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

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

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

Loïc Liégeois, Carole Etienne, Christophe Benzitoun and Christophe Parisse

Abstract

In the field of Linguistics, data sharing has become increasingly important over the past few decades. In order to overcome technical difficulties that arise when using on the same data more than one piece of software, there is an ever-growing need for a common format that researchers can use to share data. In this paper, we present a synthesis of work undertaken within the framework of the IRCOM consortium (Research Infrastructure for Oral and Multimodal Corpora) by members of the "Interoperability" workgroup. The aim of this paper is twofold. Firstly, we present a solution for structuring, storing and sharing diverse oral/multimodal language corpora based on the TEI format. When necessary, we highlight gaps and suggest potential solutions to encode information that currently cannot be structured using TEI P5. Secondly, we offer a discussion on how the solution presented may allow different software used in the Linguistics community to become interoperable through the development of a conversion tool.

Keywords

oral and multimodal corpora, pivot format, metadata, sharing corpora, interoperability

1. Introduction

In the field of Linguistics, data sharing has become increasingly important over the past few decades. When research projects include data sharing as a project deliverable, it not only enables previous studies to be replicated but also gives an opportunity to use data for new purposes and development, thus contributing to the creation of larger and larger datasets for research and application.

With the spread of high speed Internet networks, it has become easier and easier to exchange data. While data sharing for written language data has become a relatively common practice, the sharing of oral and multimodal language material remains rare due to the size of data involved (e.g. high speed bandwidth is necessary to exchange video recordings) and a general feeling that an oral corpus that has been collected and transcribed with a specific scientific objective in mind may not be relevant for studies adopting other theoretical or analytical approaches.

Sharing oral data involves the diffusion of available and reusable corpora for different research purposes. It requires researchers to share recordings and transcripts in open formats, to define a common subset of relevant metadata to explain the context in which the data has been collected or provide information about the speakers, to deal with juridical restrictions concerning ethics and data reuse, to describe and give examples of annotations and more generally to keep in mind that all the components of an oral corpus need to be stored, described and converted into ready-to-use open formats.

Although sharing data is becoming a more widely-accepted practice in the Linguistics community, important technical difficulties persist, especially due to the wide range of research undertaken using oral and multimodal language corpora. This is partially intrinsic to the field. For example, researchers working in the branches of phonetics, gesture studies, or syntax, will often work on quite different datasets. However, even if corpora have been designed from the outset with a specific purpose or research aim in mind, it would still be of interest to work on these corpora in order to undertake phonetic, gesture, and syntactic research on the same data set and provide not only new research results but also new annotated data.

The methodological approach to share oral and multimodal corpora (or any linguistic corpora) requires a staged methodological approach that we endeavour to present in the current paper:

- a common format that is as flexible as necessary;
- a set of good practices which will not add constraints to data collection but which will allow for better reuse of the data.

1.a. IRCOM and French Huma-Num consortium

The goal of sharing data on the basis on minimal technical characteristics and information about best practices could be not being started if behind the initiative there was not a large and consensual community. This was made possible when the French national institute Huma-Num (www.huma-num.fr), supported by the CNRS (www.cnrs.fr), set up consortia devoted to developing on a national level the sharing, distribution and normalisation of digital data in a range of fields in the Humanities, including Linguistics.

IRCOM (ircom.huma-num.fr), the Infrastructure *de Recherche pour les Corpus Oraux et Multimodaux* (Research Infrastructure for Oral and Multimodal Corpora), aimed[1] at developing the use of standards, data sharing, and best practices for digital rights, long-term data archiving and the extension of the use of oral and multimodal corpora in linguistic research. IRCOM did not limit itself to one branch of Linguistics and consortium members came from several areas including child language acquisition, semantics, sociolinguistics, morphosyntax, pragmatics, multimodality studies, interaction, prosody, gesture studies, etc.

At the date of writing, IRCOM has already developed a common glossary, an oral corpora directory, a resources directory (software, events, standards, and tutorials), an individual help to finalize and make available oral corpora, and organized several training and workshops with oral corpora community. The different tasks are carried out by workgroups and one of them has been working on interoperability, and especially corpus format and metadata.

1.b. ISO TEI European group

The decision taken by the IRCOM interoperability workgroup concerning data format was to base its work on the TEI format, considering the latter was already widely used for linguistic data, and especially corpora comprising written data. For this reason, the IRCOM group has been working closely with the ISO TEI for Oral corpora initiative (ISO/DIN WG 6 PWI 24624), coordinated by Thomas Schmidt. The workgroup's main goal is to provide guidelines concerning corpora formats and metadata that conform to the requirements of the ISO group and to extend them when they can't be applied to its corpora or practices.

In following the same approach, we tried to work only within the limits of the material existing in TEI as a whole.

1.c. Policy and goals

The aim of our project is not to create a standard from scratch and later lobby to defend the proposed format. On the contrary, the group aimed to start by considering existing formats and corpora before looking at ways to create tools and standards that are tailored to the specific needs of their actual data.

A five-way process allowed the workgroup to meet this aim:

1- the IRCOM workgroup was composed of members from various research laboratories that are in different geographical regions within France but also Belgium and Switzerland. Prior to becoming members of the workgroup, the colleagues from different laboratories each had already-existing data with transcripts and metadata (even if the metadata were not always fully normalised);

2- all members of the oral and multimodal community work with existing software programs to annotate or align data with recordings, and each member had a defined format (although these ranged in the extent to which they had previously been documented and described). It is impossible when working on oral and multimodal language to edit a corpus directly in XML format (e.g. using XML editor tools like Oxygen) because when annotating data we need to work simultaneously with the audio/video recordings. XML has to be reserved for technical purposes only. Thus our goal is to provide an interchange process between these different software;

[1] The IRCOM consortium (2012-2105) goes on through other Huma-Num consortia such as for example CORLI.

3- the ORTOLANG project (www.ortolang.fr), which stands for *Outils et Ressources pour un Traitement Optimisé de la LANGue* (Open Resources and TOols for LANGuage) aims to centralise and provide the infrastructure necessary for the long-term conservation of linguistic data. One of the sub-goals of the project is to provide the largest possible access to the available data. With this purpose in mind, ORTOLANG decided to develop a tool to convert data using a common pivot format between the different transcription applications and to improve its quality by testing it on different datasets and with different uses of the same software;

4- whenever necessary (for example for multimodal data and gesture or sign language annotations), extensions were made to both this format and the TEI Guidelines. In most cases, the extensions are linked to the semantics of the data. In these cases, we tried to focus firstly on the TEI Guidelines rather than on differences between formats. This policy led us to use the TEI standard for data which had never previously been structured in TEI;

5- the workgroup wished to avoid the development, by each corpus-based project, of a) a new representation of its corpus (metadata or/and data) in TEI and b) researchers spending time representing almost the same kind of metadata or linguistic phenomena without collaboration between different members and laboratories.

2. Start point

2.a. A large data diversity…

Linguistic oral/multimodal corpora may be very diverse in nature, both in terms of scientific goals and in terms of methodologies involved. Table 1 presents an overview of some actual corpora-based research studies that focus on French language or French Sign Language. The diversity observed between the different corpus-based research projects is perceptible both where metadata and data are concerned.

With respect to the metadata, almost each project gives information on the setting, the speakers, the recordings, the access rights, how the corpus is to be released, the researchers involved in data collection and transcription. However, readers will notice that each data provider has chosen his own metadata subset, definition, value and unit. Indeed, gathering this information has been time consuming for the researchers (e.g. the ORFEO corpus-based project that includes 15 data sources).

Readers will also notice that several formats are used by different researchers for similar pieces of information such as duration, date or place. Another regular issue that the workgroup encountered was that sometimes a research project did not necessarily encode a full set of metadata for individual recordings in the corpus but rather only included this information in the overall project data. For instance, if a project focused on storytelling and involved young children, this information was not necessarily encoded for each individual recording but rather had to be extracted from the project presentation or comments. Another issue concerns a common use of evaluative values for metadata (like good/bad quality of the recording) which does not necessarily mean something specific depending on the discipline and context. In order to make data reusable, it would be more pertinent to know if a recording is more or less noisy, and whether it includes only a handful of short inaudible passages or whether the recording is affected more globally, etc.

Concerning data, we can notice that several projects presented in Table 1 include data as video recordings and study multimodality with different scientific objectives. Focusing on language acquisition, the Colaje project takes into account different semiotic modes

(including gesture, gaze or movement, for example) while the Creagest corpus only focuses on gestures during LSF (French Sign Language) interactions. Other studies based on prosody require a higher quality of sound and the use of software that are able to manage technical sound attributes. Other projects, including ESLO, include large amounts of audio data that need to be quickly aligned or inter-annotated by several colleagues.

The choice of working on audio or video recordings impacts on data processing methodologies implemented by researchers, from data collection to analysing tasks as well as the choice of transcription software. Most of the time, researchers interested in oral and multimodal data work with software dedicated to alignment and transcription tasks, including CLAN (MacWhinney, 2000), Transcriber (Barras et al., 2001), Praat (Boersma & Weenink, 2013) or ELAN (Wittenburg, et al., 2006). Moreover, verbal transcription may be completed in one software while the annotation of the visual mode may require another software. Indeed, the use of multiple transcription and annotation programs is usual. Working in this manner may cause some data/metadata to be lost, as each software does not need the same information and is unable to restitute them. Moreover, these cases of multiple uses are often documented but not provided to the community, as there is no easy way to mix data based on different formats. Most of the time, only the first version (with the starting point software) is preserved in institutional repositories.

2.b. …but a need of interoperability

Most of the time, secondary data (transcripts and annotations) are structured using the format of the programme used to transcribe and annotate the oral or multimodal primary data. For interoperability purposes, several projects have already made their data available in TEI and some solutions have been found to describe oral corpora using the TEI header and to represent transcripts in the TEI body. The ALIPE, Colaje and CLAPI projects demonstrate this way of working. Data structured within these projects respect TEI-P5 conventions, although sometimes different elements or attributes to describe almost the same metadata or oral phenomena may have been chosen, depending on the way our colleagues understood and extended the standard. The IRCOM workgroup takes advantage of what individual members have already achieved to suggest a new common proposal unified in a TEI format called TEI_CORPO.

The issue of data and metadata interoperability is central for any kind of project that aims to collect and structure a wide range of existing corpora. This is the case, for example, for the two on-going projects that we present below: the ORFEO and the CIEL-F projects.

ORFEO (Outils et Recherches sur le Français Écrit et Oral)

The ORFEO project aims to collect existing data (some have already been made available, others not). There are both written and oral components that have been transcribed and annotated using Transcriber, Praat or a word processor. Projects like ORFEO need a pivot format to manage the diversity of the data. Indeed, the main objective of this project is to create a study corpus of contemporary French. For the oral part (3 million tokens), many interaction situations are represented and the data will be released at least in two formats: their original format and a common format (TEI_CORPO).

CIEL-F (Corpus International Ecologique de la Langue Française)

The CIEL-F corpus of ecological comparable settings in French language has been collected in seventeen areas in the world and is fully described and annotated. The corpus will be stored in both the MOCA (VALIBEL research team) and CLAPI (ICAR research team) databanks in order to make the most of the different search tools available on these platforms. Thus metadata and transcripts need to be exchanged to take into account additional transcripts

or new recording formats that will become useful during analysis phases. For this purpose, the TEI-P5 standard has been chosen by the researchers because this format meets their requirements as regards metadata and transcripts.

Project name	**Description (software used, total coverage etc.)**	**Corpora state (in regards to ORTOLANG)**
ALIPE (http://hdl.handle.net/11041/alipe-000853)	Corpora of parents-child interactions (166.000 tokens). Transcription and annotation using CLAN. A TEI version of the corpora is available.	Available through ORTOLANG platform in the original format like CHAT or Transcriber and in TEI_CORPO format.
Colaje (http://hdl.handle.net/11403/colaje)	Corpora of natural interactions, based on video recordings designed to study language acquisition (1M tokens). Transcription and annotation using CLAN.	
ESLO (http://hdl.handle.net/11403/eslo1)	Corpora designed to study sociolinguistic variation (7M tokens). Transcription and annotation using Transcriber and based on audio recordings. Two corpora and two periods covered : ESLO-1 (1969-1974) and ESLO-2 (since 2008).	
CIEL-F	Corpora designed to study the diversity of French through the “Francophonie”. Transcription and annotation using Praat.	(Partially) available through a dedicated database. Deposit in ORTOLANG and TEI_CORPO conversion planned or in progress.
CLAPI	Corpora designed to study interaction, based on video and audio recordings. Transcription and annotation using CLAN, Transcriber, Praat or ELAN. A TEI version of the corpora is available.	
Creagest	Corpora designed to study interaction in FS (French Sign Language). About 300h of interaction have been collected.	
PFC	Corpora designed to study phonological variation through the “Francophonie”. Transcription and annotation using Praat.	
Rhapsodie	Corpora designed to study interfaces between prosody, syntax and discourse. Transcription and annotation using Praat and based on audio recordings.	
Transferts	Corpora designed to study problematics linked with the bilingual teachings in sub-Saharan countries (especially Mali, Burkina Faso and Niger). Transcription and annotation using CLAN.	
ORFEO	Corpus based on a collection of existing data. The objective is to create a study corpus of contemporary French in which there are 3 million tokens for the oral part of the corpus. The data will also annotated in part-of-speech and dependency.	

Table 1: Examples of corpus-based projects about French language

3. Specificity of oral and multimodal corpora

As we saw earlier, the corpora-based research projects are diverse in nature, in terms of both research goals and methodology involved. Nevertheless, all projects based on oral and multimodal corpora share some common ground that we describe in the following sections.

3.a. Transcribing and aligning the data

Aligning data with audio and video

The most specific feature of data included in oral and multimodal corpora is that the most basic data is not a linguistic transcript in a given code, but one or several media called primary data. The media are often audio or video files (there can be one or more depending on the requirement of the scientific investigation) in different formats according to the way they have been collected, such as the digital data provided by any sort of capture (body, hand, eye movements, position of the mouth or oral tract, breath, brain image, etc.). Although the technical possibilities are limitless (for example it might be possible in a semantic analysis to locate elements within the visual frame of a video rather than only in time), we limit our work to the data already produced by the tools commonly used in the IRCOM consortium (see 2.).

The ubiquitous presence of a media that is the most basic element of an oral or multimodal corpus calls for the use of a timeline, which is used as the reference to locate data. Everything in an oral and multimodal corpus is referenced to a position on a timeline, even if sometimes references might be made indirectly through another element. A timeline does not need to be expressed in time units (seconds, hours, milliseconds, etc.) as some media such as digital recordings of movement, brain, mouth movements, might not perfectly correspond to time. The format of the TEI timeline allows for such flexibility, but most of our data are actually based on ‘real’ time references.

As timelines could be represented in a chronological order, TEI-P5 gives us the opportunity to insert one or more additional timelines we will call "relative timelines" between timelines corresponding to a precise time that we will term "absolute timelines". With this notation, we will be able to link oral verbal or non-verbal phenomena to these relative timelines to respect the chronological order of events (sequentiality) even if we have no precise evaluation of them. We will use this function for instance in multimodality annotations with gestures, gaze or to represent synchronized verbal productions, for example overlapping and overlapped segments between several speakers. In the following example we only know the timing of Tx because the words are pronounced very quickly and it will cost a lot of time to identify precisely Tx+1, Tx+2 or Tx+3 (the [ ] characters represent simultaneous verbal productions):

Speaker A: [mais attends] là je suis pas d'accord [si tu veux] on prend
Speaker B: [non mais] je crois pas [hum hum]
Speaker C: [si si]

Corresponding timeline in TEI:

```
<timeline unit="s" origin="#T0">
 <when xml:id="Tx" absolute="00:00:05.26"/>
 <when xml:id="Tx+1"/>
 <when xml:id="Tx+2"/>
 <when xml:id="Tx+3"/>
 <when xml:id="Tx+4" absolute="00:00:06.00"/>
 <!-- ... -->
</timeline>
```

Example 1:Relative timelines.

Corresponding transcript in TEI:

```
<u who="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 prend
</u>
<u who="B">
<anchor synch="#Tx"/>non mais<anchor synch="#Tx+1"/>je crois pas<anchor
synch="#Tx+2"/>hum hum<anchor synch="#Tx +3"/>
</u>
<u who="C"><anchor synch="#Tx"/>si si<anchor synch="#Tx+1"/></u>
```

Example 2:The transcript using the relative timelines above to solve overlaps issue.

Representing non-verbal and non-orthographic data

As oral and multimodal corpora are ultimately based on media rather than textual transcription, this means that not every corpus has an orthographic transcript or even something that looks like actual language transcription. A first example is that of oral language corpora which corresponds to a language that does not have any kind of written form. This is the case for many languages in the world, and descriptions of these languages are often based on some type of phonological, morphological, and/or syntactic format, with equivalent glosses or transcripts in written form as dependant data rather than primary data. In this case, it will be relevant to link each annotation level directly to the media in order to bypass textual representations then gloss or written transcript will be coded as an annotation level like any other, the media becoming the basis of the annotation structure. These corpora can be represented by a phonetic representation in the utterance element instead of what is usually a written language transcription. Another solution used is to represent only the segmentation into utterances at the main level (the text), and code all data in the dependencies. A final solution is to represent the text as gloss into another language. A problem that arises often with this kind of data is that the other annotations (phonology, syntax, semantics, etc.) might not be aligned with subparts of the utterance content. In these cases, using a temporal representation is necessary.

A second example is the case of sign languages, which do not have any kind of written equivalent and where glossing might even be misleading. In this case, people often devise new coding schemes for different purposes (semantic, pragmatic, gestural, etc.) and have to define vocabularies that can be composed of any kind of signs (and often non-orthographic signs). In all these cases, sharing the structural principles underlying the transcript is often more important than sharing the transcript itself, as these principles are guided by the scientific analysis of the data. This is a function that is commonly provided in tools used for

gesture and sign language analysis (such as ELAN (Crasborn & Sloetjes, 2008) or ANVIL (Kipp, 2001)) but that is missing in the current TEI version and even in the ISO TEI extensions.

Currently, we haven't found solutions in TEI to represent coding schemes and vocabularies. So we had to use a temporary solution (see below “Template” and “Controlled vocabulary”) until a working group will be able to integrate the most appropriate elements to deal with this issue. Interestingly, the solution today is to provide better tools to organize multimodal data rather than trying to describe gesture and signs. This can be explained as these research about sign languages and gesture is still very young. Such tools that will make it easier to organize multimodal information may find applications in other fields than sign language and gesture.

3.b. Annotating the data

This section illustrates the solutions we have found to structure our annotations and encode some oral annotations.

Oral phenomena scope

The main problem we encountered when attempting to represent oral phenomena is that most element parts have no scope except for `<segment>`: they are designed to define a punctual event which occurs at a given time while we need to annotate events which start at one point and finish later and often that include insertions of other imbricated elements that XML format is not able to manage. Thus, we could try to annotate the event change and not the event itself to bypass this limitation using `<shift>` element and its attributes: @**feature** to identify the phenomena, @**new** to indicate the shortage and @synch to synchronize with timeline.

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

Example 3:A use of <shift> element.

But as several different phenomena could happen at more or less the same time, several shortages become difficult to identify. Moreover, `<shift>` element becomes obsolete in TEI standard, then using `<spanGrp>` with a customized `@type` attribute and `<span>` elements with **@from** and **@to** attributes to delimit the scope becomes a better solution. In the following example, *euh* was produced by speaker A with a lengthening and *euh ben* were produced loudly, we introduced two boundaries T2 and T3.

```
< annotationBlock start="#T1" end="#T4" who= "#A">
<u> <w>euh</w> <anchor synch="#T2" /><w>ben<w><anchor synch="#T3" /><w>voilà</w> </u>
<spanGrp type="tempo">
        <span from="#T1" to="#T2">lengthening</span>
</spanGrp>
<spanGrp type="pitch">
        <span from="#T1" to="#T3" >loud</span>
</spanGrp>
</ annotationBlock >
```

Example 4:<spanGrp> and as an alternative to <shift> element.

Alternative spelling

Some researchers try to transcribe the verbal production as it is pronounced, which means that if a speaker pronounces /bʒur/ for "*bonjour* (hello)" in French instead of /bɔ̃ʒur/, which is the most standard pronunciation, they will transcribe "*b'jour* (h'llo)". Although it is very close to the reality and useful for analyses or automatic alignment, it might pose issues if a researcher wants to use NLP process on the data because "b'jour" will not be recognized as "bonjour". There is a simple way to get around this feature by using two separates tiers: one for alternative spelling and another for standard orthography, but most part of the time the transcriber feels more at ease if he only works in alternative spelling to transcribe oral phenomena like overlaps or prosody, then alternative spelling constitutes the main level and tokens in standard orthography are generate automatically and linked to their corresponding alternative forms. In TEI, we choose a solution based on `<choice>` element including `<orig>` for alternative spelling and `<reg>` for standard orthography.

```
<choice><orig>b'jour</orig><reg>bonjour</reg></choice>
```

Structuring the annotation

Annotation might contain some main information (i.e. the utterance, which is contained in a `<u>` tag) and some free dependent information that is contained in one or several `<spanGrp>`. All this information is contained within a `<annotationBlock>` tag. The annotationBlock item bears all information that enables it to be identified in the transcript (time, position, locator). This respects the recommendations issued by the ISO TEI group.

Our proposal for `<spanGrp>` and `<span>` extends and specifies the use of these two tags but within the legal format of the TEI. Direct inclusion of a `<span>` within a `<span>` is not authorized. To do this, an intermediary `<spanGrp>` has to be used (see example below). Thus, the hierarchy should be: `<spanGrp>` contains `<span>` that can contain `<spanGrp>` that contains `<span>`, etc. All `<spanGrp>` tags have to be described using the `@type` attribute. The reason for this organisation is that it enables a researcher to be faithful to the information produced by tools such as ELAN, ANVIL and even Praat, where multiple levels of structure can correspond to different types of data. In these tools, data organisation is well described and organised. This information is available through the `@type` attributes of the `<spanGrp>` elements. It is described more fully in the template section (5b) below.

There are two types of `<span>`. The most usual one uses the `@from` and `@to` attributes to express how it is aligned with the temporal information of the timeline. The less usual one uses only the `@target` attribute that refers to the element above in the hierarchy. This corresponds to what is called symbolic subdivision or association in ELAN and ANVIL. In this case (see below the example "Letters" and compare with the example "Phones"), the elements are divided in equal parts and there is no time linking. The attribute `@target` is slightly redundant as it could be recomputed online using the structural information of XML, but this enables a better clarity of the data and also it differentiates a `<span>` with no `@from` and `@to` information – which is a time aligned `<span>` with the same extend as its parent – from a `<span>` with a `@target` which is a symbolic `<span>`.

```
<annotationBlock end="#T2" start="#T1" who="MainLine" xml:id="a1">
 <u>
  <seg>this is it!</seg>
 </u>
<spanGrp type="Syntax">
  <span target="#a1" xml:id="a9">Demonstrative
       <spanGrp type="SyntaxInformation">
     <span target="#a9" xml:id="a16">Anaphora</span>
       </spanGrp>
  </span>
  <span target="#a1" xml:id="a10">Copula</span>
  <span target="#a1" xml:id="a11">Pronoun</span>
</spanGrp>
<spanGrp type="Words">
  <span from="#T1" to="#T3" xml:id="a2">this
       <spanGrp type="Letters">
     <span target="#a2" xml:id="a17">t</span>
        <span target="#a2" xml:id="a18">h</span>
        <span target="#a2" xml:id="a19">i</span>
        <span target="#a2" xml:id="a20">s</span>
       </spanGrp>
       <spanGrp type="Phones">
     <span from="#T1" to="#T4" xml:id="a5">ð</span>
     <span from="#T4" to="#T5" xml:id="a6">ɪ
                <spanGrp type="PhonesInformation">
        <span target="#a6" xml:id="a8">nucleus</span>
                </spanGrp>
     </span>
     <span from="#T3" to="#T5" xml:id="a7">s</span>
       </spanGrp>
  </span>
  <span from="#T3" to="#T6" xml:id="a3">is</span>
  <span from="#T3" to="#T6" xml:id="a3">it</span>
</spanGrp>
</annotationBlock>
```

Example 5: <annotationBlock>and <spanGrp> elements to represent annotations layout

If oral corpora share certain specificities concerning the transcription and the annotation of the data, they also require some specific metadata presented in the next section where we make a distinction between the already existing TEI elements and the gaps we have identified.

4. Metadata in TEI Header

In order to manage storage, interrogation, diffusion and interoperability of corpora, information contained in metadata is crucial. According to the kind of analysis a researcher plans to conduct, it is particularly difficult for him to assess whether a given corpus (which he knows nothing about), matches the situation he would like to study, or not. Indeed, he needs technical information about the sound or the video quality if he wants to study prosody or gestures, information about the project goal, information about the setting to the context of the oral production, information about the speakers if he is interested in language acquisition or

sociolinguistic studies, information about access rights if latterly he would like to make available a long excerpt of audio or video recording and of course information about the annotations such as transcription conventions or the annotation scheme used in the textual representation of the recording. Bearing this in mind, the TEI format has the added advantage of providing in a single file the metadata and the transcript and offering a large set of elements and attributes that are sub-divided into different topics to describe oral data in the TEI Header and to represent oral verbal and non-verbal phenomena in the TEI body.

If the TEI-P5 standard offers a large set of possibilities to encode metadata, we have noted some elements or attributes that are missing. In the following section, we will try to identify recommended elements and to point out gaps at the end of each section inside the "Identified gap" subsection.

This section will be divided into four subparts: the nature of the primary data, the corpora categorization, the information about participants/speakers and the access rights. In each of these subparts, we wish to focus both on existing solutions and gaps. Indeed, we have identified some aspects that, from our point of view, seem to be missing in order to facilitate corpora exchange between researchers and to enable better interoperability for language data.

4.a. Metadata concerning primary data

Mostly, metadata concerning primary data are contained in the element `<sourceDesc>` and, principally, in the element `<recording>` (see example 6 below). The authorized elements especially allow a researcher to indicate the type of the source (audio or video for example) and the recording duration. Inside this element, three other types of information that are particularly important can be indicated:

- the data collection date: `<date>` element;
- the material used to collect the primary data: `<equipment>` element;
- the location where access to the primary data is available (for example an URL) as well as information about the media type (MIME type, like “audio/mpeg” or “video/mp4” for example): `<media>` element.

In the case of multiple sources for a unique TEI file, a subset of `<media>` elements can be used to encode information about each source. Let us imagine that we have collected two sources of information: a video file with a camera (in order to study speaker movements, for example) and an audio file with a lapel microphone (in order to study speaker prosody for example). Our recording is then divided in two media, as in the following example:

```
<sourceDesc>
 <recordingStmt>
  <recording dur="P32M" xml:id="example">
    <date when="2016-03-12">March 12th, 2016</date>
    <media dur="P32M" xml:id="example-mp4" url="website/example.mp4"
        mimeType="video/mp4" corresp="example-mp3"/>
    <media dur="P32M" xml:id="example-mp3" url="website/example.mp3"
        mimeType="audio/mp3" corresp="example-mp4"/>
   </recording>
  <recording>
   <equipment>
    <p xml:lang="en-GB">For these recordings, we used a digital recorder Marantz
     PMD660. This is a recorder with an omnidirectional microphone.</p>
   </equipment>
  </recording>
 </recordingStmt>
</sourceDesc>
```

Example 6:Primary data

Identified gaps

Some different media of the same recording could have different quality levels with high quality for gesture studies and low quality for faster download. Thus, a quality attribute could be added to the media element. Another gap that we have noted is the lack of a dedicated element or attribute to describe the anonymization process. Indeed, oral and multimodal recordings are often anonymized in order to distribute the corpora on the Internet or to simply share data between researchers. Although audio signals in such types of data are frequently modified, there is no way to formally encode this kind of information in metadata which is crucial to prosody or gesture studies.

4.b. Metadata concerning the corpora categorization

A large set of elements are available to categorize linguistic corpora. This categorization can be made on different levels and enables the encoding of important information necessary to define the corpus type, the way in which the corpus has been structured and its availability. To extend interoperability to the largest community possible, some metadata could be translated into several languages using the `@xml:lang` attribute like `<setting>` or `<catDesc>` in the following examples.

Concerning corpus categorization and structuration, the `<profileDesc>` element regroups some essential information including the date of creation, for example. The `<textDesc>` element provides complementary information, either as text (see `<channel>` element for example) or as arguments, such as the elements `<derivation>`, `<domain>`, `<factuality>`, `<interaction>`, `<preparedness>` and `<purpose>`. It is interesting to note that in the case of a set of linguistic corpora grouped into a single TEI file (inside a `<teiCorpus>` element), the previously mentioned elements bring together the information set that is common to all of the recordings made.

Following the `<textDesc>` element, the `<settingDesc>` element allows a user to encode complementary information concerning the oral situation. Inside the `<activity>` element, it is possible to include a `<figure>` element to represent artefacts manipulated during a meeting and a `<date>` element to indicate when a set of meetings or courses have been collected.

```
<settingDesc>
 <place type="country|city">
  <placeName type="country">France, ISO3166-1</placeName>
  <placeName type="city">Clermont-Ferrand, TGN7008202</placeName>
 </place>
 <setting xml:lang="fr-FR">
  <activity>Cet enregistrement se déroule au domicile des parents à <settlement>Clermont-Ferrand, France</settlement>. Le garçon joue avec son père avec des outils pour enfant (vis, marteau...).</activity>
  <locale>Dans la chambre de l'enfant</locale>
 </setting>
 <setting xml:lang="en-GB">
  <activity>This recording takes place at the family's home, at Clermont-Ferrand. The father and his child play with plastic screws and hammer.</activity>
  <locale>In the child's bedroom</locale>
 </setting>
 </settingDesc>
```

Example 7: Setting

If several corpora that use several data sources are gathered within the framework of a single research project, it could be useful to establish a classification with a set of criteria rather than using free textual fields to sort or select corpora. We could then define them with elements `<taxonomy>` `<category>` `<catDesc>` and use them inside `<encodingDesc>` with `<classDecl>` and `<catRef>` elements.

```
<encodingDesc>
 <classDecl>
  <catRef scheme="#RecordingQuality"
          Target="#noisy #sparsely_noisy #anechoic_chamber"/>
  <taxonomy xml:id="RecordingQuality">
  <category xml:id:="RecordingQuality.noisy">
          <catDesc xml:lang="en">noisy, disturb understanding</catDesc>
          <catDesc xml:lang="fr">bruité, gène la compréhension</catDesc>
  </category>
  <category xml:id="RecordingQuality.sparsely_noisy">
          <catDesc xml:lang="en">less than 5% inaudible</catDesc>
          <catDesc xml:lang="fr">moins de 5% inaudible</catDesc>
  </category>
  <category xml:id="RecordingQuality.anechoic_chamber">
          <catDesc xml:lang="en"> anechoic_chamber </catDesc>
          <catDesc xml:lang="fr">chambre sourde</catDesc>
  </category>
  </taxonomy>
 </classDecl>
```

Example 8:Classification

4.c. Metadata concerning participants/speakers

Metadata concerning participants/speakers are particularly important. Indeed, in a linguistic data exchange context, it is essential to be able to consult this kind of information,

especially when the study concerns issues related to Sociolinguistics and First or Second Language Acquisition.

With this in mind, the <particDesc> element can provide a set of particularly interesting solutions. Grouped together in the <listPerson> element, information concerning each participant/speaker is divided into categories. Sometimes, information is contained in very basic categories (like <persName> or <birth>, for example) whereas at other times the information is organized into several topics in order to provide further details about speakers. In the example below, this is the case for <residence> elements. Through repetition of the element, it is possible to indicate the set of the different locations of residence for each participant, as well as the duration of each period of residence. This kind of information is very important for sociolinguistic studies about regional accent, for example.

```
<particDesc>
 <listPerson>
  <person xml:id="MOT-Baptiste" sex="2" role="Mother">
   <persName>
    <name>Celine</name>
   </persName>
   <birth when="1978">
    <name type="place">Clermont-Ferrand, France</name>
   </birth>
   <langKnowledge>
    <langKnown tag="fr" level="first">French</langKnown>
   </langKnowledge>
   <residence notAfter="1999">
    <name type="place">Clermont-Ferrand, Auvergne, France</name>
   </residence>
   <residence notBefore="1999">
    <name type="place">Tours, Centre, France</name>
   </residence>
   <residence notBefore="2002">
    <name type="place">Paris, Ile-de-France, France</name>
   </residence>
   <residence notBefore="2006">
    <name type="place">Issy-les-Moulineaux, Ile-de-France, France</name>
   </residence>
   <residence notBefore="2008">
    <name type="place">Chamalières, Auvergne, France</name>
   </residence>
   <residence notBefore="2010">
    <name type="place">Royat, Auvergne, France</name>
   </residence>
   <education>Bachelor (niveau bac+5)</education>
   <occupation>Consolidation manager</occupation>
   <socecStatus>A standard is used to categorize socio-economical status: the IPSE score,
« Indice de Position SocioEconomique » (IPSE)<ref type="url">
http://www.unifr.ch/ipg/assets/files/DocGenoud/IPSE_manuel.pdf)</ref>
The score is = 77, middle/upper class</socecStatus>
  </person>
  <person xml:id="FAT-Baptiste" sex="1" role="Father">
    <!-- ... -->
  </person>
  </listPerson>
</particDesc>
```

Example 9: Speakers

If it is not possible to give such precise information in case of public situations, then we can describe them with `<personGrp>` element to gather metadata:

```
<particDesc>
 <listPerson>
  <personGrp sex="mixed" age="twenty" size="30" role="course"/>
```

```
  </listPerson>
</particDesc>
```

Example 10: Speakers as a group

Identified gaps

We have noticed some gaps in <person> section where it could be interesting to describe in a <learning> element where the non-native speaker has learnt the language, for instance:

```
<learning>
  <date from="start-time" to="end-time">a comment on the period</date>
  <settlement>a comment on the area where the French language was known</settlement>
  <community>a comment on the community: school, sport, neighborhood</community>
</learning>
```

Example 11: More information on Language knowledge

4.d. Metadata concerning access rights

Some elements like <availability> and <licence> are available in <publicationStmt> section which allow an URL link to explain rights:

```
<publicationStmt>
  <availability status="free|unkown|restricted">
    <licence target=" url explaining rights">Distributed with restrictions [...]</licence>
  </availability>
</publicationStmt>
```

Example 12:Access rights

5. Aligned transcripts and oral events representation in TEI

5.a. Usual features for oral and multimodal phenomena

If the TEI Guidelines chapter dedicated to the transcription of speech presents numerous solutions to annotate linguistic phenomena (see Chapter 8 of the TEI Guidelines), we will propose a subset of common elements to match main usual needs concerning verbal and non-verbal events.

Pauses

A pause could be identified by a @duration or a @type (short/long/..). If it happened outside an <utterance>, a timeline interval is required with **@start** and **@end** attributes and an additional **@rend** attribute is available and optional to store its original transcribed form (see examples 13 and 14 below).

```
<pause dur="PT0.61S" start="#T2" end="#T3"/>
```

Example 13: A standalone pause

```
<u><w>bon</w><pause rend="(..)" type="short"/> …. </u>
```

Example 14:A pause inside an utterance

Incident, Vocal, Kinesic

These three elements permit to encode non-verbal productions:

- **<incident>** is dedicated to an external event which disturbs the recording and is not attributed to a speaker but could happen during a speaker production. It can be encoded inside an <utterance>, or inside an <annotationBlock>, or at the same level than <**annotationBlock**> with mandatories **@start** and **@end** properties;
- **<kinesic>** and **<vocal>** elements are attributed to a speaker and can be encode inside an >utterance>, inside an <**annotationBlock**> with **@start** and **@end** attributes, or outside <**annotationBlock**> with **@who, @start** and **@end** attributes.

```
<u who="#A" start"#T1" end="#T2">
  <w>bon</w> <w>donc</w> <w>on</w> <w>a</w>
  <incident><desc>moving chairs</desc></incident>
</u>
```

Example 15: <incident> element inside an utterance

```
<annotationBlock who="#A" start"#T1" end="#T3">
  <u><w>bon</w> <w>donc</w><anchor synch="#T2"/><w>on</w> <w>a</w></u>
  <incident start="#T2" end="#T3"><desc>moving chairs</desc></incident>
<annotationBlock/>
```

Example 16: <incident> element at the same level than an utterance

```
<vocal who="#B" start="#T1" end="#T3">
  <desc>a longer whistle</desc>
</vocal>
```

Example 17: <vocal> element

Unclear, gap

In case of doubt when the recording is more or less audible:

```
<unclear><w>quand</w><w>même</w></unclear>
```

Example 18: <unclear> element

Sometimes, the doubt implies the transcription of two alternative utterances:

```
<unclear>
  <choice>
    <seg exclude="#u1-b" xml:id="u1-a">
      <w>quand</w>
      <w>même</w>
    </seg>
    <seg exclude="#u1-a" xml:id="u1-b">
      <w>bien</w>
      <w>même</w>
    </seg>
  </choice>
</unclear>
```

Example 19:Using <unclear> element to represent alternative transcriptions

In the case of a noisy recording, some words may not be recognized; the **@extent** attribute allows a user to give information about the length of the inaudibility using syllables as the measure, or more precisely with duration:

```
<w>des</w><w>expériences</w><gap extent="3 sylls"/>
```

```
<w>des</w><w>expériences</w><gap dur="PT1.5S"/>
```

Example 20: a gap in the transcription

An optional **@reason** attribute could be used with values like "external noise" which avoids to encode an **<incident>** element to explain this gap. As usual, **<gap>** could stand inside an <**utterance**> element, inside or outside an <**annotationBlock**> with **@start** and **@end** attributes if outside.

5.b. Complementary features for oral and multimodal transcripts

If the TEI Guidelines chapter dedicated to the transcription of speech presents numerous solutions to annotate linguistic phenomena (see 5.a.), some phenomena cannot be annotated using the pre-existing elements. In this case, TEI offers two possibilities, which we present in this section. The first is a Features Structure declaration (see chapter 18 of the TEI Guidelines). The second is a controlled vocabulary definition.

Features Structure

The use of a Features Structure requires a prior definition in the teiHeader section. This definition is particularly important at two levels. First, the features used are controlled, in the sense that the annotator can use only the values declared in the teiHeader. Secondly, the presence inside the file of the precise definition of the features structure allows for better interoperability and favors data exchange between researchers. Indeed, the annotation that is then added to the corpus will be easier to understand for a researcher who retrieves some data that he did not collect himself. Furthermore, the fact that the definition is present in the metadata makes the structure of the feature interpretable by NLP software such as a corpora conversion program, for example.

The definition of the features structure used in a corpus is made inside the `<fsDecl>` element (in the `<encodingDesc>` element) by means of a gloss describing the phenomena annotated and a list of the features used to account for the phenomena. Each of the features used is defined by an `<fDecl>` element, which also contains a gloss allowing a user to characterize the feature. All the values supported by the annotation feature are listed inside this element. Below, we present one example that is the annotation of the speech addressee. This is currently not possible to include using an existing element (see examples above). As indicated in the gloss, this annotation aims to characterize an utterance indicating that it is addressed to a child (Child Directed Speech, CDS) or to an adult (Adult Directed Speech, ADS). This information, often present in child-parents interaction corpora, is very important in order to study input effect, for example.

```
<fsDecl type="addressee">
  <fsDescr>The goal of this annotation is to indicate speech addressee, and more precisely if the parent talks to her child (Child Directed Speech, CDS) or to the other parent (Adult Directed Speech, ADS).</fsDescr>
  <fDecl name="Target" xml:id="Tar">
  <fDescr>The target feature is used to indicate the person the speaker is talking to. Three values are used: “CHI”, when the mother or the father talks to her child ; “FAT”, when the mother talks to her husband ; “MOT”, when the father talks to his wife.</fDescr>
   <vRange>
    <vAlt>
    <symbol value="CHI"/>
    <symbol value="FAT"/>
    <symbol value="MOT"/>
    </vAlt>
   </vRange>
   </fDecl>
  </fsDecl>
```

Example 21: A features structure in metadata

Then, the features structure defined in the metadata can be used in the transcript in order to express the phenomena described. We present above an extract from the ALIPE corpus using the features structure defined to annotate speech addressee.

```
<u who="#FAT-Baptiste" xml:id="u3-ali-baptiste-101227-1">
  <w>qu’est-ce</w>
  <w>que</w>
  <w>tu</w>
  <w>cherches</w>
  <w>Baptiste</w>
  <fs type="addressee">
    <f name="target" fVal="CHI"/>
  </fs>
</u>
```

Example 22:Using this features structure in data

As presented above (see 3.a.), oral and multimodal transcriptions call for much broader description and transcription of the data. It is not possible to include in the TEI precise semantics such as orthographic transcripts, syntactic organization, etc. This means that the organisation of the transcript has to be registered within the transcript and not in the general guidelines of the tool or the standard. This was indeed done by most people that used tools such as ELAN and ANVIL, and these tools already provide ways to organize and structure the data. In some extreme cases, the organisation of the data is the only piece of information provided, and this forms the basis to create new data.

Template

We have used two elements to code the data generated by the ELAN and ANVIL software. The first element is termed a “template”: a description of the hierarchical organisation of the data. Each element in the data (`<spanGrp>`) is referenced by several pieces of information: code, parent (which means that the “name information” is exactly one rank below the “parent information” in the hierarchy), and type, which correspond to the organisation of the relationship between the element and the parent. It is possible to add information, such as the reference to a controlled vocabulary (see below). Currently, five

types of relationships exist, which can be temporal or structural. These five types allow the information available in ELAN, ANVIL and even Praat to be covered. As no specific structure in the TEI P5 corresponds to this type of information, for the moment it is included in the notes section. Supplementary information is available when necessary, such as the name of a controlled vocabulary (see @type='cv' below).

```
<fileDesc>
 <notesStmt>
  <note type="TEMPLATE_DESC">
   <note>
    <note type="code">MainLine</note>
    <note type="parent">-</note>
    <note type="type">-</note>
   </note>
   <note>
    <note type="code">Words</note>
    <note type="parent">MainLine</note>
    <note type="type">Time_Subdivision</note>
   </note>
   <note>
    <note type="code">Phones</note>
    <note type="parent">Words</note>
    <note type="type">Time_Subdivision</note>
    <note type="cv">Phonemes</note>
   </note>
  </note>
 </notesStmt>
</fileDesc>
```

Example 23:A software template

Controlled Vocabulary

Coding for scientific purposes has to be carefully controlled. As the information provided in oral and multimodal corpora is not limited to standard information such as orthographic transcript, or even IPA signs, this information has to be controlled within the transcript. This is what is done with the use of vocabularies such as the ones provided in ELAN. These vocabularies are produced during the transcription process and so they have to be included in the standard. This addition was facilitated by the presence of adequate structures in the TEI P5, such as the `<keywords>` structure.

```
<profileDesc>
 <textClass>
  <keywords type="" xml:id="VC-Syllabes">
   <term type="">CV</term>
   <term type="">VC</term>
   <term type="">CVC</term>
  </keywords>
 </textClass>
</profileDesc>
```

Example 24: Controlled vocabulary

Conclusion

As more and more projects gather existing corpora from different databanks and from different research teams and then, in return, produce new annotations, oral and multimodal communities need to deliver a common format proposal to organize and make available both data (i.e. transcripts and annotations) and metadata. For the TEI standard to match this need, our community had to choose a subset of TEI elements and attributes, fully documented with relevant examples, taking into account the current practices of researchers. This was presented in the current article. When necessary, we underlined current gaps in the TEI format and made some proposals to encode information that, currently, could not be structured using the TEI. Our main goal was to present our thought process which led us to a solution for the structuring, the storing and the sharing of language corpora using a standard, open and interoperable format. To deal with this ambitious aim, we benefited from collaborations between the IRCOM consortium and the ORTOLANG infrastructure that is linked to the Huma-Num Institute and from sharing our work within the ISO-TEI European group.

The aim of our workgroup also had an applied objective. Indeed, the solutions presented above were not only proposed for structuring, sharing and storing corpora, but also to manage interoperability between different applications used in our community. Thus, the format that we present has been implemented and tested in a conversion tool written in Java with the support of the ORTOLANG project. The tool has two main goals: providing a way to easily create a common representation format for the corpora stored within the ORTOLANG portal (www.ortolang.fr); and allowing different usages of the same corpus whatever software may be used without any knowledge of TEI.

The conversion tool can handle files coming from CLAN, ELAN, Praat and Transcriber. Care was taken to allow, as best as possible, backward conversion from the TEI to the original format. However, in some cases, some information might be lost, especially when programs implement some very specific features that are not shared by other tools. The conversion tools are available on the ORTOLANG web site at ct3.ortolang.fr/tei-corpo/. The conversion tool is open source and is available under a BSD-2 license. It is possible to use this tool to export data to textometric tools such as TXM (Heiden, 2010) or Lexico/Le Trameur (Lamalle et al., 2003 ; Fleury, 2009).

A complementary online tool is available at ct3.ortolang.fr/teiconvert, which provides other features such as conversion from and to Text format (Unicode), Office Open XML format used in text and spreadsheet processors.

The use of this tool is rather new so it is difficult to find yet direct applications that are already publically available. An example of unifying use of the format is provided at ct3.ortolang.fr. All the data in this site are available in TEI_CORPO format whatever their origin. The full text query tool of this website covers all the corpora and use directly the TEI format. Online corpus browsing is also available and based on the single TEI_CORPO format.

The solutions developed in this article would maybe concern other research communities working in the same way with recordings as primary data. Our work may also help to deal with the same kinds of metadata that can be of use in social or historical studies.

Bibliography

Barras, C., Geo, E. and Wu, Z. 2001). Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production. *Speech Communication* (33).: 5–22.

Boersma, P. and Weenink, D. 2013. Praat: doing phonetics by computer. [Software].

Crasborn, O. and Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In *Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation*.

Fleury, S. (2009) Le Trameur. [Software].

Heiden, S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In the proceedings of *the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC24)*, edited by K. I. Ryo, 389-398.

Kipp, M. (2001). Anvil: A Generic Annotation Tool for Multimodal Dialogue. In the *Proceedings of the 7th European Conference on Speech Communication and Technology*, 1367–1370.

Lamalle, C., Martinez, W., Fleury, S., & Salem, A. 2003. Lexico 3: Outils de statistique textuelle. SYLED-CLA2T, Université de la Sorbonne nouvelle-Paris. [Software].

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Schmidt, T. 2011. A TEI-based approach to standardising spoken language transcription. *Journal of the Text Encoding Initiative*, Issue 1.

TEI Consortium. 2013. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 2.9.1. Last updated October 15th. N.p.: TEI Consortium. http://www.tei-c.org/Vault/P5/2.5.0/doc/tei-p5-doc/en/html/.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. & Sloetjes, H. 2006. ELAN: a Professional Framework for Multimodality Research. In the *Proceedings of the Fifth International conference on Language Resources and Evaluation*, 1556–1559.

Authors

Loïc Liégeois: LLF/CLILLAC-ARP, Université Paris Diderot, SPC, Paris, France

loic.liegeois@univ-paris-diderot.fr

Loïc Liégeois holds a PhD in Linguistics from the University of Clermont-Ferrand. He is currently employed as a research engineer at the University of Paris 7 - Paris Diderot and his research interests include corpus linguistic, language acquisition, cognitive linguistic and phonology.

Carole Etienne: ICAR, CNRS, Lyon, France

carole.etienne@ens-lyon.fr

Carole Etienne is a research engineer at the ICAR laboratory in Lyon. Her research interests include spoken langage corpora, multimedia database and interoperability..

Christophe Benzitoun: ATILF, CNRS, Nancy, Université de Lorraine, France

Christophe.Benzitoun@univ-lorraine.fr

Christophe Benzitoun holds a PhD in French Linguistics from the University of Provence and he’s a lecturer at Université de Lorraine. His research interests include spoken language corpora, syntax and clause linkage.

Christophe Parisse: Modyco/INSERM, CNRS, Université Paris Ouest Nanterre, France

cparisse@u-paris10.fr

Christophe Parisse is a full-time researcher at INSERM. He works on linguistic research concerning oral language and language development. His research is rooted in Corpus and Cognitive Linguistics.