Using the TEI as a pivot format for oral and multimodal language corpora

Loïc Liégeois, Carole Etienne, Christophe Parisse, Christophe Benzitoun, Christian Chanard

- IRCOM presentation
- Some oral corpora available in TEI
- Specific needs in studies based on oral corpora
- Metadata in TEI Header
- Aligned transcripts and oral events representation in TEI

IRCOM Presentation
http://ircom.huma-num.fr

- a french infrastructure to share informations and solutions on oral corpora studies and … in the end propose good practice
- an interdisciplinary approach: acquisition, semantic, sociolinguistic, morphosyntax, pragmatic, multimodality, prosody, sign language
- a common glossary
- an oral corpora directory
- a ressources directory (software, events, standards, tutorials…)
- an individual help to finalize and make available oral corpora
- several training and workshops with oral corpora community
- 6 workgroups dealing with a main topic: corpus-based research, i**nteroperability**, plurilingual and multilingual corpora, multimodality, juridical issues, archive

➔ Contribution **to ISO TEI European group**, coordinated by Thomas Schmidt

Some oral corpora from IRCOM directory, already available in TEI

- ALIPE: This corpus includes approximately 30 hours of verbal interactions between three children and their parents, recorded in natural settings.
- CLAPI: a databank of Spoken French corpora of around 60 hours of audio or video recordings in various natural settings : private, commercial, professionnal, medical…
- COLAJE (*): This corpus contains about 200 hours of child language interaction in natural settings between the children and the family members present during the recording.
- ESLO: This is a sociolinguistic study performed in the 70s (ESLO 1) and since the 2000s (ESLO2). About 7 million words are currently available. This is an ongoing project.
- Ongoing projects:
  - CIEL-F (*): an ecological corpora of French Spoken in 17 area in the world => metadata collection exchange in TEI between MOCA and CLAPI databanks.
  - ORFEO (*): a study corpus on French Language, gathering data from 20 different corpora with a common subset of metadata and semi-automatic annotations (PoS, Lemma, Syntaxic…).

(*) ANR projects

Specific needs in studies based on oral corpora

- **Transcripts made with dedicated software like transcriber, praat,clan, elan, … ➔ graphic representation ➔ script to generate TEI files (no direct entry in TEI)**
- Audio or video signal
- Transcript aligned with the signal including verbal and non-verbal productions (prosody, gesture…)
- Different levels of (in)dependant annotations aligned with the signal like "a music score": translation, gesture or gaze, macro-syntax, interaction, prosody, …
- Oral phenomena linked to a set of words or a signal extract, with (most part of the time) a given scope
- A set of metadata including the type of oral corpora, the speakers, the signals and the setting

Metadata in TEI Header (bold = main missing elements, properties)

- Type of oral corpora: a complexfull personnalized classification using element <taxonomy> <catDesc>
- The access rights: <availability> and <licence>
- The speakers: <person> <occupation> <education> <birth> <langknowledge> **<learning> (place and context)**
- The signals:

  <recording> is not well adapted ➔ ISO-TEI workgroup add a **<media>** element inside <recording> for the several signals (formats , quality, audio/video)

  in place of <ab> or <p>: **<anonymisation type=""><quality level="">**
- The setting: <setting> <activity>

  <target> inside activity for artifact manipulated during the setting,

  a **date property** in <activity> to order a set of meetings, lessons.

Aligned transcripts and oral events representation in TEI

- **Annotations: guided/controlled vocabulary to fixed annotations, type attribute ➔ example of recording quality**

```
<encodingDesc>
        <classDecl>
         <catRef scheme="# qualiteEnregistrement " target="#bruité #peu bruité #chambre sourde "/>
        <taxonomy xml:id="qualiteEnregistrement" >
                <category xml:id="bruité">
                        <catDesc> bruité, gène la compréhension</catDesc>
                        <category xml:id="bruité-gene"><catDesc>gène</catDesc></category>
                        <category xml:id="bruité-masque"><catDesc>masque</catDesc></category>
                </category>
                 <category xml:id="peu bruité">
                        <catDesc xml:lang="fr" >moins de 5% inaudible</catDesc>
                        < catDesc xml:lang="en" >less than 5% inaudible</catDesc>
                 </category>
                <category xml:id="chambre sourde">
                        <catDesc> chambre sourde </catDesc>
                </category>
        </taxonomy>
```

Aligned transcripts and oral events representation in TEI

- **Annotations: guided/controlled vocabulary to fixed annotations, type attribute ➔ example of recording quality**

```
<sourceDesc>
   <recordingStmt>
      <recording>
        <media dur="PT37M32S"  mimeType="video/mp4"
url=file://Clapi_Signal_aperitif__pois_Aperitif_pois_a52474e3b8.mp4
      ana="#peu bruité">
       </recording>
    </recordingStmt>
```

Aligned transcripts and oral events representation in TEI

- **Annotations: guided/controlled vocabulary to fixed annotations, type attribute ➔ example of annotation of liaison (phonological phenomena)**

<editorialDecl>

<segmentation>

** **Le texte a été segmenté pour la liaison. Ainsi, nous retrouvons au sein de ces balises l'ensemble du contexte de liaison avec son codage. Lorsque plusieurs contextes de liaison s'enchaînent (par exemple : "nous en avons"), l'ensemble des contextes de liaison se retrouve au sein de ces balises.** **

</segmentation>

<fsDecl type="liaison>

<fsDescr xml:lang="fr-FR"> La liaison est un phénomène phonologique impliquant une consonne de liaison (CL). Graphiquement, cette CL est présente à la fin du premier mot mis en jeu dans la liaison (mot1) **</fsDescr>**

<fsDescr xml:lang="fr-FR"> …**</fsDescr>**

Aligned transcripts and oral events representation in TEI

- **Annotations: guided/controlled vocabulary to fixed annotations, type attribute ➔ example of annotation of liaison (phonological phenomena)**

```
<fDecl name="SyntacticContext" xml:id="SC">
    < fDescr> Cette partie du codage concerne le contexte syntaxique du contexte de liaison. En nous référant aux travaux précédemment cités (<bibl corresp="#Du1"/> <bibl corresp="#Ma1"/>), nous avons défini douze contextes. Généralement, les quatre premiers correspondent à des contextes de liaisons catégoriques (contextes A, B, C et D) alors que les six suivants correspondent à des contextes de liaisons variables (contextes E, F, G, H, I, J).
    </fDescr>
<vRange >
    <vAlt>
        <symbol value="A"/>
        <symbol value="B"/>
    <vAlt>
<vRange>
```

Aligned transcripts and oral events representation in TEI

- **Timelines**: mainly absolute but sometimes **relative**

<timeline unit="s" origin="#T0"> <when xml:id="Tn-1" absolute="00:00:05.26"/> **<when xml:id="Tn" />** <when xml:id="Tn+1" absolute="00:00:06.00"/>	<timeline unit="s" origin="#T0 »> <when xml:id="Tn-1" interval="5.26" since="#T0"/> **<when xml:id="Tn" interval="unknown" since="#T0"/>** <when xml:id="Tn+1" interval=" 6.00" since="#T0"/>

<u who="#A" **start="#Tx" end="#Ty"** >

and/or **inside** an utterance

<u who="#A" **start="#Tx" end="#Ty"** > … <**anchor synch="#Tn"/> …. </u>**

Aligned transcripts and oral events representation in TEI

- Pauses: **<pause dur**="PT0.61S" **start**="#T2" **end**="#T3 »/>

- **Overlapped segments** between different speakers: frequent cases in interacting conversations or meetings, relative timings in a very short extract ➔ **many timelines !**

  Speaker *A* *[mais attends]* *là je suis pas d'accord* *[si tu veux]* *on prends*

  Speaker *B* *[non mais]* *je crois pas* *[hum hum]*

  Speaker *C* *[si si]*

A **<anchor synch="#Tx"/>** *mais attends* **<anchor synch="#Tx+1"/>** *là je suis pas d'accord* **<anchor synch="#Tx+2"/>** *si tu veux* **<anchor synch="#Tx+3"/>** *on prends*

B **<anchor synch="#Tx"/>** *non mais* **<anchor synch="#Tx+1"/>** *je crois pas* **<anchor synch="#Tx+2"/>** *hum hum* **<anchor synch="#Tx +3"/>**

C **<anchor synch="#Tx"/>** *si si* **<anchor synch="#Tx+1"/>**

Aligned transcripts and oral events representation in TEI

- Main elements: **<choice> <shift> <vocal> <incident> <incident> <unclear> <gap>**

  - **<choice>:** an alternative form like **alternative spelling** *vs* **orthographic word**

```
<choice>  <orig>c`ui</orig> <reg>celui</reg> </choice>
```

  - **<shift>:** ponctual phenomena**: prosody, intonative, lengthening**

```
<w>euh  <shift feature="tempo" new="rall"/> </w>
<w>ça</w> <w>va <shift feature="pitch" new="asc"/> </w>
```

  - **<vocal>:** a specificic **vocal but non verbal** production **attributed to a speaker**

```
<vocal who="#B"> <desc> a longer whistle </desc> </vocal>
```

Aligned transcripts and oral events representation in TEI

- Main elements: **<choice> <shift> <vocal> <incident> <incident> <unclear> <gap>**

- **<incident>:** if signal loss → external noise, other participant…

  <incident> <desc> *moving chairs* **</desc> </ incident> </vocal>**

- **<unclear>:** doubt, alternative

  <unclear><w>quand</w><w>même</w>**</unclear>**

- **<gap>:** unknown

  <w >*des*</w><w>*expériences*</w>**<gap extent="2 sylls"/>**

Aligned transcripts and oral events representation in TEI

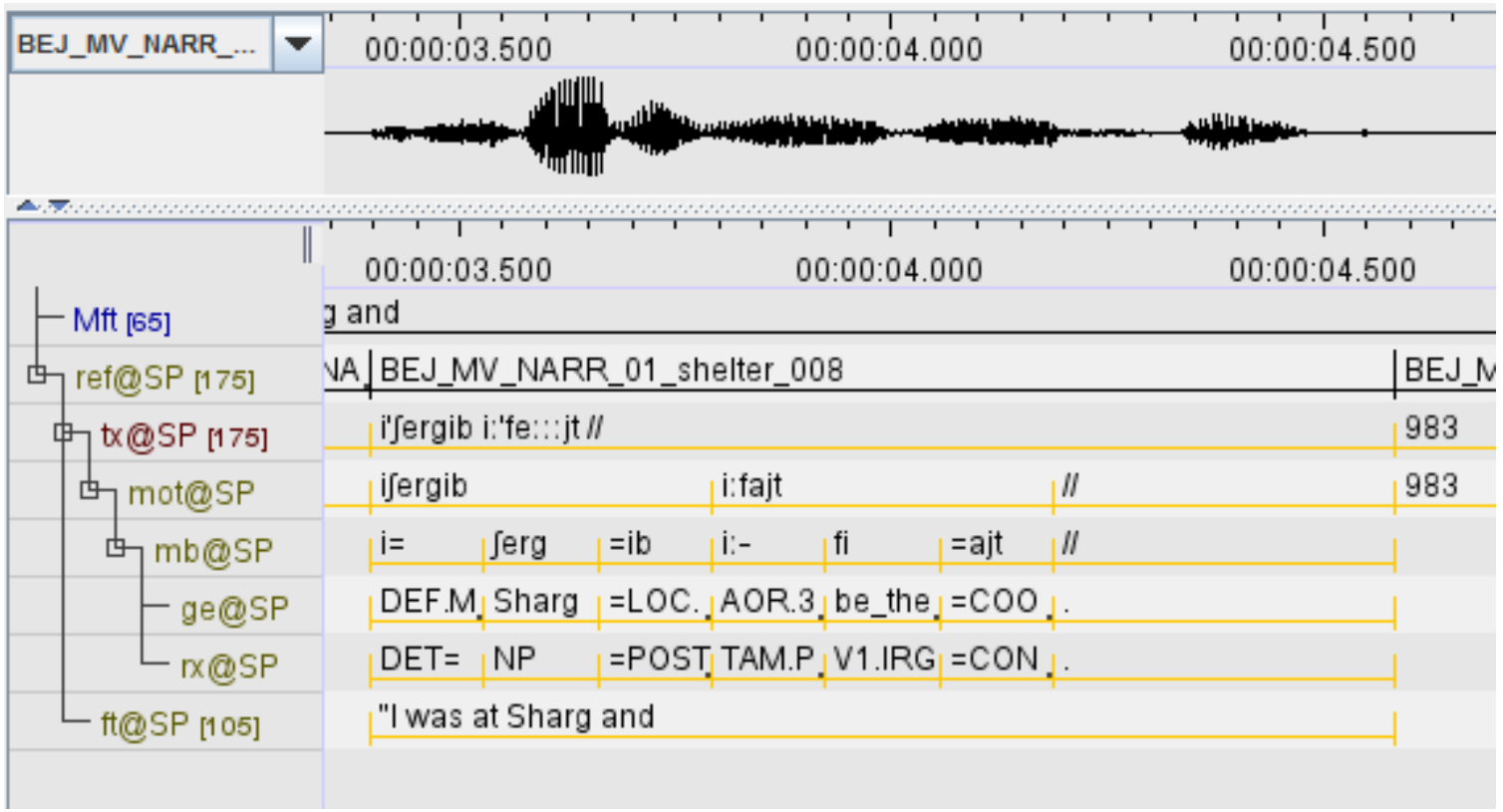
- **Annotations : guided/controlled vocabulary to fixed annotations, type attribute ➔ example of annotation of liaison (phonological phenomena)**

Utterance : "Regarde ça c'est un petit scooter"

```
<w>regarde</w> <w>ça</w>
<seg>
        <w>c'est</w>
                <fs type="liaison">
                        <f name="Word1" fVal="c'est"/>
                        <f  name="Word2" fVal="un"/>
                        <f name="SynctacticContext" fVal="H"/>
                        <f name="ExpectedConsonnant" fVal="t"/>
                        <f name="ProducedConsonnant" fVal="t"/>
                        <f name="ObligatoryOptional" fVal="0"/>
                </fs>
        <w>un</w>
</seg>
<w>petit</w> <pause/> <w>scooter</w>
```

Aligned transcripts and oral events representation in TEI

- **Rare languages: tree-level of annotations with different scopes**



Aligned transcripts and oral events representation in TEI

- **Rare languages: tree-level of annotations with different scopes**

```
<annotatedGrp who="SP" start="ts18" end="ts19">
<u xml:id="tx_8" xml:lang="bej" rend="phonology">iˈʃergib iːˈfeːjt //</u>
<seg type="reference">
  <seg xml:id="ann66" type="ref">BEJ_MV_NARR_01_shelter_008</seg>
  <seg type="texte">
    <seg xml:id="ann77" xml:lang="bej" type="tx">iˈʃergib iːˈfeːjt //</seg>
    <w xml:id="mot_14"  type="mot">
      <seg xml:id="a136" xml:lang="bej" type="mot">iʃergib</seg>
      <m xml:id="mb_13"  type="morpheme">
        <seg xml:id="a565en" xml:lang="bej" type="mb">i-</seg>
        <seg xml:id="a1227en" xml:lang="en" type="ge">DEF.M-</seg>
        <seg xml:id="a1889en" xml:lang="en" type="rx">DET-</seg>
      <m xml:id="mb_14"  type="morpheme">…</m>
      <m xml:id="mb_15"  type="morpheme">…</m>
    </w><w>…</w>
    <seg xml:lang="en" type="ft">"I was at Sharg and</seg>
 </seg></seg>
</annotatedGrp>
```

Coding the transcription

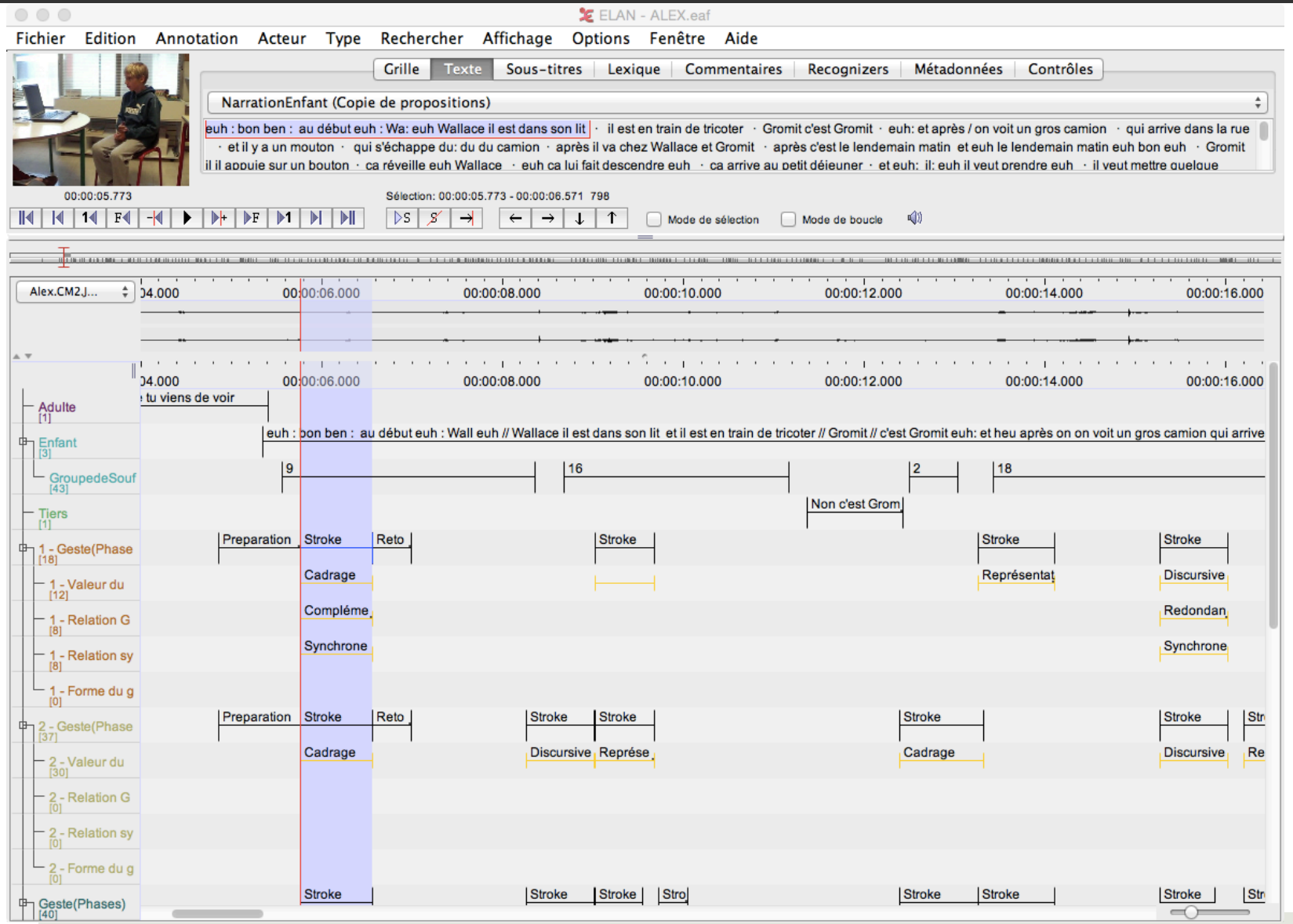
Transcriber

CLAN

```
– <annotationGrp end="#T44" start="#T40" wh="spk2" xml:id="au27">
  – <u type="spk2">
      <seg>un quart d'heure par jour à peu près ? </seg>
    </u>
  </annotationGrp>
– <annotationGrp end="#T45" start="#T46" wh="spk1" xml:id="au28">

– <annotationGrp end="#T171" start="#T170" wh="CHI" xml:id="au256">
  – <u type="CHI">
      <seg>celui+là #. </seg>
    </u>
  – <spanGrp>
    – <span type="pho">
        <seg>sqila</seg>
      </span>
    – <span type="act">
        <seg>CHI prend une figurine d'animal</seg>
      </span>
    </spanGrp>
  </annotationGrp>
```

Gestures


ELAN - ALEX.eaf
Fichier Edition Annotation Acteur Type Rechercher Affichage Options Fenêtre Aide
Grille Texte Sous-titres Lexique Commentaires Recognizers Métadonnées Contrôles
NarrationEnfant (Copie de propositions)
euh : bon ben : au début euh : Wa: euh Wallace il est dans son lit · il est en train de tricoter · Gromit c'est Gromit · euh: et après / on voit un gros camion · qui arrive dans la rue
· et il y a un mouton · qui s'échappe du: du du camion · après il va chez Wallace et Gromit · après c'est le lendemain matin et euh le lendemain matin euh bon euh · Gromit
il il appuie sur un bouton · ca réveille euh Wallace · euh ca lui fait descendre euh · ca arrive au petit déjeuner · et euh: il: euh il veut prendre euh · il veut mettre quelque
00:00:05.773
Sélection: 00:00:05.773 - 00:00:06.571 798
Mode de sélection
Mode de boucle
Alex.CM2J...
00:00:06.000
00:00:08.000
00:00:10.000
00:00:12.000
00:00:14.000
00:00:16.000
Adulte [1]
Enfant [3]
GroupedeSouf [43]
Tiers [1]
1 - Geste(Phase [18]
1 - Valeur du [12]
1 - Relation G [8]
1 - Relation sy [8]
1 - Forme du g [0]
2 - Geste(Phase [37]
2 - Valeur du [30]
2 - Relation G [0]
2 - Relation sy [0]
2 - Forme du g [0]
Geste(Phases) [40]
tu viens de voir
euh : bon ben : au début euh : Wall euh // Wallace il est dans son lit et il est en train de tricoter // Gromit // c'est Gromit euh: et heu après on on voit un gros camion qui arrive
9
16
2
18
Non c'est Grom
Preparation
Stroke
Reto
Cadrage
Compléme
Synchrone
Représentat
Discursive
Redondan
Discursive
Représe
Cadrage
Stro

Tei version for gestures

```
– <annotationGrp end="#ts21" start="#ts18" wh="1 - Geste(Phases)" xml:id="a126">
  – <u type="1 - Geste(Phases)">
      <seg>Stroke</seg>
    </u>
  – <spanGrp>
    – <span target="a126" type="1 - Relation synchronique">
        <seg xml:id="a165">Synchrone</seg>
      </span>
    – <span target="a126" type="1 - Relation Geste/Parole">
        <seg xml:id="a156">Complément</seg>
      </span>
    – <span target="a126" type="1 - Valeur du Geste
        <seg xml:id="a143">Cadrage</seg>
      </span>
  </spanGrp>
</annotationGrp>
```

Symbolic subdivision

```
– <annotationGrp end="#ts680" start="#ts10" wh="Enfant" xml:id="a324">
  – <u type="Enfant">
    – <seg>
        euh : bon ben : au début euh : Wall euh // Wallace il est dans son lit et il
        Wallace euh ça lui fait descendre euh // ça (a)rrive au p(e)tit déjeuner //
        de plante qui: euh: // qui es(t) enlevé qui a été mangé // bon après il ente
        est dans le: // dans l(e) salon i(l) regarde son journal et y a un trou au mi
      </seg>
    </u>
  – <spanGrp>
    – <span type="GroupedeSouffle">
        <anchor synch="#ts13"/>
        <seg xml:id="a367">9</seg>
        <anchor synch="#ts30"/>
        <anchor synch="#ts31"/>
        <seg xml:id="a368">16</seg>
        <anchor synch="#ts52"/>
        <anchor synch="#ts66"/>
        <seg xml:id="a369">2</seg>
        <anchor synch="#ts67"/>
```

Temporal subdivision

Information about structure linked to software

Template (structure) of the document avoid having to compute this information at conversion time
Provides information that can be shared in a set of transcriptions
➔ How to represent in TEI complementary information without <note> <p> elements ? semantic contrainsts ?


```
<template code="Valeur du Geste" cv_id="#FonctionGeste" parent="Geste(Phases)
scribe="" subtype="Valeur du geste" type="Symbolic_Association" xml:lang="fr"/>
<template code="1 - Geste(Phases)" cv_id="#Phases" parent="-" scribe=""
subtype="PhaseGeste" type="-" xml:lang="fr"/>
```

```
<listForest>
 <forest type="derivation-syntactic">
  <tree ord="true">
   <root children="#fgex1 #fgex2"/>
<!-- ... -->
  </tree>
<!-- ... -->
 </forest>
 <forest type="derivation-prosodic">
  <tree ord="true">
   <root children="#fgex3 #fgex4"/>
<!-- ... -->
  </tree>
<!-- ... -->
 </forest>
</listForest>
```

TEI version for gestures

- Gesture Studies and LSF:
  - Adding an image to split a gesture in different stages:

```
<figure>
        <graphic url="fig1.png"/>
        <head>Figure One: tittle</head>
        <figDesc>A short descriptions</figDesc>
</figure>
```

- Other particular cases:
  - Prosody : intonational patterns, f0, ...
  - Parallel corpus
  - Eye tracker, motion sensor or other experimental process
  - …

Workgroup: Interoperability

- As a conclusion…
  - on going work based on few experiences to **solve a subset of common issues** in our corpora
  - IRCOM as a french infrastructure to explain and give informations on "TEI for oral corpora"
    - organization of several meetings
    - collaboration to ISO-TEI European group
    - on-line good practice
    - a future training on Tei: principles, subset of TEI elements (metadata and transcripts), examples from our corpora
  - ANR projects t conduce research teams to make available and reusable fully annotated data in Tei format (Ciel-F, Colaje, Orféo, …) and not only archive format (Dublin Core)