15 Simple Analytic Mechanisms

This chapter describes a tag set for associating simple analyses and interpretations with text elements. We use the term *analysis* here to refer to any kind of semantic or syntactic interpretation which an encoder wishes to attach to all or part of a text. Examples discussed in this chapter include familiar linguistic categorizations (such as "clause", "morpheme", "part-of-speech" etc.) and characterizations of narrative structure (such as "theme", "reconciliation" etc.). The mechanisms presented in this chapter offer simpler but less powerful than those described in chapter 16 *Feature Structures*.

Section 15.1 *Linguistic Segment Categories* introduces a tag set for characterizing text segments according to the familiar linguistic categories of *sentence* or *s-unit*, *clause*, *phrase*, *word*, *morpheme*, and *character*. These elements represent special cases of the generic <seg> element described in section 14.3 *Blocks, Segments and Anchors*.

Section 15.2 *Global Attributes for Simple Analyses* introduces an additional global attribute which allows passages of text to be associated with specialised elements representing their interpretation. These 'interpretative' elements (and <interp>) are described in detail in section 15.3 *Spans and Interpretations*. They allow the encoder to specify an analysis as a series of names and associated values,[122] each such pair being linked to one or more stretches of text, either directly, in the case of spans, or indirectly, in the case of interpretations.

Finally section 15.4 *Linguistic Annotation* revisits the topic of linguistic analysis, and illustrates how these interpretative mechanisms may be used to associate simple linguistic analysis with text segments.

The following DTD fragments show the overall organization of the class of analytic elements discussed in the remainder of this chapter. File *teiana2.ent* defines the additional global attribute made available by this tag set.

```
<!-- 15.: Modifications to TEI class system for analysis-->
<!--declarations from 15.2: Global attribute for analysis inserted here -->
<!-- end of 15.-->
```

File *teiana2.dtd* contains declarations for elements used to represent simple analyses or interpretations of portions of a text.

```
<!-- 15.: Simple analytic mechanisms-->
<!--Text Encoding Initiative Consortium:
Guidelines for Electronic Text Encoding and Interchange.
Document TEI P4, 2002.
Copyright (c) 2002 TEI Consortium. Permission to copy in any form
is granted, provided this notice is included in all copies.
These materials may not be altered; modifications to these DTDs should
be performed only as specified by the Guidelines, for example in the
chapter entitled 'Modifying the TEI DTD'
These materials are subject to revision by the TEI Consortium. Current versions
are available from the Consortium website at http://www.tei-c.org-->
<!--We declare the various elements, group by group.-->
<!--declarations from 15.3: Spans inserted here -->
<!--declarations from 15.1: Linguistic Segment Categories inserted here -->
<!-- end of 15.-->
```

This tag set is selected as described in 3.3 *Invocation of the TEI DTD*; in a document which uses the markup described in this chapter, the document type declaration should contain the following declaration of the entity *TEI.analysis*, or an equivalent one:

```
<!ENTITY % TEI.analysis 'INCLUDE'>
```

The entire document type declaration for a document using this additional tag set together with that for linking and alignment and the base tag set for prose might look like this:

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

[122] Or, as they are widely known, *attribute-value pairs*; this term should not be confused, however, with SGML or XML attributes and their values, which are similar in concept but distinct in their formal definitions.

```
    <!ENTITY % TEI.analysis 'INCLUDE' >
]>
```

15.1 Linguistic Segment Categories

In this section we introduce specialized *linguistic segment category* elements which may be used to represent the segmentation of a text into the traditional linguistic categories of *sentence*, *clause*, *phrase*, *word*, *morpheme*, and *characters*.

<s> contains a sentence-like division of a text.

<cl> represents a grammatical clause.

<phr> represents a grammatical phrase.

<w> represents a grammatical (not necessarily orthographic) word. Attributes include:

lemma identifies the word's lemma (dictionary entry form).

Values a string of characters representing the spelling of the word's dictionary entry form.

<m> represents a grammatical morpheme. Attributes include:

baseform identifies the morpheme's base form.

Values a string of characters representing the spelling of the morpheme's base form.

<c> represents a character.

As members of the seg class, these elements share the following attributes:

type characterizes the type of segment.

function characterizes the function of the segment.

The <s> element may be used simply to segment a text end-to-end into a series of non-overlapping segments, referred to here and elsewhere as *s-units*, or *sentences*.

```
<p>
  <s>Nineteen fifty-four, when I was eighteen years old,
    is held to be a crucial turning point in the history of
    the Afro-American &mdash; for the U.S.A. as a whole &mdash; the
    year segregation was outlawed by the U.S. Supreme Court.</s>
  <s>It was also a crucial year for me because on June 18,
    1954, I began serving a sentence in state prison for
    possession of marijuana.</s>
</p>
```

The <s> may be thought of as providing an abbreviated version of the tag <seg type='s-unit'>, with the important additional proviso that (unlike <seg> elements) <s> elements may not be nested within each other. The type attribute of the <s> element corresponds to the subtype attribute on the <seg> element, that is, a tag <s type="xxx"> should be thought of as synonymous with a tag <seg type="s-unit" subtype="xxx">. Similar considerations apply to the <cl> and <phr> elements , which can be thought of as short for <seg type="clause"> and <seg type="phrase">, respectively.

The <s> element may be further subdivided into *clauses*, marked with the <cl> element, as in the following example:

```
<p>
  <s>
    <cl>It was about the beginning of September, 1664,
      <cl>that I, among the rest of my neighbours,
        heard in ordinary discourse
        <cl>that the plague was returned again to Holland; </cl> </cl> </cl>
    <cl>for it had been very violent there, and particularly at
      Amsterdam and Rotterdam, in the year 1663, </cl>
    <cl>whither, <cl>they say,</cl> it was brought,
      <cl>some said</cl> from Italy, others from the Levant, among some goods
      <cl>which were brought home by their Turkey fleet;</cl> </cl>
    <cl>others said it was brought from Candia;
      others from Cyprus. </cl>
  </s>
  <s>
    <cl>It mattered not <cl>from whence it came;</cl> </cl>
```

```
    <cl>but all agreed <cl>it was come into Holland again.</cl> </cl>
  </s>
</p>
```

Clauses may be further divided into <phr> elements in the same way. A text may be segmented directly into clauses, or into phrases, with no need to include segmentation at a higher level as well.

For verse texts, the overlapping of metrical and syntactic structure requires that special care be given to representing both using an element hierarchy. One simple approach is to split the syntactic phrases into fragments when they cross verse boundaries, reuniting them with the part attribute:

```
<div type="stanza">
  <l><cl part="I">Tweedledum and Tweedledee</cl></l>
  <l><cl part="F">Agreed to have a battle;</cl></l>
  <l><cl part="I">For Tweedledum said <cl part="I">Tweedledee</cl></cl></l>
  <l><cl part="F"><cl part="F">Had spoiled his nice new rattle.</cl></cl></l></div>
<div type="stanza">
  <l><cl part="I">Just then flew down a monstrous crow,</cl></l>
  <l><cl part="F">As black as a tar barrel;</cl></l>
  <l><cl part="I">Which frightened both the heroes so,</cl></l>
  <l><cl part="F"><cl>They quite forgot their quarrel.</cl></cl></l></div>
```

Another approach is to use the next and prev attributes defined in the additional tag set for linking (chapter 14 *Linking, Segmentation, and Alignment*):

```
<l><cl next="c5" id="c3" part="I">For Tweedledum said
   <cl next="c6" id="c4" part="I">Tweedledee</cl></cl></l>
<l><cl prev="c3" id="c5" part="F">
   <cl prev="c4" id="c6" part="F">Had spoiled his nice new rattle.</cl></cl></l>
```

Other methods are also possible; for discussion, see chapter 31 *Multiple Hierarchies*.

The type attribute on linguistic segment categories can be used to provide additional interpretative information about the category. The function attribute on the <cl> and <phr> elements can be used to provide additional information about the function of the category. Legal values for these two attributes are not defined by these Guidelines, but should be documented in the <segmentation> element of the <encodingDesc> element within the document's header. A general approach to the encoding of linguistic categories assigned to parts of a text is discussed in section 15.4 *Linguistic Annotation* below.

Using traditional terminology, these attributes provide a convenient way of specifying, for example, that the clause 'from whence it came' is a relative clause modifying another, or that the phrase 'by the U.S. Supreme Court' is a prepositional post-modifier:

```
<cl>It mattered not
  <cl type="relative" function="clause modifier">from whence it came;</cl>
</cl>
<phr type="NP">the year segregation</phr>
<phr>was outlawed</phr>
<phr type="PP" function="postmodifier (agent)">by the U.S. Supreme Court.</phr>
```

Segmentation into clauses and phrases can, of course, be combined. Such detailed encodings as the following may require careful formatting if they are to be easily readable however.

```
<p>
  <s>
 <cl type="finite declarative" function="independent">
      <phr type="NP" function="subject">Nineteen fifty-four,
  <cl type="finite relative declarative" function="appositive">
    when <phr type="NP" function="subject">I</phr>
    <phr type="VP" function="predicate">was eighteen years old</phr>
  </cl></phr>,
      <phr type="VP" function="predicate">
  <phr type="V" function="main verb">is held</phr>
  <phr type="NP" function="complement">
    <cl type="nonfinite" function="predicate nom.">
   <phr type="V" function="copula">to be</phr>
   <phr type="NP" function="predicate nom.">a crucial turning point
```

```
    <phr type="PP" function="postmodifier">in
   <phr type="NP" function="prep.obj.">the history
     <phr type="PP" function="postmodifier">of the Afro-American</phr>
   </phr>
    </phr>
    &mdash;
    <phr type="PP" function="appositive postmodifier">for
   <phr type="NP" function="prep.obj.">the U.S.A.
     <phr type="PP" function="postmodifier">as a whole</phr>
   </phr>
    </phr>
  </phr>
  &mdash;
  <phr type="NP" function="appositive pred.nom.">the year
    <cl type="finite relative" function="adjectival">
   <phr type="NP" function="subject">segregation</phr>
   <phr type="VP" function="predicate">
     <phr type="V" function="main verb">was outlawed</phr>
     <phr type="PP" function="postmodifier">by the U.S. Supreme Court</phr>
   </phr></cl></phr></cl></phr></phr>.</cl></s>
 <s>
<cl type="finite declarative" function="independent">
     <phr type="NP" function="subject">It</phr>
     <phr type="VP" function="predicate">
 <phr type="V" function="main verb">was</phr>
 also
 <phr type="NP" function="predicate nom.">a crucial year for me</phr>
     </phr>
     <cl type="finite declarative" function="dependent causative">because
 <phr type="PP" function="sentence adverb">on June 18, 1954</phr>,
 <phr type="NP" function="subject">I</phr>
 <phr type="VP" function="predicate">
   <phr type="V" function="main verb">began serving</phr>
   <phr type="NP" function="complement">a sentence in state prison
  <phr type="PP" function="complement">for possession of marijuana</phr>
   </phr></phr></cl></cl></s>.</p>
```

This style of markup, however, introduces spurious new lines and blanks into the text, which could make restoring the text to its original layout problematic. If the original layout is important, the original line breaks and font shifts should be recorded using <lb> elements, the global rend attribute, etc.

The <w>, <m> and <c> elements are also identical in meaning to the <seg> element with a type attribute of “w”, “m”, or “c”, and may occur wherever <seg> is permitted to occur. However, they have more restricted content models than does <seg>: for example, the <w> element can only contain <w>, <m> and <c> elements, and parsed character data; the <m> element can only contain <c> elements and parsed character data; the <c> element can only contain parsed character data, and will often contain only a single character. Consequently, while <m> et al. can be translated directly into typed <seg> elements, the reverse is not necessarily the case.

The restriction on the content of the <w> attribute in particular requires that a certain care must be exercised when using it, especially in relation to the use of other tags that one may think of as *word level*, but which are in fact defined as *phrase level*. Consider the problem of segmenting an occurrence of the <mentioned> element as a word.

```
<mentioned>grandiloquent</mentioned>
```

The first of the following two encodings is legitimate; the second is not, since the <mentioned> element is not part of the content model of the <w> element:

```
<!-- This is all right. -->
<mentioned><w>grandiloquent</w></mentioned>

<!-- This is NOT all right! -->
<w><mentioned>grandiloquent</mentioned></w>
```

On the other hand, both of the following encodings *are* legitimate:

```
<mentioned>
   <phr>grandiloquent speech</phr>
</mentioned>

<phr>
   <mentioned>grandiloquent speech</mentioned>
</phr>
```

The first encoding describes the citing of a phrase. The second describes a phrase which consists of something mentioned.

The <w> and <m> elements carry additional attributes which may be of use in many indexing or analytic applications. The lemma attribute may be used to specify the *lemma*, that is the head- or base- form of an inflected verb or noun, for example:

```
<s lang="la">
   <w lemma="timeo" >timeo</w>
   <w lemma="danaii">Danaos</w>
   <w lemma="et"    >et</w>
   <w lemma="donum" >dona</w>
   <w lemma="fero"  >ferentes</w>
</s>
```

Similarly, the baseform attribute may be specified for the <m> element, to indicate the 'base form' of a transformed morpheme:

```
<w type="adjective">
   <m type="prefix" baseform="con">com</m>
   <m type="root">fort</m>
   <m type="suffix">able</m>
</w>
```

The <w>, <m>, and <c> elements can be used together to give a fairly detailed low-level grammatical analysis of text. For example, consider the following segmentation of the English S-unit 'I didn't do it'.

```
<w>I</w>
<w>
   <w>did</w>
   <m>n't</m>
</w>
<w>do</w>
<w>it</w>
<c>.</c>
```

This segmentation, crude as it is, succeeds in representing the idea that 'did' occurs as a word inside the word 'didn't'. A further advantage of segmenting the text down to this level is that it becomes relatively simple to associate each such segment with a more detailed formal analysis. This matter is taken up in detail in section 15.4 *Linguistic Annotation*.

The <s>, <cl>, <phr>, <w>, <m>, and <c> elements are formally declared as follows:

```
<!-- 15.1: Linguistic Segment Categories-->
<!ELEMENT s %om.RR;  %phrase.seq;>
<!ATTLIST s
      %a.global;
      %a.seg;
      TEIform CDATA 's'  >
<!ELEMENT cl %om.RR;  %phrase.seq;>
<!ATTLIST cl
      %a.global;
      %a.seg;
      TEIform CDATA 'cl'  >
<!ELEMENT phr %om.RR;  %phrase.seq;>
<!ATTLIST phr
      %a.global;
      %a.seg;
      TEIform CDATA 'phr'  >
<!ELEMENT w %om.RR; (#PCDATA | seg | w | m | c | %m.Incl;)*>
<!ATTLIST w
```

```
      %a.global;
      %a.seg;
      lemma CDATA #IMPLIED
      TEIform CDATA 'w'  >
<!ELEMENT m %om.RR; (#PCDATA | seg | c | %m.Incl;)*>
<!ATTLIST m
      %a.global;
      %a.seg;
      baseform CDATA #IMPLIED
      TEIform CDATA 'm'  >
<!ELEMENT c %om.RR;  (#PCDATA)>
<!ATTLIST c
      %a.global;
      %a.seg;
      TEIform CDATA 'c'  >
<!-- end of 15.1-->
```

15.2 Global Attributes for Simple Analyses

When the tag set described by this chapter is selected, an additional attribute is defined for all elements:

ana indicates one or more elements containing interpretations of the element on which the ana attribute appears.

The ana attribute may be specified for any element. Its effect is to associate the element with one or more others representing an analysis or interpretation of it. Its target should be one of the elements described in the section 15.3 *Spans and Interpretations* below, or some other interpretative element such as <note>, on which see section 6.8 *Notes, Annotation, and Indexing* or <fs>, on which see chapter 16 *Feature Structures*.

The ana attribute is formally declared as follows:

```
<!-- 15.2: Global attribute for analysis-->
<!ENTITY % a.analysis '
      ana IDREFS #IMPLIED'>
<!-- end of 15.2-->
```

15.3 Spans and Interpretations

The simplest mechanisms for attaching analytic notes in some structured vocabulary to particular passages of text are provided by the empty and <interp> elements, and their associated grouping elements <spanGrp> and <interpGrp>.

**** associates an interpretative annotation directly with a span of text. Attributes include:

- **value** identifies the specific phenomenon being annotated.
  - *Values* Any string of characters.
- **from** specifies the beginning of the passage being annotated; if not accompanied by a to attribute, then specifies the entire passage.
  - *Values* The identifier of the element which occurs at the beginning of the passage.
- **to** specifies the end of the passage being annotated.
  - *Values* The identifier of the element which occurs at the end of the passage.

<spanGrp> collects together tags.

<interp> provides for an interpretative annotation which can be linked to a span of text. Attributes include:

- **value** identifies the specific phenomenon being annotated.
  - *Values* Any string of characters.

<interpGrp> collects together <interp> tags.

These elements are all members of the class interpret, and thus share the following attributes:

resp indicates who is responsible for the interpretation.

type indicates what kind of phenomenon is being noted in the passage. Possible values are:

- image identifies an image in the passage.
- character identifies a character associated with the passage.

`theme` identifies a theme in the passage.
`allusion` identifies an allusion to another text.
`(discourse type)` specifies that the passage is of a particular discourse type.
inst points to instances of the analysis or interpretation represented by the current element.

The type and value attributes of the and <interp> elements may be used to associate an interpretive name, type, and value with a specific stretch (or *span*) of text. In the case of the element, the span of text being annotated is indicated by values of the from and to attributes, the value of each being a pointer. If the optional to attribute is omitted, the span consists just of the element pointed at by the obligatory from attribute. In the case of <interp> (see below), the span is indicated by a pointer from a <link> element or some similar mechanism. Here is an example of the element.

```
<p id="MQp1s2p114">
<s id="MQp1s2p114s1">There was certainly a definite point at which the
  thing began.</s>
<sa id="MQp1s2p114s2">It was not; then it was suddenly inescapable,
  and nothing could have frightened it away.</s>
<s id="MQp1s2p114s3">There was a slow integration, during which she,
  and the little animals, and the moving grasses, and the sun-warmed
  trees, and the slopes of shivering silvery mealies, and the great
  dome of blue light overhead, and the stones of earth under her feet,
  became one, shuddering together in a dissolution of dancing
  atoms.</s>
<s id="MQp1s2p114s4">She felt the rivers under the ground forcing
  themselves pain&shy;fully along her veins, swelling them out in an
  unbearable pressure; her flesh was the earth, and suffered growth
  like a ferment; and her eyes stared, fixed like the eye of the
  sun.</s>
<s id="MQp1s2p114s5">Not for one second longer (if the terms for time
  apply) could she have borne it; but then, with a sudden movement
  forwards and out, the whole process stopped; and <emph
  rend="italic">that</emph> was <soCalled rend='dquo'>the
  moment</soCalled> which it was impossible to remember
  afterwards.</s>
<span resp="DTL"
      value='the moment'
      from="MQp1s2p114s3"
      to=  "MQp1s2p114s5" />
<s id="MQp1s2p114s6">For during that space of time (which was
  timeless) she understood quite finally her smallness, the
  unimportance of humanity.</s>
<!-- ... -->
</p>
```

The element may, as in this example, be placed in the text near the textual span it is associated with, or it may be placed outside the text enclosed within a <spanGrp> element as follows.

```
<spanGrp resp="DTL">
  <span value='the moment' from="MQp1s2p114s3" to="MQp1s2p114s5"/>
</spanGrp>
```

As may be seen, the type attribute may be omitted in order to associate a span of text simply with a descriptive name.

Spans may also be used to represent the structural divisions assigned to the narrative by an interpreter. Consider the following narrative:

> Sigmund, the son of Volsung, was a king in Frankish country. Sinfiotli was the eldest of his sons, the second was Helgi, the third Hamund. Borghild, Sigmund's wife, had a brother named — But Sinfiotli, her stepson, and — both wooed the same woman and Sinfiotli killed him over it.[123] And when he came home, Borghild asked him to go away, but Sigmund offered her weregild, and she was obliged to accept it. At the funeral feast Borghild was serving beer. She took poison, a big

[123] The rule marks spaces left for the missing name in the manuscript.

drinking horn full, and brought it to Sinfiotli. When Sinfiotli looked into the horn, he saw that poison was in it, and said to Sigmund “This drink is cloudy, old man.” Sigmund took the horn and drank it off. It is said that Sigmund was hardy and that poison did him no harm, inside or out. And all his sons could tolerate poison on their skin. Borghild brought another horn to Sinfiotli, and asked him to drink, and everything happened as before. And a third time she brought him a horn, and reproachful words as well, if he didn’t drink from it. He spoke again to Sigmund as before. He said “Filter it through your mustache, son!” Sinfiotli drank it off and at once fell dead.

Sigmund carried him a long way in his arms and came to a long, narrow fjord, and there was a small boat there and a man in it. He offered to ferry Sigmund over the fjord. But when Sigmund carried the body out to the boat, it was fully laden. The man said Sigmund should go around the fjord inland. The man pushed the boat out and then suddenly vanished.

King Sigmund lived a long time in Denmark in the kingdom of Borghild, after he married her. Then he went south to Frankish lands, to the kingdom he had there. Then he married Hiordis, the daughter of King Eylimi. Their son was Sigurd. King Sigmund fell in a battle with the sons of Hunding. And then Hiordis married Alf, the son of King Hialprec. Sigurd grew up there as a boy.

Sigmund and all his sons were tall and outstanding in their strength, their growth, their intelligence, and their accomplishments. But Sigurd was the most outstanding of all, and everyone who knows about the old days says he was the most outstanding of men and the noblest of all the warrior kings.

A structural analysis of this text, dividing it into narrative units in a pattern shared with other texts from the same literature, might look like this:

```
<p id="P1">
<s id="S1">Sigmund ... was a king in Frankish country.</s>
<s id="S2">Sinfiotli was the eldest of his sons.</s>
<s id="S3">Borghild, Sigmund's wife, had a brother ...</s>
<s id="S4A">But Sinfiotli ... wooed the same woman</s>
<s id="S4B">and Sinfiotli killed him over it.</s>
<s id="S5">And when he came home, ... she was obliged to accept it.</s>
<s id="S6">At the funeral feast Borghild was serving beer.</s>
<s id="S7">She took poison ... and brought it to Sinfiotli.</s>
<!-- ... -->
<s id="S17">Sinfiotli drank it off and at once fell dead.</s>
<anchor id="NIL1" />
</p>
<p id="P2">Sigmund carried him a long way in his arms ... </p>
<p id="P3">King Sigmund lived a long time in Denmark ... </p>
<p id="P4">Sigmund and all his sons were tall ... </p>
<!-- ... -->
<span resp="TMA" type="structural unit" value="introduction"   from="S1" to="S3"  />
<span resp="TMA" type="structural unit" value="conflict"       from="S4a"         />
<span resp="TMA" type="structural unit" value="climax"         from="S4b"         />
<span resp="TMA" type="structural unit" value="revenge"        from="S5" to="S17" />
<span resp="TMA" type="structural unit" value="reconciliation" from="NIL1"        />
<span resp="TMA" type="structural unit" value="aftermath"      from="P2" to="P4"  />
```

Note the use of an empty <anchor> element to provide a target for the ‘reconciliation’ unit which is normally part of the narrative pattern but which is not realized in the text shown.

If groups of elements with the same resp or type are used, as in this example, they may be grouped together inside a <spanGrp> element, with the values of the common attribute(s) inherited from the higher element, as follows.

```
<spanGrp resp="TMA" type='structural unit'>
 <span value='introduction'   from="S1"   to="S3" />
 <span value='conflict'       from="S4a"          />
 <span value='climax'         from="S4b"          />
 <span value='revenge'        from="S5"   to="S17"/>
 <span value='reconciliation' from="NIL1"         />
```

```
 <span value='aftermath'      from="P2"   to="P4" />
</spanGrp>
```

The same analysis may be expressed with the <interp> element instead of the element; this element provide attributes for recording an interpretive category and its value, as well as the identity of the interpreter, but does not itself indicate which passage of text is being interpreted; the same interpretive structures can thus be associated with many passages of the text. The association between text passages and <interp> elements must be made either by pointing from the text to the <interp> element with the ana attribute defined in section 15.2 *Global Attributes for Simple Analyses*, or by pointing at both text and interpretation from a <link> element, as described in chapter 14 *Linking, Segmentation, and Alignment*.

To encode the first example above using <interp>, it is necessary to create a text element which contains — or corresponds to — the the third, fourth, and fifth orthographic sentences (S-units) in the paragraph. This can be done either with the seg element, described in 14.3 *Blocks, Segments and Anchors*, or the join element, described in 14.7 *Aggregation*. The resulting element can then be associated with the <interp> element using the ana attribute described in section 15.2 *Global Attributes for Simple Analyses*. We illustrate using the <seg> element.

```
<p id='MQp1s2p114'>
<s id='MQp1s2p114s1'>There was certainly a definite point ... </s>
<s id='MQp1s2p114s2'>It was not; then it was suddenly inescapable ... </s>
<seg id='MQp1s2p114s3-5' ana='moment'>
<s id='MQp1s2p114s3'>There was a slow integration ... </s>
<s id='MQp1s2p114s4'>She felt the rivers under the ground ... </s>
<s id='MQp1s2p114s5'>Not for one second longer ... </s>
</seg>
<s id='MQp1s2p114s6'>For during that space of time ... </s>
<!-- ... -->
</p>
<!-- ... -->
<interp id='MOMENT' resp='DTL' value='the moment'/>
```

The second example above can be recoded using <interp> and <interpGrp> tags in a similar manner. The interpretation itself can be expressed in an <interpGrp> element, which would replace the <spanGrp> in the example shown above:

```
<interpGrp resp='TMA' type='structural unit'>
 <interp id='INTRO'    value='introduction'  />
 <interp id='CONFLICT' value='conflict'      />
 <interp id='CLIMAX'   value='climax'        />
 <interp id='REVENGE'  value='revenge'       />
 <interp id='RECONCIL' value='reconciliation'/>
 <interp id='AFTERM'   value='aftermath'     />
</interpGrp>
```

This <interpGrp> element would be linked to the text either by means of the ana attribute, or by means of <link> elements. Using the ana attribute (on <seg> elements introduced specifically for this purpose), the text would be encoded as follows:

```
<p id='P1'>
<seg id='S1-S3' ana='INTRO'>
<s id='S1'>Sigmund ... was a king in Frankish country.</s>
<s id='S2'>Sinfiotli was the eldest of his sons.</s>
<s id='S3'>Borghild, Sigmund's wife, had a brother ... </s>
</seg>
<s id='S4A' ana='CONFLICT'>But Sinfiotli ... wooed the same woman</s>
<s id='S4B' ana='CLIMAX'>and Sinfiotli killed him over it.</s>
<seg id='S5-S17' ana='REVENGE'>
<s id='S5'>And when he came home, ... she was obliged to accept it.</s>
<s id='S6'>At the funeral feast Borghild was serving beer.</s>
<!-- ... -->
<s id='S17'>Sinfiotli drank it off and at once fell dead.</s>
</seg></p>
<anchor id='NIL1' ana='RECONCIL'/>
<p id='P2'>Sigmund carried him a long way in his arms ... </p>
```

```
<p id='P3'>King Sigmund lived a long time in Denmark ... </p>
<p id='P4'>Sigmund and all his sons were tall ... </p>
<!-- ... -->
<join id='P2-P4' targets='P2 P3 P4' ana='AFTERM'/>
```

The linkage may also be accomplished using a <linkGrp> element, whose content is a set of <link> elements which point to each interpretive element and its corresponding text unit. This method does not require the use of the ana attribute on the text units.

```
<linkGrp resp='TMA' targFunc='text interpretation'>
  <link targets='INTRO    S1-S3' />
  <link targets='CONFLICT S4A'   />
  <link targets='CLIMAX   S4b'   />
  <link targets='REVENGE  S5-S17'/>
  <link targets='RECONCIL NIL1'  />
  <link targets='AFTERM   P2-P4' />
</linkGrp>
```

One obvious advantage of using <interp> rather than elements for the Sigmund text is that the <interp> elements can be reused for marking up other texts in the same document, whereas the elements cannot. Another is that the <interp> element can be used to provide interpretations for discontinuous text elements (represented by <join> elements). On the other hand, the use of <interp> elements may require the creation of special text elements not otherwise needed (e.g. the <seg> and the <join> in the revised encoding of the text), whereas the the use of elements does not.

The formal declarations for the , <spanGrp>, <interp> and <interpGrp> elements are:

```
<!-- 15.3: Spans-->
<!ELEMENT span %om.RO; EMPTY>
<!ATTLIST span
      %a.global;
      %a.interpret;
      value CDATA #REQUIRED
      from IDREF #REQUIRED
      to IDREF #IMPLIED
      TEIform CDATA 'span'  >
<!ELEMENT spanGrp %om.RR; ((span)*)>
<!ATTLIST spanGrp
      %a.global;
      %a.interpret;
      TEIform CDATA 'spanGrp'  >
<!ELEMENT interp %om.RO; EMPTY>
<!ATTLIST interp
      %a.global;
      %a.interpret;
      value CDATA #REQUIRED
      TEIform CDATA 'interp'  >
<!ELEMENT interpGrp %om.RR; ((interp)*)>
<!ATTLIST interpGrp
      %a.global;
      %a.interpret;
      TEIform CDATA 'interpGrp'  >
<!-- end of 15.3-->
```

15.4 Linguistic Annotation

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*.

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 language 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*.

As one example of such types of analysis, consider the following sentence, taken from the Lancaster/IBM Treebank Project.[124]

> 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.[125] 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:[126]

```
<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;.&PUN;</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 a system processing the document may be adjusted by modifying the entity declarations to suit the convenience of whatever analytic software is to be used. For example, if a parser operating on this sentence uses a set of entity declarations in the following form, then the word class tags will simply disappear from the output.

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

Alternatively, suppose the entity set in use follows the following pattern:

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

Then the sample sentence will be processed as if it began:

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

[124] See G. N. Leech and R. G. Garside, *Running a Grammar Factory*, in *English Computer Corpora: Selected Papers and Research Guide*, S. Johansson and A.-B. Stenstr⊘m: ed. (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.

[125] 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).

[126] We have replaced the Claws code `$` for the ''s' morpheme by `GEN`, as in the tag set used by the British National Corpus (see 16.10 *Two Illustrations*), and the code `.` for the final full stop by `PUN`.

It would be more useful if the replacement texts for each entity were a code of some significance to a particular analysis program. If the codes are considered to be *atomic*, then one of the mechanisms based on the <interp> element described in section 15.3 *Spans and Interpretations* is sufficient. If the codes are considered to be compositional (for example that NN1 and NN2 have something in common, namely their *noun-ness*, which they do not share with, say, VVD), then this compositionality may be most clearly expressed using a mechanism based on the <fs> element defined in chapter 16 *Feature Structures*. For a detailed example, see 16.10 *Two Illustrations*.

One such replacement for the word-class entity references above is a set of empty <ptr> elements bearing target attributes as described in section 6.6 *Simple Links and Cross References*. The required entity definitions would look as follows.

```
<!ENTITY   AT   "<ptr target='AT'/>" >
<!ENTITY   NN1  "<ptr target='NN1'/>">
<!ENTITY   GEN  "<ptr target='GEN'/>">
<!-- ... -->
```

Then the text would be expanded to read:

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

The <ptr> elements are designed to point to elements with unique identifiers. But we have yet to specify what those elements are. Suppose we say that they are <interp> elements whose values are the same as their identifiers. That is, we provide an <interpGrp> element as follows:

```
<interpGrp type="word classes">
   <interp id="at" value="AT"/>
   <interp id="nn1" value="NN1"/>
   <interp id="gen" value="GEN"/>
   <!-- ... -->
</interpGrp>
```

Although common practice, this (or any similar) method of relating text to interpretation is seriously flawed. The interpretations are related not to text elements, but to points in the text, namely those that are occupied by the <ptr> elements. In order to relate the interpretation to the appropriate text units, a uniform convention needs to be applied; for example, that an interpretation relates to all the text material preceding the <ptr> element that points to it up to the immediately preceding <ptr>, or up to the <s> that delimits the S-unit containing that <ptr> element, whichever is nearer. While this convention works with texts that are marked up solely with <ptr> elements that point to interpretation elements, it does not work with texts with additional markup, for example <ptr> elements that are used for some other purpose. In addition, the convention fails for any markup in which interpretations are intended to be associated with nested text elements.

None of these difficulties arise if the text is fully segmented, using the linguistic segment elements described in section 15.1 *Linguistic Segment Categories*, and the ana attribute to point to the interpretations that are associated with each such segment, as follows:

```
<s type="sentence">
   <w ana="at">The</w>
   <w ana="nn1">victim</w>
   <m ana="gen">'s</m>
   <w ana="nn2">friends</w>
   <w ana="vvd">told</w>
   <w ana="nn2">police</w>
   <w ana="cst">that</w>
   <w ana="np1">Krueger</w>
   <w ana="vvd">drove</w>
   <w ana="ii">into</w>
   <w ana="at">the</w>
```

```
   <w ana="nn1">quarry</w>
   <w ana="cc">and</w>
   <w ana="rr">never</w>
   <w ana="vvd">surfaced</w>
   <c ana="pun">.</c>
</s>
```

Analysis into phrase and clause elements can be superimposed on the word and morpheme tagging in the preceding illustration. For example, Claws provides the following constituent analysis of the sample sentence (the word class codes have been deleted):

```
[N [G The victim's G] friends N] [V told [N police N]
[Fn that [N Krueger N] [V [V& drove [P into [N the quarry N]P]V&]
and [V+ never surfaced V+]V]Fn]V]
```

Treating the labels on the brackets as phrase or clause interpretations, this analysis of the structure of the example sentence can be combined with the word class analysis and represented as follows (the symbol V& representing the first part of a coordinate phrase, has been replaced by V1, and V+, representing the second part, has been replaced by V2).

```
<s type="sentence">
   <phr ana="n">
      <phr ana="g">
         <w ana="at">The</w>
         <w ana="nn1">victim</w>
         <m ana="gen">'s</m>
      </phr>
      <w ana="nn2">friends</w>
   </phr>
   <phr ana="v">
      <w ana="vvd">told</w>
      <phr ana="n">
         <w ana="nn2">police</w>
      </phr>
      <cl ana="fn">
         <w ana="cst">that</w>
         <phr ana="n">
            <w ana="np1">Krueger</w>
         </phr>
         <phr ana="v">
            <phr ana="v1">
               <w ana="vvd">drove</w>
               <phr ana="p">
                  <w ana="ii">into</w>
                  <phr ana="n">
                     <w ana="at">the</w>
                     <w ana="nn1">quarry</w>
                  </phr>
               </phr>
            </phr>
            <w ana="cc">and</w>
            <phr ana="v2">
               <w ana="rr">never</w>
               <w ana="vvd">surfaced</w>
            </phr>
         </phr>
      </cl>
   </phr>
   <c ana="pun">.</c>
</s>
```

A representation using the <linkGrp> element can be obtained by supplying each linguistic segment with its own id attribute, removing its ana attribute, and putting each segment-interpretation pair into a <link> element inside the <linkGrp> element.

Each linguistic 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 state of affairs does not of course apply in all types of analysis, and these Guidelines provide a number of mechanisms to support the representation of discontinuities or multiple analyses. A brief overview of these facilities is provided in chapter 31 *Multiple Hierarchies*; also see 14 *Linking, Segmentation, and Alignment*. These mechanisms all depend to a greater or lesser 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 mechanisms proposed in this chapter may also be used to encode analyses of an entirely different kind, for example discourse function. Here is an application of the span technique to record details of a sales transaction in a spoken text.

```
<u id="u1" who="p1">Can I have ten oranges and a kilo of bananas please?</u>
<u id="u2" who="p2">Yes, anything else?</u>
<u id="u3" who="p1">No thanks.</u>
<u id="u4" who="p2">That'll be dollar forty.</u>
<u id="u5" who="p1">Two dollars</u>
<u id="u6" who="p1">Sixty, eighty, two dollars. Thank you.</u>
<spanGrp type="transactions">
   <span value="sale request" from="u1"/>
   <span value="sale compliance" from="u2" to="u3"/>
   <span value="sale" from="u4"/>
   <span value="purchase" from="u5"/>
   <span value="purchase closure" from="u6"/>
</spanGrp>
```

For further discussion of the <u> (utterance) element and other elements recommended for transcriptions of spoken language, see chapter 11 *Transcriptions of Speech*.