

CIInt Guidelines (draft version)

Marta Vila*

marta.vila@ub.edu

Montse Nofre**

montsenofre@ub.edu

(*) CLiC-UB (Centre de Llenguatge i Computació)

(**) STeL-UB (Servei de Tecnologia Lingüística)

XML tags

<transcriptions> is the upper category in the hierarchy. It contains the different transcriptions (each one tagged as **<transcription>**).

Each **<transcription>** contains a **<teiHeader>** and a **<text>** tags.

The **<teiHeader>** contains information about the **<text>**: a file description **<fileDesc>**, an encoding description **<encodingDesc>**, a text profile **<profileDesc>**, and a revision history **<revisionDesc>**.

The **<fileDesc>** contains information about the current transcription file: its transcriber **<respStmt>**, the creation date **<date>**, and its source **<sourceDesc>**. The **fileDesc** also has an attribute indicating its code. This code consists of "CIInt" plus underscore (_) plus the recording code (see **<recording>** tag below) (e.g. CIInt_ARMO_01).

The **<respStmt>** contains information about the responsible person/people for the file, that is, the transcriber(s) **<resp>**.

<resp> contains the name of the transcriber of the file in this format *Name_Surname* (e.g. Marta_Vila). If there is more than one transcriber, a **<resp>** tag must be used for each one.

<date> contains the date when the file was created in the format *Month_day_year* (e.g. January_15_2009).

The **<sourceDesc>** contains information about the source recording(s) **<recordingStmt>**.

The **<recordingStmnt>** contains the recording(s) from which the transcription is derived <recording>.

<recording> contains the recording date <date>. If there is more than one source recording, a <recording> tag must be used for each. <recording> also has three attributes indicating the code, the type, and the duration of the recording. The code consists of four capital letters indicating the doctor to whom the recording corresponds (e.g. ARMO) plus an underscore (_) plus a two-digit number that differentiates between the different recordings that correspond to the same doctor (e.g. code= "ARMO_01"). type indicates that it is an "audio" (e.g. type= "audio"). Finally, duration indicates the minutes and the seconds the audio lasts in the format *minutes:seconds* (e.g. duration= "22:32").

<date> contains the recording date in the format *Month_day_year* (e.g. October_08_2008).

The **<encodingDesc>** contains a link to this guide.

The **<profileDesc>** contains descriptive and contextual information about the text: the language used <langUsage> and the participants <particDesc>.

<langUsage> contains information about the language(s) used in the interaction <language>.

<language> contains (one of) the language(s) used in the interaction (e.g. Spanish). If more than one language is used, a <language> tag must be used for each one. <language> also has an attribute, lang, to indicate the official code of the language (e.g. lang= "SP").

<partDesc> contains information about the participants in the interaction <listPerson>.

<listPerson> contains the people participating in the interaction <person>.

<person> contains the role of the persons participating in the interaction (e.g. doctor). A

<person> tag must be used for each one. <person> also has two attributes: who and sex. Who indicates the participant's code (e.g. who= "AA"). Doctor's code is a letter: A, for example. This doctor's patients are AA, AB, AC, etc. If other health professionals participate in the interaction, they are encoded as AAX and AAY (for patient AA) or ABX and ABY (for patient AB). If someone accompanies patient AA, this person (these people) is AAW (and AAZ). Sex specifies whether this person is a man (by default), a woman, or a child (e.g. sex= "woman"). For the cases in which we don't know the person whom the utterance corresponds to (see <u> tag below), the <person> tag can also contain "unknown". In this case, <person> only has an attribute for the code, which is, again, "unknown".

The **<revisionDesc>** contains information about the changes made posterior to the definition of these guidelines.

<text> contains the transcription itself distributed in different utterances <u> and with inserted incidents <incident>. <text> also has an attribute, lang, that indicates the transcription language (e.g. lang= "SP").

<u> is a stretch of speech preceded and followed by a silence or by a change of speaker. Every vocalized production (neither necessarily lexicalized nor necessarily communicative) by one of the participants implies a new utterance. <u> can contain vocal but non-lexical sounds <vocals>, pauses <pause>, incidents <incident>, written text which is read <shift>, emphasized elements <emph>, gaps <gap>, unclear elements <unclear>, foreign words <foreign>, wrong words <choice> and long sounds <long>. <u> also has two attributes: who and trans. who contains the code of the person whom the utterance corresponds to (who= "AA"). If we don't know, we must write "unknown". The trans attribute is provided as a means of characterizing the transition from one utterance to the next. The specified value applies to the transition from the preceding utterance to the utterance bearing the attribute. The possible transitions are: smooth (by default), pause, overlap, no trans (the first utterance of the transcription, and the ones that appear after an incident) (e.g. trans= "overlap"). As we have seen, an utterance may contain running text, or text within which other basic structural elements are nested. Where such nesting occurs, the who

attribute is considered to be inherited by the nested elements, except for incident, which is not necessarily tied to any participant. Finally, when a person's surname, address, or telephone number are revealed, they appear as SURNAME¹, ADDRESS, or TELEPHONE NUMBER (in capital letters) in the transcription.

<vocal> contains vocalized but non-lexical and not necessarily communicative sounds. These sounds are tagged as description <desc> or sound <sound>.

<desc> contains the description of a non-lexical vocal sound (e.g cough).

<sound> contains the transcription of a semi-lexical sound (e.g. eh).

<pause/> is an empty element that indicates a silence in the interaction.

<incident> includes any phenomenon or occurrence, which is not vocalized and not necessarily communicative, for example incidental noises or other events affecting communication. It contains a description tag <desc>.

<desc> description of the incident (e.g. A² is typing).

<shift> and its new attribute, always with the "reading" value, are used to specify when a passage of written text is read to participants (e.g. new= "reading").

<emph> is used to indicate that the words are uttered in an emphatic way.

<gap/> is an empty element used to indicate that a piece of the audio has not been understood by the transcriber.

<unclear> is used to mark words which the transcriber has included although s/he is unsure about their accuracy.

<foreign> indicates that there is a change in the language defined in the lang attribute of the <text> tag. <foreign> also has a lang attribute indicating the code of the new language that is being used (e.g. lang= "CAT").

¹ If it is only the name, it is transcribed. If it is name together with the surname, only the name is transcribed (e.g. "Juan SURNAME").

² A is a doctor's code.

<choice> indicates that there is an erroneous word or a wrongly used word. It contains the original form **<orig>** and the regularized one **<reg>**.

<orig> contains a wrong word (e.g. *apática* in the sense of *asmática*).

<reg> contains the regularized form (e.g. *asmática*)

<long> indicates that the sound is longer than it should be.

<incident> is an incident that occurs between two utterances.

<desc> is a description of the incident.

Here there is a sum-up table that shows the hierarchy of tags, whether they have PCDATA (which the transcriber has to fill), and/or whether they have attributes:

<transcriptions>			
<transcription>			
<teiHeader>			
<fileDesc> code= "CIInt_ARMO_01"			
<respStmt>			
<resp> PCDATA			
<date> PCDATA			
<sourceDesc>			
<recordingStmt>			
<recording> code= "ARMO_01"; type= "audio"; duration= "22 :32"			
<date> PCDATA			
<encodingDesc> PCDATA			
<profileDesc>			
<langUsage>			
<language> lang= "SP" PCDATA			
<partDesc>			
<listPerson>			
<person> who= "AA"; sex= "woman" PCDATA			
<revisionDesc> PCDATA			
<text> lang = "SP"			
<u> who= "AA"; trans= "overlap" PCDATA			
<vocal>			
<desc> PCDATA			
<sound> PCDATA			
<pause/>			
<incident>			
<desc> PCDATA			
<shift> new= "reading" PCDATA			
<emph> PCDATA			
<gap/>			
<unclear> PCDATA			
<foreign> lang= "SP" PCDATA			

				<choice>
				<orig> PCDATA
				<reg> PCDATA
				<long> PCDATA
				<incident>
				<desc> PCDATA

References

TEI (Text Encoding Initiative)

→ P5: Guidelines for Electronic Text Encoding and Interchange

→ 2 The TEI Header

<<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD4>>

→ 8 Transcriptions of Speech

<<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html>>

NOTE: Some tag definitions are (partially) extracted from these sources.