Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5692–5695
Marseille, 20-25 June 2022

Atril: an XML Visualization System for Corpus Texts

Andressa Gomide, Conceição Carapinha, Cornelia Plag
CELGA-ILTEC - Universidade de Coimbra
Faculdade de Letras - 3004-530, Coimbra - Portugal
{andressa.gomide, mccarapinha, cornelia.plag}@fl.uc.pt

Abstract

This paper presents Atril, an XML visualization system for corpus texts, developed for, but not restricted to, the project *Corpus de Audiências* (CorAuDis), a corpus composed of transcripts of sessions of criminal proceedings recorded at the Coimbra Court. The main aim of the tool is to provide researchers with a web-based environment that allows for an easily customizable visualization of corpus texts with heavy structural annotation. Existing corpus analysis tools such as SketchEngine, TEITOK and CQPweb offer some kind of visualization mechanisms, but, to our knowledge, none meets our project's main needs. Our requirements are a system that is open-source; that can be easily connected to CQPweb and TEITOK, that provides a full text-view with switchable visualization templates, that allows for the visualization of overlapping utterances. To meet those requirements, we created Atril, a module with a corpus XML file viewer, a visualization management system, and a word alignment tool.

Keywords: xml visualization, corpus software, utterance alignment

1. Introduction

Many corpora are heavily marked with structural tags, normally adopting Extensible Markup Language (XML) formatting. This type of annotation benefits different types of corpora. It can mark boundaries within texts, allowing for searches within specific sections, as is the case with The English Scientific Text Corpus (SciTex) (Degaetano-Ortlieb et al., 2013). It also allows for the creation of subcorpora derived from these identified boundaries. For instance, for spoken corpora with rich speaker and text metadata such as the Spoken BNC 2014 (Love et al., 2017) you can create subcorpora by filtering the utterances by speakers' profile and conversation context. XML mark-up is also useful to indicate events and situations occurring in parallel to the discourse, such as pauses in a speech and unintelligible words in a document.

In the context of Corpus Linguistics (CL), this structural annotation, also known as markdown, is normally the first of two steps of text annotation. In small or carefully prepared corpora, this annotation tends to be done manually or semi-automatically. The semi-automatic way is often done by automatizing the identification of patterns and their respective replacement with XML tags.

The markdown process is normally followed by a second and automatic step for the annotation of each token, being lemma and parts-of-speech the most common types of annotation. Some corpus analysis software has an annotation built-in system and this step is done concurrently with the corpus loading or installation, as it is the case with SketchEngine (Kilgarriff et al., 2014).

Once the processes above are done, the usage and visualization of the XML tags is subject to the functions offered by the chosen corpus software. This restriction yields, especially for richly marked texts, two main problems. First, when all mark-ups come together, the richness of information negatively affects the user interaction with the text itself. Each type of information requires careful thought on what is the best user-friendly way of rendering the extralinguistic information encoded in the XML tag. However, most corpus tools do not offer the possibility to customize how each tag is stylized and displayed.

Another issue is that for corpora with several XML tag types, the user is more likely to want to browse the entire text while performing corpus analysis. However, widely used corpus tools do not offer a full view of each text.

2. Motivation

2.1 CorAuDis

This paper presents a system developed within the context of *Corpus de Audiências* (CorAuDis), a singular case where European Portuguese is concerned.

The sessions constituting CorAuDis were recorded at the Coimbra Court in 2016 and 2017 and involve criminal proceedings. At present, the corpus is still under construction and will comprise a set of 17 sessions, corresponding to 84 hours of recording. The audio material was subject to transcription and anonymisation of critical data. Participants are only identified through their interactional roles.

The transcriptions gathered in this corpus are presented in full text format and allow research concerning different linguistic aspects. Since the data have not been organised for a single specific purpose, they allow for different approaches supported by different theoretical frameworks such as Discourse Analysis, Conversation Analysis, and Pragmatics (Carapinha and Plag, 2018:177).

Due to the complex transcription process and the wide range of oral phenomena to be considered (e.g. different types of vocalizations, inaudible sequences, overlapping utterances, backchannelling), for the project we needed a tool that offered (a) a simple customization tool for the visualization of the XML tags; (b) a full-text view; and (c) a tool for the alignment of co-occurring tokens on its exact matching point.

2.2 Existing Tools

CorAuDis' first version was made available via TEITOK (Janssen, 2016). TEITOK is a web application that allows for the creation, edition and publishing of corpora. It does allow for an easy full-text view and it offers some pre-defined visualization for XML tags, following the Text Encoding Initiative (TEI). Users can also import their customized stylesheet to edit and to personalize the project, but that requires some coding skills.

Similar skills are necessary to customize the XML visualization for corpus texts in CQPweb (Hardie, 2012). Additionally, this function is only available for users granted with administrator permission. Another downside

of CQPweb when XML visualization is concerned is that a full-text view is not possible. However, CQPweb offers statistics tools (e.g. collocations and distribution) that can be applied to sequences of tokens within a given XML tag, which can come in handy when doing corpus analysis. Similarly, SketchEngine provides users with powerful corpus analysis tools (e.g. word sketch, n-grams), but does not allow for full-text view or easy style customization of XML tags.
In terms of alignment of speeches, EXMARaLDA Partitur Editor (Schmidt and Wörner, 2014) allows users to align transcriptions in multiple layers and offers a full display of the text. It does not offer; however, robust corpus analysis like the ones available at CQPweb and SketchEngine.
The pieces of software above are powerful tools for corpus creation, edition, analysis, and distribution. However, they still lack some elements necessary to our project (2.1).

tools	Text view	simplified	
		XML editing	Word Align
EXMARaLDA	yes	no	yes
TEITOK	yes	no	no
CQPweb	no	no	no
SketchEngine	no	no	no

Table 1: XML editing in corpus tools.

3. Atril

3.1 Assumptions and Technical Aspects

To completely fulfil our needs, we developed Atril, a system with an XML file viewer, a visualization management system, and a word alignment tool. To develop the system, we attempted to follow the Unix tools philosophy of writing programs (i) that do just one thing, but do it well; (ii) that can work together with other tools; and (iii) that add complexities only when extremely necessary (Raymond, 2003:12).
With that in mind, we wrote Atril in XSLT and JavaScript, so it can be easily implemented into web-based applications such as CQPweb and TEITOK. We also aimed at designing an interface as clean as possible, to prevent the display of unnecessary information.

3.2 Functions

3.2.1 Main Page and Visualization Page

Atril will later be implemented as an extra module to TEITOK and CQPweb. For now, it works on its own and is accessed via the web browser. The first page (figure 1) lists all the texts available and clicking on a text leads to its visualization page. Once Atril is implemented to CQPweb and TEITOK, the visualization page will also be retrieved via a hyperlink displayed on the results of concordance lines and whenever the text metadata is provided. Clicking on the text name or ID will take the user to the visualization page, which offers two main functionalities: *Vis Edit* and *Vis Switch*.

3.2.2 Vis Edit

Vis Edit allows the user to customize the visualization of the XML tags by simply choosing and clicking on the options provided by Atril.

To keep the interface as clean as possible, the options for tag visualization edition appear conditionally, following the scheme in figure 2, so that only relevant information is displayed. This means that editing options are presented hierarchically and sub-options will appear if and only if they are possible. This not only prevents the tool from displaying too much information at the same time, but it also works as a validation system for each input.

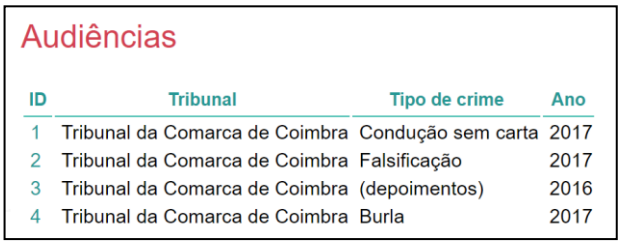

Figure 1: Main page.

Clicking on the **Edit** button will return a dropdown list with all XML tag names existent in all the corpus files. When a tag is chosen, another dropdown button appears listing three editing options: *Simple*, *Conditional* and *Align*.

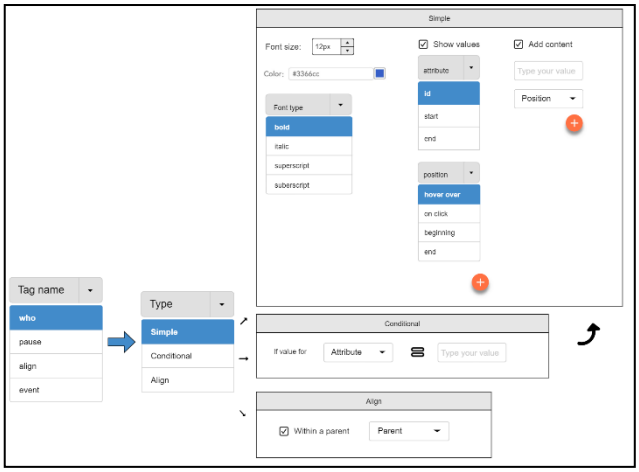

Figure 2: Visualization Scheme.

The Simple option is used to make changes to the visualization of all occurrences of the chosen tag in the entire corpus, regardless its attribute values. When the simple visualization is chosen, a pop-up window (figure 3) will appear, allowing the user to make three types of changes. The first option, *style*, allows the user to edit the font size, colour and type of the text within the XML tags. The *attribute* option offers the user the possibility to add to the visualization any attribute values included in the XML tag. The user can choose where the value will be placed (in the beginning, at the end, when hovering or when clicking on the text). The final option, *extra*, is similar to the previous one, with the difference that instead of using the attribute values, the users can add their own texts. Both last options have a plus icon that, when clicked, allows for the entry of new items.
The Conditional option, as the name says, performs changes to the visualization only when a condition is matched. If this option is chosen, a dropdown box with the attribute list and a text box will be displayed. If there is a

match for the condition inserted by the user, the same options for the simple visualization will appear.

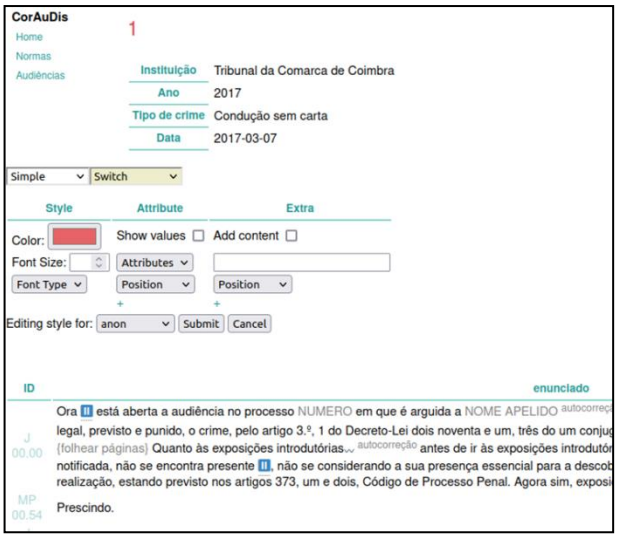

Figure 3: Simple Edit.

The Align option was developed to align utterances when there is an overlap of speech. This option will only be displayed if the corpus texts have tags named align (figure 4), with at least two attributes, one named *ref* and another *loc*. Ref should have a unique id for each group of tags that should be aligned together. The loc value should be a numerical value indicating in which position the tags should be aligned. This position is the index value for the character on which the alignment should occur, as illustrated on figure 5. The align visualization will display each text within the tag in a new line, aligning them accordingly.

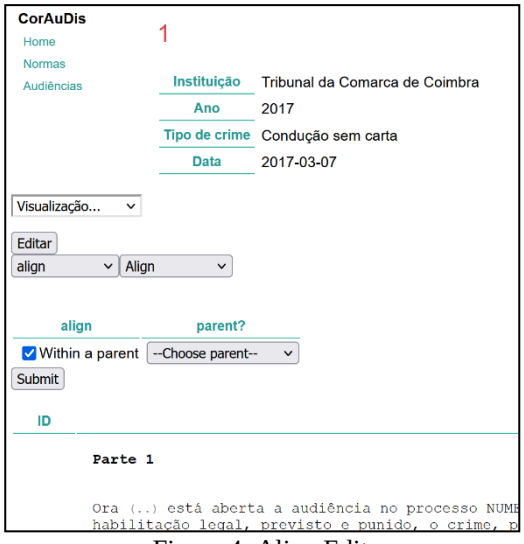

Figure 4: Align Edit.

If the align tag is nested in a hierarchically higher tag (i.e. if the align tag has a parent or any other ancestor tag), the option "display inside parent tag" will be shown. If the box is ticked, a dropdown menu with all the ancestors' tags for the align tag will be displayed. This option allows the user to display the token inside the align tag with other content, as illustrated in figure 6.

Although it was developed to meet the specific goal of this project, i.e. to graphically indicate when two utterances overlap, the align function can be applied to any cooccurring events in a corpus text.

Figure 5: Word alignment rationale.

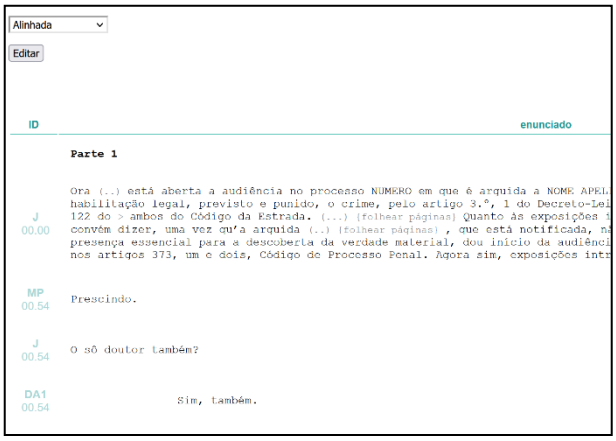

Figure 6: Word alignment display.

3.2.3 Vis Switch

As the visualization scheme becomes richer, or the corpus starts reaching a wider audience, different forms of visualizing the same text might be necessary. The Vis Switch functions allows the users to navigate between already existing templates.

In this current version of Atril, there are four different possibilities, designed specifically for the CorAuDis project. The default option is the display of the text and the *customized visualization* of the XML tags. This template was especially designed to render all the text mark-ups. Because many prospective users might be more interested in the text than on its annotation, a second template was created to display only the text content. To allow for the verification of the data and to present the corpus structure, a third template presents the raw XML file, with text and tags. Finally, a fourth template was created to apply alignment visualization to the overlapping utterances. In this visualization, the text does not break in lines according to the size of the web browser, but only when the line break is specified by the user and the alignment scheme. Although this visualization is useful to see the exact moment of overlapping words, long utterances without line breaks make the reading of the dialogue uncomfortable for the reader.

The four templates were designed to meet the needs of this current moment of the project. If other user profiles emerge, new stylesheets can be included in system, feeding the Vis Switch repertoire.

4. Further Steps

In this paper, we presented a brief description of Atril purpose and its version operation scheme. We are currently working on the improvement of the system, making the functions Vis Edit and Vis Switch more robust.

At the current development stage, the use of Atril is restricted to our team with a single administrator account. This means that we still do not have the possibility to preserve individual user customization. Another drawback is that once one change is made to the visualization the

previous version is lost. These issues will be addressed with the connection module of Atril to both CQPweb and TEITOK. This module will associate each created and saved template to their author, allowing for future editions. It will also include a versioning control history, which will allow the users to keep track of the changes made to the template and navigate through and select different versions. Once Atril's evaluation phase is complete and all necessary adjustments are made, we will launch it as an open-source module with a connection module to both CQPweb and TEITOK.

Although Atril was created to display transcripts of a Portuguese Spoken corpus, the system was created in a way that it could be used for different types of corpora of any language.

We believe Atril is a tool that is potentially useful for researchers working on projects of any language, and who share our need for easily customizable, free, and open-source resources for corpus creation and analysis.

5. References

Carapinha, C. and Plag, C. A. (2018) interação verbal em sala de audiências. In *Actas do XIII Congreso Internacional de Lingüística Xeral*, Vigo 2018, pp. 175-182.

Degaetano-Ortlieb, S., Kermes, H., Lapshinova-Koltunski, E., & Teich, E. (2013). SciTex-a diachronic corpus for analyzing the development of scientific registers. *New methods in historical corpus linguistics*, 3, pp. 93-104.

Janssen, M. (2016). TEITOK: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4037-4043.

Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, 17(3), pp. 380-409.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7-36.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), pp. 319-344.

Raymond, E. S. (2003). The art of Unix programming. Addison-Wesley Professional.

Schmidt, T., & Wörner, K. (2014). EXMARaLDA. In J. Durand, U. Gut & G. Kristoffersen (Eds.), The Oxford Handbook of Corpus Phonology. Oxford: Oxford University Press, pp. 402-419.