #IMPLIED
          code               IDREF              #IMPLIED      >
<!ELEMENT socecStatus   - o  (%phrase.seq)                   >
<!ATTLIST socecStatus        %a.global;
          scheme             IDREF              #IMPLIED
          code               IDREF              #IMPLIED      >
<!ELEMENT particLinks   - o  (p+ | relation+)                >
<!ATTLIST particLinks        %a.global;                      >
<!ELEMENT relation      - o  EMPTY                           >
<!ATTLIST relation           %a.global;
          type               CDATA              personal
          desc               CDATA              #IMPLIED
          active             IDREFS             #IMPLIED
          passive            IDREFS             #IMPLIED
          mutual             (Y | N)            Y             >
```

26.2.3 The Setting Description

The **<settingDesc>** element is used to describe the setting or settings in which language interaction takes place. It may contain a prose description, analogous to a stage description at the start of a play, stating in broad terms the locale, or a more detailed description of a series of such settings. Individual settings may be associated with particular participants by means of the optional *who* attribute if, for example, participants are in different places. This attribute identifies one or more individual participants or participant groups, as discussed earlier in section 26.2.2, “The Participants Description,” on page 14. If this attribute is not specified, the setting details provided are assumed to apply to all participants represented in the language interaction.[4]

&@fn@sep

[4] The present proposals do not support the encoding of different settings for the same participant. This is a subject for further work.

Each distinct setting is described by means of a **<setting>** element, which contains either a prose description or a combination of the other elements listed below:

<setting> : describes one particular setting in which a language interaction takes place. Attributes include:
who : supplies the identifiers of the participants at this setting.
<place> : contains the name of, or a reference to, a place.
<date> : contains a date in any format. Attributes include:
value : gives the value of the date in some standard form, usually yyyy-mm-dd.
<time> : contains a phrase defining a time of day in any format. Attributes include:
value : gives the value of the time in a standard form.
<locale> : contains a brief informal description of the nature of a place for example a room, a restaurant, a park bench etc.
<activity> : contains a brief informal description of what a participant in a language interaction is doing other than speaking, if anything.

The following example demonstrates the kind of background information often required to support transcriptions of language interactions, first encoded as a simple prose narrative:

```
<settingDesc>
<p>The time is early spring, 1989.  P1 and P2 are
 playing on the rug of a suburban home in Bedford.
 P3 is doing the washing up at the sink.  P4 (a
 radio announcer) is in a broadcasting studio in London.
</settingDesc>
```

The same information might be represented more formally in the following way:

```
<settingDesc>
<setting who="P1 P2">
 <place>Bedford, UK</>
 <date value=1989>early spring, 1989</>
 <locale>rug of a suburban home</>
 <activity>playing</>
</setting>
<setting who=P3>
 <place>Bedford, UK</>
 <date value=1989>early spring, 1989</>
 <locale>at the sink</>
 <activity>washing-up</>
</setting>
<setting who=P4>
 <place>London, UK</>
 <time>unknown</>
 <locale>broadcasting studio</>
 <activity>radio performance</>
</setting>
</settingDesc>
```

The elements discussed in this section have the following formal definitions:

```
<!-- 26.2.3:  Header extensions for Corpus Texts (cont'd)     -->
<!-- (continuation of sec. 26.2)                              -->
<!ELEMENT setting       - -  (p+ | (place?, time?, locale?,
                              activity?))                       >
<!ATTLIST setting            %a.global;
          who                IDREFS             #IMPLIED        >
<!ELEMENT locale        - o  (%phrase.seq)                      >
<!ATTLIST locale             %a.global;                         >
<!ELEMENT activity      - o  (%phrase.seq)                      >
<!ATTLIST activity           %a.global;                         >
<!-- PLACE, DATE, and TIME are defined in the core.          -->
```

26.3 Associating Contextual Information with a Text

This section discusses the assocation of the contextual information held in the header with the individual elements making up a TEI text or corpus. Contextual information is held in elements of various kinds within the TEI header, as discussed elsewhere in this section and in chapter 5, “The TEI Header,” [in separate fascicle]. Here we consider what happens when different parts of a document need to be associated with different contextual information of the same type, for example when one part of a document uses a different encoding practice from another, or where one part relates to a different setting from another. In such situations, there will be more than one instance of a header element of the relevant type.

The TEI DTDs allow for the following possibilities:

- A given element may appear in the corpus header only, in the header of one or more texts only, or in both places
- There may be multiple occurrences of certain elements in either corpus or text header.

To simplify the exposition, we deal with these two possibilities separately in what follows; however, they may, of course, be combined as desired.

26.3.1 Combining Corpus and Text Headers

A TEI conformant document may have more than one header only in the case of a TEI corpus, which must have a header in its own right, as well as the obligatory header for each text. Every element specified in a corpus-header is understood as if it appeared within every text header in the corpus. An element specified in a text header but not in the corpus header supplements the specification for that text alone. If any element is specified in both corpus and text headers, the corpus header element is over-ridden for that text alone.

The **<titleStmt>** for a corpus text is understood to be prefixed by the **<titleStmt>** given in the corpus header. All other optional elements of the **<file-**

Desc> should be omitted from an individual corpus text header unless they differ from those specified in the corpus header. All other header elements behave identically, in the manner documented below.

This facility makes it possible to state once for all in the corpus header each piece of contextual information which is common to the whole of the corpus, while still allowing for individual texts to vary from this common denominator.

For example, the following schematic shows the structure of a corpus comprising three texts, the first and last of which share the same encoding declaration. The second one has its own encoding declaration

```
<TEI.corpus.2>
<TeiHeader>
    <!-- contains declarations common to the whole corpus -->
    <fileDesc> ... </>
    <encodingDesc> ...
        <!-- for example, this encodingDesc is common to
             whole corpus. --> ... </>
    <revisionDesc> ... </>
</TeiHeader>

<TEI.2>
<TeiHeader>
   <fileDesc>
        <!-- details peculiar to this text --> ... </>
</TeiHeader>
<text>
 ...
</TEI.2>

<TEI.2>
<TeiHeader>
   <fileDesc>
        <!-- details peculiar to this text --> ... </>
   <encodingDesc>
        <!-- encoding description
             peculiar to this text --> ... </>
</TeiHeader>
<text>
 ...
</TEI.2>

<TEI.2>
<TeiHeader>
   <fileDesc><!-- details peculiar to this text --> ... </>
</TeiHeader>
<text>
 ...
</TEI.2>
</TEI.corpus.2>
```

26.3.2 Declarable Elements

Certain of the elements which can appear within a TEI Header are known as *declarable elements*. These elements have in common the fact that they may be linked explicitly with a particular part of a text or corpus by means of a *decls* attribute. This linkage is used to over-ride the default association between declarations in the header and a corpus or corpus text. The only header elements which may be associated in this way are those which would not otherwise be meaningfully repeatable. An alphabetically ordered list of declarable elements follows:

<analysis> : describes scope of any analytic or interpretive information added to the text in addition to the transcription.
<bibl> : contains a loosely-structured bibliographic citation of which the sub-components may or may not be explicitly tagged.
<bibl.full> : contains a fully-structured bibliographic citation, in which all components of the TEI file description are present.
<bibl.struct> : contains a structured bibliographic citation, in which only bibliographic subelements appear and in a specified order.
<broadcast> : describes a broadcast used as the source of a spoken text.
<correction> : states how and under what circumstances corrections have been made in the text.
<editorialDecl> : provides details of editorial principles and practices applied during the encoding of a text.
<equipment> : provides technical details of the equipment and media used for an audio or video recording used as the source for a spoken text.
<hyphenation> : summarizes the way in which hyphenation in a source text has been treated in an encoded version of it.
<langUsage> : describes the languages, sublanguages, registers, dialects etc. represented within a text.
<list.bibl> : contains a list of bibliographic citations of any kind.
<normalization> : indicates the extent of normalization or regularization of the original source carried out in converting it to electronic form.
<particDesc> : describes the identifiable speakers, voices or other participants in a linguistic interaction.
<projectDesc> : describes in detail the aim or purpose for which an electronic file was encoded, together with any other relevant information concerning the process by which it was assembled or collected.
<quotation> : specifies editorial practice adopted with respect to quotation marks in the original.
<recording> : details of an audio or video recording event used as the source of a spoken text, either directly or from a public broadcast.
<samplingDecl> : contains a prose description of the rationale and methods used in sampling texts in the creation of a corpus or collection.
<scriptStmt> : contains a citation giving details of the script used for a spoken text.
<segmentation> : describes the principles according to which the text has been segmented, for example into sentences, tone-units, graphemic strata, etc.
<sourceDesc> : supplies a bibliographic description of the copy text(s) from which an electronic text was derived or generated.

<stdVals> : specifies the format used when standardized date or number values are supplied.
<textClass> : groups information which describes the nature or topic of a text in terms of a standard classification scheme, thesaurus, etc.
<textDesc> : provides a description of a text in terms of its *situational parameters*.

All of the above elements may be multiply defined within a single header, that is, there may be more than one instance of any declarable element type at a given level. When this occurs, the following rules apply:

- every declarable element must bear a unique identifier
- for each different type of declarable element which occurs more than once within the same parent element, exactly one element must be specified as the default

In the following example, an editorial declaration contains two possible **<correction>** policies, one identified as C1 and the other as C2. Since there are two, one of them (in this case C1) must be specified as the default:

```
<editorialDecl>
    <correction id=C1 default=Y> ... </correction>
    <correction id=C2> ... </correction>
    <normalization id=N1>
    <p> ...
    <p> ...
</editorialDecl>
```

For texts associated with the header in which this declaration appears correction method C1 will be assumed, unless they explicitly state otherwise. Here is the structure for a text which does state otherwise:

```
<text>
    ...
    <div1 n=d1> ... </div1>
    <div1 n=d2 decls=C2> ... </div1>
    <div1 n=d3> ... </div1>
    ...
</text>
```

In this case, the contents of the divisions D1 and D3 will both use correction policy C1, and those of division D2 will use correction policy C2.

The *decls* attribute is defined for any element which is a member of the class *declaring*. This includes the major structural elements **<text>**, **<group>**, and **<div>**, as well as smaller structural units, down to the level of paragraphs in prose, individual utterances in spoken texts, or entries in dictionaries. However, TEI recommended practice is to limit the number of multiple declarable elements used by a document as far as possible, for simplicity and ease of processing.

The identifier or identifiers specified by the *decls* attribute are subject to two further restrictions:

- An identifier specifying an element which contains multiple instances of one or more other elements should be interpreted as if it explicitly identified the elements identified as the default in each such set of repeated elements
- Each element specified, explicitly or implicitly, by the list of identifiers must be of a different type.

To demonstrate how these rules operate, we now expand our earlier example slightly:

```
<encodingDesc>
       ...
       <editorialDecl  id=ED1 default=Y>
           <correction id=C1a default=Y>
           <correction id=C1b>
           <normalization id=N1>
           <p> ...
           <p> ...
       </editorialDecl>
       <editorialDecl id=ED2>
           <samplingDecl id=SAMP2>
           <correction id=C2a default=Y>
           <correction id=C2b>
           <normalization id=N2a>
           <normalization id=N2b default=Y>
           <p>  <p>
       </editorialDecl>
       ...
</encodingDesc>
```

This encoding description now has two editorial declarations, identified as ED1 (the default) and ED2. For texts not specifying otherwise, ED1 will apply. If ED1 applies, correction method C1a and normalization method N1 apply, since these are the specified defaults within ED1. In the same way, for a text specifying *decls* as “ED2”, correction C2a, sampling SAMP2 and normalization N2b will apply.

A finer grained approach is also possible. A text might specify **<text decls='C2b N2a'>**, or even **<text decls='C1a N2a SAMP2'>**, to “mix and match” declarations as required. A tag such as **<text decls='ED1 ED2'>** would (obviously) be illegal, since it includes two elements of the same type; a tag such as **<text decls='ED2 C1a'>** is also illegal, since in this context “ED2” is synonymous with the defaults for that editorial declaration, namely “SAMP2 C2a N2b”, resulting in a list that identifies two correction elements (C1a and C2a).

26.3.3 Summary

The rules determing which of the declarable elements are applicable at any point may be summarized as follows:

1. If there is a single occurrence of a given declarable element in a corpus header, then it applies by default to all elements within the corpus.
2. If there is a single occurrence of a given declarable element in the text header, then it applies by default to all elements of that text irrespective of the contents of the corpus header.
3. Where there are multiple occurrences of declarable elements within either corpus or text header,
   - each must have a unique value specified as the value of its *id* attribute;
   - one only must bear a *default* attribute with the value "YES".
4. It is a semantic error for an element to be associated with more than one occurrence of any declarable element.
5. Selecting an element which contains multiple occurrences of a given declarable element is semantically equivalent to selecting only those contained elements which are specified as defaults.
6. An association made by one element applies by default to all of its descendants.

26.4 Linguistic Annotation of Corpora

Language corpora often include analytic encodings or annotations, designed to support a variety of different views of language. The present Guidelines do not advocate any particular approach to linguistic annotation (or "tagging"); instead a number of general analytic facilities are provided which support the representation of most forms of annotation in a standard and self-documenting manner. Analytic annotation is of importance in many fields, not only in corpus linguistics, and is therefore discussed in general terms elsewhere in the Guidelines.[5] The present section presents informally some particular applications of these general mechanisms to the specific practice of corpus linguistics.

26.4.1 Levels of Analysis

By *linguistic annotation* we mean here any annotation determined by an analysis of linguistic features of the text, excluding as borderline cases both the formal structural properties of the text (e.g. its division into chapters or paragraphs) and descriptive information about its context (the circumstances of its production, its genre or medium). The structural properties of any TEI-conformant text should be represented using the structural elements discussed elsewhere in this chapter and in chapters 6, "Tags Available in All TEI DTDs," [in separate fascicle], 7, "Default Text Structure," [in separate fascicle], and the various chapters of Part III (on base tag sets). The contextual properties of a TEI text are fully documented in the TEI Header, which is discussed in chapter 5, "The TEI Header," [in separate fascicle], and in section 26.2, "Contextual Information," on page 7 of the present chapter.

[5] See in particular chapters 16, "Segmentation and Alignment," [in separate fascicle], 17, "Simple Analytic Mechanisms," [in separate fascicle], and 18, "Feature Structure Analysis," [in separate fascicle].

Other forms of linguistic annotation may be applied at a number of levels in a text. A code (such as a word-class or part-of-speech code) may be associated with each word or token, or with groups of such tokens, which may be continuous, discontinuous or nested. A code may also be associated with relationships (such as cohesion) perceived as existing between distinct parts of a text. The codes themselves may stand for discrete non-decomposable categories, or they may represent highly articulated bundles of textual features. Their function may be to place the annotated part of the text somewhere within a narrowly linguistic or discoursal domain of analysis, or within a more general semantic field, or any combination drawn from these and other domains.

The manner by which such annotations are generated and attached to the text may be entirely automatic, entirely manual or a mixture. The ease and accuracy with which analysis may be automated may vary with the level at which the annotation is attached. The method employed should be documented in the **<analysis>** element within the encoding description of the TEI Header, as described in section 5.3.3, "The Editorial Practices Declaration," [in separate fascicle]. Where different parts of a corpus have used different annotation methods, the *decls* attribute may be used to indicate the fact, as further discussed in section 26.3, "Associating Contextual Information with a Text," on page 20.

26.4.2 An Extended Example

As one example of such types of analysis, consider the following sentence, taken from the Lancaster/IBM Treebank Project:[6]

> The victim's friends told police that Kruger drove into the quarry
> and never surfaced.

Our discussion focuses on the way that this sentence might be analysed using the Claws system developed at the University of Lancaster, but exactly the same principles may be applied to a wide variety of other systems.[7] Output from the system consists of a segmented and tokenized version of the text, in which word class codes have been associated with each token. For our example sentence, we might conveniently represent these codes using entity references:

---

[6] See G. N. Leech and R. G. Garside, "Running a Grammar Factory," in *English Computer Corpora: Selected Papers and Research Guide,* ed. S. Johansson and A.-B. Stenström (Berlin: de Gruyter; New York: Mouton, 1991), pp. 15-32. This sentence and its analysis are reproduced by kind permission of the University of Lancaster's Unit for Computer Research on the English Language.

[7] For the word-class tagging method used by Claws see I. Marshall , "Choice of Grammatical Word Class without Global Syntactic Analysis: Tagging Words in the LOB Corpus," in *Computers and the Humanities* 17 (1983): 139-50.. For an overview of the system see R. G. Garside, G. N. Leech, and G. R. Sampson, *The Computational Analysis of English: a Corpus-Based Approach* (Oxford: Oxford University Press, 1991).

```
<s>The&AT victim&NN1 's&GEN friends&NN2 told&VVD police&NN2
that&CST Krueger&NP1  drove&VVD into&II the&AT
quarry&NN1  and&CC  never&RR surfaced&VVD
</s>
```

The names used for these entity references have some significance for the human reader (“AT” for article, “NN1” for singular noun, “NN2” for plural noun, etc.), but their representation in the output from an SGML system processing the document may be chosen to suit the convenience of whatever analytic software is to be used. For example, if the SGML parser operating on this sentence uses a set of entity declarations in the following form:

```
<!ENTITY AT "">
<!ENTITY NN1 "">
<!-- ... -->
```

then the wordclass tags will simply disappear from the output. Alternatively, if the entity set in use follows the following pattern:

```
<!ENTITY AT  "[definite article]">
<!ENTITY NN1 "[singular noun]">
<!ENTITY gen "[genitive suffix]">
<!-- ... -->
```

then our sample sentence will be processed by an SGML-aware processor as if it began

```
The[definite article] victim[singular noun] 's[genitive suffix] ...
```

More usefully, the replacement texts for each entity will be a code of some significance to a particular analysis program. In this case, it might be argued that some replacement should be chosen which will not obscure the fact that (for example) “NN1” and “NN2” have something in common, namely their “nounness”, which they do not share with (say) “VVD”. It is a matter for the analyst to decide whether or not the word class categories represented by these codes are decomposable into bundles of more primitive analytic features. In cases where they are, such analytic features should be defined as a set of *feature structures*, as described in chapters 18, “Feature Structure Analysis,” [in separate fascicle] and 29, “Feature System Declaration,” [in separate fascicle]. A suitable replacement for the word-class entity references above would then be an empty **<anchor>** element bearing a *struct* attribute (for which see chapter 18, “Feature Structure Analysis,” [in separate fascicle]).

```
<s>The <anchor struct=AT> victim <anchor struct=NN1>
's <anchor struct=amp> friends <anchor struct=NN2>
told <anchor struct=VVD> police <anchor struct=NN2>
that <anchor struct=CST> Krueger <anchor struct=NP1>
drove <anchor struct=VVD> into <anchor struct=II>
the <anchor struct=AT> quarry <anchor struct=NN1
and <anchor struct=CC>  never <anchor struct=RR>
surfaced <anchor struct=VVD>
</s>
```

The **<anchor>** element simply marks a place in the text. The *struct* attribute associates that place with the relevant feature structure by specifying its unique identifier. This solution requires that there be somewhere a formal definition for what exactly is meant by analytic code "NN1", in terms as simple or as complex as seems appropriate to the analyst. For example, we might represent the constituent features of the codes AT and NN1 using feature structure notation as follows:

```
<fs id=AT>
   <f name=formClass><sym value=article>
</fs>
<!-- ... -->
<fs id=NN1>
   <f name=formClass><sym value=noun>
   <f name=proper><minus>
   <f name=number><sym value=singular>
</fs>
```

For a more detailed presentation of this method of representing the analytic features underlying these word class code, refer to chapter 18, "Feature Structure Analysis," [in separate fascicle].

Although common practice, this method of presenting a token-level analysis requires application of an unstated convention about how the input stream is to be tokenized. The interpretive structures are associated not with a distinct stretch of text, but with a point in it, by convention immediately following the token to which they "really" apply. This convention makes it impossible to associate interpretive structures directly with sequences longer than a single token.

Neither of these inconveniences applies if the text is fully segmented, using the general purpose **<s>** element both to represent each token and the outer segment containing the whole sentence:

```
<s>
<s struct=AT  >The</s>
<s struct=NN1 >victim</s>
<s struct=amp >'s</s>
<s struct=NN2 >friends</s>
<s struct=VVD >told</s>
<s struct=NN2 >police</s>
<s struct=CST >that</s>
<s struct=NP1 >Krueger</s>
<s struct=VVD >drove</s>
<s struct=II  >into</s>
<s struct=AT  >the</s>
<s struct=NN1 >quarry</s>
<s struct=CC  >and</s>
<s struct=RR  >never</s>
<s struct=VVD >surfaced</s>
</s>
```

As this example demonstrates, the **<s>** element has scope and may thus be used to represent segmentation carried out at several levels, provided that the whole of the segmentation is describable by a single hierarchic tree. For example, the **<s>** element might be used to mark the boundaries of a high-level unit roughly corresponding with the traditional notion of sentence or "s-unit", as discussed in section 6.11, "Segmentation," [in separate fascicle] as well as its individual tokens (as above) together with a third, intermediate, level of analysis, corresponding with the traditional notion of phrase. It might also be used to indicate a lower-level morphological level of analysis. To distinguish amongst these various uses of the **<s>** element, the *type* attribute may be used, as in the following example:

```
<s type=sent>
   <s type=phrasal>
      <s type=lexical struct=AT>The</s>
      <s type=lexical struct=NN1>
         <s type=morph struct=NN1>victim</s>
         <s type=morph struct=GEN>'s</s>
      </s>
      <s type=lexical struct=NN2>friends</s>
   </s>
   <s type=phrasal>
      <s type=lexical struct=VVD>told</s>
   </s>
   <s type=phrasal>
      <s type=lexical struct=NN2>police</s>
   </s>
      <!-- ... -->
</s>
```

However, in projects where annotation is routinely carried out at multiple levels it will often be felt preferable to define an element appropriate to each level of analysis. For example, Claws provides the following constituent analysis of the sample sentence:

> [N [G The_AT victim_NN1 's_$ G] friends_NN2 N] [V told_VVD [N police_NN2 N] [Fn that_CST [N Krueger_NP1 N] [V [V& drove_VVD [P into_II [N the_AT quarry_NN1 N]P]V&] and_CC [V+ never_RR surfaced_VVD V+]V]Fn]V] ._.

Using **<c>** for **<s type=constituent>**, this analysis of the structure of our example sentence might be represented as follows:

```
<s type=sent>
<c struct=N>
   <c struct=G> The victim's </c>
   friends
</c>
<c struct=V>
   told
   <c struct=N> police </c>
   <c struct=Fn>
      that
```

```
          <c struct=N> Krueger </c>
          <c struct=V>
              <c id=c1 struct=VX next=c2> drove
                 <c struct=P> into
                    <c struct=N> the quarry </c>
                 </c>
              </c>
              and
              <c id=c2 struct=VX prev=c1> never surfaced </c>
          </c>
       </c>
    </c>
    </s>
```

In the above analysis, the value borne by each *struct* attribute corresponds with one of a set of feature structures describing the constitutive function of each sequence of tokens. For example, *N* is a "noun phrase", *G* a "genitive", *FN* a "noun clause", and *V* a "verb phrase". The feature structures indicated will contain further information about the classification used, in the same way as for the word-level tagging described above. Alternatively, if no further specification for the labels attached to constituents is required, then a *type* attribute might be defined for the **<c>** element. In such a case, the significance of the values specified should be specified in the **<encodingDecl>** element of the header.

26.4.3 Non-nesting Structures

Each analytic segment so far discussed has been well-behaved with respect to the basic document hierarchy, having only a single parent. Moreover, the segmentation has been complete, in that each part of the text is accounted for by some segment at each level of analysis, without discontinuities or overlap. This happy state of affairs does not of course apply in all types of analysis, and these Guidelines therefore provide a number of mechanisms to support the representation of discontinuities or multiple analyses. A brief overview of these facilities is provided in chapter 36, "Markup for Non-Hierarchical Phenomena," [in separate fascicle].

These mechanisms all depend to a greater or less degree on the ability to associate a unique identifier with any element in a TEI-conformant text, and then to specify that identifier as the target of a pointing element of some kind. The previous section included one example of such a linkage in its use of the *next* and *prev* attributes to represent the fact that the two *conjunct* verbal phrases in the example may be aggregated together. A similar technique might be used to represent anaphoric links, for example the fact that "the victim" and "Krueger" in this example both refer to the same person:

```
<c struct=N corresp=N1>The victim</c> 's
   friends
<c struct=V>
   told
   <c struct=N> police </c>
   <c struct=Fn>
```

```
        that
        <c struct=N id=N1> Krueger </c>
  <!-- ... -->
```

These attributes, and a number of other related topics, are further discussed in section 16.4, "Aggregation and Virtual Elements," [in separate fascicle] and 16.3, "Correspondence and Alignment," [in separate fascicle].

26.4.4 Further Examples

The same mechanisms may also be used to encode analyses of an entirely different kind. Here, for example, the feature structures being associated with the various segments of the text are concerned with their discoursal function:

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

The associated feature structures will document any internal substructure the analyst wishes to associate with each category assigned to the various parts of the discourse above, or simply document them further; in this case specifying that *SR* is a "sale request", *SC* a "sale compliance", *PC* a "purchase closure" and so on. For further discussion of the **<u>** (utterance) element and other elements recommended for transcriptions of spoken language, see chapter 10, "Base Tag Set for Transcriptions of Spoken Texts," [in separate fascicle].

26.5 Recommendations for the Encoding of Large Corpora

These Guidelines include proposals for the identification and encoding of a far greater variety of textual features and characteristics than is likely to be either feasible or desirable in any one language corpus, however large and ambitious. (The reasoning behind this catholic approach is further discussed in chapter 1, "About These Guidelines," [in separate fascicle].) For most large scale corpus projects, it will therefore be necessary to determine a subset of TEI recommended elements appropriate to the anticipated needs of the project. Mechanisms for tailoring the TEI DTDs to implement such a subset are described in chapters 3, "Structure of the TEI Document Type Declarations," [in separate fascicle] and 32, "Modifying TEI DTDs," [in separate fascicle]; they include the ability to exclude selected element types, add new element types, and change the names of existing elements. A discussion of the implications of such changes for

TEI conformance is provided in chapter 31, "TEI Conformance," [in separate fascicle].

Because of the high cost of identifying and encoding many textual features, and the difficulty in ensuring consistent practice across very large corpora, encoders may find it convenient to divide the set of elements to be encoded into the following three categories:

required: texts included within the corpus will always encode textual features in this category, should they exist in the text

recommended: textual features in this category will be encoded wherever economically and practically feasible; where present but not encoded, a note in the header should be made.

optional: textual features in this category may or may not be encoded; no conclusion about the absence of such features can be inferred from the absence of the corresponding element in a given text.

Part 7

Alphabetic Reference List of Tags

<activity> (i.e. activity)

Description: contains a brief informal description of what a participant in a language interaction is doing other than speaking, if anything.
Attributes: [None: global and inherited attributes only.]
Example:

```
<activity>driving</activity>
```

Remarks: For more fine-grained description of participant activities during a spoken text, the **<event>** element should be used.
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within: setting
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT activity      - o  (%phrase.seq)                     >
```

Attribute List Declaration:

```
<!ATTLIST activity           %a.global;                        >
```

Discussed in 26.2.3, “The Setting Description,” on page 18.
This tag is optional.

<affiliation> (i.e. affiliation)

Description: contains an informal description of a person's present or past affiliation with some organization, for example an employer or sponsor.
Attributes: [None: global and inherited attributes only.]
Example:

```
<affiliation>Junior project officer for the US
 <org>National Endowment for the Humanities</org>
</affiliation>
```

Remarks: If included, the name of the organization should be tagged using the **<org>** element.
Part: auxiliary tag set for corpora and collections.

Member of classes:
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within:
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT affiliation   - -  (%phrase.seq)                    >
```

Attribute List Declaration:

```
<!ATTLIST affiliation        %a.global;                       >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

Class Name: **agent** (i.e. agent (individual or corporate body))
Description: defines a group of elements which contain names of individuals or corporate bodies.
Attributes: [None: global and inherited attributes only.]
Remarks: This class is used in the **<resp>** element, to allow a statement of responsibility to apply to an individual or a body.
Part: type=base name=core>.
Member of classes:
DTD file:
Discussed in 6.4, “Names, Numbers, Dates, and Abbreviations,” [in separate fascicle].

<birth> (i.e. Birth details)

Description: contains information about a person's birth, such as its date and place.
Attributes:

date: specifies the date of birth in a ISO standard form (yyyy-mm-dd).
Data type: %ISO-date
Value: a date in ISO standard form (yyyy-mm-dd).
Default value: #IMPLIED
This attribute is optional.

Example:

```
<birth>Before 1920, Midlands region.

<birth date=1960-12-10>In a small cottage near
<place>Aix-la-Chapelle</place>, early in the morning of <date>10 Dec
1960</date>
```

Remarks: Dates and places, if included in the content of this element, should be tagged using the general **<date>** and **<place>** element.
Part: auxiliary tag set for corpora and collections.
Member of classes: demog -->
DTD file:
Data description:
May occur within:
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT birth         - -  (%phrase.seq)                     >
```

Attribute List Declaration:

```
<!ATTLIST birth              %a.global;
          date               %ISO-date          #IMPLIED       >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<channel> (i.e. primary channel)

Description: describes the medium or channel by which a text is delivered or experienced. For a written text, this might be print, manuscript, e-mail, etc.; for a spoken one, radio, telephone, face-to-face, etc.
Attributes:

mode: specifies the mode of this channel with respect to speech and writing.
Data type: (s | w | ws | sw | m | x)
Legal values are:
s : spoken
w : written
sw : spoken to be written (e.g. dictation)
ws : written to be spoken (e.g. a script)
m : mixed modes
x : unknown or inapplicable

Default value: x
Example:

```
<channel mode=s>face-to-face conversation
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description:
May occur within: textDesc
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT channel       - o  (%phrase.seq)                      >
```

Attribute List Declaration:

```
<!ATTLIST channel            %a.global;
          mode               (s | w | ws | sw | m | x)
                                                    x               >
```

Discussed in 26.2.1, “The Text Description,” on page 9.
This tag is optional.

Class Name: **chunk**

Description: includes all elements which can occur between, but not within, paragraphs and other chunks.
Attributes: [None: global and inherited attributes only.]
Remarks: This element class does not contain all those elements which can appear between chunks: the class *inter* contains a set of elements which can appear either within or between chunks. Unlike elements of that class, chunks cannot occur within chunks. In prose, this means the elements in this class can appear between but not within paragraphs.
Part: type=base name=core>.
Member of classes: common
DTD file:
Discussed in 3.4, “Element Classes and Other Parameter Entities,” [in separate fascicle].

Class Name: **common**

Description: common chunk- and inter-level elements.
Attributes: [None: global and inherited attributes only.]
Remarks: This class defines the set of chunk- and inter-level elements available in all bases; it is used in defining the standard models *chunk.seq* and *specialPara* in the general and mixed bases.
Part: type=base name=core>.

Member of classes:
DTD file:
Discussed in 3.4, “Element Classes and Other Parameter Entities,” [in separate fascicle].

<constitution>

Description: describes the internal composition of a text or text sample, for example as fragmentary, complete, etc.
Attributes:

type: specifies how the text was constituted.
Data type: (single | composite | frags | unknown)
Legal values are:
single : a single complete text
composite : a text made by combining several smaller items, each individually complete
frags : a text made by combining several smaller, not necessarily complete, items
unknown : composition unknown or unspecified

Default value: single
Example:

```
<constitution type=frags>Prologues only.
```

Remarks: The function of this element seems to overlap with both the ORG attribute on DIVs and the samplingDecl in the encodingDesc.
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: contains any sequence of phrase level data
May occur within: textDesc
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT constitution  - o  (%phrase.seq)                    >
```

Attribute List Declaration:

```
<!ATTLIST constitution       %a.global;
          type               (single | composite | frags |
                             unknown)            single         >
```

Discussed in 26.2.1, “The Text Description,” on page 9.
This tag is optional.

Class Name: **data**

Description: phrase-level elements containing names, dates, numbers, measures, and similar data.
Attributes: [None: global and inherited attributes only.]
Remarks:
Part: type=base name=core>.
Member of classes: phrase
DTD file:
Discussed in 6.4, “Names, Numbers, Dates, and Abbreviations,” [in separate fascicle].

<date>

Description: contains a date in any format.
Attributes:

calendar: indicates the system or calendar to which the date belongs.
Data type: CDATA
Value: Recommended values include: *Gregorian*, *Julian*, *Roman*, *Mosaic*, *Revolutionary*, *Islamic*.
Default value: #IMPLIED
Example:

```
He was born on <date calendar=Gregorian>Feb. 22, 1732</date>
(<date calendar=Julian value='1732-2-22'>Feb. 11, 1731/32,
O.S.</date>).
```

This attribute is optional.

value: gives the value of the date in some standard form, usually yyyy-mm-dd.
Data type: CDATA
Value: Any string representing a date in standard format; recommended form is “yyyy-mm-dd”, as defined by ISO 8601: 1988, *Data elements and interchange formats — Information interchange — Representation of dates and times.*
Default value: #IMPLIED
Example:

```
This list begins in the year 1632, more precisely
on Trinity Sunday, i.e. the Sunday after Pentecost,
in that year the <date calendar=Julian value=1632-6-6>
27th of May (old style)</date>.
```

Remarks: For simple dates, the value should give the Gregorian date in the form (yyyy-mm-dd) specified by ISO 8601. More complicated dates or special applications may require another calendar or another form; these should be documented in the **<std.vals>** element in the TEI Header.
This attribute is optional.

certainty: indicates the degree of certainty attributed to the date.
Data type: CDATA
Value: Any appropriate value, e.g. *ca.*, *approx*, *after*, *before*.
Default value: #IMPLIED
This attribute is optional.

Example:

```
<date value='1980-02-21'>21 Feb 1980</date>

Given on the <date value='1977-06-12'>Twelfth Day of June
in the Year of Our Lord One Thousand Nine Hundred and
Seventy-seven of the Republic the Two Hundredth and first
and of the University the Eighty-Sixth.</date>

<date value='1990-09'>September 1990</date>
```

Remarks:
Part: additional tag set for common core features.
Member of classes: data date -->
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within:
May contain: %phrase.seq;
Element Declaration:

```
<!ELEMENT date          - -  (%phrase.seq;)                     >
```

Attribute List Declaration:

```
<!ATTLIST date                %a.global;
          calendar            CDATA               #IMPLIED
          value               CDATA               #IMPLIED
          certainty           CDATA               #IMPLIED      >
```

Discussed in 6.4.3, “Dates,” [in separate fascicle].
This tag is optional.

Class Name: **declarable**

Description: elements which may be independently selected (using the special purpose *decls* attribute) from a candidate list of declarations within a TEI header.
Attributes:

default: indicates whether or not this element is selected by default when its parent is selected.
Data type: (YES|NO)

YES : This element is selected if its parent is selected
NO : This element can only be selected explicitly, unless it is the only one of its kind, in which case it is selected if its parent is selected.

Default value: NO
This attribute is mandatory when applicable.

Remarks: The rules governing the association of declarable elements with individual parts of a TEI text are fully defined in chapter 26.3, "Associating Contextual Information with a Text," on page 20. Only one element of a particular type may carry the value "default=yes".
Part: type=aux name=hdr>.
Member of classes:
DTD file:
Discussed in 26.3, "Associating Contextual Information with a Text," on page 20.

Class Name: **declaring**
Description: elements which may be independently associated with a particular declarable element within the header, thus overriding the inherited default for that element.
Attributes:

decls: identifies one or more *declarable elements* within the header, which are understood to apply to the element bearing this attribute and its content.
Data type: IDREFS
Value: must identify a set of declarable elements of different types.
Default value: #IMPLIED
This attribute is mandatory when applicable.

Remarks: The rules governing the association of declarable elements with individual parts of a TEI text are fully defined in chapter 26.3, "Associating Contextual Information with a Text," on page 20.
Part: type=aux name=hdr>.
Member of classes:
DTD file:
Discussed in 26.3, "Associating Contextual Information with a Text," on page 20.

Class Name: **demographic**
Description: elements describing demographic characteristics of the participants in a linguistic interaction.
Attributes: [None: global and inherited attributes only.]
Remarks:
Part: type=aux name=hdr>.
Member of classes:

DTD file: TEIcorp2
Discussed in 26.2.2, “The Participants Description,” on page 14.

<derivation>

Description: describes the nature and extent of indebtedness or derivativeness of this text with respect to others.
Attributes:

type: categorizes the derivation of the text.
Data type: CDATA
Sample values include:
original : text is original
revision : text is a revision of some other text
translation : text is a translation of some other text
abridgment : text is an abridged version of some other text
plagiarism : text is plagiarized from some other text
traditional : text has no obvious source but is one of a number derived from some common ancestor

Default value: #IMPLIED
Example:

```
<derivation type=original>
```

Remarks: For derivative texts, details of the ancestor may be included in the source description.
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description:
May occur within: textDesc
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT derivation    - o  (%phrase.seq)                      >
```

Attribute List Declaration:

```
<!ATTLIST derivation         %a.global;
          type               CDATA               #IMPLIED       >
```

Discussed in 26.2.1, “The Text Description,” on page 9.
This tag is optional.

<domain> (i.e. Domain of use)

Description: describes the most important social context in which the text was realized or for which it is intended, for example private vs. public, education, religion, etc.

Attributes:

type: categorizes the domain of use.
Data type: CDATA
Sample values include:
art : art and entertainment
domestic : domestic and private
religious : religious and ceremonial
business : business and work place
education : education
govt : government and law
public : other forms of public context

Default value: #IMPLIED

Example:

```
<domain type=dom>
<domain type=rel>religious broadcast</domain>
```

Remarks: The list presented here is primarily for illustrative purposes.
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: Usually empty, unless some further clarification of the type attribute is needed, in which case it may contain running prose.
May occur within: textDesc
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT domain        - o  (%phrase.seq)                     >
```

Attribute List Declaration:

```
<!ATTLIST domain             %a.global;
          type               CDATA             #IMPLIED       >
```

Discussed in 26.2.1, “The Text Description,” on page 9.
This tag is optional.

<education> (i.e. Education)

Description: contains a brief prose description of the educational background of a participant.
Attributes: [None: global and inherited attributes only.]
Example:

```
<education>Left school at age 16
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes: demog -->
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within:
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT education     - o  (%phrase.seq)                    >
```

Attribute List Declaration:

```
<!ATTLIST education          %a.global;                        >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<factuality>

Description: describes the extent to which the text may be regarded as imaginative or non-imaginative, that is, as describing a fictional or a non-fictional world.
Attributes:

type: categorizes the factuality of the text.
Data type: (fiction|fact|mixed|inapplicable)
Legal values are:
fiction : the text is to be regarded as entirely imaginative
fact : the text is to be regarded as entirely informative or factual
mixed : the text contains a mixture of fact and fiction
inapplicable : the fiction/fact distinction is not regarded as helpful or appropriate to this text

Default value: #IMPLIED
Example:

```
<factuality type=fiction>

<factuality type=mixed>contains a mixture of gossip and
  speculation about real people and events
```

Remarks: For many literary texts, a simple binary opposition between “fiction” and “fact” is naï;ve in the extreme; this parameter is not intended for purposes of subtle literary analysis, but as a simple means of characterising the claimed fictiveness of a given text. No claim is made that works characterised as “fact” are in any sense “true”.

Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: Usually empty, unless some further clarification of the type attribute is needed, in which case it may contain running prose
May occur within: textDesc
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT factuality    - o  (%phrase.seq)                     >
```

Attribute List Declaration:

```
<!ATTLIST factuality         %a.global;
          type               (fiction|fact|mixed|inapplicable)
                                                   #IMPLIED       >
```

Discussed in 26.2.1, “The Text Description,” on page 9.
This tag is optional.

<firstLang> (i.e. First language)

Description: specifies the first language of a participant.
Attributes: [None: global and inherited attributes only.]
Example:

```
<firstLang>French
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes: demog -->
DTD file:
Data description:
May occur within:
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT firstLang     - o  (%phrase.seq)                     >
```

Attribute List Declaration:

```
<!ATTLIST firstLang          %a.global;                        >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<group>

Description: contains the body of a composite text, grouping together a sequence of distinct texts (or groups of such texts) which are regarded as a unit for some purpose, for example the collected works of an author, a sequence of prose essays, etc.
Attributes: [None: global and inherited attributes only.]
Example:

Remarks:
Part: auxiliary tag set for common core features.
Member of classes: decling -->
DTD file:
May occur within: group text
May contain: %m.divbot; %m.divtop; group text
Element Declaration:

```
<!ELEMENT group        - O  ((%m.divtop;)*, (text | group)+,
                            (%m.divbot;)*)                    >
```

Attribute List Declaration:

```
<!ATTLIST group             %a.global;
                            %a.declaring;                    >
```

Discussed in 26.1, "Varieties of Composite Text," on page 4.
Discussed in 7, "[Groups and Composite Texts]," [in separate fascicle].
This tag is mandatory when applicable.

Class Name: **inter**

Description: elements of the intermediate (inter-level) class, which can occur both within paragraphs and between paragraphs.
Attributes: [None: global and inherited attributes only.]
Remarks: This element class contains a subset of those elements which can appear in the unstructured "soup" with which paragraph and other elements at the lowest level of crystal structures are filled: specifically all the elements which can also occur as structural elements in their own right. In prose, this means the elements in this class can appear both within and between paragraphs. This class is thus distinct from the purely phrase-level elements which can appear only within soup, and not on their own; the latter class, in keeping with this metaphor, is called "broth"; it is represented by the class *phrase*. Cf. also the class *chunks*.
Part: type=base name=core>.
Member of classes:
DTD file:
Discussed in 3.4, "Element Classes and Other Parameter Entities," [in separate fascicle].

<interaction>

Description: describes the extent, cardinality and nature of any interaction among those producing and experiencing the text, for example in the form of response or interjection, commentary etc.

Attributes:

type: specifies whether or not there is any interaction between active and passive participants in the text.

Data type: (none|partial|complete|inapplicable)

Legal values are:

none : no interaction of any kind, e.g. a monologue

partial : some degree of interaction, e.g. a monologue with set responses

complete : complete interaction, e.g. a face to face conversation

inapplicable : this parameter is inappropriate or inapplicable in this case

Default value: #IMPLIED

active: specifies the number of active participants (or *addressors*) producing parts of the text.

Data type: CDATA

Legal values are:

singular : a single addressor

plural : many addressors

corporate : a corporate addressor

unknown : number of addressors unknown or unspecifiable

Default value: #IMPLIED

passive: specifies the number of passive participants (or *addressees*) to whom a text is directed or in whose presence it is created or performed.

Data type: CDATA

Suggested values include:

self : text is addressed to the originator e.g. a diary

single : text is addressed to one other person e.g. a personal letter

many : text is addressed to a countable number of others e.g. a conversation in which all participants are identified

group : text is addressed to an undefined but fixed number of participants e.g. a lecture

world : text is addressed to an undefined and indeterminately large number e.g. a published book

Default value: #IMPLIED

Example:

```
<!-- An informal conversation:  ->
<interaction type=complete active=plural passive=many>
```

```
<!-- A formal lecture:-->
<interaction type=none active=singular passive=group>
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description:
May occur within: textDesc
May contain: %phrase.seq;
Element Declaration:

```
<!ELEMENT interaction   - o  (%phrase.seq;)                   >
```

Attribute List Declaration:

```
<!ATTLIST interaction        %a.global;
          type               (none|partial|complete|
                             inapplicable)       #IMPLIED
          active             CDATA               #IMPLIED
          passive            CDATA               #IMPLIED      >
```

Discussed in 26.2.1, “The Text Description,” on page 9.
This tag is optional.

<langKnown> (i.e. linguistic competence)

Description: contains an informal description of a person's competence in different languages, dialects, etc.
Attributes: [None: global and inherited attributes only.]
Example:

```
<langKnown>Spoken French
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within:
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT langKnown     - -  (%phrase.seq)                    >
```

Attribute List Declaration:

```
<!ATTLIST langKnown          %a.global;                          >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<locale> (i.e. locale)

Description: contains a brief informal description of the nature of a place for example a room, a restaurant, a park bench etc.
Attributes: [None: global and inherited attributes only.]
Example:

```
<locale>a fashionable restaurant
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within: setting
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT locale        - o  (%phrase.seq)                       >
```

Attribute List Declaration:

```
<!ATTLIST locale             %a.global;                          >
```

Discussed in 26.2.3, “The Setting Description,” on page 18.
This tag is optional.

Class Name: names

Description: elements with proper nouns as content.
Attributes:

key: provides an alternative identifier for the object being named, such as a database record key.
Data type: CDATA
Value: any string
Default value: #IMPLIED
Example:

```
<name type='place'>Montaillou</name>
is not a large parish.
At the time of the events which led to <name type='person' key=BXII>
Fournier's</name> investigations,
the local population consisted of between 200 and
250 inhabitants.
```

Remarks: The value may be a unique identifier from a database, or simply a more explicit name for the referent. Its purpose is only to record an identification; if the analysis leading to the identification is to be recorded as well, the analytic tags described in chapter 18, “Feature Structure Analysis,” [in separate fascicle] should be used in addition or instead.
This attribute is optional.

reg (i.e. regularization): gives a normalized or regularized form of the proper name
Data type: CDATA
Value: Any string of characters.
Default value: #IMPLIED
Example:

```
At the time of the events which led to
<name type=person reg='Benedict XII, Pope of Avignon
(Jacques Fournier)'>Fournier's</name> investigations,
the local population consisted of between 200 and
250 inhabitants.
```

Remarks: In providing a “regularized” form, no claim is made that the form in the source text is incorrect; the regularized form is simply that chosen as the main form for purposes of unifying variant forms under a single heading.
This attribute is optional.

Remarks:
Part: type=base name=core>.
Member of classes:
DTD file:
Discussed in 6.4.1, “Names,” [in separate fascicle].

<occupation>

Description: contains an informal description of a person's trade, profession or occupation.
Attributes:

scheme: identifies the classification system or taxonomy in use by supplying the identifier of a **<taxonomy>** element elsewhere in the header.
Data type: IDREF

Value: must identify a **<taxonomy>** element
Default value: #IMPLIED
This attribute is optional.

code: identifies an occupation code defined within the classification system or taxonomy defined by the *source* attribute.
Data type: IDREF
Value: Must identify a **<category>** element
Default value: #IMPLIED
This attribute is optional.

Example:

```
<occupation>accountant</occupation>

<occupation source=RG code=ACC>

<occupation source=RG code=ACC>
      accountant with specialist knowledge of oil industry
</occupation>
```

Remarks: The content of this element may be used as an alternative to the more formal specification made possible by its attributes; it may also be used to supplement the formal specification with commentary or clarification.
Part: auxiliary tag set for corpora and collections.
Member of classes: demog -->
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within:
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT occupation    - -  (%phrase.seq)                     >
```

Attribute List Declaration:

```
<!ATTLIST occupation         %a.global;
          scheme             IDREF              #IMPLIED
          code               IDREF              #IMPLIED       >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<particDesc> (i.e. participation description)

Description: describes the identifiable speakers, voices or other participants in a linguistic interaction.
Attributes: [None: global and inherited attributes only.]
Example:

```
<particDesc>
<participant id=P1 sex=F age=42>
  <p>Female informant, well-educated, born in Shropshire
   UK, 12 Jan 1950, of unknown occupation.
   Speaks French fluently. Socio-Economic status B2.
</participant>
<participant id=P2 sex=M age=43>
<particLinks>
   <relation active="P1 P2" desc="spouse">
</particLinks>
</particDesc>
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes: declabl -->
DTD file: TEIcorp2
Data description: May contain a prose description organized as paragraphs, or a structured list of participants and participant groups, with an optional formal specification of any relationships amongst them.
May occur within:
May contain: p participant particGrp particLinks
Element Declaration:

```
<!ELEMENT particDesc    - o  (p+ | ( (participant | particGrp)+,
                             particLinks?) )                   >
```

Attribute List Declaration:

```
<!ATTLIST particDesc         %a.global;
                             %a.declarable;                   >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<participant>

Description: describes a single participant in a language interaction.
Attributes:

role: specifies the role of this participant in the group.
Data type: CDATA
Value: a set of keywords to be defined
Default value: #IMPLIED

sex: specifies the sex of the participant.
Data type: (m | f | u)

m : male
f : female

u : unknown or inapplicable

Default value: #IMPLIED

age: specifies the age group to which the participant belongs.
Data type: CDATA
Value: suggested values are to be supplied
Default value: #IMPLIED

Example:

```
<participant id=P1 sex=F age=42>
  <p>Female informant, well-educated, born in Shropshire
   UK, 12 Jan 1950, of unknown occupation.
   Speaks French fluently. Socio-Economic status B2.
</participant>
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: May contain a prose description organized as paragraphs, or any sequence of demographic elements in any combination.
May occur within: particDesc
May contain: %m.demog p
Element Declaration:

```
<!ELEMENT participant   - o  (p+ | (%m.demog)* )                >
```

Attribute List Declaration:

```
<!ATTLIST participant        %a.global;
          role               CDATA              #IMPLIED
          sex                (m | f | u )       #IMPLIED
          age                CDATA              #IMPLIED       >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<particGrp> (i.e. participant group)

Description: describes a group of individuals treated as a single participant.
Attributes:

role: specifies the role of this group of participants in the interaction.
Data type: CDATA
Value: a set of keywords to be defined
Default value: #IMPLIED

sex: specifies the sex of the participant group.

Data type: (m | f | u | x)

m : male
f : female
u : unknown
x : mixed

Default value: #IMPLIED
This attribute is optional.

age: specifies the age group of the participants.
Data type: CDATA
Value: suggested values are to be supplied
Default value: #IMPLIED

size: specifies the size or approximate size of the group.
Data type: CDATA
Value: may contain a number and an indication of accuracy, e.g. 'approx 200'
Default value: #IMPLIED

Example:

```
<particGrp role=audience sex=x size='approx 50' id=PG1>
```

Remarks: The global *id* attribute should be used to identify each speaking participant in a spoken text if the *who* attribute is specified on individual utterances.
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: May contain a prose description organized as paragraphs, or any sequence of demographic elements in any combination.
May occur within: particDesc
May contain: %m.demog p
Element Declaration:

```
<!ELEMENT particGrp     - o  (p+ | (%m.demog)* )                >
```

Attribute List Declaration:

```
<!ATTLIST particGrp          %a.global;
          role               CDATA              #IMPLIED
          sex                (m | f | u | x)    #IMPLIED
          age                CDATA              #IMPLIED
          size               CDATA              #IMPLIED       >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<particLinks> (i.e. participant relationships)

Description: describes the relationships or social links existing between participants in a linguistic interaction.
Attributes: [None: global and inherited attributes only.]
Example:

```
<particLinks>
  <relation desc='parent' active="P1 P2"
         passive="P3 P4" mutual=N>
  <relation desc='spouse' active="P1 P2">
  <relation class='social' desc='employer'
          active=P1 passive='P3 P5 P6 P7' mutual=N>
</particLinks>
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: May contain a prose description organized as paragraphs, or a sequence of **<relation>** elements.
May occur within: particDesc
May contain: p relation
Element Declaration:

```
<!ELEMENT particLinks   - o  (p+ | relation+)                  >
```

Attribute List Declaration:

```
<!ATTLIST particLinks        %a.global;                        >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<person>

Description: contains a name of a person, or a reference to a person.
Attributes: [None: global and inherited attributes only.]
Example:

Remarks: This element may be used either for proper nouns only (in which case it is synonymous with **<name type='person'>**) or for any reference to persons it is felt desirable to tag.[1]
Part: additional tag set for common core features.

[1] This ambiguity may need to be removed. Use broader definition only and expect **<name>** to nest within **<person>** if the tagger wants it that way? -MSM I think it shd be restricted to names -LB

Member of classes: data agent names demog -->
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within:
May contain: %phrase.seq;
Element Declaration:

```
<!ELEMENT person        - -  (%phrase.seq;)                    >
```

Attribute List Declaration:

```
<!ATTLIST person             %a.global;
                             %a.names;                        >
```

Discussed in 6.4.1, “Names,” [in separate fascicle].
This tag is optional.

Class Name: **phrase**

Description: includes elements which can occur at the level of individual words or phrases.
Attributes: [None: global and inherited attributes only.]
Remarks: This class of elements can occur only within larger elements of the class *inter* or *chunk*. In prose, this means these elements can occur within paragraphs, list items, lines of verse, etc.
Part: type=base name=core>.
Member of classes:
DTD file:
Discussed in 3.4, “Element Classes and Other Parameter Entities,” [in separate fascicle].

<place>

Description: contains the name of, or a reference to, a place.
Attributes: [None: global and inherited attributes only.]
Example:

Remarks: This element may be used either for proper nouns only (in which case it is synonymous with **<name type='place'>**) or for any reference to places it is felt desirable to tag.[2]

[2] This ambiguity may need to be removed. Use broader definition only and expect **<name>** to nest within **<place>** if the tagger wants it that way? -MSM I think it shd be restricted to names -LB

Part: additional tag set for common core features.
Member of classes: data demog names -->
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within: setting
May contain: %phrase.seq;
Element Declaration:

```
<!ELEMENT place          - -  (%phrase.seq;)                    >
```

Attribute List Declaration:

```
<!ATTLIST place               %a.global;                        >
```

Discussed in 6.4.1, “Names,” [in separate fascicle].
This tag is optional.

<preparedness>

Description: describes the extent to which a text may be regarded as prepared or spontaneous.
Attributes:

type: a keyword characterizing the type of preparedness.
Data type: CDATA
Sample values include:
none : spontaneous or unprepared
scripted : follows a script
formulaic : follows a predefined set of conventions
revised : polished or revised before presentation

Default value: #IMPLIED

Example:

```
<preparedness type='none'>
```

Remarks: Difficulty in forming a consensus here means that I have allowed only a single attribute: possibly a degree attribute should be added (LB)
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description:
May occur within: textDesc
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT preparedness  - o  (%phrase.seq)                    >
```

Attribute List Declaration:

```
<!ATTLIST preparedness       %a.global;
          type               CDATA               #IMPLIED      >
```

Discussed in 26.2.1, “The Text Description,” on page 9.
This tag is optional.

<purpose>

Description: characterizes a single purpose or communicative function of the text.
Attributes:

type: specifies a particular kind of purpose.
Data type: CDATA
Suggested values include:
persuade : didactic, advertising, propaganda etc
express : self expression, confessional etc
inform : convey information, educate etc
entertain : amuse, entertain etc

Default value: #IMPLIED

degree: specifies the extent to which this purpose predominates.
Data type: (high | medium | low | unknown)
Legal values are:
high : this purpose is predominant
medium : this purpose is intermediate
low : this purpose is weak
unknown : extent unknown

Default value: #IMPLIED
Example:

```
<purpose type=persuade degree=high>
<purpose type=entertain degree=low>
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: Usually empty, unless some further clarification of the type attribute is needed, in which case it may contain running prose
May occur within: textDesc
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT purpose       - o  (%phrase.seq)                      >
```

Attribute List Declaration:

```
<!ATTLIST purpose            %a.global;
          type               CDATA              #IMPLIED
          degree             (high | medium | low | unknown)
                                                #IMPLIED       >
```

Discussed in 26.2.1, “The Text Description,” on page 9.
This tag is optional.

<relation> (i.e. relationship)

Description: describes any kind of relationship or linkage amongst a specified group of participants.

Attributes:

type: categorizes the relationship in some respect, e.g. as social, personal or other.
Data type: CDATA
Suggested values include:
social : relationship concerned with social roles
personal : relationship concerned with personal roles, e.g. kinship, marriage, etc.
other : other kinds of relationship
Default value: personal

desc: briefly describes the relationship.
Data type: CDATA
Value: an open list of application-dependent keywords
Default value: #IMPLIED

active: identifies the “active”participants in a non-mutual relationship, or all the participants in a mutual one.
Data type: IDREFS
Value: a list of identifier values for participant or participant groups
Default value: #IMPLIED

passive: identifies the “passive” participants in a non-mutual relationship.
Data type: IDREFS
Value: a list of identifier values for participant or participant groups
Default value: #IMPLIED

mutual: indicates whether the relationship holds equally amongst all the participants.
Data type: (Y | N)
Legal values are:

Y : the relationship is mutual
N : the relationship is directed
Default value: Y

Example:

```
<relation type=social desc="supervisor"
     active=P1 passive="P2 P3 P4" mutual=N>

<relation type=personal desc="friends"
     active="P2 P3 P4" mutual=Y>
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes:
DTD file:
Data description: Empty
May occur within: particLinks
May contain: EMPTY
Element Declaration:

```
<!ELEMENT relation      - o  EMPTY                              >
```

Attribute List Declaration:

```
<!ATTLIST relation           %a.global;
          type               CDATA              personal
          desc               CDATA              #IMPLIED
          active             IDREFS             #IMPLIED
          passive            IDREFS             #IMPLIED
          mutual             (Y | N)            Y              >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<residence> (i.e. residence)

Description: describes a person's present or past places of residence.
Attributes: [None: global and inherited attributes only.]
Example:

```
<residence>Childhood in East Africa, long term
  resident of Glasgow, Scotland.
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes: demog -->
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within:

May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT residence     - o  (%phrase.seq)                      >
```

Attribute List Declaration:

```
<!ATTLIST residence          %a.global;                         >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<settingDesc> (i.e. setting description)

Description: describes the setting or settings within which a language interaction takes place, either as a prose description or as a series of **<setting>** elements.
Attributes: [None: global and inherited attributes only.]
Example:

```
<settingDesc>
<p>Texts recorded in the Canadian Parliament building
in Ottawa, between April and November 1988
</settingDesc>
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes: declabl -->
DTD file: TEIcorp2
Data description: May contain a prose description organized as paragraphs, or a series of **<setting>** elements.
May occur within:
May contain: p setting
Element Declaration:

```
<!ELEMENT settingDesc   - o  (p+ | setting+)                    >
```

Attribute List Declaration:

```
<!ATTLIST settingDesc        %a.global;
                             %a.declarable;                    >
```

Discussed in 26.2.3, “The Setting Description,” on page 18.
This tag is optional.

<setting> (i.e. setting)

Description: describes one particular setting in which a language interaction takes place.

Attributes:

who: supplies the identifiers of the participants at this setting.
Data type: IDREFS
Value: must correspond with ID values of **<participant>** or **<participant.grp>** elements in the current document.
Default value: #IMPLIED

Example:

```
<setting who=P1>
<place>New York City, US
<time>1989
<locale>on a park bench
<activity>feeding birds
</setting>
```

Remarks: If the *who* attribute is not supplied, the setting is assumed to be that of all participants in the language interaction.
Part: auxiliary tag set for common core features.
Member of classes:
DTD file:
Data description:
May occur within: settingDesc
May contain: activity locale p place time
Element Declaration:

```
<!ELEMENT setting       - -  (p+ | (place?, time?, locale?,
                             activity?))                       >
```

Attribute List Declaration:

```
<!ATTLIST setting            %a.global;
          who                IDREFS             #IMPLIED       >
```

Discussed in 26.2.3, “The Setting Description,” on page 18.
This tag is optional.

<socecStatus> (i.e. socio-economic status)

Description: contains an informal description of a person's perceived social or economic status.

Attributes:

scheme: identifies the classification system or taxonomy in use.
Data type: IDREF
Value: Must identify a **<taxonomy>** element

Default value: #IMPLIED
This attribute is optional.

code: identifies a status code defined within the classification system or taxonomy defined by the *source* attribute.
Data type: IDREF
Value: Must identify a **<category>** element
Default value: #IMPLIED
This attribute is optional.

Example:

```
<socecStatus scheme=RG code=AB1>

<socecStatus>Status AB1 in the RG Classification scheme
```

Remarks: The content of this element may be used as an alternative to the more formal specification made possible by its attributes; it may also be used to supplement the formal specification with commentary or clarification.
Part: auxiliary tag set for corpora and collections.
Member of classes: demog -->
DTD file:
Data description: May contain character data and phrase-level elements.
May occur within:
May contain: %phrase.seq
Element Declaration:

```
<!ELEMENT socecStatus   - o  (%phrase.seq)                    >
```

Attribute List Declaration:

```
<!ATTLIST socecStatus        %a.global;
          scheme             IDREF              #IMPLIED
          code               IDREF              #IMPLIED       >
```

Discussed in 26.2.2, “The Participants Description,” on page 14.
This tag is optional.

<TEI.corpus.2> (i.e. TEI corpus)

Description: contains the whole of a TEI encoded corpus, comprising a single corpus header and one or more TEI.2 elements, each containing a single text header and a text.
Attributes: [None: global and inherited attributes only.]
Example:

```
<TEI.corpus.2>
    <TeiHeader> ...
    <!-- header for corpus-level information -->
    </TeiHeader>
    <TEI.2 id=T1>
         <TeiHeader> ...
         <!-- header for text-level information -->
         </TeiHeader>
         <text> ... </text>
    </TEI.2>
    <TEI.2 id=T2>
         <TeiHeader> ... </TeiHeader>
         <text> ... </text>
    </TEI.2>
    <!-- etc. -->
</TEI.corpus.2>
```

Remarks: This element is mandatory when applicable.
Part: auxiliary tag set for common core features.
Member of classes:
DTD file:
Data description: Must contain one TEI header for the corpus, and a series of **<TEI.2>** elements, one for each text.
May occur within:
May contain: tei.2 TeiHeader
Element Declaration:

```
<!ELEMENT TEI.corpus.2  - O  (TeiHeader, tei.2+)               >
```

Attribute List Declaration:

```
<!ATTLIST TEI.corpus.2      %a.global;                         >
```

Discussed in 3.2, “Base Tag Sets and Additional Tag Sets,” [in separate fascicle].
Discussed in 26.1, “Varieties of Composite Text,” on page 4.
This tag is mandatory when applicable.

<TEI.2> (i.e. TEI 2 document)

Description: contains a single TEI-conformant document, comprising a TEI header and a text, either in isolation or as part of a Teicorpus.
Attributes: [None: global and inherited attributes only.]
Example:

```
<tei.2>
  <teiHeader>
    <fileDesc>
       <titleStmt>
          <title>The shortest TEI Document Imaginable</title>
       <publicationStmt>
          <p>Published as part of TEI P2.
       </publicationStmt>
    <sourceDesc>
          <p>No source: this is an original work.
    </sourceDesc>
  </teiHeader>
  <text>
    <body>
     <p>This is about the shortest TEI document imaginable.
    </body>
  </text>
</tei.2>
```

Remarks: This element is required.
Part: auxiliary tag set for common core features.
Member of classes:
DTD file:
Data description: Must contain one TEI header and one text.
May occur within:
May contain: text TeiHeader
Element Declaration:

```
<!ELEMENT TEI.2         - O  (TeiHeader, text)                 >
```

Attribute List Declaration:

```
<!ATTLIST TEI.2              %a.global;                        >
```

Discussed in 3.2, “Base Tag Sets and Additional Tag Sets,” [in separate fascicle].
This tag is required.

<text>

Description: contains a single text of any kind, whether unitary or composite, for example a poem or drama, a collection of essays, a novel, or a corpus sample.
Attributes: [None: global and inherited attributes only.]
Example:

```
<text>
<front><docTitle>Autumn Haze</docTitle>
<body>
<l>Is it a dragonfly or a maple leaf
<l>That settles softly down upon the water?
</body>
</text>
<!-- Amy Lowell, in Imagist Poetry, ed Jones, (penguin 1972) -->

<text>
<front>
<group>
<text>
<text>
</group>
</text>
```

Remarks:
Part: base tag set for common core features.
Member of classes: decling chunk -->
DTD file:
May occur within: group TEI.2
May contain: back body front group
Element Declaration:

```
<!ELEMENT text          - -  (front?, (body | group), back?)    >
```

Attribute List Declaration:

```
<!ATTLIST text               %a.global;
                             %a.declaring;                    >
```

Discussed in 7, “Base Tag Set for Prose,” [in separate fascicle].
This tag is required.

<textClass> (i.e. text classification)

Description: groups information which describes the nature or topic of a text in terms of a standard classification scheme, thesaurus, etc.
Attributes: [None: global and inherited attributes only.]
Example:

```
<textClass>
  <catRef target=TC1>
  <classCode source=DD12>001.9
  <keywords>
    <list><item>End of the world
    <item>History - philosophy
    </list>
```

```
</textClass>
```

Remarks:
Part: auxiliary tag set for TEI headers.
Member of classes: declabl -->
DTD file:
May occur within:
May contain: catRef classCode keywords
Element Declaration:

```
<!ELEMENT textClass     - -  ((classCode | catRef | keywords)*
                             )                                  >
```

Attribute List Declaration:

```
<!ATTLIST textClass          %a.global;
                             %a.declarable;                    >
```

Discussed in 5.4.3, “The Text Classification,” [in separate fascicle].
This tag is optional.

<textDesc> (i.e. text description)

Description: provides a description of a text in terms of its *situational parameters*.
Attributes: [None: global and inherited attributes only.]
Example:

```
<textDesc id=t1 n='Informal domestic conversation'>
  <channel mode=s>
  <constitution type=single>
  <derivation type=original>
  <domain type=domestic>
  <factuality type=mixed>
  <interaction type=complete active=plural passive=many>
  <preparedness type='spontaneous'>
  <purpose type=entertain degree=high>
  <purpose type=inform degree=medium>
</textDesc>
```

Remarks:
Part: auxiliary tag set for corpora and collections.
Member of classes: declabl -->
DTD file:
Data description:
May occur within:
May contain: channel constitution derivation domain factuality interaction preparedness purpose
Element Declaration:

```
<!ELEMENT textDesc      - o  (channel, constitution, derivation,
                              domain, factuality, interaction,
                             preparedness, purpose+)           >
```

Attribute List Declaration:

```
<!ATTLIST textDesc           %a.global;                        >
```

Discussed in 26.2.1, “The Text Description,” on page 9.
This tag is optional.

<time>

Description: contains a phrase defining a time of day in any format.
Attributes:

value: gives the value of the time in a standard form.
Data type: CDATA
Value: Any string representing a time in standard format; recommended form is “hhmm”, using the 24 hour clock.
Default value: #IMPLIED
Example:

```
<time value=1600>four o' clock</time>
```

Remarks: If the normalized form of time is not given using the twenty-four hour clock, this fact should be documented in the **<stdVals>** element in the TEI Header.
This attribute is optional.

Example:

```
As he sat smiling, the quarter struck &mdash <time value=1145>
the quarter to twelve</time>.
```

```
Given on the <time value='1977-06-12'>Twelfth Day of June
in the Year of Our Lord One Thousand Nine Hundred and
Seventy-seven of the Republic the Two Hundredth and first
and of the University the Eighty-Sixth.</time>

<time value='1990-09'>September 1990</time>
```

Remarks:
Part: additional tag set for common core features.
Member of classes: data -->
DTD file:
Data description: May contain character data and phrase-level elements.

May occur within: setting
May contain: %phrase.seq;
Element Declaration:

```
<!ELEMENT time          - -  (%phrase.seq;)                    >
```

Attribute List Declaration:

```
<!ATTLIST time               %a.global;
          value              CDATA              #IMPLIED       >
```

Discussed in 6.4.3, “Dates,” [in separate fascicle].
This tag is optional.

Usability Report and Comment

If you have comments of any kind on the current draft of the TEI Guidelines, please complete this form and return it to one of the following addresses:

C. M. Sperberg-McQueen, Computer Center (M/C 135), University of Illinois at Chicago, 1940 W. Taylor Room 124, Chicago IL 60612-7352, USA. Bitnet: U35395@UICVM, Internet: U35395@uicvm.uic.edu

Lou Burnard, Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, UK. Fax: +44 (865) 273275, Janet: LOU@UK.AC.OXFORD.VAX Internet: LOU@vax.oxford.ac.uk

Further copies of this form may be downloaded from LISTSERV @ UICVM (or listserv @ uicvm.uic.edu). Send a mail message containing the line 'GET TEIURC MEMO TEI-L'

Your name:
Your postal address:

Your e-mail address (if any):
Your occupation:
Your academic background:
Your immediate reactions to the Guidelines (please tick or cross one only):

Relevance to your present concerns or interests: high medium low
Importance to your future research plans: high medium low
Comprehensibility/usability of current draft: high medium low

Detailed technical comment: [Please feel free to comment on any part of the Guidelines or in general. Comments relating to specific parts of the text are most helpful if they include a section or chapter number]

Thank you!

Temporary Index