TEITOK: Text-Faithful Annotated Corpora

Maarten Janssen
CLUL, Universidade de Lisboa
Lisboa, Portugal
Maarten.Janssen@campus.ul.pt

Abstract

TEITOK is a web-based framework for corpus creation, annotation, and distribution, that combines textual and linguistic annotation within a single TEI based XML document. TEITOK provides several built-in NLP tools to automatically (pre)process texts, and is highly customizable. It features multiple orthographic transcription layers, and a wide range of user-defined token-based annotations. For searching, TEITOK interfaces with a local CQP server. TEITOK can handle various types of additional resources including Facsimile images and linked audio files, making it possible to have a combined written/spoken corpus. It also has additional modules for PSDX syntactic annotation and several types of stand-off annotation.

Keywords: annotated corpora, document transcription, TEI

1. Introduction

Corpora that are based on manuscripts typically come in two flavours. On the one hand there are those created by philologists, which focus on faithfully representing the content of the manuscript and include line breaks, typesetting information (colour, italics), changes of hand, deleted fragments, etc. And on the other hand there are those created by (corpus) linguists, which capture linguistic meta-information including POS tags, lemmas, normalized orthography, semantic classifications, grammatical parses, etc.

There are few corpora that include both, which can be attributed largely to three factors. Firstly, the two types of corpora are created by different audiences that often don't see the value of the other type of information: philologists see little reason for including POS tags, while corpus linguists tend to view typesetting information as ephemeral. Secondly, both types of corpora are labour-intensive, and combining both types of information makes it even more so. And thirdly, there are no or hardly any tools that facilitate the creation of such combined corpora.

The first point is mostly due to a mutual misunderstanding. On the one hand, using linguistic annotations without access to the textual information can often lead to conclusion that have nothing to do with the actual manuscript, but more with the choices by the transcriber: manuscript are often hard to read and degraded, and transcription often involves reconstructions and interpretations. And from the other side, linguistic annotations are not merely useful for grammatical or statistical analysis, but also provide richer search options on the manuscript transcriptions, for instance by searching lemma rather than (often deviant) orthography.

The fact that a combined corpus is labour intensive is very true, but existing linguistically annotated manuscript-based corpora are almost always based on prior (textually annotated) transcriptions. Most linguistically annotated corpora do not keep the textual annotation, however, due to the lack of appropriate tools. TEITOK is an online tool that aims to provide a solution to that: a tool that allows adding layers of linguistic annotations to textually annotated texts. This results in a corpus combining both types of information, which adds value in all those contexts where there is additional information in the source document or where easy reading is advantageous. This not only includes corpora of old manuscripts, but also for instance of learner texts, texts from less resourced languages, critical editions, and texts used for educational purposes.

TEITOK was developed at the CLUL institute in Lisbon, and is currently a (for security reasons) private repository on GitLab that can be obtained by contacting the author. The repository is frequently updated with bug fixes as well as new features. More information about the framework can be found on the project web-site: http://teitok.corpuswiki.org

2. Basic Design

TEITOK is a framework for creating, maintaining, and publishing annotated corpora. It is a web-based environment written mostly in a combination of PHP and Javascript. In TEITOK, a corpus consists of a collection of XML files, each in the Text Encoding Initiative (TEI) format (Burnard and Bauman, 2013), with a slightly modified tokenization system (see section 2.1.). The system makes it easy to display each XML file (see section 2.2.), edit metadata (section 2.3.) and individual tokens (section 3.3.), and search through the corpus (section 3.2.).

The basic functionality of TEITOK makes it most compatable with other systems for creating TEI documents and publishing them online, such as TXM (Heiden, 2010). However, for many users it will be mostly a way to publish and search an online CQP corpus, making it more comparable to for instance CQPWeb (Hardie, 2012). And the modular design with various options allowing for syntactic annotation, error annotation, etc. make it more like a suite of tools for linguistic corpora comparable to for instance FoLiA (van Gompel and Reynaert, 2013).

TEITOK is built upon the older projects CorpusWiki (Janssen, 2012) and Spock (Janssen and Freitas, 2008). It is written as a multipurpose environment that can be used in a variety of different types of projects, including historical corpora, learner corpora, and spoken corpora. Given the

different needs of the different types of corpora, TEITOK is highly customizable. Apart from an extensive settings file that allow projects to customise the system in a wide variety of ways, the system itself has a modular design in which there is a range of scripts that directly interact with the XML files. It is possible to add additional scripts to incorporate new functionalities into the system, or even overwrite existing functionality on a per-project basis.

2.1. Tokenization

Tokenization in TEITOK is added inline to the TEI document. The linguistic annotation in TEITOK deviates from the standard TEI practice in several ways, but is both TEI compatible, and easy to convert into a pure TEI format. In TEITOK, tokens are represented by a <tok> element, and all linguistic annotation is represented as attributes over these elements placed around each word. In relatively rare cases, the tokenizer has to split existing XML tags, for instance when a word-and-a-half in the text were underlined. The inline tokenization creates a straight-forward merge of the textual and the linguistic annotation: the textually annotated corpus is everything minus the <tok> elements, whereas the linguistic corpus is the sequence of <tok> elements, which can be rendered in a verticalized format when needed. Both annotations can be edited relatively independently, meaning that the source text can be edited without affecting the linguistic annotation, which is often required in the type of corpora TEITOK was designed for.

The <tok> element is largely identical to the <w> tag in TEI, but is named differently in part since punctuation marks are considered tokens, but are not typically considered words.

One of the specific features of manuscript corpora is that there are various ways to encode the orthography of the word, which are represented in TEI by a <choice> element: abbreviations (frequent in manuscripts) can be kept or expanded, deviant forms can be maintained or normalized. All these different forms are relevant for different purposes and should ideally all be kept. There are however some additional differences.

In TEITOK, the <choice> element is not used, but rather <tok> elements can have multiple orthographic forms for the same word. These different forms are modeled as attributes rather than as XML element, to assure that they are string-based elements that can be used in NLP processing. So where TEI uses a structure like the one in figure 1, TEITOK uses a single token with different attributes like the one in figure 2. The TEITOK representation can easily be converted into <choice> elements, although the reverse is not true, since <choice> can be on many other levels apart from the word, for instance providing alternatives for entire sections.

Where there is a limited number of options in a <choice> element, there can be as many orthographic attributes on a <tok> as needed. This because the reality of manuscripts can become rather complicated: in a Ladino corpus currently being developed in TEITOK, there is the original orthography, and expanded form, and the normalized form in current spelling. However, since the orignal documents are mostly in Hebrew characters, it is very useful for accessibility to have a romanized orthography. And to reach a wider audience and given the high similary with Spanish, a variant of each token in current Spanish spelling is kept as well. This hence leads to a total of five different orthographic realisations, several of which are not foreseen in the TEI format.

```
<w>
   <choice>
      <org>ob-<lb/>scuras</org>
      <reg>oscuras</reg>
   </choice>
</w>
```

Figure 1: A TEI <choice> example.

```
<tok form="obscuras" nform="oscuras">
   ob-<lb/>scuras
</tok>
```

Figure 2: A TEITOK <tok> example.

Although many orthographies can be needed, in most cases all different forms will be identical for the majority of words. Keeping a number of copies of the same form would not only be inefficient, but also hard to maintain: if there is an error in one, all of them would have to be corrected. Therefore, in TEITOK there is an inheritance hierarchy for the different forms, where (for instance) in those cases where there is no explicit normalized form, it is assumed to be identical to the written form. In the interface, it is possible to use these multiple orthographic forms to switch between different orthographic realizations of the same text, for instance switching between the original orthography and the normalized version. Missing or wrong normalization are easy to spot when presented as running text, and wrong normalizations are often an indication of other errors in POS and lemma.

One of the eternal problems in annotated corpora is how to handle contractions: whether to treat them as one token or two. That is why TEITOK takes a mixed approach: the <tok> elements are roughly speaking orthographic words. In the case of contractions, two or more grammatical words are inserted as children of the <tok>, called <dtok>. In this manner, it is possible to associate POS and lemma to the grammatical words, but associate the normalized orthography to the orthographic word.

2.2. Display

The annotated text is rendered directly in the browser. That is to say, the XML body of the text is inserted into the HTML page, where (project-specific) CSS rules define how to display the different XML elements in the text - for instance by using a greyed-out strikethrough for <del> elements.

When there are facsimile images in the text, each page can (optionally) display the image next to the text. And when there are various orhographic forms it is possible to switch between the various orthographical realizations of the text

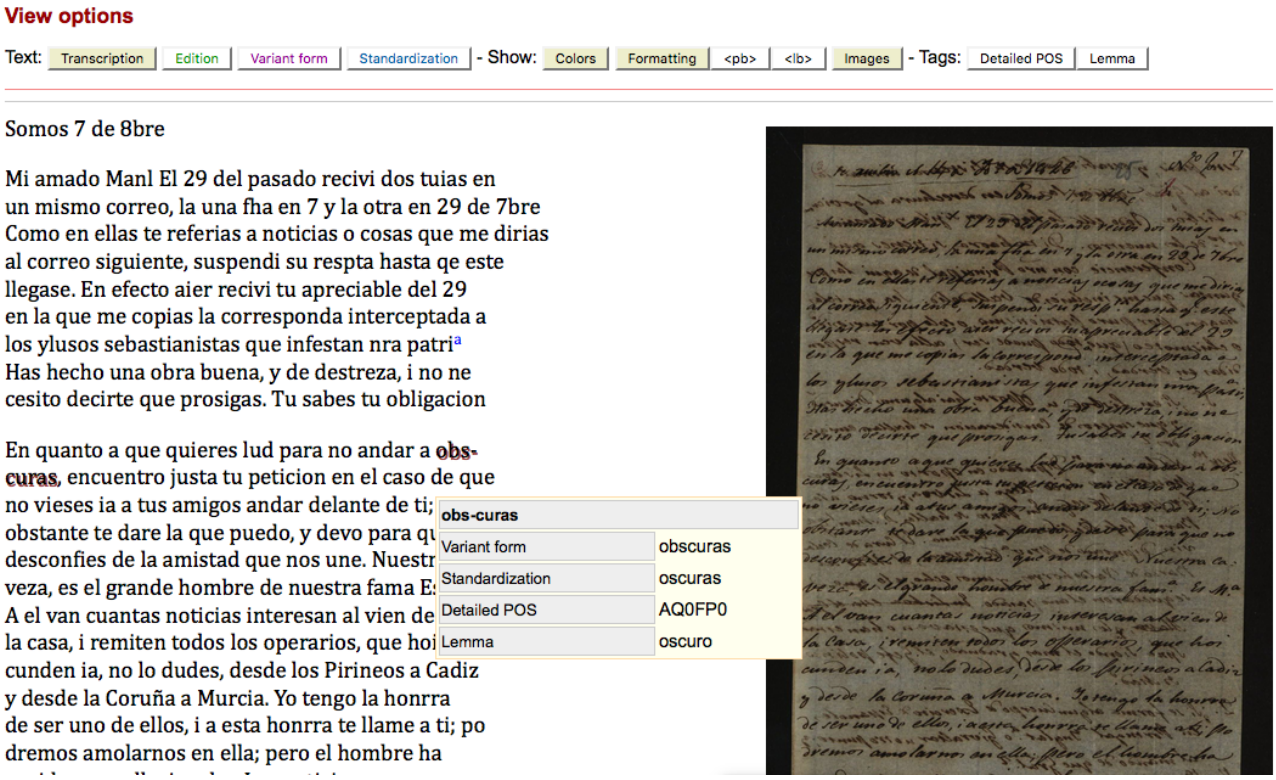

Figure 3: A transcribed letter from 1826 from the Post Scriptum project

by clicking the buttons on top. Figure 3 shows a manuscript of the Post Scriptum project (see section 4.1.) with the facsimile image next to it and the text in original orthography. The pop-up window show the annotation data for the word *obscuras* (obscure), with its original orthography (split across a linebreak), the modern orthography, the POS tag, and the lemma. Clicking on "Standardization" will replace the word *obscuras* with its current spelling, *oscuras*, as well as changing all other tokens with their normalized orthography.

The CSS rendering creates visually attractive, easy to read representations of the text, making the texts usable for a wide range of purposes, not only for linguistic studies, but also for use in classrooms, use in historic or sociological studies, etc. Furthermore, the graphical interface makes it easy to spot potential errors in the transcription, which with the easy token-based editing option explained in the next section makes TEITOK an efficient editing environment for manually transcribed corpora.

2.3. Metadata

One of the typical obstacles for using TEI is its complicated metadata system. To make it easier to work with metadata, TEITOK provides the option to define a set of metadata that are relevant to the project, and build a edit table out of it. An edit table is an HTML table that describes all the relevant fields, with XPath definitions alongside of them that specify a specific field in the teiHeader. TEITOK then uses this table to generate a simple HTML form that replaces all XPaths with HTML input fields, and does a lookup in the XML file for the corresponding value, allowing the user to edit or add information. After clicking save, the information is written back into the XML file in the appropriate location, where nodes are created when they do not exist yet. That way, sorting out the correct representation of all relevant information in the teiHeader, and define that in terms of XPath has to be done only once, after which all (administrative) users can simply edit the simple HTML table without having to have intricate knowledge of either TEI or XML.

The same type of XPath-based table is also used to define which information from the teiHeader should be displayed on top of each file in the interface, where it is possible to define both a short and a long version of the header.

3. Automatic Processing

TEITOK is intended as an online environment with a low threshold, making it as easy as possible for people to work. As such, the framework provides one-click options to apply automatic methods to the TEI documents, such as tokenization, POS tagging, lemmatization, and creating a CQP corpus. It is also possible to add custom (command line) script under a button, and more scripts and modules are added frequently.

3.1. POS Tagging

For POS tagging, TEITOK uses NeoTag (Janssen, 2012), a language-independent HMM tagger which uses the internal word structure for tagging OOV items. NeoTag can not only be used to tag texts in the corpus, but can furthermore use the corpus itself as a training corpus, to build a tagger that is highly specialized to the kind of texts in the corpus,

and will improve as the corpus grows. Both tagging and training are done by a simple click, and tagging is done directly on the XML file.

It is hard to say anything generic about the accuracy of NeoTag since the accuracy too much depends on the language in question, the tagset used, and the size and quality of the training corpus. But especially with a sizeable training corpus the accuracy of NeoTag is comparable to that of other recent taggers, and often better when the tagger is trained on the corpus itself (especially for corpora with a particular style).

To tag either languages for which no NeoTag parameters are (yet) available, or where the use of other existing taggers is preferable, it is easy to use most line-based taggers using a simple script. TEITOK comes with an example script to use Freeling, where the text is first exported to a verticalized format, then ran through the tagger, and finally the resulting tagged text is imported back into the XML.

3.2. Querying the Corpus

Although it is possible to search through XML files directly, XML based search methods are cumbersome and slow for linguistic purposes. Therefore, a CQP version (Christ et al., 1999) of the corpus is built automatically, and can be queried using CQL directly on the website. The query is run through the CQP engine, and the results are rendered in the browser, showing the KWIC line for each result, making a system comparable to for instance CQPWeb (Hardie, 2012) or Bwananet (Vivaldi, 2009).

In the export to CQP, it is necessary to decide which orthographic forms to put in the CQP corpus, and whether to use the `<tok>` or `<dtok>` in the case of contractions, where by default the grammatical words are exported. When various orthographic forms are exported, say the original as well as the normalized orthography, it becomes possible to use CQP to search for orthographic changes or errors (depending on the corpus), for instance one can study the development of the word-initial *h* in Spanish by searching for all words that used to be written with an *h* but no longer are.

Although TEITOK originally worked in much the same way as CQPWeb rendering CQP results in the browser, it now works in a somewhat more involved way: the CQP corpus is created by a dedicated tool called *tt-cwb-encode*, which like *cwb-encode* builds CQP corpus files, except that it builds them directly from the XML files. And while building the CQP files, it also keeps a file that indicates the offset position of each token in the XML file so that it can be rapidly recovered.

This more direct way has two upshots. The first is tt-cwb-encode can include additional levels of s-attributes from the XML files, such as utterances, sentences, rendering elements, etc. and even include stand-off annotation from external files (see section 4.2.).

The second upshot is that rather than displaying strings in the KWIC list, the results directly display an XML fragment from the original file. This in turn has a number of positive consequences: it means that all elements that are not exported to the CQP corpus, such as deleted words and typographic information, are displayed in the result nevertheless. It also means that just like in the XML display it is possible to switch between say the original and the normalized orthography. And it means that contractions are shown as contractions and not as separated words. For instance, for a contraction such as the Spanish *del*, you search in CQP for the grammatical words *de* and *el*, but see the actual spelling *del* in the result.

While CQP is meant for corpus linguistic queries, the interface also allows searching for documents: by not selecting any token-based query, but only ask for documents of a specific date, location, language, or other metadata criteria, the result will not be a KWIC list or context display, but rather a list of documents with the appropriate metadata characteristics. This makes TEITOK an accessible framework for for instance historians looking for specific manuscripts rather than for linguistic constructions.

3.3. Editing Tokens

Editing the linguistic annotation is very easy in TEITOK: in the text-view of the text, all annotations on a words are show on mouse-over. Any token that needs to be corrected can be edited by simply clicking on the word, which will open an HTML form representation of the content of the token. Hitting save will directly update the content of the token in the XML file. So when for instance normalizing a text, you just have to read the text and correct any token that is not normalized yet.

For more structural editing sessions, editing can also be done in a verticalized table format, where you first define which `<tok>` attributes you want to see, and which ones you want to edit, and then system then builds a table with one token per row, with editable boxes in the appropriate cells.

However, for real structural changes, especially in larger corpora, it is often necessary to first define exactly which tokens you want to edit, and the most efficient way to express that is using the CQP query language. It is possible to run a CQP search, open the context of all results that need to edited, and click on the token to edit it. But doing so is very slow and labour intensive. Therefore, it is also possible in TEITOK to run a CQP query, and correct all results directly. For instance we can search for all occurrences of *obscura* in the corpus, and set the normalized form to *oscura* for all of them in one go.

4. Example Projects

Given that TEITOK is a relatively young framework, there are not that many projects using it that are already sufficiently developed to be accessible online. However, apart from a growing number of projects in the development phase, there are several projects that are already available online. Three of these projects are described below, all of them internal projects from the CLUL institute. Apart from demonstrating the versatile use of the TEITOK environment, it also illustrates some of the custom modules that have been added to TEITOK.

4.1. Post Scriptum

Post Scriptum (CLUL, 2014) is a European project consisting of unpublished epistolary writings written by authors from different social backgrounds written between 1500

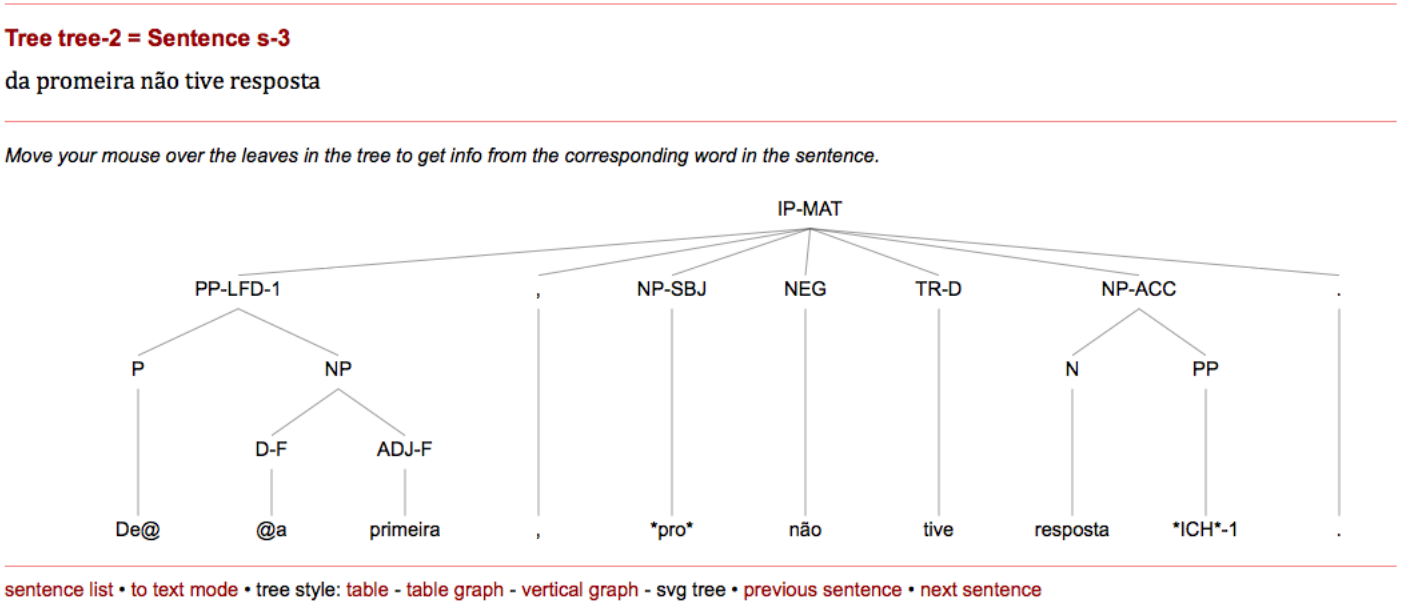

Figure 4: A syntactic tree for a sentence from Post Scriptum

and 1900. The corpus is of specific interest since by the nature of the letters the corpus is as close to a spoken corpus from before there were audio recordings as possible. The corpus currently consists of 1520 letters written in Spanish, with a total of 584k tokens, and 1456 letter written in Portuguese, counting 579k tokens.

Each letter is transcribed in TEI transcribing all typographic marks such as underlinings, capitalizations, linebreaks and pagebreaks, and each page is provided with a facsimile image. An example of a letter was shown in figure 3.

Each token contains the original orthography, the expanded text in the original orthography in the case of abbreviation, as well as a normalised orthography in modern spelling. Also, each token is provided with a manually verified POS tag using the EAGLES tagset for Portuguese and Portuguese, and a lemmatized form. On top of the transcription itself, each letter is adorned with a rich set of metadata, including biographic data about the author and recipient of the letter, a description of the historic context, a set of keywords, and geographical data about the origin of the letter.

In Post Scriptum, the TEI transcription files are first generated outside of TEITOK using Oxygen, and when the transcription is done, all corrections, normalizations, and annotations are done using TEITOK. For the first hundred or so texts, POS tagging was done on the manually normalized text using the Spanish and Portuguese parameter files of Freeling (Padró et al., 2010) using a set of scripts to export a verticalized version of the text, tag it, and import the tags back into the XML file. Since the texts are of a particular style with a reduced lexicon, NeoTag trained on the already tagged files relatively quickly outperformed Freeling and all texts since have been tagged directly with NeoTag. An additional advantage of using NeoTag directly is that TEITOK can be told to use the teiHeader to determine whether to use the Spanish or Portuguese parameter set.

One of the modules that was developed specifically for the PS project is a module for syntactic annotation in PSDX (the XML variant of the Penn Treebank format). The PSDX files are kept outside the corpus XML files in order to speed up querying, and they are aligned on token-level to the TEI XML. Each tree can be displayed in a number of different views, an example is given in figure 4. When displaying the tree, all token-based annotations are shown when moving the mouse over a leaf, and the textually annotated version of the text is shown above the tree. Trees can be searched through online using XPath search queries, and the site shows a list of typical queries pre-defined in XPath.

Another recently added module is a module for a *document map*: each letter is provided with the geographic coordinates of where it was sent from, and these coordinates can be plotted onto the world map, making it easy to find documents from a specific region.

4.2. COPLE2

The *Corpus de Português Língua Estrangeira / Língua Segunda* - COPLE2, (CLUL, 2015) is an Portuguese second/foreign language learner corpus, consisting of written and spoken texts produced by students of Portuguese language courses in the Instituto de Cultura e Língua Portuguesa (ICLP) and the Centro de Avaliação de Português Língua Estrangeira (CAPLE) at FLUL. The corpus currently consists of 978 texts from students of 14 different native languages, with a total of 182k tokens.

Each text in COPLE2 contains metadata about the student, including his level of proficiency in Portuguese, his native tongue, and his age, etc. Written texts contain all correction, deletions, and additions by both the student and the teacher. Teacher corrections are displayed in different colour from corrections by the student himself. In the CQP corpus, both the form written by the student and the corrected form provided by the teacher are searchable, making

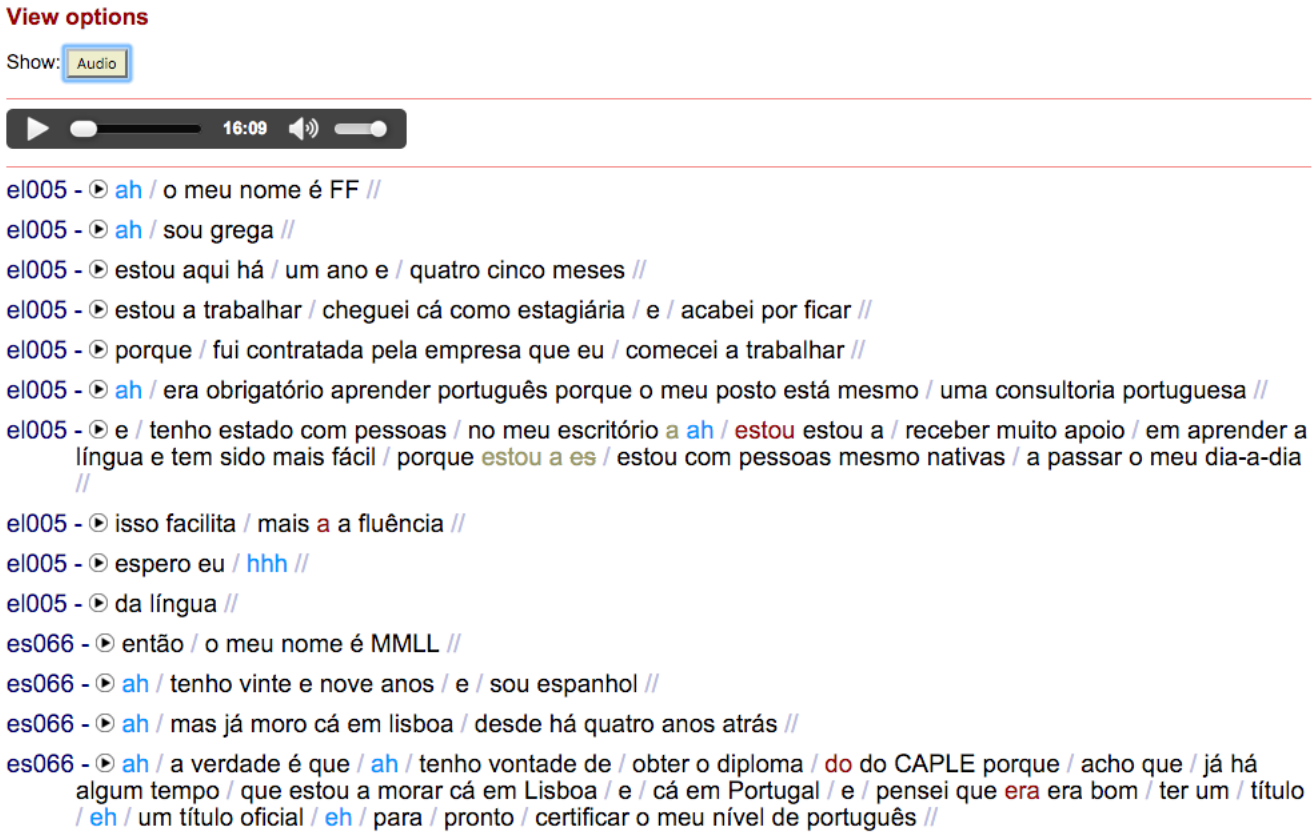

Figure 5: A spoken document from the COPLE project

it possible to search for various types of orthographic errors.

COPLE2 contains a section of spoken data, which can be handled alongside of written data in TEITOK in a mixed modal corpus. Spoken data in TEITOK are not displayed in tiers as is common in frameworks such as ELAN, Praat, or EXMARaLDA, but rather in lists of utterances (as is also done in the web version EXMARaLDA for instance). Just like with written texts, spoken texts are transcribed in TEI where CSS takes care of rendering the text including pauses and other non-token based annotations. The sound file is displayed on top of the text, and if the utterances are time-aligned, each utterance has a play button enabling listening to each individual utterance. An example of a spoken text can be seen in figure 5. When using an only spoken corpus, it is possible to set up the CQP search in such a way that each query displays the context utterance instead of a KWIC list, and each utterance can be directly listened to from the CQP result list.

In COPLE2, the TEI files for written texts were created in Oxygen, while the spoken data were transcribed and aligned in EXMARaLDA, and then converted to the TEITOK format using a simple script. The way information is stored in TEI and EXMARaLDA is fundamentally different, but the script is reversible as long as there are only orthographic tiers in the original. POS tagging was done directly in TEITOK using a parameter set in NeoTag that was trained on the CRPC corpus (Bacelar do Nascimento, 2000).

One of the modules that was developed specifically for the COPLE2 project is a module for adding layers of standoff annotation. The stand-off annotations are kept in a separate file, linked to the XML file by token IDs. The annotated segments can be displayed next to the original text, and part of the text will highlight when moving the mouse over the annotation. In COPLE2, these stand-off annotations are used to mark errors in the texts - annotations that can not be added directly to the XML since they often cross with typographic XML tags. However, the stand-off annotation module is not specific for error anntation, but can be used for any type of annotation, with a central definition of the type of annotation and the attribute to store for each annotated segment.

4.3. EFFE

Escreves como Falas - Falas como Escreves - EFFE, (Rodrigues et al., 2015) is an L1 Portuguese learner corpus, consisting of descriptions of image stimuli. The metadata of the corpus consist of the age and class level of the pupil, as well as a description of the task they performed. To identify whether orthographic mistakes in the texts were orthographic or phonetic, the pupils were furthermore asked to read out some of the words. The texts contain the original spelling as well as a normalized orthography, a POS tag and a lemma. The corpus currently consists of 295 texts written by student between the age of 7 and 9, with a total of 46.127 tokens.

In EFFE, the TEI documents were written directly in

TEITOK, where the headers were created using the simple metadata tables, and the transcriptions were done in the built-in XML editor (Ace). The tagging and lemmatization was done using Neotag with the parameter files from the Post Scriptum project.

In the TEITOK interface, each text comes with the original image that was described, the image of the text itself, the sound file containing some of the words of the text, as well as the transcription of the text, where the words spoken out loud are coloured differently in the text. In this way, corpora with texts that are first written down and then recorded orally (or the other way around) can have both types of media support. For a small number of texts the audio has been aligned with the words in the text they represent, and play button can be displayed before the corresponding words.

5. Conclusion

The development of TEITOK started not long ago, although it is built on the basis of the older CorpusWiki system, sharing much of its infrastructure. It is clear by the growing number of users that TEITOK fills a need that many projects face: to be able to work with linguistic annotation while keeping other types of annotation as well.

Apart from the projects listed in section 4., there are other projects under development in TEITOK that are likely to be release in the near future, which include historic corpora (for Portuguese, Ladino, Spanish, Galician, and Slovenian), spoken corpora (for African varieties of Portuguese and for Brazilian Portuguese), learner corpora of various types, and a reference corpus. For several of these projects, adaptations were made to the platform, to make TEITOK work well with texts ranging from single sentences (in psycholinguistic experiments) to manuscripts of hundreds of pages, with spoken texts and written texts, with texts with virtually no mark-up as well as with texts with heavy textual annotation.

Practice has shown that there are currently two large hurdles for users of TEITOK. The first is setting up the system: TEITOK is so customizable that it becomes hard to see what needs to be defined and where to put the definitions. We are constantly looking into ways to help this process along, but given the large amount of options it will always involve some getting used to. The other hurdle is that TEITOK does not provide a WYSIWYG editing environment: although HTML and CSS work very nicely for visualizing XML, it is a bad environment for editing XML since the browser Dom engine modifies the XML - adding and deleting white spaces, changing self-closing tags, etc. For that reason, for graphically editing TEI documents third-party tools have to be used, after which the XML files can be uploaded to TEITOK. It is however more recommendable to write the XML directly, either in a program like Oxygen, or directly in TEITOK. Both hurdles have proven easy enough to overcome, and once users start really using the system, it has proven to be intuitive to use.

In general, TEITOK has proven a valuable tool for a wide range of corpus types where data are hard to come by and/or labour intensive to create. And the resulting corpora are not only appreciated by corpus linguists, but also by the larger academic community including psycholinguists, language teachers, and historians.

6. Bibliographical References

Bacelar do Nascimento, M. F. (2000). Corpus de référence du portugais contemporain. In M. Bilger, editor, *Corpus, Méthodologie et Applications Linguistiques*, pages 25–30. H. Champion et Presses Universitaires de Perpignan, Paris.

Burnard, L. and Bauman, S. (2013). TEI P5: Guidelines for electronic text encoding and interchange. Technical report, Text Encoding Initiative Consortium, Charlottesville, Virginia.

Christ, O., Schulze, B., Hofmann, A., and Koenig, E. (1999). The IMS corpus workbench: Corpus query processor (CQP): User's manual. Technical report, IMS, University of Stuttgart.

Hardie, A. (2012). CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17 (3).

Heiden, S. (2010). The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme. In *24th Pacific Asia Conference on Language, Information and Computation*, Sendai, Japan.

Janssen, M. and Freitas, T. (2008). Spock: a spoken corpus klient. In *Proceedings of LREC 2008*, Marrakech.

Janssen, M. (2012). NeoTag: a POS tagger for grammatical neologism detection. In *Proceedings of LREC 2012*, Istanbul.

Padró, L., Collado, M., Reese, S., Lloberes, M., and Castellón, I. (2010). Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10)*, La Valletta, Malta, May.

van Gompel, M. and Reynaert, M. (2013). FoLiA: A practical xml format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, 12/2013.

Vivaldi, J. (2009). Corpus and exploitation tool: IULACT and bwanaNet. a survey on corpus-based research. In *Actas del I Congreso Internacional de Lingüística de Corpus (CICL-09)*, pages 224 – 239, Murcia.

7. Language Resource References

CLUL. (2014). *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. Universidade de Lisboa, CLUL, http://ps.clul.ul.pt.

CLUL. (2015). *Learner Corpus of Portuguese L2 - COPLE2*. Universidade de Lisboa, CLUL, http://alfclul.clul.ul.pt/teitok/learnercorpus.

C. Rodrigues and M. C. Lourenço-Gomes and I. Alves and M. Janssen and I. L. Gomes. (2015). *EFFE-On - Escreves como falas - Falas como escreves?* Universidade de Lisboa, CLUL, http://alfclul.clul.ul.pt/teitok/effe.