17 Certainty and Responsibility

Encoders of text often find it useful to indicate that some aspects of the encoded text are problematic or uncertain, and to indicate who is responsible for various aspects of the markup of the electronic text. These Guidelines provide three methods of recording uncertainty about the text or its markup:

- the <note> element defined in section 6.8 *Notes, Annotation, and Indexing* may be used with a value of certainty for its type attribute.
- the <certainty> element defined in this chapter may be used to record the nature and degree of the uncertainty in a more structured way.
- the <alt> element defined in the additional tag set for linking and segmentation may be used to provide alternative encodings for parts of a text, as described in section 14.8 *Alternation*.

There are three methods of indicating responsibility for different aspects of the electronic text:

- the TEI header records who is responsible for an electronic text by means of the <respStmt> element and other more specific elements (<author>, <sponsor>, <funder>, <principal>, etc.) used within the <titleStmt>, <editionStmt>, and <revisionDesc> elements.
- the <note> element may be used with a value of resp or responsibility in its type attribute.
- the <respons> element defined in this chapter may be used to record fine-grained structured information about responsibility for individual tags in the text.

No special steps are needed to use the <note> and <respStmt> elements, since they are defined in the core tag set and header respectively. The <alt> element is only available when the additional tag set for linking has been selected, as described in chapter 14 *Linking, Segmentation, and Alignment*. To use the <certainty> and <respons> elements, the additional tag set for certainty and responsibility must be selected; this is done by defining the parameter entity TEI.certainty with the value INCLUDE, as shown in the example below:

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

17.1 Levels of Certainty

Many types of uncertainty may be distinguished. The <certainty> element is designed to encode the following sorts:

- a given tag may or may not correctly apply (e.g. a given word may be a personal name, or perhaps not)
- the precise point at which an element begins or ends is uncertain
- the value to be given for an attribute is uncertain
- content supplied by the encoder (such as the expansion of an abbreviation marked by the <abbr> tag) is uncertain
- the transcription of a source text is uncertain, perhaps because it is hard to read or hard to hear; this sort of uncertainty is also handled by the <unclear> element in section 18.2.3 *Damage, Illegibility, and Supplied Text*

The following types of uncertainty are *not* indicated with the <certainty> element:

- a number or date is imprecise
- the text is ambiguous, so a given passage has several possible interpretations
- a transcriber, editor, or author wishes to indicate a level of confidence in a factual assertion made in the text

- an author is not sure if the sentence she has chosen to start a paragraph is really the one she wants to retain in the final version

Precision of numbers and dates is discussed in section 6.4 *Names, Numbers, Dates, Abbreviations, and Addresses*; well-defined ambiguity is handled with alternations in feature-structure values in chapter 16 *Feature Structures*. Uncertainty about the truth of assertions in the text and other sorts of authorial and editorial uncertainty about whether the content is satisfactory are not handled by the <certainty> element, though they may be expressed using the <note> element.

17.1.1 Using Notes to Record Uncertainty

The simplest way of recording uncertainty about markup is to attach a note to the element or location about which one is unsure. In the following (invented) paragraph, for example, an encoder might be uncertain whether to mark "Essex" as a place name or a personal name, since both might be plausible in the given context:

> Elizabeth went to Essex. She had always liked Essex.

Using <note>, the uncertainty here may be recorded quite simply:

```
<persName>Elizabeth</persName> went to <placeName>Essex</placeName>.
She had always liked <placeName>Essex</placeName>.<note
type="uncertainty" resp="MSM">It is not
clear here whether <mentioned>Essex</mentioned>
refers to the place or to the nobleman. -MSM</note>
```

Using the normal mechanisms, the note may be associated unambiguously with specific elements of the text, thus:

```
<persName>Elizabeth</persName> went to <placeName id="p1">Essex</placeName>.
She had always liked <placeName id="p2">Essex</placeName>.<note
type="uncertainty" resp="MSM" target="p1 p2">It
is not clear here whether <mentioned>Essex</mentioned>
refers to the place or to the nobleman. If the latter,
it should be tagged as a personal name. -MSM</note>
```

The advantage of this technique is its relative simplicity. Its disadvantage is that the nature and degree of uncertainty are not conveyed in any systematic way and thus are not susceptible to any sort of automatic processing.

17.1.2 Structured Indications of Uncertainty

To record uncertainty in a more structured way, susceptible of at least simple automatic processing, the <certainty> element may be used:

<certainty> indicates the degree of certainty or uncertainty associated with some aspect of the text markup. Attributes include:

target points at the elements whose markup is uncertain.
Values one or more valid identifiers, separated by white space.

locus indicates the precise location of the uncertainty in the markup: applicability of the element, precise position of the start- or end-tag, value of a specific attribute, etc.
Suggested values include:

- #gi uncertain whether the element used actually applies to the passage.
- #startloc start-tag may not be correctly located.
- #endloc end-tag may not be correctly located.
- #location both the start-tag and the end-tag may not be correctly located.
- name the value given for the attribute name is uncertain.
- #transcribedContent the content of the element may not be a correct transcription of the source text.
- #suppliedContent the content of the element may not have been correctly supplied by the reader, e.g. as in the cases of corr and abbrev elements.

given indicates conditions assumed in the assignment of a degree of confidence.
Values a characterization of the conditions which are assumed in the assignment of a degree of confidence. This may be in prose.

degree indicates the degree of confidence assigned to the aspect of the markup named by the locus attribute.

Values Values of degree might be yes or no, the reals between 0 and 1, or traditional characterizations such as 'doubtful', 'circa', etc. Generally we recommend decimal numbers between 0 and 1, where larger numbers denote a greater degree of confidence in the assertions; 0 representing 'certainly false' and 1 representing 'certainly true'.

assertedValue provides an alternative value for the aspect of the markup in question—an alternative generic identifier, transcription, or attribute value, or the identifier of an <anchor> element (to indicate an alternative starting or ending location). If an assertedValue is given, the confidence level specified by degree applies to the alternative markup specified by assertedValue; if none is given, it applies to the markup in the text.

Values generic identifier, attribute value, location (e.g. indicated by a reference to an <anchor> element or to an <xptr> element), or other appropriate alternative value.

desc further describes the uncertainty in prose, perhaps indicating its nature, cause, or the justification for the degree of confidence asserted.

Values a prose description of how and why the markup is uncertain.

Returning to the example, the <certainty> element may be used to record doubts about the proper encoding of "Essex" in several ways of varying precision. To record merely that we are not certain that "Essex" is in fact a place name, as it is tagged, we use the target attribute to identify the element in question, and the locus attribute to indicate what aspect of the markup we are uncertain about (in this case, whether we have used the correct element type):

```
Elizabeth went to
<placeName id="p1">Essex</placeName>.
<!-- ... elsewhere in the document ... -->
<certainty target="p1" locus="#gi" desc="possibly not a placename"/>
```

Because it is linked to the location of the uncertainty by a reference, the <certainty> element will typically be included in the same document as its target. It may be placed adjacent to the target element, or elsewhere in the document.

To record the further information that we estimate, subjectively, that there is a 60 percent chance of "Essex" being a place name here, we can add a value for our degree of confidence (usually a number between 0 and 1, representing the estimated probability):

```
Elizabeth went to
<placeName id="p1">Essex</placeName>.
<!-- ... -->
<certainty target="p1" locus="#gi" desc="possibly not a placename" degree="0.6"/>
```

According to one expert, there is a 60 percent chance of "Essex" being a place name here, and a 40 percent chance of its being a personal name. We use two <certainty> elements to indicate the two probabilities independently. Both elements indicate the same location in the text, but the second provides an alternative choice of generic identifier (in this case <persName>) is given as the value of the assertedValue attribute:

```
Elizabeth went to
<placeName id="p1">Essex</placeName>.
<!-- ... -->
<certainty target="p1" locus="#gi"
   desc="probably a placename, but possibly not" degree="0.6"/>
<certainty target="p1" locus="#gi" assertedValue="persName"
   desc="may refer to the Earl of Essex" degree="0.4"/>
```

Finally, we may wish to make our probability estimates contingent on some condition. In the passage "Elizabeth went to Essex; she had always liked Essex," for example, we may feel there is a 60 percent chance that the county is meant, and a 40 percent chance that the earl is meant. But the two occurrences of the word are not independent: there is (we may feel) no chance at all that one occurrence refers to the county and one to the earl. We can express this by using the given attribute to list the identifiers of <certainty> elements.

```
Elizabeth went to <placeName id="p1">Essex</placeName>.
She had always liked <placeName id="p2">Essex</placeName>.
<!-- ... -->
<!-- 60% chance that P1 is a placename,
     40% chance a personal name. -->
<certainty id="cert-1" target="p1" locus="#gi"
   desc="probably a placename, but possibly not" degree="0.6"/>
<certainty id="cert-2" target="p1" locus="#gi"
   desc="may refer to the Earl of Essex" assertedValue="persName" degree="0.4"/>
<!-- 60% chance that P2 is a placename,
     40% chance a personal name.
    100% chance that it agrees with P1. -->
<certainty target="p2" locus="#gi" given="cert-1"
   desc="if P1 is a placename, P2 certainly is" degree="1.0"/>
<certainty target="p2" locus="#gi" assertedValue="persName" given="cert-2"
   desc="if p1 refers to the Earl of Essex, so does P2" degree="1.0"/>
```

When given conditions are listed, the <certainty> element is interpreted as claiming a given degree of confidence in a particular markup given the assertional content of the <certainty> elements indicated—that is, *if the markup described in the indicated <certainty> elements is correct.*

Conditional confidence may be less that 100 percent: given the sentence “Ernest went to old Saybrook”, we may interpret “Saybrook” as a personal name or a place name, assigning a 60 percent probability to the former. If it is a place name, there may be a 50 percent chance that the place name actually in question is “Old Saybrook” rather than “Saybrook”, while if it is correctly tagged as a personal name, it is much more likely (say, 90 percent certain) that the name is “Saybrook”. This state of affairs can be expressed using the <certainty> element thus:

```
Earnest went to <anchor id="a1"/> old <persName id="p1">Saybrook</persName>.

<certainty id="c1" target="p1" locus="#gi" degree="0.6"/>
<certainty target="p1" locus="startloc" given="c1" degree="0.9"/>

<certainty id="c2" target="p1" locus="#gi" assertedValue="persName" degree="0.4"/>
<certainty target="p1" locus="startloc" given="c2" degree="0.5"/>
<certainty id="c3" target="p1" locus="startloc" assertedValue="a1" given="c1" degree="0.5"/>
```

In this case, the assertedValue on <certainty> element c3 is a reference to an <anchor> element at the alternative starting point for the element.

Multiplying the numeric values out, this markup may be interpreted as assigning specific probabilities to three different ways of marking up the sentence:

```
Earnest went to old <persName>Saybrook</persName>.    (0.6 * 0.9, or 0.54)
Earnest went to old <placeName>Saybrook</placeName>.  (0.4 * 0.5, or 0.20)
Earnest went to <placeName>old Saybrook</placeName>.  (0.4 * 0.5, or 0.20)
```

The probabilities do not add up to 1.00 because the markup indicates that if “Saybrook” is (part of) a personal name, there is a 10 percent likelihood that the element should start somewhere other than the place indicated, without however giving an alternative location; there is thus a 6 percent chance (0.1×0.6) that none of the alternatives given is correct.

If an attribute value is uncertain, the locus attribute takes as its value the name of the attribute in question. In this example, there is only a 50 percent chance that the question was spoken by participant A:

```
<u id="u1" who="a">Have you heard the election results?</u>
<!-- ... -->
<certainty target="u1" locus="who" degree="0.5"/>
```

Doubts about whether the transcription is correct may be expressed by assigning to locus the value ‘#transcribedContent’. For example, if the source is hard to read and so the transcription is uncertain:

```
I have a <emph id="p1">gub</emph>.
<certainty target="p1" locus="#transcribedContent" degree="0.5"/>
```

Degrees of confidence in the proper expansion of abbreviations may also be expressed, by using the value ‘#suppliedContent’:

```
You will want to use <expan id="e1" abbr="SGML">Standard
Generalized Markup Language</expan> ...
<!-- ... -->
<certainty target="e1" locus="#suppliedContent" degree="0.9"/>
```

The assertedValue attribute should be used to provide an alternative value for whatever aspect of the markup is in doubt: an alternative generic identifier, or the identifier of an alternative starting or ending point, as already shown, an alternative attribute value, or alternative element content, as in this example:

```
I have a <emph id="p1">gub</emph>.
<certainty target="p1" locus="#transcribedContent" asserted-
Value="gun" desc="a gun makes more sense in a holdup" degree="0.8"/>
```

Since attribute values have no internal substructure, the assertedValue attribute is useful for specifying alternative transcriptions only in relatively restricted circumstances (specifically, when the alternative reading has no elements nested within it). More robust methods of handling uncertainties of transcription are the <unclear> element and the <app> and <rdg> elements described in chapter 19 *Critical Apparatus*. The <certainty> element allows for indications of uncertainty to be structured with at least as much detail and clarity as appears to be currently required in most ongoing text projects. It is expected that in the future more adequate systems for expressing uncertainty will be developed. These may extend the <certainty> element or they may make use of the feature-structure encoding mechanisms described in chapter 16 *Feature Structures*.

The <certainty> element and the other TEI mechanisms for indicating uncertainty provide a range of methods of graduated complexity. Simple expressions of uncertainty may be made by using the <note> element. This is simple and convenient, and can accommodate either a discursive and unstructured indication of uncertainty, or a complex and structured but probably project-specific expression of uncertainty. In general, however, unless special steps are taken, the <note> element does not provide as much expressive power as the <certainty> element, and in cases where highly structured certainty information must be given, it is recommended that the <certainty> element be used.

The <certainty> element may be used for simple unqualified indications of uncertainty, in which case only the locus and target attributes might be specified. In more complex cases, the other attributes may be used to provide fuller information. While these attributes may take any string of characters as value, the recommended values should be used wherever possible; if they are not appropriate in a given situation, encoders should provide their own controlled vocabulary and document it in the <encodingDesc> or <tagUsage> elements of the TEI header.

The <certainty> element has the following formal declaration:

```
<!-- 17.1.2: Certainty and uncertainty-->
<!--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-->
<!ELEMENT certainty %om.RO;  EMPTY>
<!ATTLIST certainty
      %a.global;
      target IDREFS #REQUIRED
      locus CDATA #REQUIRED
      assertedValue CDATA #IMPLIED
      desc CDATA #IMPLIED
      given CDATA #IMPLIED
      degree CDATA #IMPLIED
      TEIform CDATA 'certainty'  >
<!--declarations from 17.2: Responsibility for markup inserted here -->
<!-- end of 17.1.2-->
```

17.2 Attribution of Responsibility

In general, attribution of responsibility for the transcription and markup of an electronic text is made by <respStmt> elements within the header: specifically, within the title statement, the edition statement(s), and the revision history.

In some cases, however, more detailed element-by-element information may be desired. For example, an encoder may wish to distinguish between the individuals responsible for transcribing the content and those responsible for determining that a given word or phrase constitutes a proper noun. Where such fine-grained attribution of responsibility is required, the <respons> element can be used:

<respons> identifies the individual(s) responsible for some aspect of the markup of particular element(s). Attributes include:

target gives the identifier(s) of the element(s) for which some aspect of the responsibility is being assigned.
Values one or more valid identifiers, separated by white space.

locus indicates the specific aspect of the markup for which responsibility is being assigned.
Suggested values include:
#gi responsibility for the claim that the element is of the type indicated by the markup
#location responsibility for the claim that the element begins and ends where indicated
#startloc responsibility for the claim that the element begins where indicated
#endloc responsibility for the claim that the element ends where indicated
name responsibility for the claim that the name attribute has the value given in the markup
#transcribedContent responsibility for the transcription of the element content
#suppliedContent responsibility for the contents supplied by the encoder (corrections, expansions of abbreviations, etc.)

resp identifies the individual or agency responsible for the indicated aspect of the electronic text.
Values any string of characters, typically the initials of an individual, the acronym of an agency, the name of a computer program, etc.

desc (description) gives a brief prose note supplying any additional information which should be recorded
Values any string of characters, typically a phrase or sentence in a natural language.

This element allows one or more aspects of the markup to be attributed to a given individual. The target and locus attributes function as they do on the <certainty> element described in section 17.1 *Levels of Certainty*: the target attribute points at a particular element (or set of elements), and locus indicates the particular aspect of the encoding of those elements for which responsibility is to be assigned. The suggested values may be combined as appropriate. For example, to indicate that RC is responsible for transcribing an illegible word, and that AR is responsible for identifying that word as a proper noun, the text might be encoded thus:

```
Earnest went to old <persName id="p1">Saybrook</persName>.
<!-- ... -->
<respons target="p1" locus="#transcribedContent" resp="RC"/>
<respons target="p1" locus="#gi #location" resp="AR"/>
```

Some elements bear specialized resp or agent attributes, which have specific meanings that vary from element to element; the <respons> element should be reserved for the general aspects of responsibility common to all text transcription and markup, and should not be confused with the more specific attributes on individual elements.

The formal declaration of the <respons> element is this:

```
<!-- 17.2: Responsibility for markup-->
<!ELEMENT respons %om.RO; EMPTY>
<!ATTLIST respons
      %a.global;
      target IDREFS #REQUIRED
      locus CDATA #REQUIRED
      resp CDATA #REQUIRED
      desc CDATA #IMPLIED
      TEIform CDATA 'respons'  >
<!-- end of 17.2-->
```