New work item proposal

Transcription of Spoken Language

Coordinating author: Thomas Schmidt ([thomas.schmidt@ids-mannheim.de](mailto:thomas.schmidt@ids-mannheim.de))

Document history:

22-October-2012: 1st draft for ISO/DIN Meeting in Berlin, based on jTEI Paper

18-March-2013: 2nd draft for ISO/DIN Meeting in Potsdam, modifications and additions

according to discussion in Berlin and comments from the French group

# Table of Contents

[Table of Contents](#h.hagawfj907ym)

[Experts](#h.7degqrm52wou)

[Introduction](#h.yysh5i7gq8pk)

[1 Scope](#h.qyn0r4puh88a)

[2 Normative references](#h.yfg9qbw4veac)

[3 Terms and definitions](#h.k8lhntxmgnlv)

[4 General document structure](#h.mxd7fdb9hf8f)

[5 Metadata for the transcription of multimodal corpora](#h.tp11p16vrk1o)

[5.1. fileDesc](#h.g4ls38l1aug)

[5.1.1. publicationStmt](#h.k3tmktpev7ms)

[5.1.2 recordingStmt](#h.u8r4mr3tw2on)

[5.2. profileDesc](#h.7e144u94vjs1)

[5.2.1. particDesc](#h.m6wjj4g2ebgj)

[5.2.2. settingDesc](#h.jkdr948sajnl)

[5.3. encodingDesc](#h.g58o0g7l3can)

[6 Macrostructure](#h.jokwrjg8j1vs)

[6.1 Characterisation in terms of annotation graphs](#h.dfbr5b22grbz)

[6.2 Timeline (<timeline>)](#h.dfbr5b22grbz)

[6.3 Utterances (<u>)](#h.31tvgqh77j5)

[6.4 Dependent annotations (<spanGrp>)](#h.4v12ridd6jc5)

[6.5 Grouping of utterances and dependent annotations (<div>)](#h.80clzsvnajnb)

[6.6 Independent elements outside utterances](#h.dfbr5b22grbz)

[7 Microstructure](#h.cch0wsj32244)

[7.1 Words](#h.lxgbc8fwubin)

[7.1.1 Characterisation](#h.aqf1hmp37zy)

[7.1.2 Representation as <w>](#h.1eypzalj545n)

[7.1.3 Further constraints](#h.7kfy6iwvrix9)

[7.1.4 Examples](#h.sdziubp2fac9)

[7.2 Pauses](#h.z3o3yr42f496)

[7.2.1 Characterisation](#h.x2cshg87ru4t)

[7.2.2 Representation as <pause>](#h.33gkhoe4mvlj)

[7.2.3 Further constraints](#h.b8a9llpufup4)

[7.2.4 Examples](#h.1yclhttdt8hs)

[7.3 Audible non-speech events](#h.bbkr4lki56qj)

[7.3.1 Characterisation](#h.bplr5t6cwkf)

[7.3.2 Representation as <vocal>, <kinesic> or <incident>,](#h.jk6li6g4qn6v)

[7.3.3 Further constraints](#h.9lqsmgrbsjcc)

[7.3.4 Examples](#h.o0t7l0xrs9ed)

[7.4 Punctuation](#h.vdzrkhs7l0s2)

[7.4.1 Characterisation](#h.2xy9getlagg5)

[7.4.2 Representation as <c> (or <pc>?)](#h.wbf984n48hgc)

[7.4.3 Further constraints](#h.h0bz0y1zdq7z)

[7.4.4 Examples](#h.bm4hbmm13f0v)

[7.5 Uncertainty and incomprehensible passages](#h.19xwnk4qpcrs)

[7.5.1 Characterisation](#h.zigkq2ao9brr)

[7.5.2 Representation as <unclear>](#h.9qjwz3kvltva)

[7.5.3 Further constraints](#h.3qi69r7pwi08)

[7.5.4 Examples](#h.8jyhcow71dsl)

[7.6 Units above the word and below the <u> level](#h.gvsavu68whzo)

[7.6.1 Characterisation](#h.sv2zrql6ctw5)

[7.6.2 Representation as <seg>](#h.87orfhd8c8ry)

[7.6.3 Further constraints](#h.vyllcplp9hbj)

[7.6.4 Examples](#h.1u333zjazr4s)

[8 Bibliographical reference](#h.mr3zkmijs0p1)

[9 Annex](#h.wn7z4lufnxr7)

[10 Annex - fully encoded example](#h.yaezeltxcvgx)

[11 Annex(es)](#h.y08vdpyda3nq)

[12 Annex(es)](#h.t304ji4idlqo)

[Minutes ISO/DIN Meeting Berlin, 22-October-2012](#h.lw3w0tgd5m6q)

[Comments Carol Etienne (by e-mail 11-February-2013)](#h.lidhcauvwusg)

# 

# 

# Experts

Core Group:

* Lou Burnard
* Bertrand Gaiffe
* Laurent Romary
* Thomas Schmidt
* Andreas Witt

EIT / MMI Group:

* Nadia Mana from FBK (Trento, Italy)
* Tatjana Scheffler (DFKI, Germany)
* Khiet Truong (Univ of Twente)
* Benjamin Weiss (TU Berlin)
* Mathias Wilhelm (DAI Labor)
* Felix Burkhardt (Deutsche Telekom's Innovation Lab)

Working group 'Interopérabilité' of the Consortium ‘Corpus Oraux et multimodaux (IRCOM)’

* Coordinators: Christophe Parisse, Carole Etienne, Bernard Bel.
* Participants : Sarra El Ayari, Nicolas Ballier, Christophe Benzitoun, Philippe Blache, Annelies Braffort, Christian Chanard, Dominique Fohr, Séverine Guillaume, Clément Planck.

Tool developers (not yet contacted):

* Michael Kipp (FH Augsburg, developer of ANVIL)
* Han Sloetjes (MPI Nijmegen, developer of ELAN)
* Kai Wörner (Hamburg, co-developer of EXMARaLDA)
* (To be contacted: Praat, Paul Boersms, Transcriber: Edouard Geoffrois

Other experts (not yet contacted):

* Ulrike Gut (Münster, Corpus Phonology Network)
* Florian Schiel, Christoph Draxler (München, Bavarian Archive for Sound Signals)

# 

# 

# Introduction

Context, multiple tools, multiple formats, multiple transcription conventions cf. Intro TS-jTEI

Complementarity with specific annotation levels MAF, SynAF, SemAF

Joint ISO-TEI initiative (cf. MoU)

Legacy data vs. new data/tools

# 1 Scope

This standard aims at enabling and facilitating the interchange of transcriptions of spoken language between different computational tools and environments for creating, editing, publishing and exploiting such data. Typically, transcription of spoken language here means an orthography-based transcription of verbal activity as recorded in an audio or video recording of a natural interaction. The description of activity in other modalities (e.g. gestures, mimics) may be part of a spoken language transcription, but the standard departs from the assumption that the verbal dimension is the primary focus of a spoken language transcription.

The standard takes into account data models and encoding practices supported by widely used transcription editors (such as Transcriber, ELAN, EXMARaLDA etc.) and aims at being compatible with the formats produced by these tools.

# 2 Normative references

XML

Maybe only ISO documents to be cited - TEI P5 (version?)

# 3 Terms and definitions

transcription, annotation, transcription convention, transcription tool, tier, …

# 4 General document structure

One, multiple documents? – TEICorpus?

primary - subordinate

tier

metadata

# 5 Metadata

The TEI Guidelines formulates extensive suggestions for encoding metadata inside different subsections of the <teiHeader> element. In the following, only those pieces of metadata are addressed which are crucial for ensuring the interpretability and exchangeability of spoken language transcriptions in general. This does not preclude the possibility of or necessity for encoding further metadata inside the <teiHeader> element.

## 5.1. fileDesc

Text

### 5.1.1. publicationStmt

The <publicationStmt> element inside the <fileDesc> section of the <teiHeader> should be used to record information about access rights and contact information for the transcription in question.

|  |
| --- |
| <publicationStmt>  <authority>Hamburger Zentrum für Sprachkorpora</authority>  <availability>Available free for research and teaching purposes</availability>  <distributor>Hamburger Zentrum für Sprachkorpora</distributor>  <address>  <street>Max Brauer-Allee 60</street>  <postCode>22765</postCode>  <placeName>Hamburg</placeName>  <country>Germany</country>  </address>  </publicationStmt> |

### 5.1.2 recordingStmt

The <publicationStmt> element inside the <fileDesc> section of the <teiHeader> should be used to record information about the transcribed recording(s). A <media> element should be used to refer to the corresponding digital file. Where two or more files are derived from the same master recording, these should be represented as different <media> elements inside the same <recording> element, rather than as different <recording> elements.

Comment CE: <recording> and <equipment> to define media files quality, a set of extracts of the same record in different format or quality, available streaming

|  |
| --- |
| <recordingStmt>  <recording type="video">  <!-- element from TEI P5, but not allowed there as a child of recording -->  <media mimeType="video/mpeg" url="Beckhams.mpg"/>  <media mimeType="audio/wav" url="Beckhams.wav"/>  <broadcast>  <ab>Parkinson Talkshow on BBC, broadcast on 02 November 2007</ab>  </broadcast>  <!-- information about the equipment used for creating the recording -->  <!-- where recordings are made by the researcher, this would be -->  <!-- place to specify the recording equipment (e.g. Camcorder) -->  <equipment>  <ab>Video excerpt downloaded from YouTube with aTube-Catcher, converted  into MPG format with Adobe Premiere</ab>  <ab>Audio extracted from video with Audacity 1.3 beta</ab>  </equipment>  </recording>  </recordingStmt> |

## 5.2. profileDesc

### 5.2.1. particDesc

The participants of the transcribed interaction should be described in <person> elements inside the <particDesc> section of a <profilDesc> element. Using the <abbr> element to define an abbreviated code for the respective participant can be crucial for many processing purposes. <u> elements inside the body of the transcription refer to the @id attribute of a <person> element which should therefore always be provided.

Comment TS: The question came up what to do when it is necessary to identify speakers across individual documents (i.e. where one and the same speaker figures in different transcriptions of the same corpus). Is it a reasonable suggestions to ensure that, in these cases, ids of a given speaker should be identical across documents?

|  |
| --- |
| <particDesc>  <person xml:id="SPK0" sex="1">  <persName>  <abbr>DS</abbr>  </persName>  <!-- possibly further descriptive elements -->  </person>  <person xml:id="SPK1" sex="0">  <persName>  <abbr>FB</abbr>  </persName>  </person>  </particDesc> |

### 5.2.2. settingDesc

The <settingDesc> elements should be used to provide information about the spatial organization, artifacts etc. of the transcribed interaction.

Comment CE: <setting> and <activity> to describe spatial organization, artefacts, participants tasks (no link with their usual occupation defined in <person>)

|  |
| --- |
| <settingDesc>  <place>BBC studio London</place>  <setting>  <activity>Talkshow host Michael Parkinson interviewing David and Victoria  Beckham about their relationship</activity>  </setting>  </settingDesc> |

## 5.3. encodingDesc

The <encodingDesc> element should be used to record information about the tool which created the transcription and about the convention which was used in transcribing the data.

Comment TS: Does TEI provide a more formalized way of referring to transcription conventions? I think it would be better to use something more structured and machine-readable for this purpose.

|  |
| --- |
| <encodingDesc>  <appInfo>  <!-- information about the application with which -->  <!-- the transcription was created -->  <application ident="EXMARaLDA" version="1.5.1">  <label>EXMARaLDA Partitur-Editor</label>  <desc>Transcription Tool providing a TEI Export</desc>  </application>  </appInfo>  <!-- information about the transcription convention used -->  <!-- might be better to define a more machine-friendly tag -->  <!-- for this, e.g. in analogy to appInfo -->  <!-- <transcriptionConvention ident="HIAT" version="2004" -->  <p>Orthographic transcription according to HIAT</p>  </encodingDesc> |

# 6 Macrostructure

## 6.1 Characterisation in terms of annotation graphs

(see also transcription graphs in Schmidt 2005 and the discussion in Schmidt et al. 2009)

Question TS: Is this section useful? Does it make sense in a standardisation document to relate the markup standardisation to a more abstract algebraic framework? If yes, I will elaborate this section. If no, I will remove it.

Annotation graphs as in Bird/Liberman (2001) plus:

* graph must be *fully anchored*
* nodes must be *fully ordered*
* arcs can (but need not) be assigned to one member of a set of *speakers*
* each arc must be assigned to exactly one member of a set of *categories*
* exactly one category is assigned to *type* T(ranscription), the remaining categories are assigned to either type A(nnotation) or D(escription)
* arcs assigned to categories of types T or A must also be assigned to a speaker
* partition of arcs:
  + all arcs assigned to the same speaker and the category of type ‘T’ → main tier for that speaker, constraint: no overlapping arcs (typically: orthographic transcription)
  + all arcs assigned to the same speaker and one of the categories of type ‘A’ → dependent tier(s) for that speaker, constraint: corresponding arcs in main tier of the same speaker (typically: linguistic annotation)
  + all arcs assigned to the same speaker and one of the categories of type ‘D’ → independent secondary tier(s) for that speaker (typically: descriptions of non-verbal behaviour)
  + a (maximally?) contiguous set of arcs in a main tier → segment chain → <u> element

## 6.2 Timeline (<timeline>)

Comment/Question TS: change in revision #2: use relative offsets (interval) instead of absolute. I am still not sure how to define the first <when> element which represents the start of the recording.

<when> elements inside a <timeline> element should be used to define points in the recording which are then referred to by @start, @end and @synch attributes of other elements of the trancription to represent its temporal strcuture. It is therefore obligatory to provide an @id attribute for each <when> element. <when> elements must be in the same order as the timepoints they refer to. Specifying an absolute offset into the recording via an @interval attribute is optional, but very useful for many processing purposes.

|  |
| --- |
| <timeline unit="s" origin="#T0">  <when xml:id="T0" absolute="00:00:00"/>  <when xml:id="T1" interval="02.13" since="#T0"/>  <when xml:id="T2" interval="03.74" since="#T0"/>  <when xml:id="T3" interval="04.71" since="#T0"/>  <when xml:id="T4" interval="unknown" since="#T0"/>  <when xml:id="T5" interval="08.53" since="#T0"/>  <when xml:id="T6" interval="11.36" since="#T0"/>  <when xml:id="T7" interval="13.91" since="#T0"/>  <when xml:id="T8" interval="15.47" since="#T0"/>  <!-- [...] more when elements -->  </timeline> |

## 6.3 Utterances (<u>)

The <u> element is the fundamental unit of organization for a transcription, roughly comparable to a paragraph (<p> element) in a written document. It corresponds to a contiguous stretch of speech of a single speaker. A more exact definition and delimitation of a <u> is not in the scope of this document; the TEI definition characterising a <u> as “often preceded by a silence or a change of speaker” should be viewed as a suggestion only, i.e. it is permissible to use a more refined definition for a <u>.

<u> elements must be assigned to a single speaker by providing a value for the @who attribute which points to the @id of a <person> element defined in the header.

<u> elements must be assigned to the timeline by providing values for the @start and @end attributes pointing to the the @id of a <when> element defined in the timeline.

Further temporal structure can be recorded by inserting <anchor> elements at appropriate places inside the content of a <u> element.

The preferred mechanism for representing overlap is to encode it implicitly through the appropriate use of @start and @end attributes and <anchor> elements. Other mechanisms, such as a @trans=’overlap’ attribute for the <u> element can be used instead or in addition, but can not be processed in an appropriate manner by many of the widely used annotation tools.

Comment CE: In a record including a lot of overlaps (800 overlaps in an hour in some settings), is it possible to define two scales to avoid splitting timeline :

. a timeline for timecode with @absolute as 6.2 example and in utterances an <anchor synch="#Tnn">

- an anchor identified with <anchor xml:id="XX"> at the starting point of an overlapped segment assigned to a speaker A and a corresponding anchor tag <anchor synch="#XX"> at the starting point of the overlapping segment assigned to a speaker B to synchronize or link the two segments without time identification if we haven't time-aligned each overlap

|  |
| --- |
| <!-- u with start and end attributes only (minimal temporal strcuture) -->  <u who="#SPK1" start="#T0" end="#T1">Good morning! </u>  <!-- u with embedded anchor elements (additional temporal structure) -->  <u who="#SPK0" start="#T1" end="#T4">  Okay. <anchor synch="#T2"/>Très bien, <anchor synch="#T3"/>très bien.  </u>  <!-- two u’s with partial overlap →  <u who="#SPK0" start="#T0" end="#T2">Do not <anchor synch="#T1"/> interrupt me!</u>  <u who="#SPK1" start="#T1" end="#T3">Sorry, <anchor synch="#T2"/> mate!</u> |

## 6.4 Dependent annotations (<spanGrp>)

Whereas <u> contains the basic orthographic transcription, <span> elements should be used to represent additional annotations (e.g. POS tagging, prosodic annotation, translation) on that basic transcription. Annotations of the same type should be grouped in a <spanGrp> element with a @type attribute specifying the annotation level.

The reference of the annotation in question can be specified using @to and @from attributes in two ways:

* the values of @to and @from can point to the @id attributes of other elements (e.g. a <u> or a <w>) of the transcription
* the values of @to and @from can point to the @id attributes of <when> elements from the timeline

If the latter mechanism is used, <spanGrp> elements must be grouped with the <u> element they refer to by using an <annotatedU> element (see below).

Comment TS: My original proposal was to always group spanGrp with the <u> element they belong to. In the Berlin meeting, it was felt that this should be made optional. However, it now occurs to me that, if there is no such grouping, providing only @from and @to attributes pointing to the timeline may not be enough to identify the annotated part of the transcription, because (in the case of simultaneous speech) there may be more than one candidate. An alternative to making the grouping obligatory would be to provide a @who element in <spanGrp> or <span>, but this is not allowed according to the guidelines.

The use of further annotation techniques (e.g. via feature structures) is not precluded, but not in the scope of this document.

|  |
| --- |
| <!-- part-of -speech annotations -->  <!-- using a reference to ids of <w> elements -->  <spanGrp type="pos">  <span from="#w148" to="#w148">PersPron</span>  </spanGrp>  <!-- annotations from a sup (=suprasegmentals) tier -->  <!-- using a reference to the timeline -->  <spanGrp type="sup">  <span from="#T2" to="#T4">faster</span>  </spanGrp>  <!-- annotations from an en (=English translation) tier -->  <!-- using a reference to the timeline -->  <spanGrp type="en">  <span from="#T1" to="#T2">Okay. </span>  <span from="#T2" to="#T4">Very good, very good.</span>  </spanGrp> |

## 6.5 Grouping of utterances and dependent annotations (<annotatedU>)

<u> elements and the annotations referring to it can be grouped under a <annotatedU> element. This has the advantage of creating “local” annotated environments each (succession) of which can be treated as an independent transcription in its own right (“tesselation” of the transcription document)

|  |
| --- |
| <annotatedU>  <!-- the transcribed text from the primary tier -->  <u who="#SPK0">  <!-- [...] (see above) -->  </u>  <!-- additional annotations from a sup (=suprasegmentals) tier -->  <spanGrp type="sup">  <!-- [...] (see above) -->  </spanGrp>  <!-- additional annotations from an en (=English translation) tier -->  <spanGrp type="en">  <!-- [...] (see above) -->  </spanGrp>  </div> |

## 6.6 Independent elements outside utterances

<pause> and <incident> elements should be used to represent non-verbal phenomena which cannot be attributed to a speaker. In the document, these elements appear on the same hierarchical level as <annotatedU> elements. In order to fit them into the temporal structure they must have @start and @end attributes pointing to the timeline.

|  |
| --- |
| <annotatedU>  <!-- [...] u and spanGrp elements, see above -->  </annotatedU>  <!-- an incident from a nv (=nonverbal) tier describing nonverbal behaviour -->  <incident who="#SPK0" type="nv" start="#T3" end="#T6">  <desc>right hand raised</desc>  </incident>  <annotatedU>  <!-- [...] u and spanGrp elements, see above -->  </annotatedU> |

# 7 Microstructure

## 7.1 Words

### 7.1.1 Characterisation

Most transcription conventions do not provide an exact and comprehensive definition of the unit *word*. Rather, they depart from the word definition of standard written orthography and supplement this with rules for a selected number of special cases (e.g. words specific to spoken language like ‘ehm’, abbreviations, spellings etc.). A more precise definition should and need not be attempted in this document - the decision of what is to be treated (i.e. marked up) as a word can be left to the individual transcription system. → Simple Analytic Mechanisms

Some transcription conventions have methods for representing an *assimilation*, i.e. a blending, of two or more words into one. Also common are methods for characterising a word as *incomplete* (cut-off or initialising a self repair sequence). Where certain syllables are pronounced more lengthened than in standard pronunciation, some transcription conventions mark this syllable accordingly.

pause inside w

trunc element (speech management) <gap>???

### 7.1.2 Representation as <w>

Words (as defined by the transcription system used) should be encoded as <w> elements underneath a <u> element. In order to make words referenceable in annotations, the use of an @id attribute is recommended. A @type attribute can be used to represent special features of a word, especially when the corresponding distinction is an integral part of the transcription system. For instance, the following distinctions made by several widely used transcription systems can be encoded in a @type attribute of a <w> element:

* @type=’assimilated’ on the later word for assimilated words
* @type=’truncated’ for truncated words
* @type=’repetition’ for repeated words
* ???here new ‘restart’???

Alternatively, this kind of information can be recorded as an annotation in a <span> element (see above).

Beneath the level of words, many transcription conventions contain instructions for marking a given syllable as accentuated or a given sound as lengthened. To delimit such units below the word level, a <seg> element can be used and either be characterised as an accentuated syllable or lengthened sound by an appropriate @type attribute or, again, by referencing the <seg> element from a <span> via its @id attribute.

Pauses inside words can occur and should be encoded as <pause> elements as described below.

### 

### 7.1.3 Further constraints

Since overlaps starting or ending inside a word occur, <w> must allow <anchor> as a child.

### 7.1.4 Examples

|  |
| --- |
| <!-- an utterance divided into words -->  <u>  <!-- [...] -->  <w xml:id="w148">I</w>  <w xml:id="w149">am</w>  <w xml:id="w150">very</w>  <w xml:id="w151">much</w>  <w xml:id="w152">aware</w>  <w xml:id="w153">of</w>  <w xml:id="w154">that</w>  </u>  <!-- a word with a time anchor inside -->  <w xml:id="w152">a<anchor synch="#T3"/>ware</w>  <!-- a word with an accentuated syllable -->  <w xml:id="w152"><seg xml:id="seg152a"/>awe</seg>some</w>  <span from="seg152a" to="seg152a">accentuated</span> |

## 7.2 Pauses

### 7.2.1 Characterisation

Most transcription systems distinguish measured pauses and typed pauses, the latter being typically divided into four types like ‘micro’, ‘short’, ‘medium’ and ‘long’. Pauses can occur outside speaker’s utterances and between or inside words attributed to a <u> element.

### 7.2.2 Representation as <pause>

All pauses should be represented as <pause> elements. For measured pauses, the length should be provided in a @dur attribute, for typed pauses, the type should be provided in a @type attribute.

### 7.2.3 Further constraints

Temporal plausibility - length of measured pause should not contradict temporal information as encoded in timeline references

### 7.2.4 Examples

|  |
| --- |
| <!-- measured pause -->  <pause dur="PT1.2S"/>  <!-- typed pause -->  <pause type="micro"/>  <!-- measured pause outside <u>, with its own start and end attributes -->  <pause dur="PT0.61S" start="TLI\_10" end="TLI\_11"/> |

## 7.3 Audible non-speech events

### 7.3.1 Characterisation

breathing, laughing, coughing etc.

noises not attributable to a speaker (e.g. telephone rings)

visible non-speech events (e.g. nods)? → multimodal, not in the scope of this document?

### 7.3.2 Representation as <vocal>, <kinesic> or <incident>,

### 7.3.3 Further constraints

order of contributions and elements on the same level:

1. ascending by position of @start in <timeline>
2. descending by position of @end in <timeline> (when @start are equal)
3. ascending by position of @who in <particDesc> (when @start and @end are equal)

### 7.3.4 Examples

|  |
| --- |
| <!-- coughing encoded as incident between words and anchors of a u -->  <u>  <anchor synch="#T4"/>  <w>dépend</w>  **<incident>**  **<desc>cough</desc>**  **</incident>**  <anchor synch="#T5"/>  <w>un</w>  <w>peu</w>  <anchor synch="#T6"/>  </seg>  </u> |

## 7.4 Punctuation

### 7.4.1 Characterisation

punctuation according to orthography (e.g. period at the end of a grammatical sentence or comma introducing a subordinate clause in German) vs. puncutation representing properties of speech (e.g. comma representing a rising tone movement, slash representing the initialisation of a repair sequence) - the latter should ideally be mapped to a corresponding annotation element, but this is not always feasible

N.B.: In contrast to other elements, punctuation does usually not directly correspond to some event occuring in time → no start and end attribute possible

### 7.4.2 Representation as <pc>

### 7.4.3 Further constraints

### 7.4.4 Examples

|  |
| --- |
| <!-- punctuation represented as pc elements -->  <seg function="utterance">  <w xml:id="w330">No</w>  <pc>,</pc>  <w xml:id="w331">I</w>  <w xml:id="w332">mean</w>  <w xml:id="w333">I</w>  <w xml:id="w334">knew</w>  <pc>.</pc>  </seg> |

## 7.5 Uncertainty and incomprehensible passages

### 7.5.1 Characterisation

### 7.5.2 Representation as <unclear>

<unclear> with PCDATA represents uncertainty

alternatives should be each marked as unclear and grouped with a <choice> element

completely uncomprehensible passages should be represented by a <gap> element. <gap reason=”passing truck”/>

### 7.5.3 Further constraints

### 7.5.4 Examples

|  |
| --- |
| <!-- uncertain passage -->  <w>you</w>  <unclear>  <w>should</w>  </unclear>  <w>let</w>  <!-- uncertain passage with alternatives -->  <w>you</w>  <choice>  <unclear>  <w>should</w>  </unclear>  <unclear>  <w>could</w>  </unclear>  </choice>  <w>let</w> |

# 7.6 Units above the word and below the <u> level

### 7.6.1 Characterisation

Speaker’s contributions can often be subdivided in chunks comprising more than one word and/or pauses and/or non-audible speech events. These are the “sentence equivalents” of spoken language.

How these chunks are defined, distinguished and delimited varies greatly between different conventions (and is hotly debated). Two popular approaches: use pragmatic/syntactic criteria (--> notion of an utterance, e.g. HIAT/CHAT) vs. use prosodic critertia (--> notion of an Intonation phrases, e.g. GAT, DT)

### 7.6.2 Representation as <seg>

### 7.6.3 Further constraints

### 7.6.4 Examples

|  |
| --- |
| <!-- u divided into two seg elements -->  <div>  <u who="#SPK0">  <anchor synch="#T40"/>  <seg function="utterance" type="declarative">  <w xml:id="w319">And</w>  <incident>  <desc>unv.</desc>  </incident>  <w xml:id="w320">disappointed</w>  <w xml:id="w321">when</w>  <w xml:id="w322">you</w>  <w xml:id="w323">got</w>  <w xml:id="w324">  to  <anchor synch="T41"/>  gether  </w>  </seg>  <anchor synch="T42"/>  <seg function="utterance" type="interrogative">  <incident>  <desc>unv.</desc>  </incident>  <w xml:id="w325">you</w>  <pc>,</pc>  <w xml:id="w326">Victoria</w>  </seg>  <anchor synch="#T43"/>  </u>  </div> |

# 

# 8 Bibliographical reference

**Bird, S. & Liberman, M. (2001).** A formal framework for linguistic annotation. In: Speech Communication (33), 23-60.

**Romary Laurent, Witt Andreas (2012)**. Data formats for phonological corpora. Handbook of Corpus Phonology Oxford University Press (Ed.) [<http://hal.inria.fr/inria-00630289>]

**Schmidt, Thomas (2005)** Computergestützte Transkription - Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln. Frankfurt a. M.: Peter Lang.

**Schmidt, Thomas (2011).** A TEI-based Approach to Standardising Spoken Language Transcription. Journal of the Text Encoding Initiative [Online], Issue 1 | June 2011, URL : [[http://jtei.revues.org/142]](http://jtei.revues.org/142) ; DOI : 10.4000/jtei.142

**Schmidt, T.; Duncan, S.; Ehmer, O.; Hoyt, J.; Kipp, M.; Magnusson, M.; Rose, T. & Sloetjes, H. (2009)** An Exchange Format for Multimodal Annotations. In: Michael Kipp, Jean-Claude Martin, P. P. & Heylen, D. (eds.): Multimodal Corpora, Lecture Notes in Computer Science 207-221. Springer.

**Schmidt, Thomas, Elenius, Kjell & Trilsbeek, Paul (2010).** Multimedia Corpora (Media encoding and annotation). Draft submitted to CLARIN WG 5.7. as input to CLARIN deliverable D5.C-3 "Interoperability and Standards" [http://www.clarin.eu/system/files/clarin-deliverable-D5C3\_v1\_5-finaldraft.pdf]

see also: <http://www1.uni-hamburg.de/exmaralda/files/CLARIN_Standards.pdf>

# 

# 

# 9 Annex

ODD spec

# 10 Annex - fully encoded example

# 11 Annex(es)

Mappings - macrostructure

# 12 Annex(es)

Mappings - microstructure

## Minutes ISO/DIN Meeting Berlin, 22-October-2012

(Lou Burnard, see also <http://www.tei-c.org/Activities/Council/Working/tcw25.xml>)

EIT MMI Meeting, Berlin 22 oct 2012  
  
As noted at the last FTF, Laurent Romary in his capacity as ISO TC7 WG3 chair has proposed a new ISO/TEI joint activity in the area of speech transcription, which comes with the slightly obscure label of EIT MMI: the last part of which is short for “multimodal interaction”, although it seems the activity is really only concerned with speech transcription. I was invited to attend the third EIT MMI workshop, held at the DIN's offices in Berlin.

Prime movers in the activity, apart from Laurent, appear to be Thomas Schmidt and Andreas Witt from the Institut fur Deutsche Sprache in Mannheim, but a number of other European research labs, mostly concerned with analysis of corpora of human computer interaction, were also represented; specifically: Nadia Mana from FBK (Trento, Italy); Tatjana Scheffler (DFKI, Germany); Khiet Truong (Univ of Twente) ; Benjamin Weiss (TU Berlin); Mathias Wilhelm (DAI Labor); Bertrand Gaiffe (ATILF, Nancy). This being an ISO activity, the real world of commerce and industry was also represented by Felix Burkhardt from Deutsche Telekom's Innovation Lab.  
Related ISO activity mentioned by Laurent included the work on Discourse Relations led by Harry Bunt, and the long-awaited MAF (morpho-syntactic annotation framework) which are both due to appear Real Soon Now. A quick tour de table confirmed my impression that most of the attendees were primarily researchers in Human Computer Interaction with little direct experience of the construction or encoding of spoken corpora, but Thomas Schmidt more than made up for that.

The main business of the day was to go through his preliminary draft working document, the objective of which is to confer ISO authority on a subset of the existing TEI proposals for spoken text transcription, with some possible modification. The underlying work is well described in Schmidt's recent excellent article in TEIJ, so I won't repeat it: essentially, it consists of a close look at the majority of transcription formats used by the relevant research community/ies and tools, a synthesis of what they have in common, and suggestions of how that synthesis maps to TEI. This is to a large extent motivated by concerns about preservation and migration of data in “legacy” formats.  
  
The discussion began by establishing boundaries: despite my proposal to the contrary, it seems there was little appetite to extend the work into the area of truly multimodal transcriptions, which was still generally felt to be insufficiently understood for a practice-based standard to be appropriate. Concern was expressed that we should not make ad hoc premature suggestions. So the document really only concerns transcribed   
speech. There was no disagreement with the general approach which is to distinguish a small number of macro-structural features provide guidelines about how to mark up specific units of analysis at the micro-structural level, using a subset of the TEI. I was also much cheered by two further remarks he made the graph-based “annotation framework” formalisation proposed by Bird and Liberman was theoretically complete but so generic as to be practically useless (I paraphrase) at the micro level, everything you need is there in the TEI (I quote)  
  
Discussion focussed on the following points raised by the working document:

* **Tiers:** Many existing tools organise transcriptions into “tiers” of annotation.   
  These seem to be purely technical artefacts, which can be addressed more   
  exactly by used of XML markup. Unlike “levels” of annotation, they have   
  no semantics. It's doubtful that we need a <tier> element.
* **Metadata -1:** How many of the (very rich) TEI proposals should be included, or mentioned? And how should the three things Thomas had found missing be supplied? I suggested that <appinfo> was an appropriate way to record information about the transcription tool used; that the definition of the transcription system used belonged in the <encodingDesc>; and agreed that there was nothing specifically provided for recording pointers or links to the original video or audio transcribed. In the meeting, I   
  speculated that maybe there was scope for extending (or misusing) <facsimile> for this last purpose; another possibility which occurs to me as I type these notes is that one could also extend <recordingDesc>.
* **Timing:** The timeline is fundamental to the macrostructure of a transcript. Thomas' examples all used absolute times for its <when>s, but I suggested that relative ones might be easier. The document ordering both of <when>s and of transcribed speech should reflect the temporal order as far as possible; this would allegedly facilitate interoperability
* **Metadata-2:** What metadata was needed, required, recommended for the description of participants? (@sex raised its ugly head here). Could we use <person> to refer to artificial respondents in MMI experiments? (yes, if they have person-like characteristics; no otherwise).   
  It was noted that almost any personal trait or state might be crucial to the analysis of some corpora. We noted that CMDI now recommended using the ISOCAT data category registry as an independent way of defining metadata terminology; also that ISOCAT was now available within the TEI scheme (though whether it fits into personal metadata I am less sure). There was (I think) general agreement that we'd reference the various options available in the TEI but not incorporate all of them. We agreed that the principles underlying a given transcription should be clearly documented, either in associated articles, in the formal specification for an encoding, or in the header of individual documents.
* **Utterances:** Several people disliked the expanded element name <u> and its definition, for various theoretical reasons. Its definition should be modified to remove the implication that it necessarily followed a silence, though we seemed to agree that a <u> could only contain a stretch of speech from a single speaker.  
  The temporal alignment of a <u> can be indicated either by @start and @end or by nested <anchor/>s : the standard should probably recommend use of one or the other methods but not both. We discussed whether or not the fact that existing tools did not support the (even simpler) use of @trans to indicate overlap should lead us not to recommend it.
* **U-plus:** Thomas wanted some method of associating with a <u> the whole block of   
  annotations made on it (represented as one or more <interpGrp>s). His document suggested using <div> for this purpose. A lighter-weight solution might be to include <interpGrp> within <u>, or to propose a new wrapper <annotatedU> element.
* **Tokenization:** Laurent noted that MAF recommended use of <w> for individual tokens; we didn't need to take a stand on the definition of “word” but could simply refer to MAF. We needed some way of signalling the things that older transcription formats had found important, e.g. words considered incomplete, false starts, repetitions, abbreviations etc. so we needed to choose an appropriate TEI construct for them, even if we thought the concept was not useful or ill-defined. The general purpose <seg> element might be the simplest solution, but some diplomacy would be needed about how to define its application and its possible @type or @function values.

**Conclusions**  
  
This workgroup will probably produce a useful document describing an important use case for the TEI recommendations on spoken language. It is currently a Google Doc which the group has agreed to share with the Council. I undertook to help turn this into an ODD, which could eventually become one of our Exemplars. Work on standardising other aspects of transcribed multimodal interactions probably needs to be deferred to a later stage.

## Comments Carol Etienne (by e-mail 11-February-2013)

Les membres du groupe de travail 'Interopérabilité' de l'IRCOM, IR Corpus Oraux et Multimodaux, sont intéressés par l'utilisation de la TEI pour les corpus oraux, dans les projets regroupant plusieurs "sources" de corpus ou bien en tant que format d'échange de données entre les logiciels d'annotation comme Transcriber, Clan, Elan ou Anvil majoritairement utilisés dans notre communauté.

Les besoins en terme de métadonnées comme en terme de transcription sont différents et plus ou moins complexes suivant les objets d'étude de chaque équipe mais les besoins suivants émergent :

1) des **métadonnées pour les locuteurs** : données sociolinguistiques, relation entre les utilisateurs, langue maternelle et autres langues, langue des parents, situation professionnelle, locuteurs génériques (client/commerçant, hotline, élève/enseignant, ...) La Tei fournit un ensemble important de balises dans <person> qui devraient couvrir la majorité des besoins

2) des **métadonnées pour l'enregistrement audio ou vidéo** : type, qualité, anonymisation, url, ...

La balise <recording> reste à ma connaissance bien pauvre et <equipment> peu formatée

3) des **métadonnées pour les langues utilisées** : segments clairement identifiées vs segments difficiles à catégoriser, code-switching, langues de contact, ...

Les balises <langUsage> et <language> reposent sur une identification de la langue , une langue principale mais gèrent assez mal les segments non attribués à moins de définir une typologie de langue ...

4) des **métadonnées pour identifier la nature de l'oral**, le nombre de locuteurs, ... : récit, enseignement, apprentissage, conversation, … La balise <setting> peut répondre mais elle est peu formatée

5) dans la transcription, plusieurs niveaux d'annotations sont réalisés et ne correspondent pas forcément au même découpage temporel de cette transcription, le non-verbal n'a par exemple pas les mêmes bornes que le verbal, la traduction s'entend sur un segment de plusieurs mots alors que les lemmes se rapportent à un mot précis, ...

Je pense que nos collègues travaillant sur le syntaxe seront aussi intéressés par un **jeu d'étiquettes liées au mot**.

Pour ce point, les balises <spanGrp> et <anchor> de la TEI peuvent répondre à cette question avec une définition de plusieurs bornes (repère temporel ou lien direct à l'id d'un mot ou un anchor "éléctron libre" sans lien avec le timeline ou le mot) pour identifier la portée de ces annotations.