

A Christmas carol

ELTeC-por report

Diana Santos

d.s.m.santos@ilos.uio.no

Distant  Reading

7 December 2020

Organization

Four responsables

- Raquel Amaro
- Paulo da Silva Pereira
- Isabel Araújo Branco
- Diana Santos

Several workers for correcting the OCR

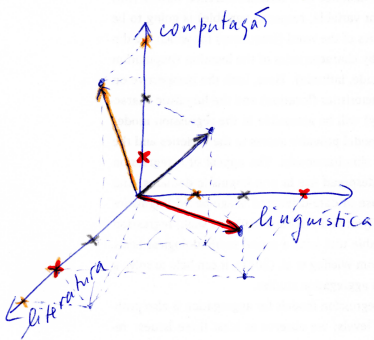
- Adeliana Silva
- Inês Lucas
- Aline Baldé
- Vanda Morgado
- Madalena Rato

A status page on the Web, giving a snapshot of the situation.



Columbano Bordalo Pinheiro (1857-1929):
O Grupo do Leão, 1885

Current state of the collection



- 100 novels
- E5C: 94.62 mainly due to only 18 long novels :-(and difference 37 vs 13 for T2 vs T1
- lack of thorough epigraph treatment
- no page numbers
- different ortographies
- not foolproof reprint count

We want to get 20 long novels... so two of the current ones will be changed.

Text selection: Opportunistic, opportunistic, opportunistic

- As many as possible woman writers: lists produced by Paulo, but most works could not be found – or were not novels, or were too short. So we have only 17 novels written by 15 women.
- The National Library of Portugal sent us a list/catalogue of all works published in the COST period. We then selected 10 and then 10 works for them to digitize, prioritizing again women, and selecting works with the word *romance* in the title.
- In addition to Gutenberg, we browsed sites claiming to have Portuguese books, and then we searched archive.org intensely, selecting common Portuguese names (COST period), and scrutinizing what seemed to be a novel.
- Approaching the end, we were more concerned about the balance, and had to reshuffle the works – 17 in the ELTeC-por-ext collection.
- As to the repeated authors, almost all are canonical (otherwise we would not have enough canonical works in the collection)
- We chose *novelas* and *romances* (novellas and novels), and used novels published initially in newspapers.

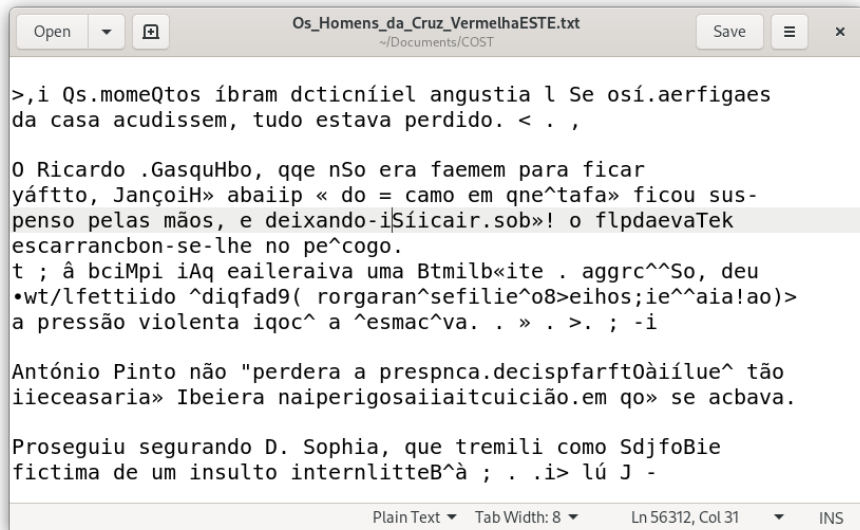
- We started from simple text with footnote, verse and foreign text encoding in pseudo-XML
- We then ran a simple script that TEI-encoded (light) for chapters
- Headers were manually created, and the above text was included in the final file
- TEI-conformance was checked with `xmllint` and, in the most complex cases, with `Oxygen`, until it was OK. The very little one needed to learn from TEI was learned in practice.

At the same time we included them in github, we added them to our searchable online literary corpus, *Literateca*, fully parsed and semantically annotated.

Problems? All sorts of problems

- I would think there would be much more digitized works in Portuguese: it was a shock to see how few
- Most of the digitization efforts had used terrible OCR software, which meant that a book could require 20 or more hours to clean. Not only that, suddenly pages were missing or repeated.
- After having included ca. 20 canonical works from the Vercial project, we realized that we had never got explicit authorization from them, and decided to replace them all by Gutenberg editions
- Very little guidance from literary experts on how to choose non-canonical novels
- A lot of consequences on what had been decided on the first meetings only got understood in later meetings (for example that it was more important to have length balance than 100 novels)
- Not easy to measure canonicity even if operationalized (Worldcat not reliable)

Some pages...



The screenshot shows a text editor window with the title bar "Os_Homens_da_Cruz_VermelhaESTE.txt" and a path "~/Documents/COST". The window contains three paragraphs of text that appear to be heavily garbled or encoded. The first paragraph starts with ">,i Qs.momeQtos íbram dcticníiel angustia l Se osí.aerfigaes da casa acudissem, tudo estava perdido. < . ,". The second paragraph starts with "O Ricardo .GasquHbo, qqe nSo era faemem para ficar yáftto, JançoiH» abaiip « do = camo em qne^tafa» ficou suspenso pelas mãos, e deixando-iSíicair.sob»! o flpdaevaTek escarrancbon-se-lhe no pe^cogo." The third paragraph starts with "António Pinto não "perdera a prespnca.decispfarft0àiílu^ tão iieceasaria» Ibeiera naiperigosaiiaitcuicião.em qo» se acabava." The fourth paragraph starts with "Proseguiu segurando D. Sophia, que tremili como SdjfoBie fictima de um insulto internlitteB^à ; . .i> lú J -". The status bar at the bottom indicates "Plain Text", "Tab Width: 8", "Ln 56312, Col 31", and "INS".

```
>,i Qs.momeQtos íbram dcticníiel angustia l Se osí.aerfigaes
da casa acudissem, tudo estava perdido. < . ,

O Ricardo .GasquHbo, qqe nSo era faemem para ficar
yáftto, JançoiH» abaiip « do = camo em qne^tafa» ficou sus-
penso pelas mãos, e deixando-iSíicair.sob»! o flpdaevaTek
escarrancbon-se-lhe no pe^cogo.
t ; â bciMpi iAq eaileraiva uma Btmilb«ite . aggrc^^So, deu
•wt/lfettiido ^diqfad9( rorgaran^sefilie^o8>eihos;ie^^aia!ao)>
a pressão violenta iqoc^ a ^esmac^va. . » . >. ; -i

António Pinto não "perdera a prespnca.decispfarft0àiílu^ tão
iieceasaria» Ibeiera naiperigosaiiaitcuicião.em qo» se acabava.

Proseguiu segurando D. Sophia, que tremili como SdjfoBie
fictima de um insulto internlitteB^à ; . .i> lú J -
```

Christmas in ELTeC-por: some initial distant reading



Machado de Castro (1731-1822): Presépio, Basílica da Estrela (1781-1786), detail

- Only 17 works mention at all *Natal* (Christmas), and only two novels describe events in Christmas (eve/day)
- *Páscoa* (Easter) is only mentioned in 15 works, and no scene happens then.
- *Carnaval* is mentioned in 10 novels, two of which have a scene in this period.

Fun fact: there are several regions called Natal (a city in Brazil, a region in South Africa, now KwaZulu-Natal) because they were discovered by Portuguese on Christmas day.

Wrap-up discussion

Do it differently?

- ideally, decide on the constitution (which books) before starting work
- not include modernized texts
- do not make canonized novels a 30% requirement: it should depend on the number of canonized authors
- no need to have three novels for 9 authors: leave that to the extension

Further development of ELTeC

- Remainder of the Action: Proceed to level 2, do some distant reading
- Include Brazilian authors, include books in Portuguese published wherever.
- Not only 100 novels, include as many as possible, and let the user choose which ones s/he wants to take for particular inquiries