Article sur TEIMETA

TeiMeta : an evolutive solution to describe oral corpora in a web interface generated from an ODD file

Abstract : The french consortium CORLI decide to promote an evolutive solution to describe oral corpora based on TEI using ODD file to deal with further changes and some specificities linked to some linguistic fields.

Keywords : oral corpora;metadata;linguistics;TEI;ODD;web;

**Needs concerning oral corpora metadata**

These last ten years, many french oral corpora became available in dedicated archives like Ortolang to be share inside research community in linguistics or in other research field. But researchers don't feel at ease to reuse these oral data for a given study because they arent't really sure of the access rights, they misunderstand the setting, they lack information on the speakers even basic descriptors like adult /child or native / non native speakers, they have questions concerning the quality of the audio or video recordings if they need to process phonological or prosodical treatements and they don't really know which annotations are already available on these data. We think then that providing a relevant subset of metadata for oral corpora will make our corpora findable, reusable and interoperable according to the FAIR (Findable Accessible Interoperable Reusable) principles as they are already accessible in our national archives..

First, CORLI consortium organise a roundtable with the main corpora providers in linguistics to collect their current practices and their needs on metadata. It appears clearly that corpora responsibles don't feel at ease with access rights and deliver metadata mainly for archive purpose using Dublin Core/OLAC descriptors which aren't enough precise for researchers needs. Thus, they find solutions adapted to each project where some relevant metadata are missing because they are included in the project itself or not necessary. For example, CFPP collect parisians audio recordings then they don't precise in each recording that it is audio source, collected in Paris with adults speakers living in Paris. Corpus gathering a set of heteregeneous settings like TCOF or CLAPI aren't in this case and needs metadata to deal with the diversity of the settings and speakers they have recorded but unfortunately they have made different choices as regard their descriptors then someone interested in both corpora has to deal with the two sets of metadata to select the most interesting settings or speakers. Another case concerns acquisition field where people used to store their corpora inside CHILDES databank specialised in acquisition then they used the same subset of metadata, available in Transcriber software, well adapted for acquisition purpose but which lacks information on the recording, the citation or XX

If we want to share the same subset of metadata, we will need guided vocabulary to fix some values in ordr to permit to use these metadata as search criteria.

All the corpora responsibles will appreciate that a common set of metadata will be decided in our community to facilitate the constitution of new corpora and the reusability of oldest. As regard standards, they think that these metadata have to be deliver in an international standard and they expected that this task could be make inside CORLI instead of spending time to do it in each lab.

Thus, we decided to create a workgroup working on metadata inside CORLI . This workgroup choose TEI as the standard because it is already use widely for writting corpora, the documentation is important with many examples for each element and the community reactive, evolutions go on in ISO-TEI workgroups answering to the new needs and moreover TEI permits to gather in the same file metadata in TeiHeader part and the transcript in text element.

**Metadata subset**

**1 Citation of the source and contributors**

A recurrent request from the corpora providers concerns the citation of the source which is missing in many articles, they like to take in consideration this point encouraging relecteurs to inform the authors. On the other end, when a corpora is not available on web site it is sometimes difficult to find out this information. In the same approach, all the contributors like collectors, transcribers, translators or people engaged in technical treatements could be cited .

**2 Corpus included in other projects, derivation or translation of other data**

In a perspective of reusability of corpora, we have to precise if a corpus is included in the main french projects like Rhapsody or CEFC corpus from Orféo ANR project to avoid to include twice this data. In the same way, we could precise if a recording is an excerpt of a longer recording with for example a precise transcript of gestures and gazes or phonological annotations concerning only this excerpt and not the whole recording.

**3 Setting**

The setting needs to be described using a common set of criteria established by our community to deal with the variety of the collected situations, as regard oral corpora we need in fact several metadata to identify the following informations :

* the domain of the setting between private or professionnal where professional means that at least one speaker is working for example when you buy some bread in a bakery the baker is working.
* the nature of the setting between interview, reading, meeeting, table talk, medical consultation, conversation, games, …
* the canal of the communication like phone call, visioconference or face-to-face if most part of the speakers, even all, are in the same place
* the context : academic, sport, association, amical, familial, commercial, …
* instructions : an additional information could be useful to precise if an interview is more or less guided by the interviewer

Another information concerning setting is the place where it happens, and of course the suburb, town, area or country and the date of the collect which could be interesting for diachronic purposes or for sociological dimension to study how an event is interpreted by the speakers.

**4 Speakers**

The nned concerning speakers is dirrectly linked to teh research field:

* in acquisition, we get interested by the precise age of the child, his scholarisation, the languages and the socioeconomic informations on their parents
* in sociolinguistic, we get interested by many informations concerning the speakers : their following occupations, their residences, the place where they learn French, their study levels, …
* in oral syntax, informations on the level of studies and knowledges of languages are necessary
* in interaction, we get interested by the relationship between the speakers, their age and their knowledges concerning languages but in some cases when we recording a group of speakers (clients in a shop, visitors, public, scholars in classroom, …) it's difficult to ask these informations before or after the recording.

Nevertheless, some informations are easier to give like the fact that speakers are adult or children, or maybe even a range of ages if we don't know the precie age, as regard knowledges in language we couldindicate if the speaker is native

**5 Recordings**

Some elementary informations on the format, the audio or video signal and the exact duration are necessary.

A problematic metadata concerns the quality of the recordings which depends on the research field a middle quality in prosodic or phonologic studies will certainly be better than in interaction where we used to record speakers in the place where the setting took place with noises linked to the activity (moving objects, noises in classroom, external noises, …). Then we will evaluate the quality on the basis of a criteria which could be the percent of the noise in the whole recording instead of just an evaluative value like bad/ middle / good .

**6 Access right**

As regard interoperability, this information is mandatory to reuse the data in research or learning environment. Corli consortium recommanded the usage of Creative Common License which are well known and fully described, and particularly the License 4.0 which permits to use the data for a non-commercial use, to modify the data to build excerpts for example, to diffuse the data in the same conditions which are to cite the source. It matches the needs of people who want to reuse data to create new annotations and make them available and it respects the preoccupation of the corpora responsibles to have their source cited.

**7 Annotations**

It is particularly difficult even if we know a corpus to identify all the annotations which are available concerning different perspectives (lemma, PoS, macrosyntactic, phonology, prosody, gestures or gazes, interaction, …) then it will be interesting to indicate in the same file all the available annotations and the convention or tools (automatic tools) which generate these annotations.

**Implementation of these metadata subset in TEI**

As the subset is more or less defined, a second step concerns its representation in TEI which has been thought to define text and not oral recording.

We will answered on the main metadata described in the previous part :

Je pourrais le remplir en rentrant mi mai et on fera des aller-retours avec Loic sur le choix des élements/attributs, ça te va Loic ?

**ODD file**

As oral corpora is a specific field far from writting texts, we think that defining a common ODD file will be interesting to precise the TEI elements and attributes we used, the articulation of these elements to provide the final teiHeader and guided vocabulary to fix the value of some criteria.

ODD specification matches this aim and will permit to illustrate somme elemnts with examples to define different levels of granularity.

**A web interface adapted to ODD file**

As researchers don't feel at ease with the installation of softwrae depending on the release of java on their labtop, we propose a web interface where people can choose their own ODD file or use the common CORLI ODD file.

**Generalisation to other ODD file and other purpose than oral corpora**

Christophe