11 Transcriptions of Speech

This chapter has not yet been revised to reflect developments in the field of speech transcription and multimodal language annotation since its first publication. It is planned that a future revision of these Guidelines will include recommendations in these areas, which will in turn imply some revision of the features discussed here.

The base tag set for transcriptions of spoken language described in this chapter is intended for use with a wide variety of transcribed spoken material. It should be stressed, however, that the present proposals are not intended to support unmodified every variety of research undertaken upon spoken material now or in the future; some discourse analysts, some phonologists, and doubtless others may wish to extend the scheme presented here to express more precisely the set of distinctions they wish to draw in their transcriptions. Speech regarded as a purely acoustic phenomenon may well require different methods from those outlined here, as may speech regarded solely as a process of social interaction.

This chapter begins with a discussion of some of the problems commonly encountered in transcribing spoken language (section 11.1 *General Considerations and Overview*). Section 11.2 *Elements Unique to Spoken Texts* describes the basic structural elements of this tag set. Finally, section 11.3 *Elements Defined Elsewhere* of this chapter reviews further problems specific to the encoding of spoken language, demonstrating how mechanisms and elements discussed elsewhere in these Guidelines may be applied to them.

The overall structure of a TEI spoken text is identical to that of any other TEI text: the <TEI.2> element for a spoken text contains a <teiHeader> element, followed by a <text> element. Even texts primarily composed of transcribed speech may also include conventional front and back matter, and may even be organized into divisions like printed texts. For simplicity's sake, therefore, the base tag set for spoken text uses the default text structure, as defined in chapter 7 *Default Text Structure*; this tag set is embedded automatically by the spoken base tag set.

To enable the base tag set for spoken texts, a parameter entity TEI.spoken must be declared within the document type declaration subset, the value of which is INCLUDE, as further described in section 3.3 *Invocation of the TEI DTD*. A document using this base tag set and no additional tag sets will thus begin as follows:

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE TEI.2 PUBLIC "-//TEI P4//DTD Main Document Type//EN" "tei2.dtd" [
    <!ENTITY % TEI.XML     'INCLUDE' >
    <!ENTITY % TEI.spoken  'INCLUDE' >
 ]>
```

This declaration makes available all of the elements and attributes discussed in the present chapter, in addition to the core elements described in chapter 6 *Elements Available in All TEI Documents*. If other elements are needed (in particular, those needed for synchronization or segmentation), additional tag sets may also be enabled in a similar way.

Two additional classes are defined by this tag set. Elements which appear only within transcribed speech constitute the comp.spoken element class. Elements with a specificable temporal duration constitute the timed element class. These classes are defined in the file teispok2.ent using the following parameter entities:

```
<!-- 11.: Class declarations for Transcribed Speech-->
<!--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-->
<!ENTITY % x.comp.spoken "" >
<!ENTITY % m.comp.spoken "%x.comp.spoken; %n.event; | %n.kinesic; |
%n.pause; | %n.shift; | %n.u; | %n.vocal; | %n.writing;">
```

```
<!ENTITY % mix.spoken '| %m.comp.spoken;' >
<!-- end of 11.-->
```

The elements of the base tag set for transcribed speech are declared in the file teispok2.dtd, which is organized as follows:

```
<!-- 11.: Base tag set for Transcribed Speech-->
<!--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-->
<!--declarations from 11.2.7: Components of Transcribed Speech inserted here -->
<!--The base tag set for transcriptions of speech uses the
standard default text-structure elements, which are embedded
here:-->
<![%TEI.singleBase;[
<!ENTITY % TEI.structure.dtd PUBLIC '-//TEI P4//ELEMENTS Default Text
Structure//EN' 'teistr2.dtd' >
%TEI.structure.dtd;
]]>
<!-- end of 11.-->
```

11.1 General Considerations and Overview

There is great variation in the ways different researchers have chosen to represent speech using the written medium.[90] This reflects the special difficulties which apply to the encoding or *transcription* of speech. Speech varies according to a large number of dimensions, many of which have no counterpart in writing (for example, tempo, loudness, pitch, etc.). The audibility of speech recorded in natural communication situations is often less than perfect, affecting the accuracy of the transcription. Spoken material may be transcribed in the course of linguistic, acoustic, anthropological, psychological, ethnographic, journalistic, or many other types of research. Even in the same field, the interests and theoretical perspectives of different transcribers may lead them to prefer different levels of detail in the transcript and different styles of visual display. The production and comprehension of speech are intimately bound up with the situation in which speech occurs, far more so than is the case for written texts. A speech transcript must therefore include some contextual features; determining which are relevant is not always simple. Moreover, the ethical problems in recording and making public what was produced in a private setting and intended for a limited audience are more frequently encountered in dealing with spoken texts than with written ones.

Speech also poses difficult structural problems. Unlike a written text, a speech event takes place in time. Its beginning and end may be hard to determine and its internal composition difficult to define. Most researchers agree that the utterances or *turns* of individual speakers form an important structural component in most kinds of speech, but these are rarely as well-behaved (in the structural sense) as paragraphs or other analogous units in written texts: speakers frequently interrupt each other, use gestures as well as words, leave remarks unfinished and so on. Speech itself, though it may be represented as words, frequently contains items such as vocalized pauses which, although only semi-lexical, have immense importance in the analysis of spoken text. Even non-vocal elements such as gestures may be regarded as forming a component of spoken text for some analytic purposes. Below the level of the individual utterance, speech may be segmented into units defined by phonological, prosodic, or syntactic phenomena; no clear agreement exists, however, even as to appropriate names for such segments.

[90] For a discussion of several of these see J. A. Edwards and M. D. Lampert, eds., *Talking Language: Transcription and Coding of Spoken Discourse* (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1993); Stig Johansson, *Encoding a Corpus in Machine-Readable Form*, in *Computational Approaches to the Lexicon: An Overview*, ed. B. T. S. Atkins et al. (Oxford: Oxford University Press, forthcoming); and Stig Johansson et al. *Working Paper on Spoken Texts*, document TEI AI2 W1, 1991.

Spoken texts transcribed according to the guidelines presented here are organized as follows. As noted above, speech is regarded as being composed of arbitrary high-level units called *texts*. A spoken <text> might typically be a conversation between a small number of people, a lecture, a broadcast TV item, or a similar event. Each such unit has associated with it a <teiHeader> providing detailed contextual information such as the source of the transcript, the identity of the participants, whether the speech is scripted or spontaneous, the physical and social setting in which the discourse takes place and a range of other aspects. For details of the header in general, refer to chapter 5 *The TEI Header*; for details of additional elements for the documentation of participant and contextual information, see section 23.2 *Contextual Information*.

Defining the bounds of a spoken text is frequently a matter of arbitrary convention or convenience. In public or semi-public contexts, a text may be regarded as synonymous with, for example, a *lecture*, a *broadcast item*, a *meeting*, etc. In informal or private contexts, a text may be simply a conversation involving a specific group of participants. Alternatively, researchers may elect to define spoken texts solely in terms of their duration in time or length in words. By default, these Guidelines assume of a text only that:

- it is internally cohesive,
- it is describable by a single header, and
- it represents a single stretch of time with no significant discontinuities.

Deviation from these assumptions may be specified (for example, the org attribute on the <text> element may take the value compos to specify that the components of the text are discrete) but is not recommended.

Within a <text> it may be necessary to identify subdivisions of various kinds, if only for convenience of handling. The neutral element discussed in section 7.1 *Divisions of the Body* is recommended for this purpose. It may be found useful also for representing subdivisions relating to discourse structure, speech act theory, transactional analysis, etc., provided that these divisions are hierarchically well-behaved. Where they are not, as is often the case, the mechanisms discussed in chapters 14 *Linking, Segmentation, and Alignment* and 31 *Multiple Hierarchies* may be used.

A spoken text may contain any of the following components:

- utterances
- pauses
- vocalized but non-lexical phenomena such as coughs
- kinesic (non-verbal, non-lexical) phenomena such as gestures
- entirely non-linguistic events occurring during and possibly influencing the course of speech
- writing, regarded as a special class of event in that it can be transcribed, for example captions or overheads displayed during a lecture
- shifts or changes in vocal quality

Elements to represent all of these features of spoken language are discussed in section 11.2 *Elements Unique to Spoken Texts* below.

An utterance (tagged <u>) may contain lexical items interspersed with pauses and non-lexical vocal sounds; during an utterance, non-linguistic events may occur and written materials may be presented. The <u> element can thus contain any of the other elements listed, interspersed with a transcription of the lexical items of the utterance; the other elements may all appear between utterances or next to each other, but except for <writing> they do not contain any other elements nor any data.

11.1.1 Divisions

A spoken text itself may be without substructure, that is, it may consist simply of units such as utterances or pauses, not grouped together in any way, or it may be subdivided into one or more divisions as described in this section.

If the notion of what constitutes a 'text' in spoken discourse is inevitably rather an arbitrary one, the notion of formal subdivisions within such a 'text' is even more debatable. Nevertheless, such divisions may be useful for such types of discourse as debates, broadcasts, etc., where structural subdivisions can easily be identified, or more generally wherever it is desired to aggregate utterances or other parts of a transcript into units smaller than a complete 'text'. Examples might include "conversations" or "discourse fragments", or more narrowly, "that part of the conversation where topic x was discussed", provided only that the set of all such divisions is coextensive with the text.

Each such division of a spoken text should be represented by the numbered or un-numbered `<div>` elements defined in chapter 7 *Default Text Structure*. For some detailed kinds of analysis a hierarchy of such divisions may be found useful; nested `<div>` elements may be used for this purpose, as in the example below.

The `<div>` element is a member of the `divn` class of structural elements, and therefore has the following attributes in common with other members of the class:

`type` specifies a name conventionally used for this level of subdivision, e.g. "act", "volume", "book", "section", "canto", etc.

`org` specifies how the content of the division is organized. Legal values are:

- `composite` composite content: i.e. no claim is made about the sequence in which the immediate contents of this division are to be processed, or their inter-relationships.
- `uniform` uniform content: i.e. the immediate contents of this element are regarded as forming a logical unit, to be processed in sequence.

`sample` indicates whether this division is a sample of the original source and if so, from which part. Legal values are:

- `initial` division lacks material present at end in source.
- `medial` division lacks material at start and end.
- `final` division lacks material at start.
- `unknown` position of sampled material within original unknown.
- `complete` division is not a sample.

`part` specifies whether or not the division is fragmented by some other structural element, for example a speech which is divided between two or more verse stanzas. Legal values are:

- `Y` the division is incomplete in some respect
- `N` either the division is complete, or no claim is made as to its completeness.
- `I` the initial part of an incomplete division
- `M` a medial part of an incomplete division
- `F` the final part of an incomplete division

The `type` attribute may be used to characterize divisions in any way that is convenient; no specific recommendations are made in these Guidelines. For example, a collection made up of transcribed 'sound bites' taken from speeches given by a politician on different occasions, might encode each extract as a distinct `<div>`, nested within a single composite `<div>` as follows:

```
<div type='soundbites' org='composite'>
   <div sample='medial'>
      <!-- ... -->
   </div>
   <div sample='medial'>
      <!-- ... -->
   </div>
   <div sample='initial'>
      <!-- ... -->
   </div>
</div>
```

As a member of the class declaring, the element may also carry a decls attribute, for use where the divisions of a text do not all share the same set of the contextual declarations specified in the TEI header. (See further section 23.3 *Associating Contextual Information with a Text*).

11.2 Elements Unique to Spoken Texts

The following elements characterize spoken texts, transcribed according to these Guidelines:

<u> a stretch of speech usually preceded and followed by silence or by a change of speaker. Attributes include:

trans (transition) indicates the nature of the transition between this utterance and the previous one.

Legal values are:

`smooth` this utterance begins without unusual pause or rapidity.
`latching` this utterance begins with a markedly shorter pause than normal.
`overlap` this utterance begins before the previous one has finished.
`pause` this utterance begins after a noticeable pause.

who supplies an identifier for the speaker or group of speakers. Its value is the identifier of a <participant> or <participantGrp> element in the TEI header.

Values Must identify a participant or participant group within the TEI Header

<pause> a pause either between or within utterances. Attributes include:

type categorizes the pause in some respect.

Values An open list

who supplies an identifier for the person or group pausing. Its value is the identifier of a <participant> or <participant.grp> element in the TEI header.

Values Must identify a participant or participant group within the TEI Header

<vocal> any vocalized but not necessarily lexical phenomenon, for example voiced pauses, non-lexical backchannels, etc. Attributes include:

who supplies an identifier for the vocalist(s). Its value is the identifier of a <participant> or <participant.grp> element in the TEI header.

Values Must identify a participant or participant group within the TEI Header

iterated (iterated) indicates whether or not the phenomenon is repeated.

Legal values are:

`y` the phenomenon is repeated.
`n` the phenomenon is atomic.
`u` unknown or unmarked.

desc (description) supplies a conventional representation for the phenomenon.

Values a description or representation of the phenomenon chosen from a semi-closed list

<kinesic> any communicative phenomenon, not necessarily vocalized, for example a gesture, frown, etc. Attributes include:

who supplies an identifier for the participant performing the gesture. Its value is the identifier of a <participant> or <participant.grp> element in the TEI header.

Values Must identify a participant or participant group within the TEI Header

iterated (iterated) indicates whether or not the phenomenon is repeated.

Legal values are:

`y` the phenomenon is repeated.
`n` the phenomenon is atomic.
`u` unknown or unmarked.

desc (description) supplies a conventional representation for the phenomenon.

Values a description or representation of the phenomenon chosen from a semi-closed list

<event> any phenomenon or occurrence, not necessarily vocalized or communicative, for example incidental noises or other events affecting communication. Attributes include:

who supplies an identifier for the agent of the event described, if any. Its value is the identifier of a <participant> or <participant.grp> element in the TEI header.
Values Must identify a participant or participant group within the TEI Header

iterated (iterated) indicates whether or not the phenomenon is repeated.
Legal values are:
- y the phenomenon is repeated.
- n the phenomenon is atomic.
- u unknown or unmarked.

desc (description) supplies a conventional representation for the phenomenon.
Values a description or representation of the phenomenon chosen from a semi-closed list

<writing> a passage of written text revealed to participants in the course of a spoken text. Attributes include:

who (who) supplies an identifier for the participant who reveals or creates the writing, if any. Its value is the identifier of a <participant> or <participant.grp> element in the TEI header.
Values Must identify a participant or participant group within the TEI Header

type (Type) categorizes the kind of writing in some way, for example as a subtitle, notice-board etc.
Values Open list

script (Script pointer) points to a bibliographic citation in the header giving a full description of the source or script of the writing.
Values Must be a valid identifier for a <script.decl> element in the TEI header

gradual (gradual) indicates whether the writing is revealed all at once or gradually.
Legal values are:
- y the writing is revealed gradually.
- n the writing is revealed all at once.
- u unknown or unmarked.

<shift> marks the point at which some paralinguistic feature of a series of utterances by any one speaker changes. Attributes include:

who supplies an identifier for the speaker or group of speakers whose shift in some feature is being noted. Its value is the identifier of a <participant> or <participant.grp> element in the TEI header.
Values Must identify a participant or participant group within the TEI Header

feature (feature) a paralinguistic feature.
Legal values are:
- tempo speed of utterance.
- loud loudness.
- pitch pitch range.
- tension tension or stress pattern.
- rhythm rhythmic qualities.
- voice voice quality.

new (new state) specifies the new state of the paralinguistic feature specified.
Values An open list (for an example of possible values, see 11.3.2 *Synchronization and Overlap*)

Each of these is further discussed and specified below in sections 11.2.1 *Utterances* to 11.2.4 *Writing*.

We can show the relationship between four of these constituents of speech using the features eventive, communicative, anthropophonic (for sounds produced by the human vocal apparatus), and lexical:

	eventive	communicative	anthropophonic	lexical
event	+	-	-	-
kinesic	+	+	-	-
vocal	+	+	+	-
utterance	+	+	+	+

The differences are not always clear-cut. Among *events* might be included actions like slamming the door, which can certainly be communicative. *Vocals* include coughing and sneezing, which are usually involuntary noises. Equally, the distinction between utterances and vocals is not always clear, although for many analytic purposes it will be convenient to regard them as distinct. Individual scholars may differ in the way borderlines are drawn and should declare their definitions in the <editorialDecl> element of the header (see 5.3.3 *The Editorial Practices Declaration*).

The following short extract exemplifies several of these elements. It is recoded from a text originally transcribed in the CHILDES format.[91] Each utterance is encoded using a <u> element (see section 11.2.1 *Utterances*). Pauses marked by the transcriber are indicated using the <pause> element (see section 11.2.2 *Pause*). Non-verbal vocal effects such as the child's meowing are indicated either with orthographic transcriptions or with the <vocal> element, and entirely non-linguistic but significant events such as the sound of the toy cat are represented by the <event> elements (see section 11.2.3 *Vocal, Kinesic, Event*).

```
<u who="mar">you never <pause/> take this cat for show&sp;and&sp;tell
<pause/> meow meow</u>
<u who="ros">yeah well I dont want to</u>
<event desc="toy cat has bell in tail which continues to make a tinkling sound"/>
<vocal who="mar" desc="meows"/>
<u who="ros">because it is so old</u>
<u who="mar">how <reg orig="bout">about</reg>
  your&stress; cat <pause/>yours is  new &stress;
  <kinesic desc="shows Father the cat"/> </u>
<u trans="pause" who="fat">thats <pause/> darling</u>
<u who="mar"> <seg>no mine&stress; isnt old</seg>
              <seg>mine is just um a little dirty</seg> </u>
```

This example also uses some elements common to all TEI texts, notably the <reg> tag for editorial regularization. Special purpose entity references have been used to indicate non-separating spaces (&sp) and unusually stressed syllables (&stress); an alternative to the latter might have been to use the core <emph> element. The <seg> element has also been used to segment the last utterance. Further discussion of all of these options is provided in section 11.3 *Elements Defined Elsewhere*.

Contextual information is of particular importance in spoken texts, and should be provided by the TEI header of a text. In general, all of the information in a header is understood to be relevant to the whole of the associated text. The elements <u> and <writing> are however members of the declaring class, and may therefore specify a different context from that of the surrounding elements within a given division or text by means of the decls attribute (see further section 23.3 *Associating Contextual Information with a Text*).

11.2.1 Utterances

Each distinct *utterance* in a spoken text is represented by a <u> element, described as follows:

<u> a stretch of speech usually preceded and followed by silence or by a change of speaker. Attributes include:

- **who** supplies an identifier for the speaker or group of speakers. Its value is the identifier of a <participant> or <participantGrp> element in the TEI header.
  Values Must identify a participant or participant group within the TEI Header
- **trans** (transition) indicates the nature of the transition between this utterance and the previous one.
  Legal values are:
  - smooth this utterance begins without unusual pause or rapidity.
  - latching this utterance begins with a markedly shorter pause than normal.
  - overlap this utterance begins before the previous one has finished.
  - pause this utterance begins after a noticeable pause.

[91] The original is a conversation between two children and their parents, recorded in 1987, and discussed in Brian MacWhinney, *CHAT Manual* ([Pittsburgh]: Dept of Psychology, Carnegie-Mellon University, 1988), pp. 87ff.

Use of the who attribute to associate the utterance with a particular speaker is recommended but not required. Its use implies as a further requirement that all speakers be identified by a <person> or <personGrp> element in the TEI header (see section 23.2.2 *The Participants Description*). Where utterances cannot be attributed with confidence to any particular participant or group of participants, the encoder may choose to define 'participants' such as all or various. For example:

```
<u who='A'>   <!-- utterance by speaker A -->        </u>
<u who='A B'> <!-- utterance by speakers A and B --> </u>
<u who='ALL'> <!-- utterance by speaker group ALL --></u>
```

The trans attribute is provided as a means of characterizing the transition from one utterance to the next at a simpler level of detail than that provided by the temporal alignment mechanism discussed in section 14.5 *Synchronization*. The value specified applies to the transition from the preceding utterance into the utterance bearing the attribute. For example:[92]

```
<u id="a1" who="a">Have you heard the</u>
<u id="b1" trans="latching" who="b">the election results? yes</u>
<u id="a2" trans="pause" who="a">it's a disaster</u>
<u id="b2" trans="overlap" who="b">it's a miracle</u>
```

In this example, utterance B1 latches on to utterance A1, while there is a marked pause between B1 and A2. B2 and A2 overlap, but by an unspecified amount. For ways of providing a more precise indication of the degree of overlap, see section 11.3.2 *Synchronization and Overlap*.

An utterance may contain either running text, or text within which other basic structural elements are nested. Where such nesting occurs, the who attribute is considered to be inherited for the elements <pause>, <vocal>, <shift> and <kinesic>; that is, a pause or shift (etc.) within an utterance is regarded as being produced by that speaker only, while a pause between utterances applies to all speakers.

Occasionally, an utterance may contain other utterances, for example where there is a change in the script associated with it. This may occur when a speaker changes script in mid-utterance. For example:

```
<!-- breakfast table conversation ... -->
<u who="a">Listen to this
    <u decls="s1" who="a">The government is confident, he
    said, that the current economic problems will be
    completely overcome by June</u>
 what nonsense</u>
```

Here speaker A's own utterance contains a second nested utterance, which is read from a newspaper. The decls attribute on the nested utterance is used to indicate that its script is S1, rather than the default. Alternatively, the embedded utterance might be regarded as a new (non-nested) one. It might also be encoded using the <writing> element described in section 11.2.3 *Vocal, Kinesic, Event* below, or the <event> element described in section 11.2.3 *Vocal, Kinesic, Event*, without transcribing the read material:

```
<u who="a">Listen to this <event desc="reads"/>
 what nonsense</u>
```

11.2.2 Pause

The <pause> empty element is used to indicate a perceived pause, either between or within utterances.

<pause> a pause either between or within utterances. Attributes include:

- **who** supplies an identifier for the person or group pausing. Its value is the identifier of a <participant> or <participant.grp> element in the TEI header.
  Values Must identify a participant or participant group within the TEI Header
- **type** categorizes the pause in some respect.
  Values An open list

[92] For the most part, the examples in this chapter use no sentence punctuation except to mark the rising intonation often found in interrogative statements; for further discussion, see section 11.3.3 *Regularization of Word Forms*.

A pause contained by an utterance applies to the speaker of that utterance. A pause between utterances applies to all speakers. The type attribute may be used to categorize the pause, for example as short, medium or long; alternatively the attribute dur may be used to indicate its length more exactly, as in the following example:

```
<u>Okay <pause dur="200"/>U-m<pause dur="75"/>the s the scene opens up
<pause dur="50"/> with <pause dur="20"/> um <pause dur="145"/> you see
a tree okay?</u>
```

If detailed synchronization of pausing with other vocal phenomena is required, the alignment mechanism defined at section 14.5 *Synchronization* and discussed informally below should be used. Note that the trans attribute mentioned in the previous section may also be used to characterize the degree of pausing between (but not within) utterances.

11.2.3 Vocal, Kinesic, Event

These three empty elements are used to indicate the presence of non-transcribed semi-lexical or non-lexical phenomena either between or within utterances.

<vocal> any vocalized but not necessarily lexical phenomenon, for example voiced pauses, non-lexical backchannels, etc. Attributes include:

who supplies an identifier for the vocalist(s). Its value is the identifier of a <participant> or <participant.grp> element in the TEI header.
Values Must identify a participant or participant group within the TEI Header

desc (description) supplies a conventional representation for the phenomenon.
Values a description or representation of the phenomenon chosen from a semi-closed list

iterated (iterated) indicates whether or not the phenomenon is repeated.
Legal values are:
- y the phenomenon is repeated.
- n the phenomenon is atomic.
- u unknown or unmarked.

<kinesic> any communicative phenomenon, not necessarily vocalized, for example a gesture, frown, etc. Attributes include:

who supplies an identifier for the participant performing the gesture. Its value is the identifier of a <participant> or <participant.grp> element in the TEI header.
Values Must identify a participant or participant group within the TEI Header

desc (description) supplies a conventional representation for the phenomenon.
Values a description or representation of the phenomenon chosen from a semi-closed list

iterated (iterated) indicates whether or not the phenomenon is repeated.
Legal values are:
- y the phenomenon is repeated.
- n the phenomenon is atomic.
- u unknown or unmarked.

<event> any phenomenon or occurrence, not necessarily vocalized or communicative, for example incidental noises or other events affecting communication. Attributes include:

who supplies an identifier for the agent of the event described, if any. Its value is the identifier of a <participant> or <participant.grp> element in the TEI header.
Values Must identify a participant or participant group within the TEI Header

desc (description) supplies a conventional representation for the phenomenon.
Values a description or representation of the phenomenon chosen from a semi-closed list

iterated (iterated) indicates whether or not the phenomenon is repeated.
Legal values are:
- y the phenomenon is repeated.
- n the phenomenon is atomic.
- u unknown or unmarked.

The who attribute should be used to specify the person or group responsible for a vocal, kinesic or event which is contained within an utterance, if this differs from that of the enclosing utterance. The attribute must be supplied for a vocal, kinesic or event which is not contained within an utterance.

The iterated attribute may be used to indicate that the vocal, kinesic or event is repeated, for example `laughter` as opposed to `laugh`. These should both be distinguished from `laughing`, where what is being encoded is a shift in voice quality. For this last case, the `<shift>` element discussed in section 11.2.6 *Shifts* should be used.

The desc attribute may be used to supply a conventional representation for the phenomenon, for example:

non-lexical burp, click, cough, exhale, giggle, gulp, inhale, laugh, sneeze, sniff, snort, sob, swallow, throat, yawn

semi-lexical ah, aha, aw, eh, ehm, er, erm, hmm, huh, mm, mmhm, oh, ooh, oops, phew, tsk, uh, uh-huh, uh-uh, um, urgh, yup

Researchers may prefer to regard some semi-lexical phenomena as 'words' within the bounds of the `<u>` element. See further the discussion at section 11.3.3 *Regularization of Word Forms* below. As for all basic categories, the definition should be made clear in the `<encodingDesc>` element of the TEI header.

Some typical examples follow (recoded from J. Maxwell Atkinson and John Heritage: eds., *Transcript notation. Structures of social action: Studies in conversation analysis,* Cambridge University Press, 1984).

```
<u who="jan">This is just delicious</u>
<event desc="telephone rings"/>
<u who="kim">I'll get it</u>
<u who="tom">I used to <vocal desc="cough"/> smoke a lot</u>
<u who="bob"><vocal desc="sniff"/>He thinks he's tough</u>
<vocal who="ann" desc="snorts"/>
```

Note that Ann's snorting could equally well be encoded as follows:

```
<u who="ann">
   <vocal desc="snorts"/>
</u>
```

The extent to which encoding of events or kinesics is included in a transcription will depend entirely on the purpose for which the transcription was made. As elsewhere, this will depend on the particular research agenda and the extent to which their presence is felt to be significant for the interpretation of spoken interactions.

11.2.4 Writing

Written text may also be encountered when speech is transcribed, for example in a television broadcast or cinema performance, or where one participant shows written text to another. The `<writing>` element may be used to distinguish such written elements from the spoken text in which they are embedded.

<writing> a passage of written text revealed to participants in the course of a spoken text. Attributes include:

who (who) supplies an identifier for the participant who reveals or creates the writing, if any. Its value is the identifier of a `<participant>` or `<participant.grp>` element in the TEI header.
Values Must identify a participant or participant group within the TEI Header

gradual (gradual) indicates whether the writing is revealed all at once or gradually.
Legal values are:
- `y` the writing is revealed gradually.
- `n` the writing is revealed all at once.
- `u` unknown or unmarked.

type (Type) categorizes the kind of writing in some way, for example as a subtitle, notice-board etc.
Values Open list

For example, if speaker A in the breakfast table conversation in section 11.2.1 *Utterances* above had simply shown the newspaper passage to her interlocutor instead of reading it, the interaction might have been encoded as follows:

```
<u who="a">look at this</u>
<writing who="a" type="newspaper" gradual="n">
   The government is confident, he said, that the
   current economic problems will be completely overcome
   by June</writing>
<u who="a">what nonsense!</u>
```

11.2.5 Temporal Information

In addition to the global attributes n, id, and lang, utterances, vocals, pauses, kinesics, events and writing elements may all take a common set of attributes providing information about their position in time. For this reason, these elements are regarded as forming a *class*, referred to here as timed. The following attributes are common to all elements in this class:

start indicates the location within a temporal alignment at which this element begins.

end indicates the location within a temporal alignment at which this element ends.

dur indicates the length of this element in time, using either specific units or the units specified on the associated temporal alignment.

Note that if start and end point to <when> elements whose temporal distance from each other is specified in a timeline, then dur is ignored.

The <anchor> element (see 14.4 *Correspondence and Alignment*) may be used as an alternative means of aligning the start and end of timed elements, and is required when the temporal alignment involves points within an element.

11.2.6 Shifts

A common requirement in transcribing spoken language is to mark positions at which a variety of prosodic features change. Many paralinguistic features (pitch, prominence, loudness, etc.) characterize stretches of speech which are not co-extensive with utterances or any of the other units discussed so far. One simple method of encoding such units is simply to mark their boundaries. An empty element called <shift> is provided for this purpose.

<shift> marks the point at which some paralinguistic feature of a series of utterances by any one speaker changes. Attributes include:

feature (feature) a paralinguistic feature.

Legal values are:

- tempo speed of utterance.
- loud loudness.
- pitch pitch range.
- tension tension or stress pattern.
- rhythm rhythmic qualities.
- voice voice quality.

new (new state) specifies the new state of the paralinguistic feature specified.

Values An open list (for an example of possible values, see 11.3.2 *Synchronization and Overlap*)

A <shift> element may appear within an utterance or a segment to mark a significant change in the particular feature defined by its attributes, which is then understood to apply to all subsequent utterances for the same speaker, unless changed by a new shift for the same feature in the same speaker. Intervening utterances by other speakers do not normally carry the same feature. For example:

```
<u who="lb"><shift feature="loud" new="f"/>Elizabeth</u>
<u who="eb">Yes</u>
<u who="lb"><shift feature="loud" new="normal"/>Come and try this <pause/>
     <shift feature="loud" new="ff"/>come on</u>
```

In this example, the word ‘Elizabeth’ is spoken loudly, the words ‘Yes’ and ‘Come and try this’ with normal volume, and the words ‘come on’ very loudly.

The values proposed here for the feature attribute are based on those used by the Survey of English Usage;[93] this list may be revised or supplemented using the methods outlined in section 29 *Modifying and Customizing the TEI DTD*.

The new attribute specifies the new state of the feature following the shift. If no value is specified, it is implied that the feature concerned ceases to be remarkable at this point: the special value `normal` may be specified to have the same effect.

A list of suggested values for each of the features proposed follows:

- tempo
  - **a** allegro (fast)
  - **aa** very fast
  - **acc** accelerando (getting faster)
  - **l** lento (slow)
  - **ll** very slow
  - **rall** rallentando (getting slower)
- loud (for loudness):
  - **f** forte (loud)
  - **ff** very loud
  - **cresc** crescendo (getting louder)
  - **p** piano (soft)
  - **pp** very soft
  - **dimin** diminuendo (getting softer)
- pitch (for pitch range):
  - **high** high pitch-range
  - **low** low pitch-range
  - **wide** wide pitch-range
  - **narrow** narrow pitch-range
  - **asc** ascending
  - **desc** descending
  - **monot** monotonous
  - **scand** scandent, each succeeding syllable higher than the last, generally ending in a falling tone
- tension:
  - **sl** slurred
  - **lax** lax, a little slurred
  - **ten** tense
  - **pr** very precise
  - **st** staccato, every stressed syllable being doubly stressed
  - **leg** legato, every syllable receiving more or less equal stress
- rhythm:
  - **rh** beatable rhythm
  - **arrh** arrhythmic, particularly halting
  - **spr** spiky rising, with markedly higher unstressed syllables

[93] For details see S. Boase, *London-Lund Corpus: Example Text and Transcription Guide* (London: Survey of English Usage, University College London, 1990).

spf spiky falling, with markedly lower unstressed syllables

glr glissando rising, like spiky rising but the unstressed syllables, usually several, also rise in pitch relative to each other

glf glissando falling, like spiky falling but with the unstressed syllables also falling in pitch relative to each other

- voice (for voice quality):

  whisp whisper
  breath breathy
  husk husky
  creak creaky
  fals falsetto
  reson resonant
  giggle unvoiced laugh or giggle
  laugh voiced laugh
  trem tremulous
  sob sobbing
  yawn yawning
  sigh sighing

A full definition of the sense of the values provided for each feature should be provided in the encoding description section of the text header (see section 5.3 *The Encoding Description*).

11.2.7 Formal Definition

The components of the tag set for transcribed speech are formally defined as follows:

```
<!-- 11.2.7: Components of Transcribed Speech-->
<!ELEMENT u %om.RR; (#PCDATA | %m.phrase; |
                  %m.comp.spoken; | %m.Incl;)* >
<!ATTLIST u
      %a.global;
      %a.timed;
      %a.declaring;
      trans (smooth | latching | overlap | pause) "smooth"
      who IDREFS %INHERITED;
      TEIform CDATA 'u'  >
<!ELEMENT pause %om.RO; EMPTY>
<!ATTLIST pause
      %a.global;
      %a.timed;
      type CDATA #IMPLIED
      who IDREF #IMPLIED
      TEIform CDATA 'pause'  >
<!ELEMENT vocal %om.RO; EMPTY>
<!ATTLIST vocal
      %a.global;
      %a.timed;
      who IDREF %INHERITED;
      iterated ( y | n | u ) "n"
      desc CDATA #IMPLIED
      TEIform CDATA 'vocal'  >
<!ELEMENT kinesic %om.RO; EMPTY>
<!ATTLIST kinesic
      %a.global;
      %a.timed;
      who IDREF %INHERITED;
      iterated ( y | n | u ) "n"
      desc CDATA #IMPLIED
      TEIform CDATA 'kinesic'  >
<!ELEMENT event %om.RO; EMPTY>
```

```
<!ATTLIST event
      %a.global;
      %a.timed;
      who IDREF %INHERITED;
      iterated ( y | n | u ) "n"
      desc CDATA #IMPLIED
      TEIform CDATA 'event'  >
<!ELEMENT writing %om.RR; %paraContent;>
<!ATTLIST writing
      %a.global;
      who IDREF %INHERITED;
      type CDATA #IMPLIED
      script IDREF #IMPLIED
      gradual ( y | n | u ) #IMPLIED
      TEIform CDATA 'writing'  >
<!ELEMENT shift %om.RO; EMPTY>
<!ATTLIST shift
      %a.global;
      who IDREF #IMPLIED
      feature (tempo | loud | pitch | tension | rhythm | voice) #REQUIRED
      new CDATA "normal"
      TEIform CDATA 'shift'  >
<!-- end of 11.2.7-->
```

11.3 Elements Defined Elsewhere

This section describes the following features characteristic of spoken texts for which elements are defined elsewhere in these Guidelines:

- segmentation below the utterance level
- synchronization and overlap
- regularization of orthography

The elements discussed here are not provided by the base tag set for spoken texts. Some of them are included in the core tag set available to all TEI documents, but others are contained in the TEI additional tag sets for linking and for analysis respectively. To enable these tag sets, the appropriate parameter entities must be declared in the document type declaration subset, as described in section 3.3 *Invocation of the TEI DTD*. For example, if a transcript using the base tag set defined in this chapter additionally wishes to make use of the <timeline> element, then the following declarations would be necessary:

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

If the complex segmentation elements defined in the additional tag set for analysis were also required, the following declarations would be needed:

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

11.3.1 Segmentation

For some analytic purposes it may be desirable to subdivide the divisions of a spoken text into units smaller than the individual utterance or turn. Segmentation may be performed for a number of different purposes and in terms of a variety of speech phenomena. Common examples include units defined both prosodically (by intonation, pausing, etc.) and syntactically (clauses, phrases, etc.) The term *macrosyntagm* has been used by a number of researchers to define units peculiar to speech transcripts.[94]

[94] The term was apparently first proposed by Bengt Loman and Nils Jørgensen, in *Manual for analys och beskrivning av makrosyntagmer* (Lund: Studentlitteratur, 1971), where it is defined as follows: "A text can be analysed as a sequence of segments which are internally connected by a network of syntactic relations and externally delimited by the absence of such relations with respect to neighbouring segments. Such a segment is a syntactic unit called a macrosyntagm" (trans. S. Johansson).

These Guidelines propose that such analyses be performed in terms of neutrally-named *segments*, represented by the <seg> element, which is discussed more fully in section 14.3 *Blocks, Segments and Anchors*. This element may take a type attribute to specify the kind of segmentation applicable to a particular segment, if more than one is possible in a text. A full definition of the segmentation scheme or schemes used should be provided in the <segmentation> element of the <editorialDecl> element in the TEI header (see 5.3.3 *The Editorial Practices Declaration*).

In the first example below, an utterance has been segmented according to a notion of syntactic completeness not necessarily marked by the speech, although in this case a pause has been recorded between the two sentence-like units. In the second, the segments are defined prosodically (an entity reference &stress; has been used to mark the position immediately following the syllable bearing the primary accent or stress), and may be thought of as 'tone units'.

```
<u who="m1">
   <seg>we went to the pub yesterday</seg>
   <pause/>
   <seg>there was no one there</seg>
</u>
<u who="f1">
   <seg>although its an old ide&stress;a</seg>
   <seg>it hasnt been on the mar&stress;ket very long</seg>
</u>
```

In either case, the <segmentation> element in the header of the text should specify the principles adopted to define the segments marked in this way.

When utterances are segmented end-to-end in the same way as the s-units in written texts, the <s> element discussed in chapter 15 *Simple Analytic Mechanisms* may be used, either as an alternative or in addition to the more general purpose <seg> element. The <s> element is available without formality in all texts, but does not allow segments to nest within each other.

Where segments of different kinds are to be distinguished within the same stretch of speech, the type attribute may be used, as in the following example. The example also shows the use of a user-specified extension to the TEI tag sets, for specifying paraphasia.

```
<u who='T1'>
<seg type='C'>I think </seg>
<seg type='C'>this chap was writing </seg>
<seg type='C'>and he <del type='repeated'>said hello</del> said </seg>
<seg type='M'>hello </seg>
<seg type='C'>and he said </seg>
<seg type='C'>I'm going to a
<paraphasia>gate</paraphasia>
            at twenty past seven </seg>
<seg type='C'>he said </seg>
<seg type='M'>ok </seg>
<seg type='M'>right away </seg>
<seg type='C'>and so <gap desc='unclear' extent='1'/> on they went </seg>
<seg type='C'>and they were <gap desc='unclear' extent='3'/>
            writing there </seg>
</u>
```

In this example, recoded from a corpus of language-impaired speech prepared by Fletcher and Garman, the speaker's utterance has been fully segmented into clausal (type="C") or minor (type="M") units. An additional element <paraphasia> has been used to define a particular characteristic of this corpus for which no element exists in the TEI scheme. See further chapter 29 *Modifying and Customizing the TEI DTD* for a discussion of the way in which this kind of user-defined extension of the TEI scheme may be performed and chapter 3 *Structure of the TEI Document Type Definition* for the mechanisms on which it depends.

This example also uses the core elements <gap> and <del> to mark editorial decisions concerning matter completely omitted from the transcript (because of inaudibility), and words which have been transcribed

but which the transcriber considers may be deleted, respectively. See further section 6.5 *Simple Editorial Changes* for a discussion of these and related elements.

It is often the case that the desired segmentation does not respect utterance boundaries; for example, syntactic units may cross utterance boundaries. For a detailed discussion of this problem, and the various methods proposed by these Guidelines for handling it, see chapter 31 *Multiple Hierarchies*. Methods discussed there include these:

- 'milestone' tags may be used; the special-purpose <shift> tag discussed in section 11.2.6 *Shifts* is an extension of this method
- where several discontinuous segments are to be grouped together to form a syntactic unit (e.g. a phrasal verb with interposed complement), the <join> element may be used
- if SGML is in use, a concurrent DTD may be defined

11.3.2 Synchronization and Overlap

A major difference between spoken and written texts is the importance of the temporal dimension to the former. As a very simple example, consider the following, first as it might be represented in a playscript:

```
Jane: Have you read Vanity Fair?
Stig: Yes
Lou: (nods vigorously)
```

Let us assume that Stig and Lou respond to Jane's question before she has finished asking it — a fairly normal situation in spontaneous speech. The simplest way of representing this *overlap* would be to use the trans attribute previously discussed:

```
<u who="jane">have you read Vanity Fair</u>
<u trans="overlap" who="stig">yes</u>
<!-- ... -->
```

However, this does not allow us to indicate either the extent to which Jane's utterance is overlapped, nor does it show that there are in fact three things which are synchronous: the end of Jane's utterance, Stig's whole utterance, and Lou's kinesic. To overcome these problems, more sophisticated techniques, employing the mechanisms for pointing and alignment discussed in detail in section 14.5 *Synchronization*, are needed. If the additional tag set for linking has been enabled (as described in section 11.3.1 *Segmentation* above), one way to represent the simple example above would be as follows:

```
<u id="u1" who="jane">have you read Vanity <anchor synch="u2 k1" id="a1"/> Fair</u>
<u id="u2" who="stig">yes</u>
<kinesic id="k1" who="lou" iterated="y" desc="nod"/>
```

For a full discussion of this and related mechanisms, section 14.5.2 *Placing Synchronous Events in Time* should be consulted. The rest of the present section, which should be read in conjunction with that more detailed discussion, presents a number of ways in which these mechanisms may be applied to the specific problem of representing temporal alignment, synchrony or overlap in transcribing spoken texts.

In the simple example above, the first utterance (that with identifier u1) contains an <anchor> element, the function of which is simply to mark a point within it. The synch attribute associated with this anchor point specifies the identifiers of the other two elements which are to be synchronized with it: specifically, the second utterance (u2) and the kinesic (k1). Note that one of these elements has content and the other is empty.

This example demonstrates only a way of indicating a point within one utterance at which it can be synchronized with another utterance and a kinesic. For more complex kinds of alignment, involving possibly multiple synchronization points, an additional element is provided, known as a <timeline>. This consists of a series of <when> elements, each representing a point in time, and bearing attributes which indicate its exact temporal position relative to other elements in the same timeline, in addition to the sequencing implied by its position within it.

For example:

```
<timeline unit='dsec' origin='P1'>
   <when id='P1' absolute="12:20:01:01 BST"/>
```

```
   <when id='P2' interval='45' since='P1'/>
   <when id='P6'/>
   <when id='P3' interval='15' since='P6'/>
</timeline>
```

This timeline represents four points in time, named P1, P2, P6, and P3 (as with all attributes named id in the TEI scheme, the names must be unique within the document but have no other significance). P1 is located absolutely, at 12:20:01:01 BST. P2 is 4.5 seconds (i.e. 45 deci-seconds) later than P2 (i.e. at 12:20:46). P6 is at some unspecified time later than P2 and previous to P3 (this is implied by its position within the timeline, as no attribute values have been specified for it). The fourth point, P3, is 1.5 seconds (15 dsec) later than P6.

One or more such timelines may be specified within a spoken text, to suit the encoder's convenience. If more than one is supplied, the origin attribute may be used on each to specify which other <timeline> element it follows. The unit attribute indicates the units used for timings given on <when> elements contained by the alignment map. Alternatively, to avoid the need to specify times explicitly, the interval attribute may be used to indicate that all the <when> elements in a time line are a fixed distance apart.

Three methods are available for aligning points or elements within a spoken text with the points in time defined by the <timeline>:

- The elements to be synchronized may specify the identifier of a <when> element as the value of one of the start, end or synch attributes
- The <when> element may specify the identifiers of all the elements to be synchronized with it using the synch attribute
- A free-standing <link> element may be used to associate the <when> element and the elements synchronized with it by specifying their identifiers as values for its target attribute.

For example, using the timeline given above:

```
<u id="u1" start="p2" end="p3">This is my <anchor synch="p6" id="p6a"/> turn</u>
```

The start of this utterance is aligned with P2 and its end with P3. The transition between the words 'my' and 'turn' occurs at point P6A, which is synchronous with point P6 on the timeline.

The synchronization represented by the preceding examples could equally well be represented as follows:

```
<timeline origin="p1" unit="dsec">
  <when id="p1" absolute="12:20:01:01 BST"/>
  <when synch="u1" id="p2" interval="45" since="p1"/>
  <when synch="x1" id="p6"/>
  <when synch="u1" id="p3" interval="15" since="p6"/>
</timeline>
<!-- ... -->
<u id="u1">This is my <anchor id="x1"/> turn</u>
```

Here, the whole of the object with identifier U1 (the utterance) has been aligned with two different points, P2 and P3. This is interpreted to mean that the utterance spans at least those two points.

Finally, a <linkGrp> may be used as an alternative to the synch attribute:

```
<timeline origin="p1" unit="dsec">
  <when id="p1" absolute="12:20:01:01 BST"/>
  <when id="p2" interval="45" since="p1"/>
  <when id="p6"/>
  <when id="p3" interval="15" since="p6"/>
</timeline><!-- ... -->
<u id="u1">
  <anchor id="u1start"/>
  This is my <anchor id="x1"/> turn
  <anchor id="u1end"/>
</u>
<linkGrp type="synchronous">
  <link targets="u1start p1"/>
  <link targets="u1end p2"/>
```

```
  <link targets="x1 p6"/>
</linkGrp>
```

As a further example of the three possibilities, consider the following dialogue, represented first as it might appear in a conventional playscript:

```
Tom: I used to smoke - -
Bob: (interrupting) You used to smoke?
Tom: (at the same time) a lot more than this.  But I never
     inhaled the smoke
```

A commonly used convention might be to transcribe such a passage as follows:

```
<1> I used to smoke [ a lot more than this ]
<2>                 [ you used to smoke ]
<1> but I never inhaled the smoke
```

Such conventions have the drawback that they are hard to generalize or to extend beyond the very simple case presented here. Their reliance on the accidentals of physical layout may also make them difficult to transport and to process computationally. These Guidelines recommend one of the courses described in what follows:

Where the whole of one or another utterance is to be synchronized, the start and end attributes may be used:

```
<u who="tom">I used to smoke <anchor id="p1"/> a lot more than this
<anchor id="p2"/>but I never inhaled the smoke</u>
<u start="p1" end="p2" who="bob">You used to smoke</u>
```

Note that the second utterance above could equally well be encoded as follows with exactly the same effect:

```
<u who="bob"><anchor synch="p1"/>You used to smoke<anchor synch="p2"/></u>
```

If synchronization with specific timing information is required, a <timeline> must be included:

```
<timeline origin="t1">
   <when id="t1"/>
<when id="t2"/>
</timeline>
<!--  .... -->
<u who="tom">I used to smoke
   <anchor synch="t1"/>a lot more than this
   <anchor synch="t2"/>but I never inhaled the smoke</u>
<u who="bob">
   <anchor synch="t1"/>You used to smoke<anchor synch="t2"/></u>
```

As above, since the whole of Bob's utterance is to be aligned, the start and end attributes may be used as an alternative to the second pair of <anchor> elements:

```
<u start="t1" end="t2" who="bob">You used to smoke</u>
```

An alternative approach is to mark the synchronization by pointing from the <timeline> to the text:

```
<timeline origin="t1">
   <when synch="n1 u2" id="t1"/>
   <when synch="n2 u2" id="t2"/>
</timeline>
<!-- .... -->
<u who="tom">I used to smoke
   <anchor id="n1"/>a lot more than this
   <anchor id="n2"/>but I never inhaled the smoke</u>
<u id="u2" who="bob">You used to smoke</u>
```

To avoid deciding whether to point from the timeline to the text or vice versa, a <linkGrp> may be used:

```
<body>
  <timeline origin='T1'>
     <when id='T1'/>
     <when id='T2'/>
  </timeline>
  <!-- ... -->
  <u who='tom'>I used to smoke
```

```
      <anchor id='N1'/>a lot more than this
      <anchor id='N2'/>but I never inhaled the smoke</u>
   <u id='U2' who='bob'>You used to smoke</u>
   <!-- ... -->
   <linkGrp type='synchronize'>
      <link targets='T1 N1 U2'/>
      <link targets='T2 N2 U2'/>
   </linkGrp>
</body>
```

Note that in each case, although Bob's utterance follows Tom's sequentially in the text, it is aligned temporally with its middle, without any need to disrupt the normal syntax of the text.

As a final example, consider the following exchange, first as it might be represented using a musical-score-like notation, in which points of synchronization are represented by vertical alignment of the text:

```
A : This is |my  |turn
B :         |Balderdash
C :         |No, |it's mine
```

All three speakers are simultaneous at the words 'my', 'Balderdash', and 'No'; speakers A and C are simultaneous at the words 'turn' and 'it's'. This could be encoded as follows, using pointers from the alignment map into the text:

```
<timeline origin="p1">
   <when synch="a1 b1 c1" id="p1"/>
   <when synch="a2 c2" id="p2"/>
</timeline>
<!-- ... -->
<u who="a">this is <anchor id="a1"/> my <anchor id="a2"/> turn</u>
<u id="b1" who="b">balderdash</u>
<u id="c1" who="c"> no <anchor id="c2"/> it's mine</u>
```

11.3.3 Regularization of Word Forms

When speech is transcribed using ordinary orthographic notation, as is customary, some compromise must be made between the sounds produced and conventional orthography. Particularly when dealing with informal, dialectal or other varieties of language, the transcriber will frequently have to decide whether a particular sound is to be treated as a distinct vocabulary item or not. For example, while in a given project 'kinda' may not be worth distinguishing as a vocabulary item from 'kind of', 'isn't' may clearly be worth distinguishing from 'is not'; for some purposes, the regional variant 'isnae' might also be worth distinguishing in the same way.

One rule of thumb might be to allow such variation only where a generally accepted orthographic form exists, for example, in published dictionaries of the language register being encoded; this has the disadvantage that such dictionaries may not exist. Another is to maintain a controlled (but extensible) set of normalized forms for all such words; this has the advantage of enforcing some degree of consistency among different transcribers. Occasionally, as for example when transcribing abbreviations or acronyms, it may be felt necessary to depart from conventional spelling to distinguish between cases where the abbreviation is spelled out letter by letter (e.g. 'B B C' or 'V A T') and where it is pronounced as a single word ('VAT' or 'RADA'). Similar considerations might apply to pronunciation of foreign words (e.g. 'Monsewer' vs. 'Monsieur').

In general, use of punctuation, capitalization, etc., in spoken transcripts should be carefully controlled. It is important to distinguish the transcriber's intuition as to what the punctuation should be from the marking of prosodic features such as pausing, intonation, etc.

Whatever practice is adopted, it is essential that it be clearly and fully documented in the editorial declarations section of the header. It may also be found helpful to include normalized forms of non-conventional spellings within the text, using the elements for simple editorial changes described in section 6.5 *Simple Editorial Changes* (see further section 11.3.5 *Speech Management*).

11.3.4 Prosody

In the absence of conventional punctuation, the marking of prosodic features assumes paramount importance, since these structure and organize the spoken message. Indeed, such prosodic features as points of primary or secondary stress may be represented by specialized punctuation marks. Pauses have already been dealt with in section 11.2.2 *Pause*; while tone units (or intonational phrases) can be indicated by the segmentation tag discussed in section 11.3.1 *Segmentation*. The <shift> element discussed in section 11.2.6 *Shifts* may also be used to encode some prosodic features, for example where all that is required is the ability to record shifts in voice quality.

For more detailed work, involving a detailed phonological transcript including representation of stress and pitch patterns, it is probably best to maintain the prosodic description in parallel with the conventional written transcript, rather than attempt to embed detailed prosodic information within it. The two parallel streams may be aligned with each other and with other streams, for example an acoustic encoding, using the general alignment mechanisms discussed in section 11.2.6 *Shifts*.

Where only a small number of phonetic or phonemic aspects are included in a transcript, it may be convenient to provide a simple set of entity declarations for the particular set of features marked. The entity references in the text may then be redefined to produce simple punctuation marks (as in the following example), or as references to bundles of phonological features, in the same way as is proposed for part of speech tags (see section 15.4 *Linguistic Annotation*).

In the following example, a small set of prosodic features are recorded throughout the transcript using a user-defined entity set such as the following:

```
<!ENTITY lf ".">    <!-- low fall intonation -->
<!ENTITY fr ",">    <!-- fall rise intonation -->
<!ENTITY lr "?">    <!-- low rise intonation -->
<!ENTITY rf "!">    <!-- rise fall intonation -->
<!ENTITY trunc "-"> <!-- truncated syllable -->
<!ENTITY long ":">  <!-- lengthened syllable -->
```

This set of entity definitions may be included directly within the document type declaration subset for the file, or more conveniently along with any other extensions or modifications within the user extensions file defined by the entity TEI.extensions.ent, as discussed in section 3.3 *Invocation of the TEI DTD*. For convenience of reading on the screen, these entity declarations will map the mnemonic entity names used in the text below to a conventional punctuation mark.

```
<div n="Lod E-03" type="exchange">
   <note>C is with a friend</note>
   <u who="cwn">
      <unclear>Excuse me&lf;</unclear> <pause/> You dont have some
      aesthetic&trunc; <pause/> <unclear>specially on early</unclear>
      aesthetics terminology &lr;</u>
   <u who="aj">
      No&lf; <pause/>No&lf; <gap extent="2"/> I'm afraid&lf;</u>
   <u trans="latching" who="cwn">
      No&lr; <unclear>Well</unclear> thanks&lr; <pause/> Oh&trunc;
      <unclear>you couldnt&trunc; can we</unclear> kind of&long;
      <pause/>I mean ask you to order it for us&long;&fr;</u>
   <u trans="latching" who="aj">
      Yes&fr; if you know the title&lf; Yeah&lf;</u>
   <u who="cwn">
      <gap extent="3"/>
      <gap extent="4"/> </u>
   <u who="aj">
      Yes thats fine. <unclear>just as soon as it comes in we'll send
      you a postcard&lf;</unclear> </u>
</div>
```

This example, which is taken from a corpus of bookshop service encounters[95] also demonstrates the use of the <unclear> and <gap> elements discussed in section 6.5 *Simple Editorial Changes*. Where words

[95] Laura Gavioli and Gillian Mansfield, *The Pixi Corpora* (Bologna: Cooperativa Libraria Universitaria Editrice, 1990), p. 74.

are so unclear that only their extent can be recorded, the empty <gap> element may be used; where the encoder can identify the words but wishes to record a degree of uncertainty about their accuracy, the <unclear> element may be used. More flexible and detailed methods of indicating uncertainty are discussed in chapter 17 *Certainty and Responsibility*.

Where a transcript includes many phonetic or phonemic aspects, it will generally be convenient to use a specialized writing system, as defined in chapters 4 *Languages and Character Sets* and 25 *Writing System Declaration*. For representation of phonemic information, the use of the International Phonetic Alphabet is recommended.

11.3.5 Speech Management

Phenomena of *speech management* include disfluencies such as filled and unfilled pauses, interrupted or repeated words, corrections, and reformulations as well as interactional devices asking for or providing feedback. Depending on the importance attached to such features, transcribers may choose to adopt conventionalized representations for them (as discussed in section 11.3.3 *Regularization of Word Forms* above), or to transcribe them using IPA or some other transcription system. To simplify analysis of the lexical features of a speech transcript, it may be felt useful to 'tidy away' many of these disfluencies. Where this policy has been adopted, these Guidelines recommend the use of the tags for simple editorial intervention discussed in section 6.5 *Simple Editorial Changes*, to make explicit the extent of regularization or normalization performed by the transcriber.

For example, false starts, repetition, and truncated words might all be included within a transcript, but marked as editorially deleted, in the following way:

```
<del type="truncation">s</del>see
<del type="repetition">you you</del>you know
<del type="falseStart">it's</del>he's crazy
```

As previously noted, the <gap> element may be used to mark points within a transcript where words have been omitted, for example because they are inaudible:

```
<gap reason="passing truck" extent="approx 10 sylls"/>
```

The <unclear> element may be used to mark words which have been included although the transcriber is unsure of their accuracy:

```
and then <unclear reason="passing truck">marbled queen</unclear>
```

Where a transcriber is believed to have incorrectly identified a word, the elements <corr> or <sic> may be used to indicate both the original and a corrected form of it:

```
<sic corr="SCSI" resp="dd">skuzzy</sic>
<corr sic="skuzzy" resp="AGB">SCSI</corr>
```

As discussed in section 6.5.1 *Correction of Apparent Errors*, the first of these would be appropriate where faithfulness to the transcribers' intuition is paramount, and the second where the editorial interpretation is felt more significant. In either case, the user of the text can perceive the basis of the choice being offered.

11.3.6 Analytic Coding

The recommendations made here only concern the establishment of a basic text. Where a more sophisticated analysis is needed, more sophisticated methods of markup will also be appropriate, for example, using stand-off markup to indicate multiple segmentation of the stream of discourse, or complex alignment of several segments within it. Where additional annotations (sometimes called 'codes' or 'tags') are used to represent such features as linguistic word class (noun, verb, etc.), type of speech act (imperative, concessive, etc.), or information status (theme/rheme, given/new, active/semi-active/new), etc., a selection from the general purpose analytic tools discussed in chapters 14 *Linking, Segmentation, and Alignment*, 15 *Simple Analytic Mechanisms*, and 16 *Feature Structures*, may be used to advantage.