

A multi-software integration platform and support for multimedia transcripts of language

Christophe Parisse* and Aliyah Morgenstern**

*Modyco, Inserm, CNRS/Paris Ouest Nanterre La Défense University

**Prismes, Paris III Sorbonne Nouvelle University

200 av de la République, 92001 Nanterre cedex, FRANCE

E-mail: cparisse@u-paris10.fr, aliyah.morgenstern@gmail.com

Abstract

Using and sharing multimedia corpora is a vital feature for research about language, but the number of different and often not easily compatible tools available makes this difficult to do. As the aims of the COLAJE project are to use multimodal linguistic data about language development in oral and sign languages, it was necessary to create a system (VICLO) that allowed sharing and using data coming from at least three different sources Clan (CHILDES), Elan (MPI) and Praat (U. of Amsterdam). For this reason, a multi-purpose storage format based on the TEI was created, which allowed us to store information coming from all (these) origins, and include every type of specific information. When part of the information is processed by a specific software, the changes are integrated later in the system without losing information specific to other software. Thus it is possible to store information shared and not shared between the different corpus editing tools. This common base allowed us to implement complementary features such as fine-grained participant and metadata information, common visualisation and data-retrieval tools. VICLO is based on XML technology and all data can be displayed using all purpose web browsers.

1. Introduction

Using and sharing multimedia data and transcripts is a vital feature for research and applications on language, especially those based on conversation analysis as well as pragmatic and semantic analyses. Recent advances in the use of video media, speed of computers and price of video-recording material have made it much easier to gather, describe and process corpora of language that include video and direct correspondence between transcription and video, what is usually called linking between transcription and video. The editing and linking process can be realised easily today thanks to a wide choice of freely available, multiplatform, and robust software such as CLAN (childes.psy.cmu.edu/clan/), ELAN (www.lat-mpi.eu/tools/elan/), PRAAT (www.fon.hum.uva.nl/praat/) and many others (Anvil, Exmaralda, Transcriber, Transana, etc. – see [http://icar.univ-](http://icar.univ-lyon2.fr/projets/corinte/confection/alignement.htm)

[lyon2.fr/projets/corinte/confection/alignement.htm](http://icar.univ-lyon2.fr/projets/corinte/confection/alignement.htm) for a more complete description of the tools). These tools are often open-source software, so it is reasonable to assume that the most commonly used ones will be maintained in the future by a large community of users and developers. This certainly helps people to invest into creating new video-linked corpora and databases.

These advances are highly useful for sign language and language acquisition, as in these domains using visual media along linguistic data is mandatory for different reasons. Sign language is obviously a visual medium of communication; for language acquisition it is virtually impossible to generate serious work about the semantic and pragmatic aspects of language interaction without visual support. The latter is also often mandatory to simply understand what very young children are saying because comprehension is very poor outside the (visual)

context. The same remarks would also apply to other fields of linguistic analyses, such as adult interaction.

A common feature between these two domains is that in both cases it is very difficult to design a single piece of software that would cover all the uses and needs of people doing research or using this material for education purposes. Another common feature is that corpus creation (recording, transcription, linking, editing) is very costly in term of human hours of work (but not in terms of material or software). Unfortunately these two common features clash (one with the other): the high costs would suggest that any data ever produced should be used and reused as much as possible whereas differences between applications make it difficult to reuse data that was produced initially using one piece of software only. For example, software such as Clan and Transcriber allow the coding of situational information (but they code it differently) and this information does not exist (yet) in Elan and Praat, so it would be lost during the conversion process. Another example is interdependence between tier levels: Elan offers a much more powerful package than other applications so this information will also be lost during conversion. A final and serious problem is the conversion between Single Timeline Multiple Tiers (STMT) organisation of data (used by Elan, Praat, Exmaralda, etc.) and Ordered Hierarchy of Content Objects (OHCO) data (used by Clan). When data created using OHCO software are converted into STMT software, elements which are not coded for time alignment (linking) in the first case have to be modified to be handled correctly in the second case. Backward conversion may not reproduce the original hierarchy.

For all these reasons, using more than one application for the same data is difficult. At first glance, as most type of applications have importing and exporting features, it

would seem that this is not a real problem. However, conversion is always performed on a common core basis. Only the features which are shared between two applications are converted. Other data are lost, so the use of multiple software is often a case of one-way conversion from one reference software (the tool the data was build with) towards another tool that has interesting complementary features but that is used only for some specific one-time feature. A good example of this procedure is conversion from Clan towards Praat. Either the “one utterance only conversion procedure” is used and the goal is only to analyse more finely this utterance with Praat, but any modification done within Praat cannot be converted back to Clan; either “the whole file procedure” is used, but then information about participants and sequences is lost so a conversion back to Clan will result in a very different data from the original one; so conversion back and forth between Clan and Praat is unlikely to happen..

2. A multi-software integration platform

We are facing a paradox: the difference between editing tools makes it difficult to use multiple tools; but this very difference is what makes it interesting to share data between tools as they have complementary features and qualities. This could also be considered an economic issue due to the cost and labor involved in corpus creation.

2.1. Goal

Our goal is to propose a solution to these limits by using a common repository which would not be based on core features of the data designed for all types of applications to be used, but on encompassing features. This means that the common format used contains recipients for all types of data for all tools, and that it is used as a pivot and common repository. This makes the preservation of specific software information possible. Data that is specific to a tool A and unused by others is kept in the repository so that it can be reintegrated when the rest of the data has been edited and modified by a tool B, and conversely. Such an integrative system offers advantages that go beyond data sharing. It makes it easier to create complementary features such as metadata and fine-grained descriptions of target participants’ behaviour because this data will benefit all corpora. It will allow us to integrate metadata from different origins, including for example OLAC (www.language-archives.org), Dublin Core (dublincore.org/documents/dcmi-terms/) or ISLE IMDI (www.mpi.nl/IMDI/). It makes it also possible to make new interrogations and to display features that could be used on data created by different tools.

The goal of the COLAJE project (financed by the ANR, France) is to create a functional platform, VICLO (French translation of Visualisation and Interrogation of Oral Language Corpora), that includes such features and is compatible with Clan, Elan and Praat and allows easy integration with other tools such as editing tools and computer linguistic tools. Compatibility with Clan

includes compatibility with the new CHILDES-XML format and the Talkbank project (www.talkbank.org). The VICLO platform is demonstrated on the COLAJE website (see www.modyco.fr/corpus/colaje/viclo/).

As the purpose of the project is to create a platform that is easy to use and to maintain, the technical solutions use only open-source and easy access software. Ready made data as visualised by the final user do not need any software installation since these data can be browsed through a web navigator such as Firefox, Safari or Chrome. Processing (converting and preparing for display) is implemented in XSLT as much as possible (any transformation uses XML data as a starting point) and in Perl for conversion starting from non-XML data. The format used for the repository is XML and is based on the TEI XML format.

3. Implementation issues

3.1. Common format

The choice of the Text Encoding Initiative (TEI) as a basis for the container format for all data is only natural as TEI is based on a reliable base (XML) and is a multi-purpose storage format for language corpora. Unfortunately, it had to be quite thoroughly extended for three reasons. First, conventions for the storage of oral language data are only general guidelines and many elements related to specific metadata and tier structured layers are not optimal. There are also issues about future implementation of structured data, decomposition into words and sub-lexical units, and description of syntactic information. Second, the multiple possibilities of various applications such as Clan, Elan and Praat were not included in the design of the TEI, although sometimes it is possible to redirect the initial purpose of some parts of the TEI (see below). Third, the TEI was obviously not designed to store data specific and software specific information, which is necessary to maintain the integrity of the original information in a software specific fashion.

3.2. General purpose additions to the TEI

Four additions were made to the TEI format, following the general structure of TEI data: participant information, tier information (especially the structural organisation of the tiers), specific vocabularies for the coding of specific tiers, and fine-grained information about participants and description of the recording session. These four types of information are stored in the description profile of the TEI header (see Table 1 above). This rich information is not part of the language corpus itself, but is of vital importance for scientific purposes because it provides information about the people involved, the coding features and the organisation of the data. This is some extended type of metadata, as demonstrated by the *textDesc* feature which is specific to the COLAJE project.

The participant information constitutes the main entry into the participant, tier and vocabulary data. Participants bear no relation one to another in the structure of the

corpus (even though they often have kinship relationships!). Participant information contains elements that are directly specific to the person involved in the corpus and is usually independent from the corpus collection purposes – age, sex, socio-economic status, etc.). Participant information is linked to tier information which contains the various levels of description of the language data: orthography, phonetics, gestures, prosody, gaze, situation, actions, etc. This information is open ended as there is no limit to what future research purposes may be. Constraints on the structure of a tier are possible through the use of vocabularies, a feature that is Elan specific (note that Clan has a similar feature but this was never translated into data constraint representations). Specific structure for specific tiers such as the orthographic and phonetic tiers is not yet implemented but may be included in the future when software such as CLAN-XML and Phon will offer this feature in their internal data, or when planned integration with lexicometric or language processing tools will be performed (see future improvements).

```
<teiHeader> ... <profileDesc> ... (TEI tags)
<participantStmnt> ...
  <!-- example generated from Elan -->
  <participant name="wit" longname="With or
without gaze" language="en" type="With or
without gaze" />
  <!-- example automatically generated from
Clan -->
  <participant name='chi' longname='Antoine'
type='participant' role='target_child'
age='2;04.03' birth='10-APR-2006' sex='male'
desc='description' />
  ...
<tierStmnt> ...
  <!-- example generated from Elan -->
  <tierDesc type="With or without gaze"
longname="With or without gaze"
parent="participant" vocid="With or without
gaze"
xgraphic_references="false"
align="true" />
  ...
<vocabularyStmnt>
  <vocabulary name="With or without gaze">
  <vocabularyDesc/>
  <token xml:id="Without gaze"> <tokenDesc/>
  ...
<textDesc>
  <sg name="saillant_features">
  <g name="motor">
  <v name="sitting_position" val="yes"></v>
  <v name="crawling" val="yes"></v>
```

Table 1 : Examples of TEI for Oral Corpus extensions

3.3. Coding of language data

Clan, Elan and Praat make different uses of the word *tier* because they have different underlying structures, OHCO for Clan and STMT for Elan and Praat. The TEI includes a mix of the two information structures but has no tier concept (neither in the sense described in 3.2 nor in the

STMT sense). It includes a ‘u’ concept which is an entry into a text part, which may or not correspond to the sentence or the turn. This concept is included along the concept of timeline which allows for representation of STMT formats. Time anchors allow to implement complex linking information using the TEI. They will be used to implement sequences of utterances. We chose to keep the concept of timeline as it was the easiest way to preserve data organisation for most oral language tools. This means that all elements in Clan need to be mapped onto a timeline point or aligned (using anchors) with other elements (this allows to time reference all elements when transformations from Clan to Elan are done). The main problem is that the mapping process is somewhat arbitrary because information about overlapping is not finely detailed in Clan, when it is described. Backward transformation is possible; however it is not yet implemented because, as Clan works quite well with overlapping timelines, this is not really an issue.

```
<u wh='chi' xml:id='id2' start='23.92'
end='24.357'>
  <tier type='ortho'>papi Michel .</tier>
  <tier type='pho'>papi mi el</tier>
  <tier type='sit'>GDF is sitting down on the
sofa</tier>
</u>
```

Table 2 : Example of TEI for Oral Corpus coding of text data

The ‘u’ format was kept because this was the TEI format but the ‘u’ for ‘utterance’ should be in fact changed to ‘e’ for ‘entry’ because nothing in the data format specifies that ‘u’ is an utterance. It can be a piece of any size of language that is produced by one speaker only. Tiers are all included in the main entry, but it is possible for individual tiers to have specific time linking, inside the duration of the main entry, which corresponds to the notion of constraint stereotype in Elan. The orthographic line is not included in the main entry, but as a separate tier, which frees the representation from a strict text oriented classical representation and makes the coding of multimodal non linguistic data possible.

3.4. Information about specific software

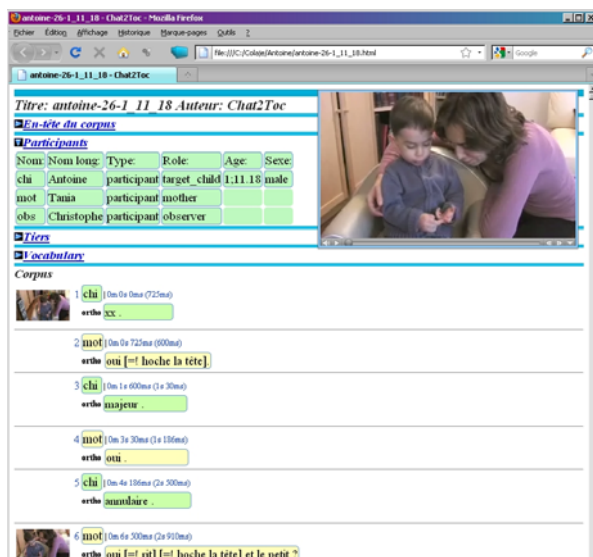
To keep specific software information in the data, and put it back when a file is changed by another type of software, it is necessary to send the changes made in a file. This is true for multi-purpose information such as the *textDesc* (see above) or more specific information about linking or tier structure. It is thus necessary to keep track of which software generated the data first, and which software it is used with later. Information about original filename, time of conversion, type and name of tools used, conversion of media software (for example from video to audio in the case of conversion from Elan to Praat) is kept in the multipurpose *notesStmnt* of the TEI header. Information related to external audio and video data files is kept in the *recordingStmnt* in the *sourcesDesc* part of the TEI header.

3.5. Editing text description data

Descriptions of the texts do not follow a fixed format common to all files, as it is usually the case for data formats, technical implementation details or even usual metadata which are normalised so as to make information available anywhere on the web. Descriptions consists for example of information about the age of acquisition of cognitive milestones for young children (when ~~did~~ they began to walk, for example) which is useful information for a researcher working on child language acquisition because child mobility may have an impact on their pragmatic contact with other adults. But this will be of no interest for people working on later acquisition of complex syntax. So the material to be edited and inserted in the corpus is prone to change. To this purpose, a specific PHP application was developed so that it was possible to set up the description and material to be edited using a single configuration file. Structure of XML data is the same for all possible descriptors as the specific information is stored only in the values of XML parameters and nodes, not the names of the nodes and parameters. A specific interrogation system for this data is under development. This interrogation will be automatically guided by the nature of the data stored in the descriptors.

3.6. Visualisation

Specific tools for visualisation of the data had to be developed in order to be able to display the specific features such as the text description (textDesc) which could not be displayed by any native tool because it is original data created in VICLO. However, having at our

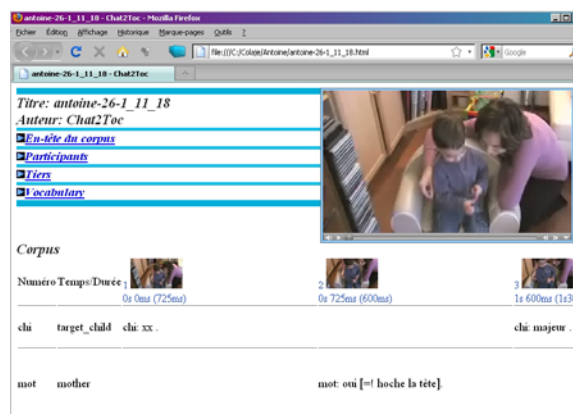


disposal several visualisation tools has other advantages. First, the data can be displayed without learning or installing a new editing tool. Cross-tools visualisation is made easier. Second, it is possible to create a large number of new display formats and respond more easily to specific requests because browser only software is easier to achieve and not constrained by editing needs. Finally, visualisation tools will become handy when one of the basic goal of the project, data interrogation will

become a reality.

Figure 1: Text presentation of data

Visualisation is implemented through web browsers and XSL transformation. It includes support of video and picture display, but does not offer the possibility of continuous playback. Such an option may be possible in the future when XML software such SMIL will be more advanced. However, real-time display of linked information is functional, so the absence of continuous playback is not so much a limitation. The direct use of a web browser (not a web server, distant or local) does not raise major issues of speed efficiency, outside a long time for the initial loading of information. On the other hand, it allows fast development and good reliability as the data generated is automatically validated thanks to the use of XML. Several format are already proposed, such as text format (Clan like, see Figure 1) and partition format (Elan like, see figure 2). Other formats are under study and our goal is to make it possible for each one to set up the system so that it meets every one's specific



needs.

Figure 2: Partition presentation of data

4. Future improvements and goals

There are a lot of possible improvements for the VICLO platform. First, it is still in its developmental, prototype phase and new technical problems may arise. Inclusion of other widely used software such as Transcriber, Anvil and Exmaralda should be undertaken. It is clear that not all specific features from all tools will be maintained at 100% in our data model. In addition, the use of tools outside the systems prevents us from guarantying full integrity of the data, although it is always possible to record changes and go back to previous versions. Second, the development of TEI extensions calls for the creation of a TEI SIG for Oral language which was one of the conclusions of the CatCod conference (Orleans, France, December 2008). Third, the major goal of VICLO is not to create a model of repository, format and tools for corpus data, but to generate new scientific results thanks to the use of new software and approaches. To this effect, one of the goals in the near future is to help researchers use their data efficiently due to better

visualisation software and improved data manipulation and mining software. For example, data could be displayed in a huge variety of formats (including utterance based formats, turn based formats, etc.). Another interesting feature is interface with data manipulation software, such as spreadsheet software (OpenCalc), lexicometric software (Lexico, Le Trameur), statistical software (R) and natural language processing tools (NLTK).

Finally, it should be stressed that the goal of VICLO is not to limit the standards and formats used by the research community but on the contrary to be opened to the rich features offered by multiplatform applications. In this sense, although VICLO is not part of the CLARIN Project (www.clarin.eu), it could perfectly fit into this project, especially during the future construction phase.

For the same reason, there is no actual plan to limit or to delve into the semantics of the transcriptions and annotations. Although we admit that a good interoperability is impossible without such common semantic grounds, we think that the semantic levels should be controlled by the existing tools such as Clan, Elan, Praat, etc. and that semantic compatibility should, at least at the beginning of our project, be assured by the user themselves. Also, our plan is not to create a new multipurpose annotation standard or format, as proposed by Bird and Liberman (2000). The TEI format is, right now, rich enough to allow coding for all the formats we have been working with. Our problem is not into creating a more powerful system, but rather in dealing with the limited features of each application (each application having its own limits and their own strengths) so that these limits will not impede the strengths of the other applications. In this sense, having a more powerful descriptive tool is not necessary at this moment.

5. Acknowledgements

The COLAJE project is funded the ANR (France), which includes three supporting laboratories, MoDyCo CNRS-Paris Nanterre University, Prismes-Sorbonne Nouvelle Paris 3 University, and LILT CNRS-Lille 3 University.

6. References

- Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Commun.*, 33(1-2), 23-60.
- Dublin Core: <http://dublincore.org/documents/dcmi-terms/>
- ELAN: Language Archiving Technology <http://www.ltm-mpi.eu/tools/elan/>
- ISLE IMDI: www.mpi.nl/IMDI/
- MacWhinney, B. (1991). *The CHILDES project - Computational tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- OLAC: <http://www.language-archives.org>
- Paul Boersma & David Weenink (2009): Praat: doing phonetics by computer (Version 5.1.05) [Computer program]. Retrieved May 1, 2009, from

<http://www.praat.org/>
PHON: <http://phon.ling.mun.ca/phontrac/wiki/>
TalkBank: <http://talkbank.org>
TEI: <http://www.tei-c.org>