RIDE – A review journal for digital editions and resources

published by the IDE

TEITOK, a visual solution for XML/TEI encoding: editing, annotating and hosting linguistic corpora

TEITOK, Maarten Janssen (ed.), 2014. http://www.teitok.org/ (Last Accessed: 24.10.2022).

Reviewed by Pilar Arrabal Rodríguez (Universidad de Granada), pilararrabal@outlook.com.

Abstract

TEITOK is a web-based system designed to bring scholarly editing and computational linguistics together with the purpose of creating and hosting online language corpora. The system offers a visually attractive environment for digital editing based on the XML/TEI standard. TEITOK consists of automatic processes for carrying out many linguistic text processing tasks and functions. It boasts an intuitive interface via which researchers and corpus creators, who are not always computer literate, can manage corpus maintenance and error correction. The tokenization strategy in TEITOK permits the linking of the different levels of editing and annotation of each word in a single XML document for subsequent retrieval of information. This method provides a tool for editing, annotating, and exploiting corpora with a powerful search engine. TEITOK stands out for its high customisation and adaptability to a wide variety of corpora. In this article we analyse its utilities oriented mainly towards the creation of historical corpora, taking for this purpose the particular case of *Oralia diacrónica del español* (ODE).

Introduction

1 A traditional form of digital editing of manuscripts focuses exclusively on the representation of the textual content[1]. Editions of this type are based on the transcription of the text itself and leave information related to paratextual elements aside, such as those linked to the writing process or the presentation of the text (crossings out, omissions and additions by the scribe, changes of hand, handwriting styles, rubrics, decorative elements, etc.), but also those related to the state of conservation of the original manuscript (deterioration of the support, stains, tears, etc.). The creation of linguistic historical corpora has traditionally been carried out on the basis of transcriptions lacking the representation of features such as those mentioned above, and with a focus on linguistic information. There are two main reasons for this. Firstly, the transcription of original sources is a complex and time-consuming task that requires paleographical skills, and, in addition, there is the problem of how to represent the non-textual elements that are part of the reality of the document. Secondly, there is a lack of tools that support the visualisation and publication of digital editions (Janssen 2016, 4037), making it difficult for philologists to manage their own corpora. This has all been manifested in the preference for linguistic corpora based on plain text to overcome aspects such as, on the one hand, the difficulty of combining the textual and extra-textual elements of the original source in the edition and, on the other hand, the often insufficient training in programming and computer languages on the part of philologists, which requires the use of tools focused on the publication of digital edition projects.

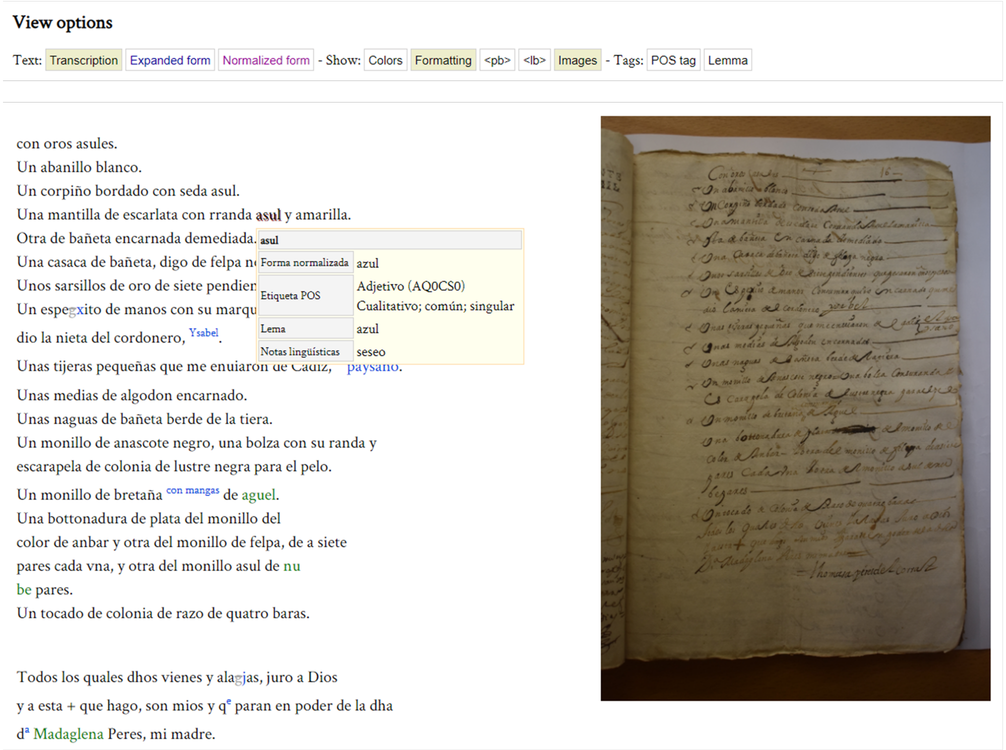
2 One solution to the first of the obstacles mentioned above, namely the combination of textual and extra-textual elements, has been the encoding of texts in the XML/TEI format. TEI provides a standard for encoding everything purely related to linguistic information, including tokenization, but also to structure, formatting, metadata, as well as extra-textual issues, such as modifications during the writing process or the state of preservation. These aspects, especially in the case of manuscripts, can affect the linguistic interpretation of the text and are as relevant as the text itself. In short, TEI proposes a standard for digital scholarly editions of texts. However, applying such a markup language requires an appropriate publishing solution.

3 The TEITOK system (Janssen 2016) solves the second problem, which is the existing gap in the digital publication of historical corpora and stands out as a particularly useful tool for outputting the visualisation and annotation of digital editions that conform

to the TEI standard. In addition, TEITOK provides computational resources for the linguistic processing of text and the creation of annotated corpora. This review examines the TEITOK system as an online tool for the editing, annotation, and exploitation of historical corpora, specifically from the perspective of the researcher[2].

Design

4 TEITOK is a web-based system for viewing digital editions, annotating linguistic corpora and retrieving data, all in a single environment. It was developed by Maarten Janssen and is an Open Source tool that can be installed on Linux and MacOS by downloading it directly from the repository on GitLab[3]. The tool is freely available, ready for production. TEITOK was initially developed at the *Centro de Linguística da Universidade de Lisboa* (CLUL) and later at CELGA-ILTEC. TEITOK is currently maintained at the *Institute of Formal and Applied Linguistics* (ÚFAL) of Charles University, Prague[4].



con oros asules.
Un abanillo blanco.
Un corpiño bordado con seda asul.
Una mantilla de escarlata con rranda asul y amarilla.
Otra de bañeta encarnada demediada.
Una casaca de bañeta, digo de felpa n
Unos sarsillos de oro de siete pendien
Un espegxito de manos con su marqu
dio la nieta del cordonero, Ysabel.
Unas tijeras pequeñas que me enuiaron de Cadiz, paysano.
Unas medias de algodon encarnado.
Unas naguas de bañeta berde de la tiera.
Un monillo de anascote negro, una bolza con su randa y
escarapela de colonia de lustre negra para el pelo.
Un monillo de bretaña con mangas de aguel.
Una bottonadura de plata del monillo del
color de anbar y otra del monillo de felpa, de a siete
pares cada vna, y otra del monillo asul de nu
be pares.
Un tocado de colonia de razo de quatro baras.

Todos los quales dhos vienes y alagjas, juro a Dios
y a esta + que hago, son mios y q^{e} paran en poder de la dha
d^{a} Madaglena Peres, mi madre.

asul	
Forma normalizada	azul
Etiqueta POS	Adjetivo (AQ0CS0) Cualitativo; común; singular
Lema	azul
Notas lingüísticas	seseo

Fig. 1: Visualisation of the paleographic transcription next to the facsimile of a document from the ODE corpus. In different colours, the textual characteristics of the manuscript (additions, deletions, sic). In the top bar it is possible to navigate between the different versions of the text and to display the morphosyntactic annotation and lemma. By hovering the mouse over the text, a pop-up window reveals detailed information for each word.

5 Making a corpus with TEITOK primarily consists of creating a root folder that will contain all the other files, templates, scripts, and resources needed for the new project. TEITOK is configured by modules that fulfil very different and varied functions. Some of these modules are common to all projects included in TEITOK and enable the basic operations. However, it is possible to add new modules by creating PHP files with different purposes. This modular configuration allows the system to be customised with different options according to the needs and characteristics of each corpus. TEITOK supports a wide range of linguistic corpora, including dictionaries and spoken corpora, learner and historical corpora. For corpora of spoken language, it allows the import of audio files and the voice to be associated with each fragment of the transcribed text (for example, the *Nurc Digital* [5]corpus). For historical corpora, TEITOK facilitates the multidisciplinary nature of these resources thanks to the different types of annotation and digital editing that can be offered for the same text. The *P.S. Post Scriptum* [6] project is a particularly good example of this (Vaamonde 2018). Following this model, the *Oralia*

diacrónica del español (*ODE*)[7] also offers a triple edition of the documents, including the image of the original facsimile next to the text. The paleographic edition presents a rigorous and faithful transcription of the manuscript, not only in terms of the textual content, but also in terms of its structure and paratextual elements, and this edition also offers the option of displaying the text with the expansion of the abbreviations. Finally, the normalised edition proposes, a transcription of the text adapted to current orthographic standards. TEITOK has an intuitive interface: it is easy to navigate between different versions of the text by clicking on them in the top menu of the document, as well as to show or hide other information related to the annotation or the layout of the text in the original manuscript. The interface also provides an attractive display and easy reading of the original manuscript features. Abbreviations, additions, deletions, scribal errata, conjectures, etc. are displayed on the web page in different colors and predefined formats with CSS code instructions (see Fig. 1).

6 TEITOK uses XML/TEI files as input to create an annotated linguistic corpus. However, it partially deviates from the TEI standard to make the set of files a searchable corpus[8]. The main difference lies in the tokenization of the texts. While TEI uses `<w>` and `<pc>` elements to differentiate words from punctuation marks, TEITOK uses a single element, `<tok>`, for all graphical forms of text. Tokenization in TEITOK is fully automatic: it adds the `<tok>` tags inline to the XML document and prepares the text for subsequent processing. The linguistic information of each word in TEITOK is added as attributes of this new `<tok>` element, which will include as many attributes as types of annotation the user wishes to offer for the same word. From this point on, the XML document is no longer valid TEI. Currently there is no ODD schema available for the TEITOK tool, but it can be found for specific projects using TEITOK such as Post Scriptum[9].

7 For example, for the form "ziud" found in an *ODE* manuscript (see Code 1), the attribute `@form` includes the paleographic transcription; `@fform` adds the expansion of the abbreviation; `@nform` collects the modernised form of the word, and `@pos` and `@lemma` include the linguistic information referring to the POS tag and the lemma of the word *ciudad* (see code example 1).

```
<tok id="w-74" form="ziud" fform="ziudad" nform="ciudad" pos="NCFS000"
lemma="ciudad">ziud</tok>
```

Code 1: Token encoding in TEITOK for the Spanish word "ciudad"

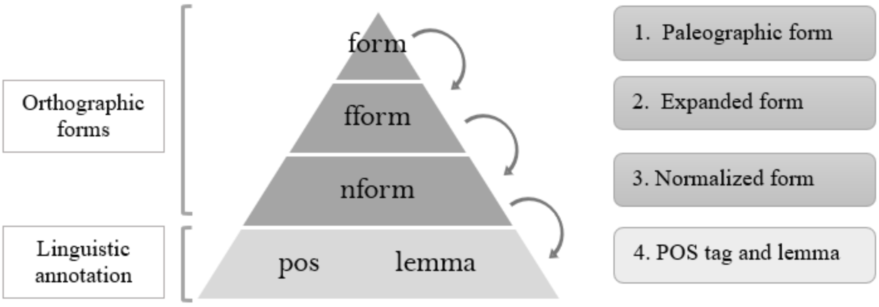
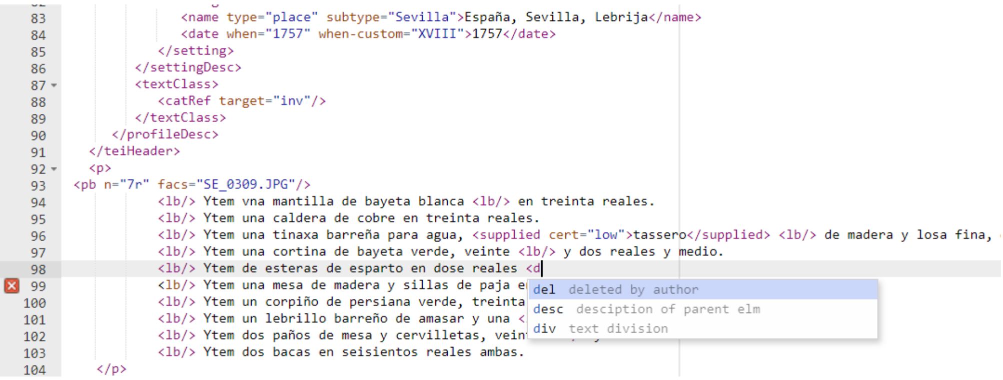


Fig. 2: Inheritance hierarchy for the different forms.

8 The design of TEITOK is based on a system of inheritance in which, starting from the paleographic form (`@form`), the rest of the different forms mentioned above will assume this form if no other is found. Likewise, for the linguistic annotation (POS and lemma), TEITOK takes the standard form (`@nform`). This inheritance hierarchy (see Fig. 2) allows the web interface to switch from one text edition to another without the need to replicate information. If this were not the case, it would be difficult to resolve annotation errors (Janssen 2016, 4038).

9 The inheritance system allows the combination of different versions of each text into a single file, and different types of linguistic information into a single token. The fact that all the data is contained in the same XML document facilitates the management and processing of files and solves the problem of the different versions in historical corpora that are not usually linked to one another (Calderón Campos 2019). In some of the historical corpora, offering two or more versions of a text translates into the overwhelming task of managing different corpora.

10 TEITOK provides an HTML form where the metadata can be filled in without accessing the XML file thanks to XPath instructions. The same system also permits the configuration of those elements of the TeiHeader that are relevant in each corpus, in order to display the metadata alongside each document in the interface.



```
                <name type="place" subtype="Sevilla">España, Sevilla, Lebrija</name>
                <date when="1757" when-custom="XVIII">1757</date>
              </setting>
            </settingDesc>
            <textClass>
              <catRef target="inv"/>
            </textClass>
          </profileDesc>
        </teiHeader>
        <p>
    <pb n="7r" facs="SE_0309.JPG"/>
              <lb/> Ytem vna mantilla de bayeta blanca <lb/> en treinta reales.
              <lb/> Ytem una caldera de cobre en treinta reales.
              <lb/> Ytem una tinaxa barreña para agua, <supplied cert="low">tassero</supplied> <lb/> de madera y losa fina,
              <lb/> Ytem una cortina de bayeta verde, veinte <lb/> y dos reales y medio.
              <lb/> Ytem de esteras de esparto en dose reales <d
              <lb/> Ytem una mesa de madera y sillas de paja e
              <lb/> Ytem un corpiño de persiana verde, treinta
              <lb/> Ytem un lebrillo barreño de amasar y una <
              <lb/> Ytem dos paños de mesa y cervilletas, vein
              <lb/> Ytem dos bacas en seisientos reales ambas.
         </p>
```

Fig. 3: The editing module directly in the TEITOK interface suggests possible completions of the tags.

11 TEITOK has recently incorporated an editing module, via which it is possible to transcribe and create new documents. To do so, the tool uses the browser-based Open Source ACE editor, which allows researchers to edit XML documents without using an external editor such as *Oxygen*. The editor in TEITOK adds tags semi-automatically while transcribing, suggesting the possible tags or checking whether or not the XML is well formed (see Fig. 3).

12 Thanks to the addition of the editing module, TEITOK can now handle the entire process of corpus elaboration, from transcription to linguistic annotation and search. As an alternative, XML files can also be imported into the platform from the users’ computer.

Automatic processing

13 TEITOK incorporates automatic services for the complete linguistic processing of texts: tokenization, standardisation and annotation of the corpus. It also includes support for the linguistic annotation of texts with the NeoTag tagger (Janssen 2012), written in Perl. The tagger uses the Viterbi algorithm to automatically assign a POS tag and a lemma to each token in the corpus on the basis of probability. The program has universal grammar statistics and applies n-grams based on word endings.

14 NeoTag works by using input data taken from parameters created from a previously annotated corpus. These parameters can either be imported or built up from the corpus itself when it reaches a considerable size, as NeoTag uses the corpus for training. The new parameters are constantly updated as the corpus grows in size, meaning the quality of the automatic annotation improves as the volume of annotated texts increases (Janssen 2016, 4039).

15 NeoTag assigns a POS tag based on a tagset and uses different rules based on linguistic analysis patterns to recognise the presence of derivational morphemes, enclitic particles, or numeric characters. It also enables the annotation of a word that is totally new in the corpus, as well as the possibility of applying probabilistic models to tag grammatical neologisms (Janssen 2012). For lemmatization, NeoTag applies textual segmentation rules to extract the lemma from the normalised form of the word. To modify the input form, the system generates a morphological parsing rule based on the final character of the word (Janssen 2012, 2).

16 There is no internal method in TEITOK to test the accuracy of the tagger. In historical corpora, the hit rate is lower than in modern text annotation and depends, above all, on the training corpus used[10]. In the particular case of ODE it has not yet been possible to estimate the accuracy of automatic annotation with statistical precision. In the initial stages, this task has required extensive manual review. Errors have decreased after the annotation of approximately 100,000 words, when NeoTag has been trained on the corpus data itself.

Search options

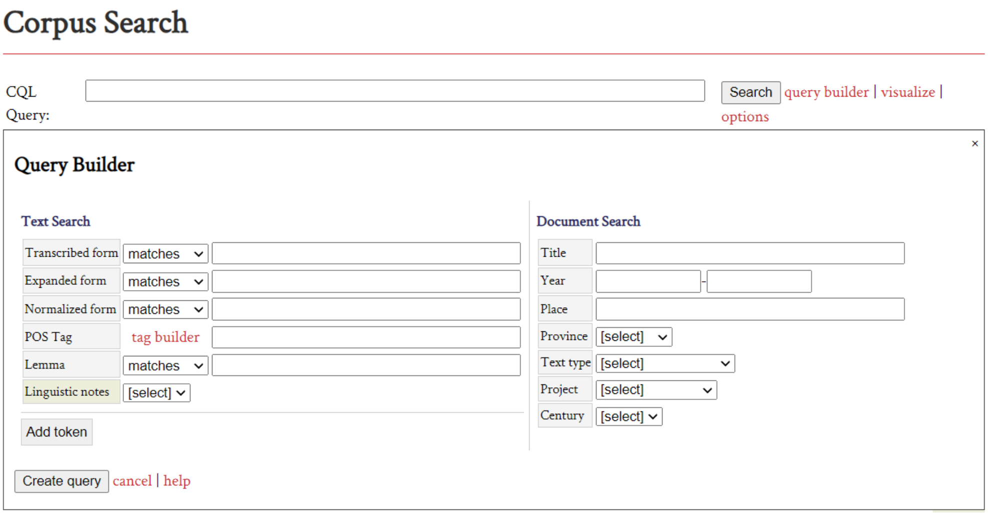
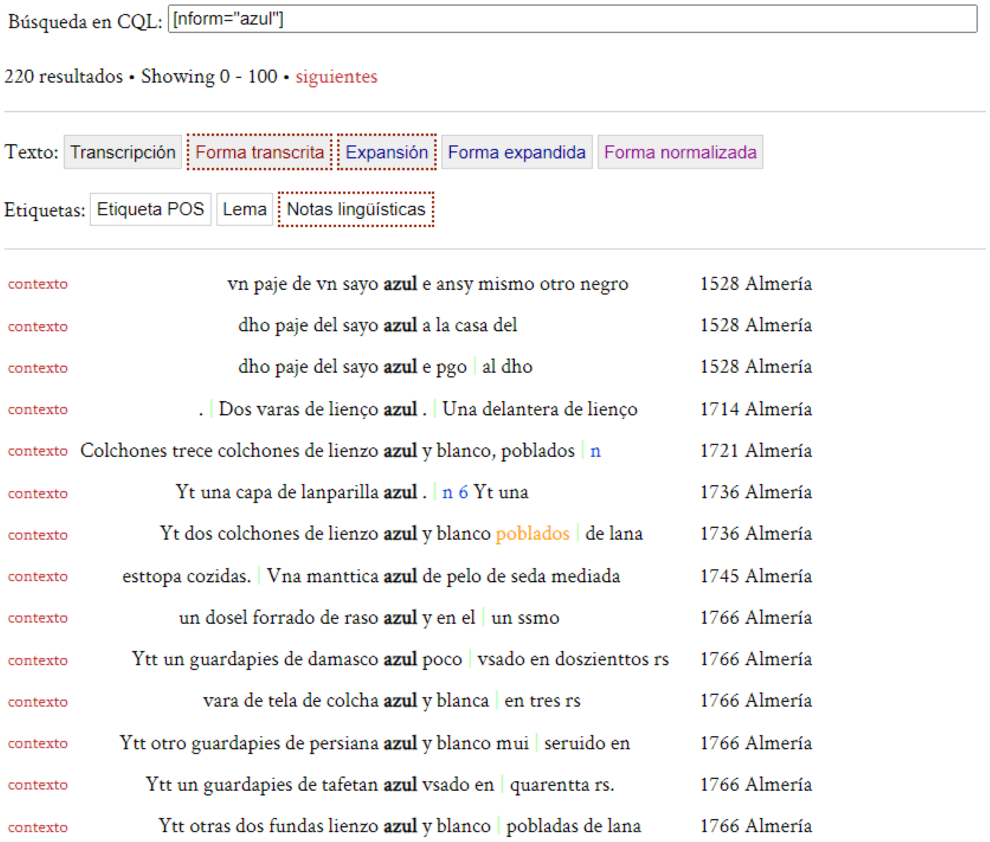


Fig. 4: TEITOK query system for ODE.

17 TEITOK integrates a CQL-based search engine for corpus data retrieval. The Corpus Query Language (CQL) enables the search of complex grammatical or lexical patterns. It has been designed for corpora with a modular structure and the token is the minimum unit for retrieving information. However, CQL searches require the use of a specific syntax, which not all users will be familiar with. In order to facilitate searches, TEITOK incorporates a query builder or Corpus Query Processor (CQP): the advanced

search graphical interface helps the user to transform instructions into CQL syntax, accessing each of the corpus editing and annotation levels (see Fig. 4).



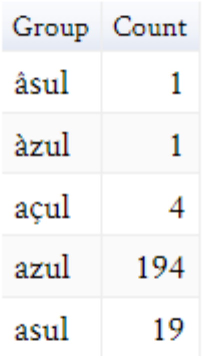
Búsqueda en CQL: [nform="azul"]

220 resultados • Showing 0 - 100 • siguientes

Texto: Transcripción Forma transcrita Expansión Forma expandida Forma normalizada
Etiquetas: Etiqueta POS Lema Notas lingüísticas

contexto	vn paje de vn sayo **azul** e ansy mismo otro negro	1528 Almería
contexto	dho paje del sayo **azul** a la casa del	1528 Almería
contexto	dho paje del sayo **azul** e pgo al dho	1528 Almería
contexto	. Dos varas de lienço **azul** . Una delantera de lienço	1714 Almería
contexto	Colchones trece colchones de lienzo **azul** y blanco, poblados n	1721 Almería
contexto	Yt una capa de lanparilla **azul** . n 6 Yt una	1736 Almería
contexto	Yt dos colchones de lienzo **azul** y blanco poblados de lana	1736 Almería
contexto	esttopa cozidas. Vna manttica **azul** de pelo de seda mediada	1745 Almería
contexto	un dosel forrado de raso **azul** y en el un ssmo	1766 Almería
contexto	Ytt un guardapies de damasco **azul** poco vsado en doszienttos rs	1766 Almería
contexto	vara de tela de colcha **azul** y blanca en tres rs	1766 Almería
contexto	Ytt otro guardapies de persiana **azul** y blanco mui seruido en	1766 Almería
contexto	Ytt un guardapies de tafetan **azul** vsado en quarentta rs.	1766 Almería
contexto	Ytt otras dos fundas lienzo **azul** y blanco pobladas de lana	1766 Almería

Fig. 5: Result from the search by normalised form "azul". In ODE the result is set to show the year and place of origin of the document next to each case.

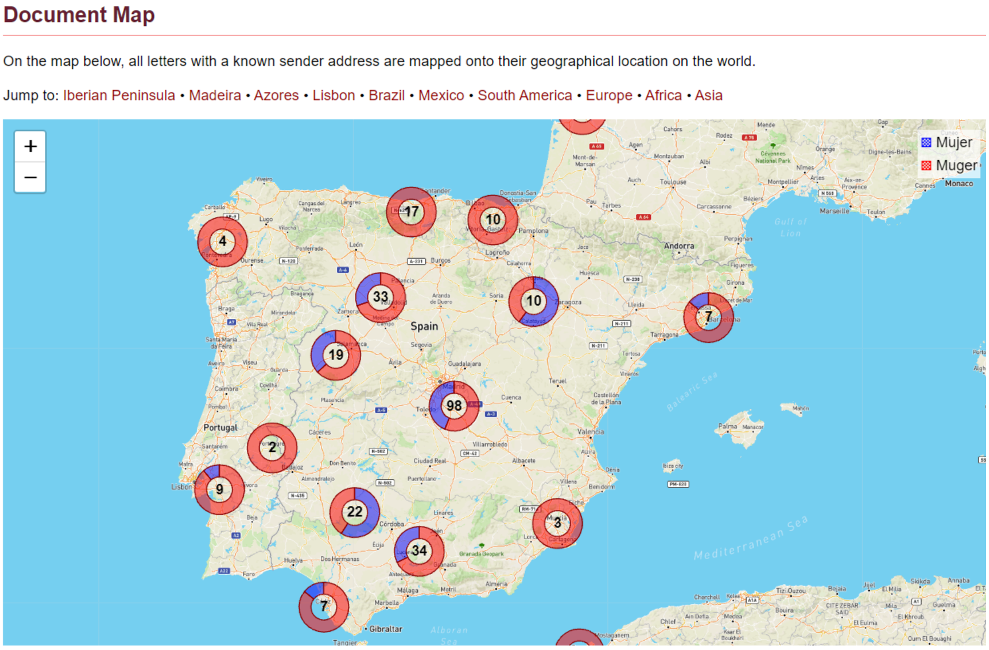


Group	Count
âsul	1
àzul	1
açul	4
azul	194
asul	19

Fig. 6: Original forms of the word "azul" in ODE

18 In addition, the advanced search allows the data to be filtered according to extratextual variables, i. e. information that is part of the metadata of the document, such as the date or the geographical location of the manuscript. In this way, the system makes it possible to combine linguistic information with the metadata of the texts in the corpus, which translates into the possibility of carrying out increasingly sophisticated queries. The result is a concordance in KWIC format, and it is possible to configure it according to the particular interests of each corpus (see Fig. 5). Not only does TEITOK provide

results in the form of concordances, but it also offers statistical frequencies of the obtained data, which can be displayed in tables or bar charts. Additionally, the results can be sorted according to different linguistic (lemma, word class, original form, etc.) and extra-linguistic (year, place, text type, etc.) criteria (see Fig. 6).



Document Map

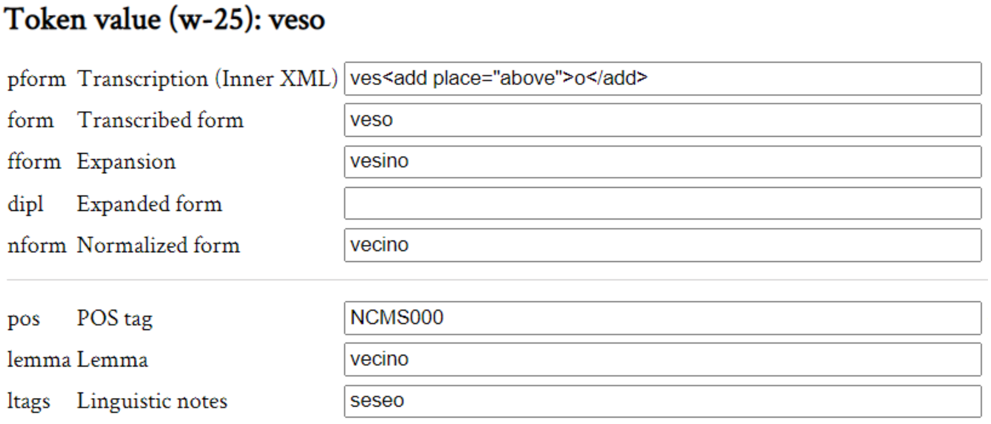
On the map below, all letters with a known sender address are mapped onto their geographical location on the world.

Jump to: Iberian Peninsula • Madeira • Azores • Lisbon • Brazil • Mexico • South America • Europe • Africa • Asia

Fig. 7: Result of a search projected on the map. Example of the corpus P.S. Post Scriptum.

19 TEITOK also allows the visualisation of comparative searches on a map. Each document in the corpus can be automatically assigned the geographical coordinates of its place of origin. This data is inserted in the header of the XML file as metadata, which allows the mapping of search results. Fig. 7 shows the result of a search in the P.S. Post Scriptum corpus for the graphic forms *mujer* and *muger* in the Iberian Peninsula.

Corpus management

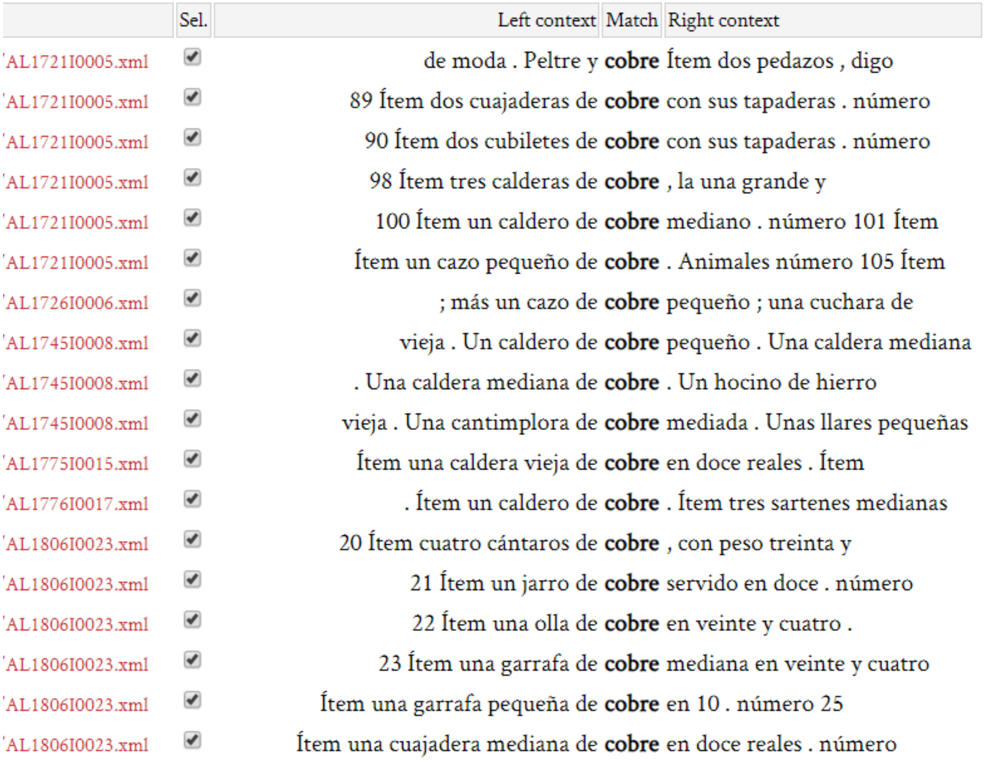


Token value (w-25): veso

pform	Transcription (Inner XML)	ves<add place="above">o</add>
form	Transcribed form	veso
fform	Expansion	vesino
dipl	Expanded form	
nform	Normalized form	vecino
pos	POS tag	NCMS000
lemma	Lemma	vecino
ltags	Linguistic notes	seseo

Fig. 8: HTML form for token editing.

20 TEITOK acts as a framework in which all editing and linguistic processing of texts can be carried out. Administrators have access to various functionalities that allow easy online maintenance of the corpus. The same interface for viewing and querying the corpus equally serves to correct annotation errors therein (Janssen et al. 2017, 3). Administrative users are able to edit words by clicking on them. This generates an HTML template in which to modify any of the token annotation levels (see Fig. 8).



	Sel.	Left context	Match	Right context
'AL1721I0005.xml	☑	de moda . Peltre y	**cobre**	Ítem dos pedazos , digo
'AL1721I0005.xml	☑	89 Ítem dos cuajaderas de	**cobre**	con sus tapaderas . número
'AL1721I0005.xml	☑	90 Ítem dos cubiletes de	**cobre**	con sus tapaderas . número
'AL1721I0005.xml	☑	98 Ítem tres calderas de	**cobre**	, la una grande y
'AL1721I0005.xml	☑	100 Ítem un caldero de	**cobre**	mediano . número 101 Ítem
'AL1721I0005.xml	☑	Ítem un cazo pequeño de	**cobre**	. Animales número 105 Ítem
'AL1726I0006.xml	☑	; más un cazo de	**cobre**	pequeño ; una cuchara de
'AL1745I0008.xml	☑	vieja . Un caldero de	**cobre**	pequeño . Una caldera mediana
'AL1745I0008.xml	☑	. Una caldera mediana de	**cobre**	. Un hocino de hierro
'AL1745I0008.xml	☑	vieja . Una cantimplora de	**cobre**	mediada . Unas llares pequeñas
'AL1775I0015.xml	☑	Ítem una caldera vieja de	**cobre**	en doce reales . Ítem
'AL1776I0017.xml	☑	. Ítem un caldero de	**cobre**	. Ítem tres sartenes medianas
'AL1806I0023.xml	☑	20 Ítem cuatro cántaros de	**cobre**	, con peso treinta y
'AL1806I0023.xml	☑	21 Ítem un jarro de	**cobre**	servido en doce . número
'AL1806I0023.xml	☑	22 Ítem una olla de	**cobre**	en veinte y cuatro .
'AL1806I0023.xml	☑	23 Ítem una garrafa de	**cobre**	mediana en veinte y cuatro
'AL1806I0023.xml	☑	Ítem una garrafa pequeña de	**cobre**	en 10 . número 25
'AL1806I0023.xml	☑	Ítem una cuajadera mediana de	**cobre**	en doce reales . número

Fig. 9: Multi-Editing tokens using CQP search.

21 This system allows individual correction of tokens in the corpus. However, TEITOK offers the possibility of making larger corrections in the annotation of texts through the CQP search. The search engine can also be used to modify multiple errors that affect a large number of tokens at the same time. The concordance permits the selection of the cases (see Fig. 9), which makes it possible to apply the same change any number of times without the need to do so document by document. This system speeds up the edition of incorrect POS tags and lemmas assigned automatically.

22 Error correction is, therefore, a task for researchers, who can manage their own corpora autonomously without external technical assistance. This means a quick and intuitive way of correcting errors detected by automatic annotation without the need to access the XML code and modify the information from the document source file.

23 Other functions related to corpus management are explained on the tool's website[11]. TEITOK has a site with a detailed description of what it is and how it works, a list of projects using the tool, publications about its development, etc. The help section, which includes a reference guide with explanations about the installation or how to customise the appearance of the website, is particularly useful. It also includes a General FAQ section and other instructions, although not all of them are equally accurate.

24 Direct contact with the developer is possible through the contact details on his personal web page[12], but the site lacks a helpdesk specifically dedicated to solving problems for TEITOK users. For community support and recent developments, there is also a Google group mailing list and a Facebook page.

Conclusions

25 In this article we review the advantages offered by TEITOK for the creation of linguistic corpora and, more specifically, of historical corpora whose primary sources are manuscripts. TEITOK adapts to the requirements and objectives of different types of corpora, as it incorporates numerous modules that can be customised, both in terms of functions and of the web interface style of each project. This means that TEITOK allows each project or corpus to adapt the design to its own look and requirements, setting up the navigation menu or modifying the text visualization with CSS.

26 TEITOK is based on XML/TEI encoded files in order to offer corpus visualisation and exploitation on a single platform. This XML/TEI markup language enables different versions of the text to be combined in the same file. The linguistic annotation of each word is added as attributes of the <tok> element. Text indexing allows the subsequent retrieval of data through advanced searches using CQP directly on the website.

27 The main difficulties in using TEITOK for researchers with a humanistic background lie in the initial phases of installation and customisation of the functions adapted to the objectives and interface of the project. Once this phase is complete, the creation of corpora on the website only requires knowledge of XML/TEI for text transcription and editing, but not of programming to carry out the rest of the tasks. This means that researchers can manage the corpus and its maintenance without requiring any other technical or computer skills. Thus, TEITOK becomes a tool that allows the combination of digital editing with the advanced features of a linguistic corpus, giving the researcher the necessary autonomy to manage the complete processing of the corpus online and to correct errors as soon as they are detected.

Notes

1. Vaamonde (2018), Kytö (2011), Claridge (2008) have insisted on this idea.

2. My academic background is mainly linguistic. In recent years I have worked on several projects within the framework of the Digital Humanities and digital publishing in which I have been part of the process of creating the *Oralia diacrónica del español* (ODE) corpus, from the first stages of transcription in XML/TEI to its linguistic annotation, always in the TEITOK web environment.

3. https://gitlab.com/maartenes/TEITOK. Accessed: September 27, 2021.

4. https://web.archive.org/web/20221110115135/https://ufal.mff.cuni.cz/.

5. https://web.archive.org/web/20221110115205/https://fale.ufal.br/projeto/nurcdigital/.

6. https://web.archive.org/web/20220308213603/http://teitok.clul.ul.pt/postscriptum/es/index.php.

7. https://web.archive.org/web/20221110115325/http://corpora.ugr.es/ode/.

8. TEITOK includes a script that facilitates the conversion of XML files to the valid TEI standard.

9. The schema can be found here: http://teitok.clul.ul.pt/postscriptum/files/ps_doc.html. Accessed: October 17, 2021. Other differences between TEI/XML and TEITOK are explained here: http://www.teitok.org/index.php?action=help&id=teixml. Accessed: October 17, 2021.

10. The accuracy of NeoTag has been tested on corpora of current Spanish with a 97% accuracy rate for morphosyntactic annotation and 95% for lemmatization (Janssen 2012).

11. https://www.teitok.org/. Accessed: September 27, 2021.

12. https://web.archive.org/web/20221110120113/http://maarten.janssenweb.net/.

References

Calderón Campos, M. 2019. "La edición de corpus históricos en la plataforma TEITOK. El caso de Oralia diacrónica del español". Chimera, 6, 21-36.

Claridge, C. 2008. "Historical corpora". In Corpus linguistics: an international handbook (Vol. 1) edited by A. Lüdeling and M. Kytö, 242-259. Berlin / New York: Walter de Gruyter.

Janssen, M., Ausensi, J. y Fontana, J. M. 2017. "Improving POS tagging in Old Spanish using TEITOK". In Proceedings of the NoDaLiDa 2017 workshop on Processing Historical Language, Gotemburg, 2-6.

Janssen, Maarten. 2016. "TEITOK: Text-Faithful Annotated Corpora". In Proceedings of the Language Resources and Evaluation Conference (LREC 2016). Portoroz: ELRA, 4037-4043.

Janssen, Maarten. 2012. "NeoTag: a POS tagger for grammatical neologism detection". In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). Estambul: ELRA, 2118-2124.

Kytö, M. 2011. "Corpora and historical linguistics". Revista Brasileira de Linguística Aplicada, Belo Horizonte, 11, vol. 2, 417-457.

Vaamonde, Gael. 2018. “La multidisciplinariedad en la creación de corpus históricos: El caso de Post Scriptum”. In Humanidades digitales: sociedades, polı̃ticas, saberes. Artnodes, 22, 118-127.

Factsheet

Resource reviewed	
Title	TEITOK
Editors	Maarten Janssen
URI	http://www.teitok.org/
Publication Date	2014
Date of last access	24.10.2022

Reviewer	
Name	Arrabal Rodríguez, Pilar
Affiliation	Universidad de Granada
Place	Granada, Spain
Email	pilararrabal (at) outlook.com

General information		
Software type	What type of software is it? (cf. Catalogue 0.1.1)	Software tool
Identification of the environment	On which platform runs the tool? (cf. Catalogue 1.4)	Web browser
Purpose	For what purpose was the tool developed? (cf. Catalogue 1.5)	developed for a specific project or materials
Funding	Which is the financial model of the tool? (cf. Catalogue 1.6)	Free/open
Maturity	What is the development stage of the tool? (cf. Catalogue 1.5)	Release
Methods and implementation		
Programming Language	Which programming languages and technologies are used? (cf. Catalogue 2.3)	C family, PHP, XSLT, Other: Javascript, Perl
Reuse	Does the tool reuse portions of other existing software? (cf. Catalogue 2.3)	no
Input format	Which input formats are supported? (cf. Catalogue 2.4)	.xml, .xml/tei

Output format	Which output formats are supported? (cf. Catalogue 2.4)	.xml, .xml/tei, .txt
Encoding	Which character encoding formats are supported? (cf. Catalogue 2.4)	utf-8
Encoding preprocessing	Is a pre-processing conversion included?	yes
Dependencies	Does the documentation list dependencies on other software, libraries or hardware? (cf. Catalogue 3.2)	yes
Dependencies installation	If yes, is the software handling the installation of dependencies during the general installation process (you don't have to install them manually before the installation)?	no
Documentation and support		
Documentation	Is documentation and/or a manual available? (tool website, wiki, blog, documentation, or tutorial) (cf. Catalogue 3.4)	yes
Documentation format	Which format has the documentation? (cf. Catalogue 3.3)	.html
Documentation parts	Which of the following sections does the documentation contain? (cf. Catalogue 3.3)	‘Getting Started’ section (installation and configuration), Examples, FAQ, Support
Documentation language	In what languages is the documentation available? (cf. Catalogue 3.3)	English
Support	Is there a method to get active support from the developer(s) or from the community? (cf. Catalogue 3.4)	yes
From of support	Which form of support is offered? (cf. Catalogue 3.4)	Mailing-list, Other: Facebook
Issue tracker	Is it possible to post bugs or issue using issue tracker mechanisms? (cf. Catalogue 3.4)	yes
Usability and sustainability		
Build and install	Grade how straightforward it is to build or install the tool on a supported platform: (cf. Catalogue 3.6)	tricky

Tests	Is there a test suite, covering the core functionality in order to check that the tool has been correctly built or installed? (cf. Catalogue 3.7)	yes
Portability and interoperability	On which platforms can the tool/software be deployed? (cf. Catalogue 3.8)	Linux/BSD/Unix, Mac OS X
Devices	On which devices can the tool/software be deployed? (cf. Catalogue 3.8)	Desktop, Laptop
Browsers	If the tool is web-based: On which browsers can the tool/software be deployed? (cf. Catalogue 3.8)	Mozilla Firefox, Google Chrome, Safari
Plugins	If the tool is web-based: Does the tool rely on browser plugins? (cf. Catalogue 3.8)	no
API	Is there an API for the tool? (cf. Catalogue 3.8)	no
Code	Is the source code open? (cf. Catalogue 3.9)	yes
License	Under what license is the tool released? (cf. Catalogue 3.9)	GNU/GPL
Credits	Does the software make adequate acknowledgement and credit to the project contributors? (cf. Catalogue 3.9)	yes
Registered	Is the tool/software registered in a software repository? (cf. Catalogue 3.9)	yes
Possible contribution	If yes, can you contribute to the software development via the repository/development platform?	no
Analysability, extensibility, reusability of the code		
Analysability	Can the code be analyzed easily (is it structured, commented, following standards)? (cf. Catalogue 3.10)	yes
Extensibility	Can the code be extended easily (because there are contribution mechanisms, attribution for changes and backward compatibility)? (cf. Catalogue 3.10)	yes

Reusability	Can the code be reused easily in other contexts (because there are appropriate interfaces and/or a modular architecture)? (cf. Catalogue 3.10)	yes
Security and privacy	Does the software provide sufficient information about the treatment of the data entered by the users? (cf. Catalogue 3.11)	yes
Supportability and maintenance	Is there information available whether the tool will be supported currently and in the future? (cf. Catalogue 3.12)	no
Citability	Does the tool supply citation guidelines (e.g. using the Citation File Format)? (cf. Catalogue 3.13)	yes
User interaction, GUI and visualization		
User profile	What kind of users are expected? (cf. Catalogue 4.1)	Humanities researcher, Digital humanist
User interaction	What kind of user interactions are expected? (cf. Catalogue 4.1)	Reading, Text editing, Text analysis, Image editing, Searching, Visualization
User Interface	What kind of interface does the tool provide? (cf. Catalogue 4.2 and 0.1.1)	Graphical User Interface (GUI)
Visualization	Does the tool provide a particular visualizations (in terms of analysis) of the input and/or the output data? (cf. Catalogue 4.3)	yes
User empowerment	Is the user allowed to customize the functioning of the tool and the output configuration? (cf. Catalogue 4.4)	yes
Accessibility	Does the tool provide particular features for improving accessibility, allowing „people with the widest range of characteristics and capabilities" to use it? (cf. Catalogue 4.5)	no
Personnel		
Editors	Janssen, Maarten	
Programmers	Janssen, Maarten	
Advisors	Janssen, Maarten	
Designers	Janssen, Maarten	

Contributors	Janssen, Maarten