TEICORPO: a conversion tool for spoken language transcription with a pivot file in TEI

Christophe Parisse, Carole Etienne, Loïc Liégeois

TEICORPO: a conversion tool for spoken language transcription with a pivot file in TEI

Christophe Parisse, INSERM, Modyco-CNRS-Nanterre University, France - cparisse@parisnanterre.fr

Carole Etienne, ICAR-CNRS, Lyon, France - carole.etienne@ens-lyon.fr

Loïc Liégeois, University of Paris, France - loic.liegeois@u-paris.fr

Abstract: CORLI is a consortium of Huma-Num, the French national infrastructure dedicated to the technical support and promotion of digital humanities. The goal of CORLI is to promote and provide tools and information for good and efficient research practices in corpus linguistics and especially spoken language corpora. Because of the time required to collect and transcribe spoken language resources, their number is limited and thus corpora need to be interoperable and reusable in order to improve research on various themes (phonology, prosody, interaction, syntax, textometry…). To help researchers reach this goal, CORLI has designed a set of tools: TEICORPO to assist in the conversion and use of spoken language corpora, and TEIMETA for metadata purposes. TEICORPO is based on the principle of an underlying common format, namely the TEI as described in its specification for spoken language use (ISO/TEI 24624:2016). This tool enables the conversion of transcriptions created with alignment software such as CLAN, Transcriber, Praat or ELAN as well as common file formats (csv, xlsx, txt or docx) and the TEI format, which plays the role of a pivot format, without losing information. Backward conversion is possible in many cases, with limitations inherent to the destination target format. TEICORPO can run the Treetagger Part of Speech tagger and the Stanford CoreNLP tools on TEI files and can export the resulting files to textometric tools such as TXM, Le Trameur, or Iramuteq, making it a tool dedicated to spoken language corpora editing as well as to various research purposes.

Keywords: TEI, transcription, oral corpora, conversion, annotationBlock

1. The CORLI consortium and corpus linguistic research

The CORLI consortium is a network of researchers and laboratories engaged in corpus linguistic research. CORLI (https://corli.huma-num.fr/en) is one of the consortia of Huma-Num, a large French infrastructure dedicated to the digital humanities (https://www.huma-num.fr/). CORLI aims to promote and assist research in corpus linguistics and is based on a network of researchers and engineers working in this field. The general policy of the consortium is to build on existing practices, research tools and material from members of the consortium, and to improve them to match the needs of all the people and laboratories interested in corpus linguistics. The main actions of the consortium are helping corpus providers to foster open science and consequently produce shareable data, providing continuous education for advanced students as well as senior researchers, and sharing good practices about data and tools. CORLI is now an official CLARIN K Centre (https://corli.huma-num.fr/en/kcentre).

1.1 The importance of data sharing

Data sharing is an absolute necessity for research and applications in spoken language, for two main reasons. First, creating spoken language corpora is very expensive. As a result, not only are existing data very frequently reused, but it is also often necessary to mix data coming from different origins when the goal is to produce very large datasets. Second, there is a very wide variety of situations for language production, in terms of speakers and production contexts. Hence, in order to create interesting sociolinguistic data, it is very often efficient to mix corpora from different sources. Moreover, the development of open science, which is a key component of research today, cannot exist without data-sharing, so that the data source of scientific results can be controlled.

1.2 Specificity of spoken language data

Spoken language corpora are different from corpora of written text. They consist of sound or video recordings that come with textual information, transcriptions and annotations. The spoken language is organized according to two dimensions, as described in Schmidt (2011): the macro structure which refers to how transcriptions, annotations, and linking are organized (where they are in the file); and the micro structure which refers to the details of how transcriptions and annotations are represented (how a transcription or an annotation is written down).

Because linking linguistic information (transcription and annotation) to audio or video data is not a straightforward process and because this information is crucial for studying spoken language, specific tools are used for creating and editing spoken language data. These tools provide the user with functionalities that are necessary for creating spoken language corpora and they also save data in controlled formats that can be used for data-sharing. Unfortunately, although all these tools manage the macro structure of spoken language corpora in a very strict fashion, this is not the case for the micro structure. Certain tools provide some micro structure information, but it is usually limited and not available for all corpora and for all types of research. There are two main reasons for these limitations. The first is that there is a wide variety of different micro structures that can be used, depending on the scientific goal and the research domain that call for corpus creation and use. The second is that controlling micro structure when creating the data is extremely time-consuming, with the result that the micro structure is often restricted to the minimal requirements of the study. Moreover, the creation of user-friendly tools for controlling micro structure requires considerable manpower, which the creators of free tools for research into spoken language seldom have. So the burden of controlling the quality of the micro structure most often falls on the shoulders of the corpus creator, and not on the tool used to create the corpus. Stability and uniformity in micro structure corpora is only found in controlled repositories that follow clear and controlled instructions – for example the TalkBank system (MacWhinney 2007, https://talkbank.org/), the Archiv für Gesprochenes Deutsch (http://agd.ids-mannheim.de) – or in specific corpus projects, which therefore have a limited coverage, even if these corpora can be very large.

1.3 The TEICORPO approach

The goal of the CORLI consortium is to make it easier to deposit, share, and reuse data. With this goal in mind, CORLI has always promoted the use of open public repositories and open formats. Our policy is to advocate the use of a common single format which can be used for data-sharing as well as for long term preservation. We encourage its use from the starting point of each new research project so as to ensure the development of open science. However, CORLI, as a consortium of linguists who share its practices, acknowledges that the most important thing for researchers remains the tools that they know and use. We know that few people will be working directly on the raw files, but through the tools that they have been trained with. So it is up to the tools to allow format and data sharing.

For this reason, it was important for CORLI not only to stress the quality of a common data format, but also to make it possible to use this format with the tools that the researchers know and regularly use. These tools include editing tools, which make it possible to create and modify the corpus transcription, and exploration tools, which are necessary in many cases to produce scientific results. The exploration tools can be the same as the editing tools, but they can also include other types, such as textometric tools (Lebart et al. 1997, Pincemin 2011, Pincemin et al. 2020) and grammatical parsers.

Finally, another constraint in the work of CORLI was to create tools that were suitable for the needs of linguists. This means that the goal of the project was not to create and promote new standards, but rather to emphasize the good practices of CORLI's members and make data sharing with other researchers easier. The goal of the TEICORPO project can thus be summarized as:

- creating a conversion tool between the different pieces of software used for spoken language corpora creation and editing. The tool should be as lossless as possible;
- using the TEI as a conversion pivot format, as the TEI is already widely used especially for written linguistic corpora and is powerful enough for our purposes;
- creating complementary conversion tools that make it easier to explore and use the corpora for research.

1.4 Similarities and differences with other approaches

Many software packages dedicated to editing spoken language transcription contain utilities that can convert many formats. This is the case for example of Exmaralda (Schmidt 2004, see https://exmaralda.org), Anvil (Kipp 2001, see https://www.anvil-software.org) and ELAN (Wittenburg et al, 2006, see https://archive.mpi.nl/tla/elan). However, in all cases, the conversions are limited to the features implemented in the tool itself, for example with a limited set of metadata, and they cannot always be used to prepare data to be used by another tool.

While our work is similar to that of Schmidt (2011), several differences make our approaches complementary. First, the two main common features are as follows:

- the tool can automatically convert data coming from different linguistic tools;
- the TEI is used as a destination format.

The list of tools that are considered in the two projects is nearly the same. The only tools missing in the TEICORPO approach are Exmaralda, and Folker (Schmidt & Schütte,

2010, see https://exmaralda.org/en/folker-en/), but this was only because the conversion tools from and to Exmaralda, Folker and the TEI already exist. They are available as XLST stylesheets in the open source distribution of Exmaralda (https://github.com/Exmaralda-Org/exmaralda). The other common point is the use of the TEI format, and especially the more recent ISO version of the TEI for spoken language (ISO/TEI see ISO 24624:2016). The TEI format produced by the Exmaralda and Folker software fit within the process chain of TEICORPO. This demonstrates the usefulness of using a well-known and efficient format such as the TEI.

There are however differences between the two projects that make them non-redundant but complementary, each project having specificities that can be useful or damaging depending on the user's needs. One minor difference is that the TEICORPO project is not a functionality of an editing tool, but is a standalone tool that targets the process of converting data between one format and another. This had certain consequences on the user interface and explains some of the choices made in the development of the two tools.

There are two major differences between TEICORPO and Schmidt's approach, which impacted both the design of the tools and the way they can be used. The first difference is that in developing TEICORPO, it was decided that the conversion between the original formats and the TEI had to be lossless (or as lossless as possible) because we wanted to offer a means to store the research data for long term conservation and dissemination in an official XML format instead of storing them in proprietary formats such as those used by CLAN (MacWhinney 2000, see http://dali.talkbank.org/clan/), ELAN, Praat (Boersma & van Heuven 2001, see http://www.fon.hum.uva.nl/praat/), and Transcriber (Barras et al. 2000, see http://trans.sourceforge.net/en/presentation.php and http://perso.ens-lyon.fr/matthieu.quignard/Transcriber/). These proprietary formats are in XML format or in Unicode text format so that they can be stored for long term conservation. However, they are not all well described or constrained, at least not in the same way as the TEI, which offers moreover a semantically relevant structure as well as an official format for long term conservation in France. Moreover, as the durability of these four pieces of software cannot be guaranteed in the long term, it does not seem safe to keep corpora in a format available only for a given tool that may disappear or fall into disuse.

The second major difference is that the TEICORPO initiative does not specifically target spoken language, but all types of annotation, including media of any type. This covers all spoken languages, vocal as well as sign languages, but also gesture and any type of multimodal coding. The goal of TEICORPO was not to advocate a linguistic mode of coding spoken data as a transcription convention does, but rather to propose a research model for storing and sharing data about language and other modalities. Consequently, the focus of the work was not on how the spoken data were coded (i.e., the micro-structure), nor on the standard that should be used for transcribing in orthographic format. Instead, the TEICORPO approach focused on how to integrate multiple pieces of information in the TEI (the macro-structure), as this is possible with tools such as ELAN or PRAAT. The goal was to be able to convert a file produced by these tools so that it can be saved in TEI format for long term conservation.

Data in PRAAT and ELAN formats can contain information that is different from what is usually present in an ISO/TEI description, but that remains nonetheless within the

structures authorized in the ISO/TEI. For example, the information is stored as described below in spanGrp, a structure available in the ISO/TEI description. This means that whenever information targets the 'classical' approach to spoken language (by 'classical', we mean approaches based on an orthographic transcription represented as a list, as in the script of a play), it will be available for further processing by using the export features of TEICORPO (see part 2.3 and further below for export functionalities) but other types of information are also available. Compared to PRAAT and ELAN, the integration of tools such as CLAN or Transcriber was much more straightforward as the organization of the files is less varied and more 'classical'.

1.4.1 Choice of the microstructure representation

Processing of the micro-structure, with the exception of information already available in the tools themselves (for example silence in Transcriber), is not done during the conversion to the TEI. In all the tools used by researchers in CORLI, the division into words or other elements such as morphemes or phonemes is not systematically done. When it exists, it is not included in the main transcription line but most often in dependent lines as it represents in fact an annotation with its own rules and guidelines to split the production into words. It is part of the linguistic analysis rather than a simple storage operation.

TEICORPO therefore preserves as long-term storage data both the original information that was created in the original software – the full unprocessed transcription –, and the other linguistically processed transcriptions and annotations. For TEICORPO, micro-structure processing, such as division into words, or text standardization when necessary, belongs to the linguistic analysis of the corpora. Hence, the TEI data file can be used both for data exploration and scientific uses. For example, when a researcher needs to parse the data, or to explore the data with textometric tools, then it is necessary to decide which type of preprocessing is necessary. As this decision is often dependent on the initial project as well as on linguistic choices, it is difficult to standardize this task.

2. The TEICORPO project

The TEICORPO project contains two different sets of tools. One set focuses on conversion between various software packages used for spoken language coding and the TEI. The second set focuses on using the TEI format for linguistic analyses (textometric or grammatical analyses).

2.1 Alignment tools

Several alignment tools are used by researchers working with spoken language corpora to align the recording, whether an audio or a video file, with its textual transcription and other specific annotations. The choice of tools is usually related to the choice of annotations, or to the common practices in a given research field, or simply to the researcher's preferences, knowledge, and requirements.

Some common practices have been identified in our community but other uses of the same software are of course possible:

- Transcriber is widely used in sociolinguistics;
- CLAN is widely used in language acquisition and especially in the Talkbank project;
- Praat is more specialized for phonetic or phonological annotations;
- ELAN is recommended for annotating video and particularly multimodality, but is often used for rare languages to describe the organisation of the segments.

It should be pointed out here that whereas Transcriber and CLAN files nearly always contain 'classical' orthographic transcriptions, this is not the case for Praat and ELAN files. As our goal is to provide a generic solution for long-term conservation and use for any type of project, conversion of all types of files produced by the four tools cited above will be possible. It is up to the user to determine and inform which part of a corpus can be used with a classical approach, and which parts should not and how they should be processed.

The list of tools reflects the uses and practices in the CORLI network, and is very similar to the list suggested by Schmidt (2011) with the exception of EXMARaLDA and FOLKER. These two tools already have built-in conversion features, so adding these tools in the TEICORPO project would be easy at a later date.

Alignment applications deal with two main types of data presentation and organization. The presentation of the data has direct consequences on how the data are exploited, and therefore on the design of the tools that are used.

- list or text format (as in the script of a play): the speakers' productions are displayed in vertical order, from the beginning to the end of the recording, in the order that they are produced, speaker by speaker, with on the left the speaker's name followed by their production. The timeline is vertical, going from top to bottom. In the result file, the speaker's productions appear inside two time points in the chronological order of the timelines, the start and the end of the production. The list format can be extended as a hierarchical format, as is the case in the TEI, for example;
- partition format (as in a musical score): each speaker has their own tier, a horizontal line representing the information pertaining to the speaker (like each instrument in a musical score). Sometimes several tiers are associated with the same speaker, for example gestures and gazes, aligned with the verbal production. The timeline is horizontal, going from left to right. The result file is not organized chronologically but is sorted by the names of the tiers (or any other order) with all the production within the same tier sorted by timeline.

No tool offers both types of presentation. ELAN offers some alternatives to editing or displaying data with the partition format, but none of them offer a fully-fledged list format editing. It is possible to represent the two structures within a similar model, as demonstrated by Bird and Liberman (2001). However this is not the case for the four tools listed above: each of them represents the data in a unique underlying data structure. Transcriber and CLAN are organized in list format; Praat and ELAN have a partition format.

Each presentation format has its own pros and cons. Because of the possibilities offered by the presentation formats, and because the same software, even within the same

presentation models, rarely provides a solution for all the needs of all users, researchers often have to use two or more pieces of software.

The use of multiple tools is quite common. For example, Praat and Transcriber cannot be used when working on video recordings because these programs are limited to audio formats. But if researchers need to conduct spectral analysis for some purpose, they will have to use the Praat software and convert not only the transcription, but also the media. In the field of language acquisition, where the CLAN software is generally used to describe both the child productions and the adult productions, when researchers are interested in gestures, they use the ELAN software, importing the CLAN file to add gesture tiers, as ELAN is more suitable for the fine-grained analysis of visual data. Another common practice consists in first doing a rapid transcription using only orthographic annotations in Transcriber and then in a second stage annotating some more interesting excerpts in greater detail including new information. In this case researchers will import the first transcription file into other tools such as Praat or ELAN and annotate them partially. It is therefore necessary to import or export files in different formats if researchers need to use different tools for different parts of their work.

Another need concerns the pooling of corpora coming from other resources or other projects. Conversions are thus necessary, and can be problematic because going from one piece of software to another often leads to a loss of information, as each tool handles corpus information differently, beyond the differences between list and partition formats. Researchers need a convenient tool to convert each file from one format to another without losing information. If this tool is not provided by the software itself, it is necessary to create a specific tool. This specific tool is generally developed in each laboratory, operates only on a given platform (Windows, MacOS, Linux, etc.), is without computer maintenance, and does not include the latest releases of each software or the requirements of the conversion process options. For these reasons, we decided in the CORLI consortium and in collaboration with the ORTOLANG infrastructure to design a common tool that could be used by the whole linguistic community. The goal was to make open source software with proper maintenance freely available on http://ct3.ortolang.fr/teicorpo/.

2.2 Conversion to and from the TEI

As explained above (see section 1.4), the conversion tools do not focus on the micro-structure of the linguistic coding as it was not the goal of CORLI to push for a unique coding of linguistic data. Nor did we split the transcription into words, as this is a research dependent process and would not preserve the content of the original files. Processing the micro-structure and splitting utterances into words is part of the second stage of our work (see section 2.3), which concerns the exploitation of the corpora. The goal of the conversion tool is to convert all the information in the metadata and all the macro-structure information into the TEI format.

2.2.1 Basic structures

Converting the metadata is straightforward as the four tools (CLAN, ELAN, Praat, Transcriber) do not enable a large amount of metadata to be edited. Most of the metadata available concerns the content of the sequence; some user metadata is also available,

especially in CLAN. The insertion of metadata follows the indications of the ISO/TEI 24624:2016 standard.

Moreover, some tools, such as Transcriber, include information about silences, pauses, and events in their XML format. This information is also processed within TEICORPO, once again following the recommendations of the ISO/TEI standard.

Conversion of the main data, the transcription and the annotations, cannot always be done on the sole basis of the description provided in the ISO/TEI guidelines. These guidelines do however suffice to fully describe the content of the CLAN and Transcriber software. We took advantage of the new *annotationBlock* element which codes several annotation levels, a function that is commonly required in spoken language annotations.

The *annotationBlock* contains two major elements: the *u* element which contains the transcription in orthographic form and the *spanGrp* elements which contain tier elements that annotate the utterance described in the *u* element. A *spanGrp* element contains as many *span* elements as required. All *span* elements have the same type of content, as indicated in the father *spanGrp* element. Figure 1 provides an example of conversion from a CLAN file to illustrate how a production annotated on different levels (orthography, morphosyntax, dependencies) is represented in TEI with a first main utterance element *u* to which two *spanGrp* are linked, one for each annotation level, in our case one *spanGrp* for morphosyntax and one s*panGrp* for dependencies (see Figure 3). A *timeline* element gives the start (T1086) and end (T1087) timecodes and an *annotatedBlock* element specifies the speaker with the *who* attribute and the *start* and *end* attributes with the timecode anchors #T1086 and #T1087. The *annotatedBlock* element includes both the utterance element and the two annotations. No decision is made about the inner content of the span elements. The content of the *type* attribute in the *spanGrp* element represents the choice of the researchers who produced the original corpus; TEICORPO imposes no constraints. The content generated is preserved as it was in the original file, making backward conversion possible. In the example in figure 1, the ‘mor’ and ‘gra’ elements represent grammatical knowledge. Using the content of these elements to produce advanced grammatical representation in more elaborate TEI and XML formats is of course possible, but this would be a tailored task which is beyond the scope of the TEICORPO project.

```
*MOT: look at the tree ! •2263675_2265197•
%mor: v|look prep|at det|the n|tree !
%gra:  1|0|ROOT 2|1|JCT 3|4|DET 4|2|POBJ 5|1|PUNCT

<timeline unit="s">
<when absolute="0" xml:id="T0"/>
<when interval="2263.675" since="#T0" xml:id="T1086"/>
<when interval="2265.197" since="#T0" xml:id="T1087"/>
<annotationBlock end="#T1087" start="#T1086" who="MOT" xml:id="au551">
        <u><seg>look at the tree ! </seg></u>
        <spanGrp type="mor">
                <span>v|look prep|at det|the n|tree ! </span>
        </spanGrp>
        <spanGrp type="gra">
                <span>1|0|ROOT 2|1|JCT 3|4|DET 4|2|POBJ 5|1|PUNCT</span>
        </spanGrp>
</annotationBlock>
```
Figure 1: CLAN representation of data (first three lines) and corresponding representation in TEI

2.2.3 Advanced structures

Although the presentation described above can represent the data of many corpora and tools, a single-level annotation structure within the *spanGrp* elements is insufficient to represent the complex organisation that can be constructed with the ELAN and Praat tools. ELAN is a tool used by many researchers to describe data of greater complexity than the data presented in the ISO/TEI guidelines. As the goal of the TEICORPO project was to convert all types of structure used in the spoken language community, including ELAN and Praat, it was necessary to extend the description method presented in section 2.2.2.

In ELAN and Praat, the multi-tiered annotations can be organized in a structured manner. These tools take advantage of the partition presentation of the data, so that the relationship between a parent tier and a child tier can be precisely organized. There are two main types of organization: the first type is purely symbolic. The elements of a child tier, C1 to Cn, can be related to an element of a parent tier P. This is called **symbolic division.** For example, a word is divided into morphemes. In figure 2, the main tier has two representations, “BEJ_MV_NARR_11_coffee_18” and “gahwat mu:nai end //”. The tier of the second level contains individual words, “gahwat”, “mu:na”, … On the third tier, the words are broken down into morphemes: “gahw”, “-t”, … There are three other levels of organization, which either follow the organization of the morphemes, or of the main tier. In all these cases, the relationships between tiers are symbolic, and could be represented by symbolic links instead of temporal links.

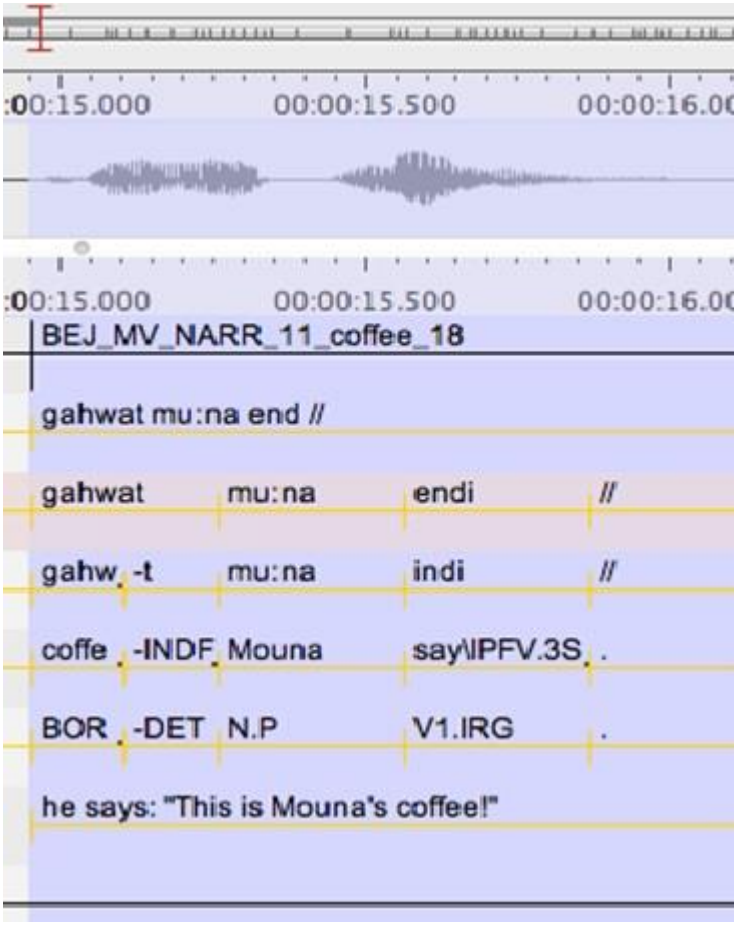


Figure 2: ELAN annotation with symbolic structures

Another type of organization is **temporal division**. In this case, the association between the main tier and the dependent tiers is described by temporal information rather than by symbolic information. An element is included in the parent if the starting and end points are within the time limits of the starting and end points of the parent tier. Figure 3 provides an example of such an organization created using Praat. In this example, the tier at the top of the representation contains phonemes which are included in the second tier that contains syllables (so for example “S” and “a” are included in “Sa”). Then the syllables are included in turn in the word level tier (bottom tier). In this example, as well as for the previous example, there are other tiers which will not be described here, but which show how complex the representation of the data can be.

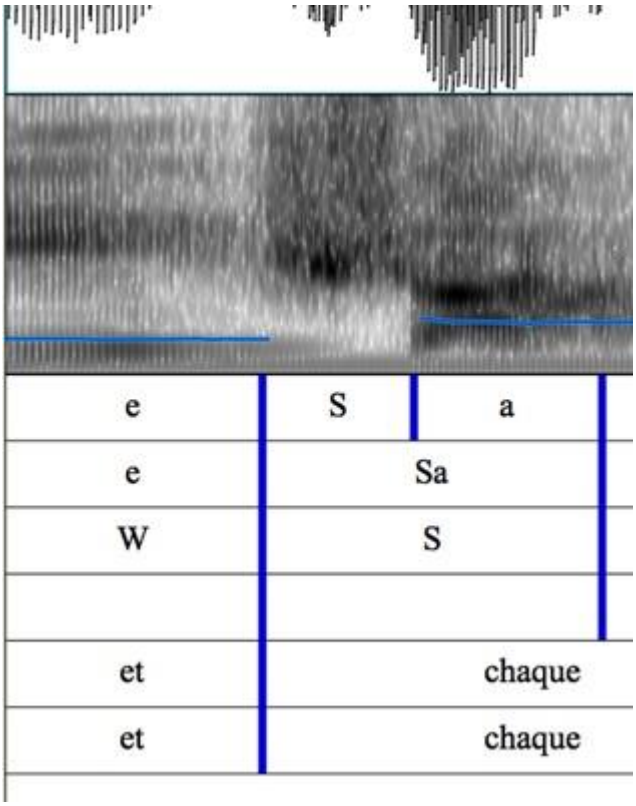


Figure 3: Praat representation of a temporal structure

2.2.4 Representation of advanced structures

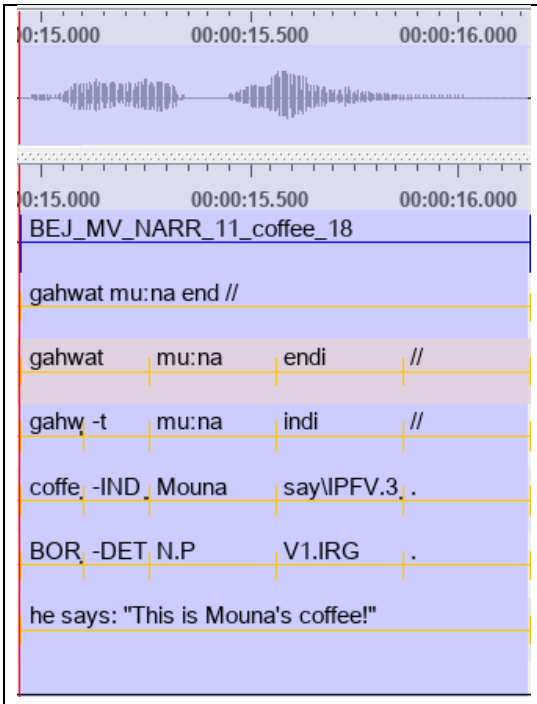
The two types of data organization presented above, i.e. multiple levels where the relation between the parent and the child tier is either symbolic or temporal (the two types of relation can appear in the same file at different levels), correspond to uses in corpus linguistics that are more advanced than the usage presented in the ISO/TEI guidelines. This type of corpus can in some cases represent a large amount of data, or it can correspond to smaller corpora with very fine-grained coding. It can also represent unusual data, or new types of corpora, such as gesture and sign language studies, for instance. In these cases, the actual data coding can be far removed from that of a usual spoken language corpus (such as those described in Schmidt, 2011). However, as this type of data is found in the production of the members of the CORLI consortium, it needs to be preserved as well as other data. As these data are coded using standard tools, coding this data is part of the TEICORPO goal.

Although this type of data is not described in the ISO/TEI guidelines, it is in fact possible to store it in TEI format using current features of the TEI. The TEI provides a general mechanism for storing hierarchically structured data by using the spanGrp/span mechanism. Moreover, the span and spanGrp tags have attributes that can point to other elements or to timelines. With this mechanism, it is therefore possible to store any type of structure, symbolic and/or temporal, that can be generated with ELAN or PRAAT, as described above.

To do this, each element which is in a symbolic or temporal relation is represented by a *spanGrp* element of the TEI. The *spanGrp* element contains as many *span* elements as

necessary to store all the elements present in the ELAN or PRAAT representation. The father element of a *spanGrp* is the main *annotationBlock* element when the division in ELAN or PRAAT is the first division of a main element. The father element is another span element when the division in ELAN or PRAAT is a sub-division of another element which is not a main element. This XML structure is complemented by explicit information as allowed in the TEI. The *span* elements are linked to the element they depend on either with a symbolic link using the *target* attribute of the span element, or with temporal links using the *from* and *to* attributes of the span element.

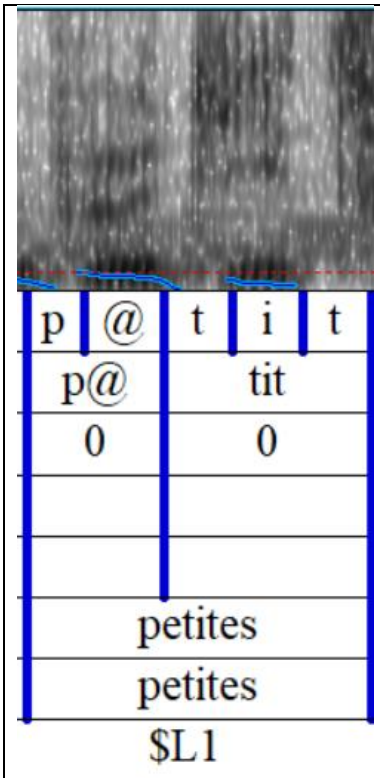
Two examples of how this is displayed in a TEI document are given below. The first example (see Figure 4) corresponds to the ELAN example above (see section 2.2.3 Figure 2). The arborescence represents the words of the sentence from left to right (from “gahwat” to “endi” in our example). The detail of the transcription is represented recursively, with *span* containing *spanGrp*, and spanGrp containing other *span* elements, until all the data are represented. This can be pursued down to any depth. In the example below (see Figure 4), this is the case with “gahw -t” which is divided into “BOR” and “-DET”.




```
<annotationBlock end="#T24" start="#T23" who="ref@SP"
                 xml:id="ann296">
  <u>
    <seg>BEJ_MV_NARR_11_coffee_18</seg>
  </u>
  <spanGrp type="tx@SP">
    <span target="#ann296" xml:id="ann304">
      gahwat mu:na end //
        <spanGrp type="mot@SP">
          <span target="#ann304" xml:id="a69">
            gahwat
            <spanGrp type="mb@SP">
              <span target="#a69" xml:id="a186">
                gahwa
                <spanGrp type="ge@SP">
                  <span target="#a186"
                                xml:id="a318">coffee</span>
                </spanGrp>
                <spanGrp type="rx@SP">
                  <span target="#a186"
                                xml:id="a450">BORR</span>
                </spanGrp>
              </span>
              <span target="#a69" xml:id="a187">
                -t
                <spanGrp type="ge@SP">
                  <span target="#a187" xml:id="a319">-
                                INDF.F</span>
                </spanGrp>
                <spanGrp type="rx@SP">
                  <span target="#a187" xml:id="a451">-
                                DET</span>
                </spanGrp>
              </span>
            </spanGrp>
          </span>
```

Figure 4: ELAN example of a symbolic division with the corresponding TEI format

The second example is structured using time references. This example (see Figure 5) corresponds to the Praat example above (see section 2.2.3 Figure 3). In this case, each part of the transcription is represented according to the timeline, but there is also a hierarchy which is represented by the *spanGrp* and *span* tags. Each *span* is part of the parent *spanGrp* with starting and ending points (which correspond to the *from* and *to* attributes in the example below). The use of *from/to* versus *target* is the only difference between the two organizations. In the example below, the syllable “p@” is divided into two phones, “p” and “@” (see xml:id s73, s74, s75).




```
<annotationBlock end="#T153" start="#T1" who="locuteur"
                 xml:id="au4763">
  <u>
  <seg>$L1</seg>
  </u>

<span from="#T21" to="#T23" xml:id="s72">
  petites
  <spanGrp type="syllabe">
    <span from="#T21" to="#T22" xml:id="s73">
      p@
      <spanGrp type="phone">
        <span from="#T21" to="#T2973" xml:id="s74">p</span>
        <span from="#T2973" to="#T22" xml:id="s75">@</span>
      </spanGrp>
    </span>
    <span from="#T22" to="#T23" xml:id="s76">
      tit
      <spanGrp type="phone">
        <span from="#T22" to="#T2974" xml:id="s77">t</span>
        <span from="#T2974" to="#T2975" xml:id="s78">i</span>
        <span from="#T2975" to="#T23" xml:id="s79">t</span>
      </spanGrp>
    </span>
  </spanGrp>
</span>
```

Figure 5: ELAN example of a temporal division with the corresponding TEI format

The *spanGrp* and *span* offer a generic representation of data coming from relatively unconstrained representations produced by partition software. The name of the tiers used in the ELAN and Praat tools is given in the content of the type attribute. These names are not used to provide structural information, the structure being represented only by the *spanGrp* and *span* hierarchy. However, the organization into *spanGrp* and *span* is not always sufficient to represent all the details of the tier organization of each software feature. This is the case for some of the ELAN structures, which can specify the nature of *span* elements further than in the TEI feature. For example, the *timediv* ELAN property specifies that only contiguous temporal division is allowed, whereas the *incl* property allows non-contiguous elements. It was therefore necessary to include the type of organization in the header of the TEI document, using the *note* structure. The note element here points to a case where dedicated tags do not currently exist in the TEI, so we used the note element as the best solution not to lose the information. This could be included in future evolutions of the TEI.

2.3 Exporting to research tools

In the TEICORPO approach, no modification is made to the original format and conversion remains as lossless as possible. This allows for all types of corpora to be stored for long-term preservation purposes. It also allows the corpora to be used with other editing tools, some of which are suited to specific processing. For example: Praat for phonetics/phonology; Transcriber/CLAN for raw transcription; ELAN for gesture and visual coding.

However, a large proportion of scientific research and applications done using corpora requires further processing of the data. For example, although querying or using raw language forms is possible, many research investigations and tools use words, part of speech,

grammatical or semantic information. This requires further processing starting with the original raw corpus form. In the case of spoken language corpora, the nature of the information inserted in the transcription has to be taken into account. This corresponds to what Schmidt (2011) calls micro-structure. This micro-structure is integrated in Schmidt's approach in which the TEI document can contain standardized information about words, specific spoken language information, and sometimes even part of speech information.

This approach was not adopted in TEICORPO for several reasons. First, we had to deal with a large variety of coding approaches which makes it difficult to conduct work similar to that done in CHILDES (MacWhinney 2000, see https://childes.talkbank.org/). Secondly, there was no consensus about the way tokenization should be performed, as many researchers consider tokenization as a choice with multiple possibilities, each with different consequences on the resulting data and grammatical analyses. In other words, tokenization is part of the linguistic analysis, so it should not be frozen at the level of data conservation.

For these reasons we decided firstly to make it possible to include in a corpus both the original raw language material and the modified tokenized and analyzed forms. This is easy to do using the TEI, as any analyzed material can be inserted in *spanGrp/span* elements, without modifying the original *u* element information. Secondly, we decided to design a second category of tools aimed at processing or making it possible to process the spoken language corpus, and to use powerful tools in corpus analysis. This part of the TEICORPO library is described in the Applications section below.

3. Applications

Two types of applications have been designed: applications that are implemented as web services and applications that require command line processing. The choice between web service and command line processing was most often a consequence of user needs. The web service interface was developed on a user request basis and further development will be done if new needs appear in the future. The web service covers only the most frequently used and basic features, such as conversion between data. It contains some specific features such as conversion from and to text, and spreadsheet and word processing formats as this was a feature required by non-advanced users.

The command line interface contains features required by more advanced users. It can handle conversion between basic formats and the TEI, and is more efficient than the web service to handle a large number of primary files. The command line interface contains other features such as syntactic analysis, metadata processing, and complex export features. Implementation of the command line parameters is less costly than the implementation of these options in a web page, so they are easier to develop first before having to be done in a more sophisticated user interface.

3.1 Basic import and export functions

The command line interface (see http://ct3.ortolang.fr/teicorpo/) can perform conversions between the TEI and the following tools: CLAN, ELAN, Praat, and Transcriber. The conversions can be performed on single files or on whole directories or on a file arborescence. The command line interface is suited to automatic processing in offline

environments. The online interface (see http://ct3.ortolang.fr/teiconvert/) can convert one or several files, selected by the user, but not whole directories. Results are in the download section.

In addition to the conversion between the alignment software, the online version of TEICORPO offers import and export in common spreadsheet formats (xlsx and csv) and word processing formats (docx and txt). Importing data is useful to create new data, and exporting is used to make reports or examples for a publication and for end-users not familiar with transcription tasks or computer software (see Figure 6).

```
*MOT: look at the tree ! •2263675_2265197•
%mor: v|look prep|at det|the n|tree !
%gra:  1|0|ROOT 2|1|JCT 3|4|DET 4|2|POBJ 5|1|PUNCT

MOT   2264   2265   look at the tree !
mor                 v|look prep|at det|the n|tree !
gra                 1|0|ROOT 2|1|JCT 3|4|DET 4|2|POBJ 5|1|PUNCT
```

Figure 6: Example of text export of part of a file originally in CHAT format

In each case, some options are available to specify how to produce the:

- timeline: at the beginning, before or after the speaker's production, timeline precision
- line number
- spoken language annotations: an export without any specific annotation is useful for automatic tools which only need an orthographic lexical transcription.

Other features are available in both types of interface (command line and web service). TEICORPO allows the user to exclude some tiers, for example adult tiers in acquisition research where the user wants to study child production only, or comment tiers which are not necessary for some studies.

3.2 Export to specialized software

Another kind of export concerns textometric software. TEICONVERT makes spoken language data available for TXM (Heiden, 2010, see http://textometrie.ens-lyon.fr), Le Trameur (Fleury & Zimina 2014, see http://www.tal.univ-paris3.fr/trameur/), and Iramuteq (see http://iramuteq.org/ and de Souza et al. 2018), providing a dedicated TEI export for these tools. For example, for the TXM software, the export includes a text element made of utterance elements including age and speaker attributes. Figure 7 presents an example for the TXM software.

```
<TEI file="/corpusformat/exemple/lily-4-00-02.tei_corpo.xml">
<teiHeader/>
<text>
   <u age="28" end="2875.100" start="2873.395" who="MOT">
      <w age="28" loc="MOT">you</w>
      <w age="28" loc="MOT">have</w>
      <w age="28" loc="MOT">to</w>
      <w age="28" loc="MOT">rest</w>
      <w age="28" loc="MOT">now</w>
      <w age="28" loc="MOT">.</w>
   </u>
   <u age="4.0" end="2875.970" start="2875.100" who="CHI">
      <w age="4.0" loc="CHI">yes</w>
      <w age="4.0" loc="CHI">.</w>
   </u>
   <u age="28" end="2877.893" start="2875.970" who="MOT">
      <w age="28" loc="MOT">from</w>
      <w age="28" loc="MOT">your</w>
      <w age="28" loc="MOT">big</w>
      <w age="28" loc="MOT">singing</w>
      <w age="28" loc="MOT">extravaganza</w>
      <w age="28" loc="MOT">.</w>
   </u>
```

Figure 7: Example of XML for the TXM software

An export has been developed for Lexico and Le Trameur textometric software with a simple sgml file without timelines (see Figure 8).

```
<file=/corpusformat/exemple/lily-4-00-02.tei_corpo.xml>
<loc=MOT>you have to rest now ?
<loc=CHI>yes .
<loc=MOT>from your big singing extravaganza ?
<loc=CHI>yes that was a party .
<loc=MOT>woof
```

Figure 8: Example of export for the Lexico or Le Trameur software

Likewise, another export concerns the textometric tool Iramuteq without timelines (see Figure 9).

```
****
-*MOT
you have to rest now ?
-*CHI
yes .
-*MOT
from your big singing extravaganza ?
-*CHI
yes that was a party .
-*MOT
woof .
-*MOT
that was a party  that sure was some party .
```

Figure 9: Example of export for the IRAMUTEQ software

In all these cases, TEICORPO is able to provide an export file and to remove unnecessary information from the TEI pivot format. This is useful, for example, with textometric software as they work only with orthographic tiers without a timeline or dependent information.

3.3 Using an automatic grammatical analyzer

Many researchers in linguistics wish to use automatic grammatical analyzers on corpora in order to improve the querying and analysis of a text, or to implement other types of linguistic research. A present difficulty with these grammatical analyzers is that most often they run only on raw orthographic material, excluding other information. Moreover, their result is not always in a format that can be used with traditional spoken language software such as CLAN, ELAN, Praat, Transcriber or of course TEI.

TEICORPO provides a way to solve this problem by running analyzers and putting the results from the analysis back in TEI format. Once the TEI format has been enriched with grammatical information, it is possible to use the results and convert them back to ELAN or Praat and use the grammatical information in these spoken language software packages. It is also possible to export to TXM and to use the grammatical information in the textometric software. Two grammatical analyzers have been implemented in TEICORPO: TreeTagger and CoreNLP.

TreeTagger

TreeTagger[1] (Schmid 1994; 1995) is a tool for annotating text with part-of-speech and lemma information. The software is freely available for research, education and evaluation. It is available in 25 languages, provides high-quality results, and can be easily improved by enriching the training set, as was done for instance by Benzitoun, Fort, and Sagot (2012) in the PERCEO project. They defined a syntactic model suitable for spoken language corpora, using the training feature of TreeTagger and an iterative process including manual corrections to improve the results of the automatic tool.

[1] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

The command line version of TEICORPO should be used to generate an annotated file with lemma and PoS based on Treetagger. TreeTagger should be installed separately. The implementation of TreeTagger in TEICORPO includes the possibility to use any syntactic model. For French data, we used the PERCEO model (Benzitoun et al. 2012).

The command line to be used is:

java -cp TEICORPO.jar fr.ortolang.TEICORPO.TeiTreeTagger filenames...

with additional parameters:

-syntaxformat “param”: “param” can take the values *w* or *ref* or *conll* (see examples below)

-model “filename”: the full name of the TreeTagger syntactic model. In our case, we use the PERCEO model.

-program “filename”: the full location of the TreeTagger program, according to the system used (Windows vs. MacOs vs. Linux).

-normalize “format”: “format” specifies the origin of the corpus. This can be useful when cleaning up source files

The environment variable TREE_TAGGER can be used to locate the model and the program. If no -program option is used, the default name for the TreeTagger program is used.

The -model parameter is mandatory.

The resulting filename ends with “.tei_corpo_ttg.tei_corpo.xml” or a specific name provided by the user (option -o).

Several “syntaxformat” parameters are available:

- word format where the utterance element *u* is divided into word elements *w* with *ana* attribute for PoS and *lemma* attribute (see Figure 10). This is the basic format in the TEI specification. This format can be used to produce data that include part of speech information which follows the standard recommendation of the TEI.

```
<annotationBlock end="#T1393" start="#T1392" who="CHI" xml:id="au1282">
   <u>
      <w lemma="thank" pos="VV">thank</w>
      <w lemma="you" pos="PP">you</w>
      <w lemma="it" pos="PP">it</w>
      <w lemma="be" pos="VBD">was</w>
      <w lemma="a" pos="DT">a</w>
      <w lemma="singing" pos="NN">singing</w>
      <w lemma="party" pos="NN">party</w>
      <w lemma="." pos="SENT">.</w>
   </u>
   <spanGrp type="xpho">
      <span>ˈθæku ɪwəzəˈsɪŋɪŋpɑrdi</span>
   </spanGrp>
</annotationBlock>
```

Figure 10: Tagging results in “w” format

- ref format with a *ref* element in a *spanGrp/span.* It is made of word elements *w*, so that the original utterance format is preserved (see Figure 11). This format is easier to process automatically than the “conll” format (see below) but contains only the word information described in the standard TEI specification (as for the “w” format).

```
<annotationBlock end="#T1393" start="#T1392" who="CHI" xml:id="au1282">
   <u>
      <seg>thank you , it was a singing party . </seg>
   </u>
   <spanGrp type="xpho">
      <span>ˈθæku ɪwəzəˈsɪŋɪŋpɑrdi</span>
   </spanGrp>
   <spanGrp inst="treetagger" type="ref">
      <span>
         <ref>
            <w lemma="thank" pos="VV">thank</w>
            <w lemma="you" pos="PP">you</w>
            <w lemma="it" pos="PP">it</w>
            <w lemma="be" pos="VBD">was</w>
            <w lemma="a" pos="DT">a</w>
            <w lemma="singing" pos="NN">singing</w>
            <w lemma="party" pos="NN">party</w>
            <w lemma="." pos="SENT">.</w>
         </ref>
      </span>
   </spanGrp>
</annotationBlock>
```

Figure 11: Tagging results in “ref” format

- CONNL[2] format with a *spanGrp* of connl *type* attribute describing each token as a table (see Figure 12). This format takes advantage of the possibilities offered by the TEI spanGrp and span elements. It is very powerful as it enables the user to insert as many description levels as necessary (ten levels exist in the current version of CONLL). The implementation of CoreNLP (see below) takes full advantage of these possibilities.

```
<annotationBlock end="#T1393" start="#T1392" who="CHI" xml:id="au1282">
  <u>
    <seg>thank you , it was a singing party . </seg>
  </u>
  <spanGrp type="xpho">
    <span>ˈθæku ɪwəzəˈsɪŋɪŋpɑrdi</span>
  </spanGrp>
  <spanGrp inst="treetagger" type="conll">
    <span>1<spanGrp type="word">
      <span>thank</span>
      </spanGrp>
      <spanGrp type="pos">
        <span>VV</span>
      </spanGrp>
      <spanGrp type="lemma">
        <span>thank</span>
      </spanGrp>
    </span>
    <span>2<spanGrp type="word">
      <span>you</span>
      </spanGrp>
      <spanGrp type="pos">
        <span>PP</span>
      </spanGrp>
      <spanGrp type="lemma">
        <span>you</span>
      </spanGrp>
    </span>
```

Figure 12: Tagging results in “conll” format

Stanford CoreNLP

The Stanford Core Natural Language Processing[3] (CoreNLP) is a suite of tools (Manning et al. 2014) that can be used under a GNU General Public License. The suite provides several tools such a tokenizer, a part of speech tagger, a parser, a named entity recognizer, temporal tagging, coreference resolution, etc. All the tools are available for English, but only some of them are available for all languages. All software libraries are integrated into Java jar files, so all that is required is to download jar files from the CoreNLP

[2] https://www.conll.org/2019

[3] https://nlp.stanford.edu/software/

website[4] to use them with TEICORPO. Using the analyzer is similar to using TreeTagger. The "-model" and "-syntaxformat" can be used in a similar way to specify the grammatical model to be used and the output format. A command line example is:

java -cp "teicorpo.jar:directory_for_SNLP/*" fr.ortolang.teicorpo.TeiSNLP -syntaxformat svalue -model filename.tei_corpo.xml

The "directory_for_SNLP" is the name of the location on a computer where all the CoreNLP jar files can be found. Note that using the CoreNLP software is heavy on the memory resources of the computer and it is necessary to indicate to the Java software that it should use large amount of memory (for example to insert parameter -mx5g before parameter -cp to indicate that 5Gb of memory will be used for a full English analysis).

The -model parameter can take three values: english (use the full English grammar), french (use the full French grammar), or the name of a CoreNLP parameter file which specifies any type of analysis that is available in CoreNLP.

The -syntaxformat parameter can take four values: *conll* (a full analysis with all possible tools: 10 levels are produced in this case), *dep* (a syntactic analysis using a dependency grammar), and *ref* or *w* (only part of speech tagging and lemma).

3.4 Exporting the grammatical analysis

The results from the grammatical analysis can be used in transcription files such as those used by Praat and ELAN. A visual presentation of data using a partition-like presentation is very handy to represent a part of speech or a "conll" result. The orthographic line will appear at the top with divisions into words, into parts of speech, and other syntactic information below. As the result of the analysis can contain a large number of tiers (each speaker will have as many tiers as there are elements in the grammatical analysis, for example word, POS, lemma for TreeTagger, and 10 tiers for CoreNlp full analysis), it is interesting to limit the number of visible tiers, either using the -a option of TEICORPO, or limiting the display with the annotation tool.

An example is presented below in the ELAN tool (see Figure 13). The original utterance was "si c'est comme ça je m'en vais (if that's how it is, I'm leaving)". It is displayed at the top in the pink line. The analysis into words (line of numbers), into lemmas (third line), into parts of speech (POS: fourth line) and into orthographic words (final line) is displayed below. So for example, word 3 has the lemma "être" (to be), the POS VER:pres (verb in the present form), and it is the word "est" (is).

[4] https://stanfordnlp.github.io/CoreNLP/index.html#download

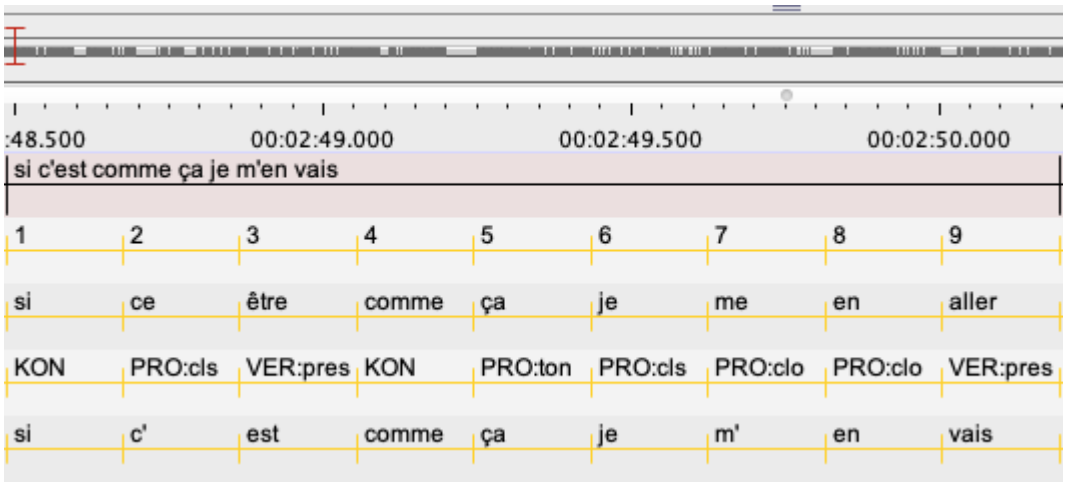


Figure 13: Example of TreeTagger analysis representation in a partition software

Export can be done towards a textometric software (see Figure 14). In this case, instead of using a partition representation, the information from the grammatical analysis is inserted at the word level in an XML structure. For example, in the case below, the TXM export includes Treetagger annotations in PoS adding lemma and pos attributes to the word element w.

```
<TEI file="/corpusformat/exemple/lil80.tei_corpo_ttg.tei_corpo.xml">
<teiHeader/>
<text>

<u age="28" end="2875.100" start="2873.395" who="MOT">
   <w age="28" lemma="you" loc="MOT" pos="PP">you</w>
   <w age="28" lemma="have" loc="MOT" pos="VHP">have</w>
   <w age="28" lemma="to" loc="MOT" pos="TO">to</w>
   <w age="28" lemma="rest" loc="MOT" pos="VV">rest</w>
   <w age="28" lemma="now" loc="MOT" pos="RB">now</w>
   <w age="28" lemma="." loc="MOT" pos="SENT">.</w>
</u>
<u age="4.0" end="2875.9700000000" start="2875.1000000000" who="CHI">
   <w age="4.0" lemma="yes" loc="CHI" pos="UH">yes</w>
   <w age="4.0" lemma="." loc="CHI" pos="SENT">.</w>
</u>
<u age="28" end="2877.8930000000" start="2875.9700000000" who="MOT">
   <w age="28" lemma="from" loc="MOT" pos="IN">from</w>
   <w age="28" lemma="your" loc="MOT" pos="PP$">your</w>
   <w age="28" lemma="big" loc="MOT" pos="JJ">big</w>
   <w age="28" lemma="singing" loc="MOT" pos="NN">singing</w>
   <w age="28" lemma="extravaganza" loc="MOT" pos="NN">extravaganza</w>
   <w age="28" lemma="." loc="MOT" pos="SENT">.</w>
</u>
```

Figure 14: Example of TreeTagger analysis representation exported in a textometric software

3.5 Comparison with other software suites

The additional functionalities available in the TEICORPO suite are close to those available in the Weblicht web services (Hinrichs & Vogel 2010). To a certain extent, the two suites of tools (Weblicht and TEICORPO) have the same purpose and functionalities. They can import data from various formats, run similar processes on the data, and export the data for scientific uses. In some cases, the services could complement each other or TEICORPO could be integrated in the Weblicht services. This is for example the case for handling the CHILDES format which in more functional in TEICORPO than in Weblicht (at the time of writing).

A major difference between the two suites is in the way they can be used and in the type of data they target. TEICORPO is intended to be used not as an independent tool, but as a utility tool that helps researchers to go from one type of data to another. For example, the syntactic analysis is intended to be used as a first step before being used in tools such as Praat, ELAN, or TXM. Our more recent developments (see Badin et al. submitted) made it possible to insert in the TEI files metadata stored in CSV files (including participant metadata). This makes it possible to achieve more powerful corpus analysis using a tool such as TXM.

Our approach is somewhat similar to what is suggested in the conclusion of Schmidt et al. (2017) where the authors describe a mechanism that makes it possible to use the power of Weblicht to process their files which are in the ISO/TEI format. A similar mechanism could

be used within TEICORPO to take advantage of the tools that are implemented in Weblicht. However, Schmidt et al. (2017) suggest in their conclusion that it would be more interesting to work directly on ISO/TEI files because they contain a richer format. This is exactly what we did in TEICORPO. Our suggestion would be to use the tools created by Schmidt et al. (2017) directly with the TEICORPO files, so that their work would complement ours. Moreover, in this way, the two projects would be compatible and provide either new functionalities when the projects have clearly different goals, or data variants when the goals are closer.

Conclusion

TEICORPO is a functional tool created by the CORLI network and ORTOLANG that converts files created by software specialized in editing spoken language data to the TEI. The result is fully compatible with the most recent developments of the TEI, especially those that concern spoken language material.

The TEI files can also be converted back to the original formats or to other formats used in spoken language editing to take advantage of their functionalities. This makes the TEI a useful pivot format. Moreover, TEICORPO allows conversion to tools dedicated to corpus exploration and browsing.

TEICORPO exists as a command line interface as well as a web service. It can thus be used by novice users as well as by advanced users or by developers of linguistic software. The tool is free and open source so it can be further used and developed in other projects.

TEICORPO is intended to be a part of a large set of tools using the TEI for linguistic corpus research. It can be used in parallel with or as a complement to other tools such as Weblicht or the Exmaralda tools (see Schmidt et al. 2017). A specificity of TEICORPO is that it is more suitable to process extended forms of TEI data (especially forms which are not inside the main ‘u’ field of the TEI). TEICORPO is also linked to TEIMETA, a flexible solution to describe spoken language corpora in a web interface generated from an ODD file (Etienne et al. under review). As the TEI enables metadata and data to be stored in the same file, sharing this format will promote metadata sharing and will keep them linked to the data during the lifecycle of the data.

Further developments can be made to provide a wider coverage of different formats such as CMDI or linked data, for edition or data exploration purposes or any other external tool such as grammatical analyzers or the visualization of multi-level annotations.

References

Badin, F., Liégeois, L, Thiberge, G., Parisse, C. (under review). Vers un outillage informatique optimisé pour corpus langagiers oraux en vue d’une exploitation textométrique : le cas des interrogatives partielles dans ESLO.

Barras, C., E. Geoffrois, Z. Wu, and M. Liberman 2000. Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication special issue on Speech Annotation and Corpus Tools, Vol 33, No 1-2, January 2000.

Benzitoun, Christophe, Karën Fort, and Benoît Sagot, 2012. TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. JEP-TALN 2012 - Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles, June 2012, Grenoble, France. pp.99-112. ⟨hal-00709187⟩

Bird, S., and Mark Liberman. 2001. “A formal framework for linguistic annotation.” Speech Communication 33: 23-60.

Boersma & van Heuven 2001. Praat, a system for doing phonetics by computer. Glot International 5(9/10): 341-345.

de Souza, Marli Aparecida Rocha, Marilene Loewen Wall, Ingrid Margareth Voth Lowen, and Aida Maris Peres 2018. "The use of IRAMUTEQ software for data analysis in qualitative research." Revista da Escola de Enfermagem da USP 52, e03353-e03353.

Etienne, C., Liégeois, L., Parisse, C. (under review). TEIMETA: An evolutive solution to describe spoken language corpora in a web interface generated from an ODD file.

Fleury Serge & Maria Zimina 2014. Trameur: A Framework for Annotated Text Corpora Exploration, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, August 2014, Dublin, Ireland, pages 57-61

Heiden, Serge. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, Yasunari Harada (Ed.), 24th Pacific Asia Conference on Language, Information and Computation - PACLIC24 (p. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University, Sendai, Japan.

Hinrichs, Erhard and Iris Vogel. 2010. CLARIN - Interoperability and Standards. In CLARIN deliverable D5.C-3. http://www-sk.let.uu.nl/u/D5C-3.pdf.

ISO 24624:2016. Language resource management – Transcription of spoken language. http://www.iso.org/iso/catalogue_detail.htm?csnumber=37338.

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

Lebart, Ludovic, André Salem, and Lisette Berry 1997. Exploring textual data. Vol. 4. Springer Science & Business Media.

MacWhinney, B. 2000. The CHILDES project: Tools for analyzing talk: Transcription format and programs (3rd ed.). Lawrence Erlbaum Associates Publishers.

MacWhinney, Brian 2007. "The TalkBank Project". In Beal, J.; Moisl, K. (eds.). Creating and Digitizing Language Corpora: Synchronic Databases, Vol.1. Houndmills, Basingstoke, Hampshire: Palgrave-Macmillan.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

Pincemin, Bénédicte 2011. "Sémantique interprétative et textométrie–Version abrégée." Corpus 10, 259-269.

Pincemin, Bénédicte, Serge Heiden, and Matthieu Decorde 2020. "Textometry on Audiovisual Corpora." In 15th International Conference on Statistical Analysis of Textual Data JADT 2020. University of Toulouse.

Schmid, Helmut, 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Schmid, Helmut, 1995. Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Schmidt, Thomas 2004. Transcribing and annotating spoken language with EXMARaLDA, In Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004. Paris ELRA.

Schmidt, Thomas, 2011. A TEI-based Approach to Standardising Spoken Language Transcription. Journal of the Text Encoding Initiative, 1, 1-22.

Schmidt, Thomas, Hanna Hedeland, and Daniel Jettka, 2017. Conversion and annotation web services for spoken language data in CLARIN. Selected papers from the CLARIN Annual Conference 2016. Linköping Electronic Conference Proceedings 136: 113–130.

Schmidt T and Schütte W 2010. "FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction", In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta, May, 2010. European Language Resources Association (ELRA).

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. 2006. ELAN: a Professional Framework for Multimodality Research. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.