23 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 up 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 23.5 *Recommendations for the Encoding of Large Corpora*). 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 chapter 7 *Default Text Structure*). 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 section 3.4 *Combining TEI Base Tag Sets*.

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 *Languages and Character Sets*); 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 *Elements Available in All TEI Documents*). 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 9 *Base Tag Set for Verse*), drama (chapter 10 *Base Tag Set for Drama*), transcriptions of spoken text (chapter 11 *Transcriptions of Speech*), etc. Chapter 3 *Structure of the TEI Document Type Definition* 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.

This chapter does however include 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 23.1 *Varieties of Composite Text*). Section 23.2 *Contextual Information* 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 23.3 *Associating Contextual Information with a Text* discusses a mechanism by which individual parts of the TEI Header may be associated with different parts of a TEI-conformant text. Section 23.4 *Linguistic Annotation of Corpora* reviews various methods of providing linguistic annotation in corpora, with some specific examples of relevance to current practice in corpus

linguistics. Finally, section 23.5 *Recommendations for the Encoding of Large Corpora* provides some general recommendations about the use of these Guidelines in the building of large corpora.

23.1 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.

<teiCorpus.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>` element.

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

type specifies the kind of document to which the header is attached.

Sample values include:

`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.

Legal values are:

`new` the header is a new header.

`update` the header is an update (has been revised).

creator identifies the creator of the TEI Header.

Values The name or initials of the person or institution responsible for creating this TEI header.

date.created indicates when the first version of the header was created.

Values A date in ISO 8601 format, generally yyyy-mm-dd.

date.updated indicates when the current version of the header was created.

Values A date in ISO 8601 format, generally yyyy-mm-dd.

<text> contains a single text of any kind, whether unitary or composite, for example a poem or drama, a collection of essays, a novel, a dictionary, 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 Definition* (for `<teiCorpus.2>` and `<TEI.2>`), chapter 5 *The TEI Header* (for `<teiHeader>`), and chapter 7 *Default Text Structure* (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 <teiCorpus.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 <teiCorpus.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:

```
<teiCorpus.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. -->
</teiCorpus.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 23.2 *Contextual Information*, as well as in chapter 5 *The TEI Header*. 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 23.3 *Associating Contextual Information with a Text*.

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 <teiCorpus.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 below and in section 7.3 *Groups of Texts*. 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 23.2.1 *The Text Description*, and those for associating declarations with corpus components described in section 23.3 *Associating Contextual Information with a Text*. 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 mis-classification 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: in both cases, the body of the text 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*.

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*.

All composite texts share the characteristic that their different component texts may be of structurally similar or dissimilar types. If all component texts may all be encoded using the same base tag set, then no problem arises. If however they require different base tag sets, then either the general or the mixed base tag set must be used, in addition to all relevant base tag sets. This process is described in more detail in section 3.4 *Combining TEI Base Tag Sets*.

23.2 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 data of a newspaper; 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 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 23.3 *Associating Contextual Information with a Text* below.

Chapter 5 *The TEI Header*, 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*); 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*); more detailed descriptive information about the creation and content of the corpus, such as the languages used within it and any descriptive classification system used (see section 5.4 *The Profile Description*); and version information documenting any changes made in the electronic text (see section 5.5 *The Revision Description*).

In addition to the elements defined by chapter 5 *The TEI Header*, 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* and are further described in the remainder of this section. They are formally defined as follows:

```
<!-- 23.2: Header extensions for Corpus Texts-->
<!ELEMENT textDesc %om.RO;  ( channel, constitution, derivation,
                     domain, factuality, interaction,
                     preparedness, purpose+ ) >
<!ATTLIST textDesc
      %a.global;
      %a.declarable;
      TEIform CDATA 'textDesc'  >
<!ELEMENT particDesc %om.RO; ( p+ |
                    ( (person | personGrp)+, particLinks? ) ) >
<!ATTLIST particDesc
      %a.global;
      %a.declarable;
      TEIform CDATA 'particDesc'  >
<!ELEMENT settingDesc %om.RO; (p+ | setting+)>
<!ATTLIST settingDesc
      %a.global;
      %a.declarable;
      TEIform CDATA 'settingDesc'  >
<!--continued in 23.2: Text description-->
<!--continued in 23.2: Participants description-->
<!--continued in 23.2: Setting description-->
<!-- end of 23.2-->
```

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 <teiCorpus.2> element), its document type declaration will resemble this:

```
<!DOCTYPE teiCorpus.2 PUBLIC
         "-//TEI P4//DTD Main Document Type//EN" "tei2.dtd" [
   <!ENTITY % TEI.XML    'INCLUDE' >
   <!ENTITY % TEI.corpus 'INCLUDE' >
]>
```

23.2.1 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*, and section 5.3.6 *The Classification Declaration*. The description is organized as a set of values and optional prose descriptions for the following 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 non-fictional 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.

Suggested values include:

`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

A TEI-conformant text description contains each of the above elements, supplied in the order specified. 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 an equivalent phrase.

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:[159]

- 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*, and elements for defining or declaring such classification schemes in section 5.3.6 *The Classification Declaration*. A second approach is to develop an application-specific set of *feature structures* and an associated *feature system declaration*, as described in chapters 16 *Feature Structures* and 26 *Feature System Declaration*.

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 5.3.6 *The Classification Declaration* 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*.

Using these situational 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>
   <derivation type="original">  </derivation>
   <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">  </interaction>
   <preparedness type="spontaneous">  </preparedness>
   <purpose type="entertain" degree="high">  </purpose>
   <purpose type="inform" degree="medium"> </purpose>
</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">     </constitution>
   <derivation type="original">     </derivation>
   <domain type="art">     </domain>
   <factuality type="fiction">     </factuality>
   <interaction type="none">     </interaction>
   <preparedness type="prepared">     </preparedness>
   <purpose type="entertain" degree="high">     </purpose>
   <purpose type="inform" degree="medium">  </purpose>
</textDesc>
```

[159] 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*, M. Kytö, O. Ihalainen, and M. Rissanen: eds. (Amsterdam: Rodopi, 1988).

The formal declarations for these elements are given below:

```
<!-- 23.2.1: Text description-->
<!ELEMENT channel %om.RO; %phrase.seq;>
<!ATTLIST channel
      %a.global;
      mode (s | w | ws | sw | m | x) "x"
      TEIform CDATA 'channel'  >
<!ELEMENT constitution %om.RO; %phrase.seq;>
<!ATTLIST constitution
      %a.global;
      type (single | composite | frags | unknown) "single"
      TEIform CDATA 'constitution'  >
<!ELEMENT derivation %om.RO; %phrase.seq;>
<!ATTLIST derivation
      %a.global;
      type CDATA #IMPLIED
      TEIform CDATA 'derivation'  >
<!ELEMENT domain %om.RO; %phrase.seq;>
<!ATTLIST domain
      %a.global;
      type CDATA #IMPLIED
      TEIform CDATA 'domain'  >
<!ELEMENT factuality %om.RO; %phrase.seq;>
<!ATTLIST factuality
      %a.global;
      type (fiction|fact|mixed|inapplicable) #IMPLIED
      TEIform CDATA 'factuality'  >
<!ELEMENT interaction %om.RO; %phrase.seq;>
<!ATTLIST interaction
      %a.global;
      type (none|partial|complete|inapplicable) #IMPLIED
      active CDATA #IMPLIED
      passive CDATA #IMPLIED
      TEIform CDATA 'interaction'  >
<!ELEMENT preparedness %om.RO; %phrase.seq;>
<!ATTLIST preparedness
      %a.global;
      type CDATA #IMPLIED
      TEIform CDATA 'preparedness'  >
<!ELEMENT purpose %om.RO; %phrase.seq;>
<!ATTLIST purpose
      %a.global;
      type CDATA #IMPLIED
      degree (high | medium | low | unknown) #IMPLIED
      TEIform CDATA 'purpose'  >
<!-- end of 23.2.1-->
```

23.2.2 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 personæ 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 <person> or <personGrp> elements, followed by an optional <particLinks> element, as described below:

<person> describes a single participant in a language interaction. Attributes include:

role specifies the role of this participant in the group.
Values a set of keywords to be defined
sex specifies the sex of the participant.
Legal values are:
`m` male
`f` female
`u` unknown or inapplicable
age specifies the age group to which the participant belongs.
Values suggested values are to be supplied

<personGrp> describes a group of individuals treated as a single person for analytic purposes. Attributes include:
role specifies the role of this group of participants in the interaction.
Values a set of keywords to be defined
sex specifies the sex of the participant group.
Legal values are:
`m` male
`f` female
`u` unknown
`x` mixed
age specifies the age group of the participants.
Values suggested values are to be supplied
size specifies the size or approximate size of the group.
Values may contain a number and an indication of accuracy, e.g. 'approx 200'

<particLinks> describes the relationships or social links existing between participants in a linguistic interaction.

Both <person> and <personGrp> elements 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 29 *Modifying and Customizing the TEI DTD*. 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 5.3.6 *The Classification Declaration*)

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

<birth> contains information about a person's birth, such as its date and place. Attributes include:
date specifies the date of birth in an ISO standard form (yyyy-mm-dd).
Values a date in ISO standard form, generally ISO 8601:2000 5.2.1.1 Complete representation, extended format (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.
Values must identify a <taxonomy> element

code identifies an occupation code defined within the classification system or taxonomy defined by the scheme attribute.
Values Must identify a <category> element

<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.
Values Must identify a <taxonomy> element
code identifies a status code defined within the classification system or taxonomy defined by the source attribute.
Values Must identify a <category> element

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

```
<person 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.</p>
</person>
```

Provided that the "PEP classification scheme" has been defined elsewhere in the heading (as a <taxonomy> element within the <textClass> element; see 5.3.6 *The Classification Declaration*), the same individual might more formally be described as follows:

```
<person id="p1" sex="f" age="mid">
   <birth date="1950-01-12">
      <date>12 Jan 1950</date>
      <name type="place">Shropshire, UK</name>
   </birth>
   <firstLang>English</firstLang>
   <langKnown>French</langKnown>
   <residence>Long term resident of Hull</residence>
   <education>University postgraduate</education>
   <occupation>Unknown</occupation>
   <socecStatus scheme="pep" code="b2"/>
</person>
```

Dates and names of persons or places, if included in the prose description, may be encoded using either the general purpose <date>, <name> or <rs> elements discussed in section 6.4 *Names, Numbers, Dates, Abbreviations, and Addresses*, or the more specialised and detailed elements provided by chapter 20 *Names and Dates*. In the latter case, the additional tag set for names and dates must be enabled together with that for language corpora.

An identified character in a drama or a novel might be defined using a subset of the same tags as follows:[160]

```
<person id="em01" sex="f" age="young">
   <p><name>Emma Woodhouse</name></p>
</person>
```

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, as described below:

<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

[160] 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 definitions for their speakers; see further section 10.2.2 *Speeches and Speakers*.

desc briefly describes the relationship.
Values an open list of application-dependent keywords

active identifies the "active"participants in a non-mutual relationship, or all the participants in a mutual one.
Values a list of identifier values for participant or participant groups

passive identifies the "passive" participants in a non-mutual relationship.
Values a list of identifier values for participant or participant groups

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 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). For non-mutual relationships, only two kinds of role are currently supported; they are named *active* and *passive*. These names are chosen to reflect the fact that 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 and passive roles of "employs".

For example:

```
<particLinks>
   <relation desc="parent" active="p1 p2" passive="p3 p4" mutual="N"/>
   <relation desc="spouse" active="p1 p2" mutual="Y"/>
   <relation type="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:

```
<!-- 23.2.2: Participants description-->
<!ELEMENT person %om.RO; (p+ | (%m.demographic;)* )>
<!ATTLIST person
      %a.global;
      role CDATA #IMPLIED
      sex ( m | f | u ) #IMPLIED
      age CDATA #IMPLIED
      TEIform CDATA 'person'  >
<!ELEMENT personGrp %om.RO; (p+ | (%m.demographic;)* )>
<!ATTLIST personGrp
      %a.global;
      role CDATA #IMPLIED
      sex ( m | f | u | x ) #IMPLIED
      age CDATA #IMPLIED
      size CDATA #IMPLIED
      TEIform CDATA 'personGrp'  >
<!ELEMENT birth %om.RR; %phrase.seq;>
```

```
<!ATTLIST birth
      %a.global;
      date %ISO-date; #IMPLIED
      TEIform CDATA 'birth'  >
<!ELEMENT firstLang %om.RO; %phrase.seq;>
<!ATTLIST firstLang
      %a.global;
      TEIform CDATA 'firstLang'  >
<!ELEMENT langKnown %om.RR; %phrase.seq;>
<!ATTLIST langKnown
      %a.global;
      TEIform CDATA 'langKnown'  >
<!ELEMENT residence %om.RO; %phrase.seq;>
<!ATTLIST residence
      %a.global;
      TEIform CDATA 'residence'  >
<!ELEMENT education %om.RO; %phrase.seq;>
<!ATTLIST education
      %a.global;
      TEIform CDATA 'education'  >
<!ELEMENT affiliation %om.RR; %phrase.seq;>
<!ATTLIST affiliation
      %a.global;
      TEIform CDATA 'affiliation'  >
<!ELEMENT occupation %om.RR; %phrase.seq;>
<!ATTLIST occupation
      %a.global;
      scheme IDREF #IMPLIED
      code IDREF #IMPLIED
      TEIform CDATA 'occupation'  >
<!ELEMENT socecStatus %om.RO; %phrase.seq;>
<!ATTLIST socecStatus
      %a.global;
      scheme IDREF #IMPLIED
      code IDREF #IMPLIED
      TEIform CDATA 'socecStatus'  >
<!ELEMENT particLinks %om.RO; (p+ |  relation+) >
<!ATTLIST particLinks
      %a.global;
      TEIform CDATA 'particLinks'  >
<!ELEMENT relation %om.RO; EMPTY>
<!ATTLIST relation
      %a.global;
      type CDATA "personal"
      desc CDATA #IMPLIED
      active IDREFS #IMPLIED
      passive IDREFS #IMPLIED
      mutual (Y | N) "Y"
      TEIform CDATA 'relation'  >
<!-- end of 23.2.2-->
```

23.2.3 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 23.2.2 *The Participants Description*. If this attribute is not specified, the setting details provided are assumed to apply to all participants represented in the language interaction.[161]

[161] 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.

Values must correspond with ID values of <person> or <personGrp> elements in the current document.

<name> contains a proper noun or noun phrase. Attributes include:

type indicates the type of the object which is being named by the phrase.

Values Values such as person, place, institution, product, acronym.

<date> contains a date in any format. Attributes include:

value gives the value of the date in some standard form, usually yyyy-mm-dd.

Values Any string representing a date in standard format; recommended form is ISO 8601:2000 5.2.1.1 Complete representation, extended format (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 some standard form, usually hh:mm.

Values Any string representing a time in standard format; recommended forms are the extended formats from ISO 8601:2000 (hh:mm, hh:mmZ, hh:mm±hh)

<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.</p>
</settingDesc>
```

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

```
<settingDesc>
   <setting who="p1 p2">
      <name type="city">Bedford</name>
      <name type="region">UK: South East</name>
      <date value="1989">early spring, 1989</date>
      <locale>rug of a suburban home</locale>
      <activity>playing</activity>
   </setting>
   <setting who="p3">
      <name type="city">Bedford</name>
      <name type="region">UK: South East</name>
      <date value="1989">early spring, 1989</date>
      <locale>at the sink</locale>
      <activity>washing-up</activity>
   </setting>
   <setting who="p4">
      <name type="place">London, UK</name>
      <time>unknown</time>
      <locale>broadcasting studio</locale>
      <activity>radio performance</activity>
   </setting>
</settingDesc>
```

For more detailed encoding of names of persons and places, the additional tag set described in chapter 20 *Names and Dates* may additionally be used; if used, however, these elements may appear only within a <p> element. The above examples assume that only the general purpose <name> element supplied in the core tag set is available. The elements discussed in this section have the following formal definitions:

```
<!-- 23.2.3: Setting description-->
<!ELEMENT setting %om.RR; (p+ | (name | time | date | locale | activity)* )>
<!ATTLIST setting
      %a.global;
      who IDREFS #IMPLIED
      TEIform CDATA 'setting'  >
<!ELEMENT locale %om.RO; %phrase.seq;>
<!ATTLIST locale
      %a.global;
      TEIform CDATA 'locale'  >
<!ELEMENT activity %om.RO; %phrase.seq;>
<!ATTLIST activity
      %a.global;
      TEIform CDATA 'activity'  >
<!-- end of 23.2.3-->
```

23.3 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*. 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 be combined as desired.

23.3.1 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 `<fileDesc>` 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

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

```
  </teiHeader>

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

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

  <TEI.2>
    <teiHeader>
      <fileDesc>
        <!-- details peculiar to this text --> ... </fileDesc>
    </teiHeader>
    <text> <!-- ... --> </text>
  </TEI.2>
</teiCorpus.2>
```

23.3.2 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:

<bibl> contains a loosely-structured bibliographic citation of which the sub-components may or may not be explicitly tagged.

<biblFull> contains a fully-structured bibliographic citation, in which all components of the TEI file description are present.

<biblStruct> 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. Attributes include:

status indicates the degree of correction applied to the text.

Legal values are:

`high` the text has been thoroughly checked and proofread.

`medium` the text has been checked at least once.

`low` the text has not been checked.

`unknown` the correction status of the text is unknown.

method indicates the method adopted to indicate corrections within the text.

Legal values are:

`silent` corrections have been made silently

`tags` corrections have been represented using editorial tags

<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. Attributes include:

eol indicates whether or not end-of-line hyphenation has been retained in a text.

Legal values are:

`all` all end-of-line hyphenation has been retained, even though the lineation of the original may not have been.

`some` end-of-line hyphenation has been retained in some cases.

`hard` all soft end-of-line hyphenation has been removed: any remaining end-od-line hyphenation should be retained.

`none` all end-of-line hyphenation has been removed: any remaining hyphenation occurred within the line.

<interpretation> describes the scope of any analytic or interpretive information added to the text in addition to the transcription.

<langUsage> describes the languages, sublanguages, registers, dialects etc. represented within a text.

<listBibl> 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. Attributes include:

source indicates the authority for any normalization carried out.

Values Should really be a bibliographic reference of some kind

method indicates the method adopted to indicate normalizations within the text.

Legal values are:

`silent` normalization made silently

`tags` normalization represented using editorial tags

<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. Attributes include:

marks indicates whether or not quotation marks have been retained as content within the text.

Legal values are:

`none` no quotation marks have been retained

`some` some quotation marks have been retained

`all` all quotation marks have been retained

form specifies how quotation marks are indicated within the text.

Legal values are:

`data` quotation marks are retained as data.

`rend` the rendition attribute is consistently used to indicate the form of quotation marks.

`std` use of quotation marks has been standardized.

`nonstd` quotation marks are represented inconsistently.

`unknown` use of quotation marks is unknown.

<recording> details of an audio or video recording event used as the source of a spoken text, either directly or from a public broadcast. Attributes include:

type the kind of recording.

Legal values are:

`audio` audio recording

`video` audio and video recording

dur the original duration of the recording.
Values Include the units, e.g. 30 min.

<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="YES">
      <p> ... </p>
   </correction>
   <correction id="c2">
      <p> ... </p>
   </correction>
   <normalization id="n1">
      <p> ... </p>
      <p> ... </p>
   </normalization>
</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>
  <body>
    <!--  ... -->
    <div1 n='d1'> <!-- ... --> </div1>
    <div1 n='d2' decls='C2'> <!-- ... --> </div1>
    <div1 n='d3'> <!-- ... --> </div1>
    <!--  ... -->
  </body>
</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, and 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='YES'>
    <correction id='C1A' default='YES'>   <p> ... </p></correction>
    <correction id='C1B'>                 <p> ... </p></correction>
    <normalization id="N1">
      <p> ... </p>
      <p> ... </p>
    </normalization>
    <p> ... </p>
    <p> ... </p>
  </editorialDecl>
  <editorialDecl id='ED2'>
    <correction id='C2A' default='YES'>   <p> ... </p></correction>
    <correction id='C2B'>                 <p> ... </p></correction>
    <normalization id='N2A'>              <p> ... </p></normalization>
    <normalization id='N2B' default='YES'><p> ... </p></normalization>
    <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, and normalization N2b will apply.

A finer grained approach is also possible. A text might specify `<text decls='C2b N2a'>`, 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 " C2a N2b", resulting in a list that identifies two correction elements (C1a and C2a).

23.3.3 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.

23.4 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.[162] The present section presents informally some particular applications of these general mechanisms to the specific practice of corpus linguistics.

23.4.1 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 *Elements Available in All TEI Documents*, 7 *Default Text Structure*, 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*, and in section 23.2 *Contextual Information* of the present chapter.

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 `<interpretation>` element within the encoding description of the TEI Header, as described in section 5.3.3 *The Editorial Practices Declaration*. 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 23.3 *Associating Contextual Information with a Text*.

An extended example of one form of linguistic analysis commonly practised in corpus linguistics is given in section 15.4 *Linguistic Annotation*.

23.5 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*. 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 dtd to implement such a subset are described in chapter 3 *Structure of the TEI Document Type Definition* and chapter 29 *Modifying and Customizing the TEI DTD*; 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 28 *Conformance*.

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:

[162] See in particular chapters 14 *Linking, Segmentation, and Alignment*, 15 *Simple Analytic Mechanisms*, and 16 *Feature Structures*.

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.

V: Auxiliary Document Types