BILLIG

Bilateral Lusophone Literature Initiative using GIS and Linguistics

Report: Second Internship in **Oslo (Online Activity)** - text analysis, computational linguistics, natural language processing and statistics.

Participant researchers:
Danielle Sanches, NOVA-FCSH

Date: 30 May 2020

In these fifteen days of training (May 7-21), I developed my skills in text analysis and understanding of computational linguistics and natural language processing. The environment for this learning was very promising because, due to the coronavirus pandemic, the training had to be carried out remotely and, for this reason, we managed to get 9 other participants to join in this training. The profiles of the 9 researchers who participated in this training were quite diverse: historians, linguists and experts in literature. This diversity of people contributed for the learning of this methodology to be carried out with greater benefit.

In this sense, we all learned to work with a set of linguistic corpus and how to make syntactic and morphological annotations and how to work with large volume of text. The points addressed in the training were as follows:

1) Presentation of the Linguateca; and what is a text corpus;
2) AC/DC from a user's point of view; regular expressions;
3) AC/DC and underlying syntactic and semantic analysis;
4) AC/DC on the side of here: the WORDS, the cut-and-sew;
5) AC/DC on the other side: the IMS-CWB and the entire processing chain;
6) Posts and NER-WORDS, other services associated with AC/DC;
7) Presentation of R, installation and basic use;
8) Visualization with R;
9) Some exploratory methods in R;
10) Recap and future.

In an online format, this training had to be adjusted as follows: 1 hour of theoretical class on the topic, and then a battery of exercises on the subject is given, to be carried out throughout the day, and at the end of the day, there is another hour of online class to solve the exercises and questions about unresolved issues. This last stage proved to be essential in the exchange of experiences between researchers, since the questions about the exercises were, at times, very different.

This provided some interesting debates among the participants. During the 15 days we had daily training, apart from the weekends, which resulted in 10 working days of training and 20 classes in total (considering the 2 times a day we talked with the trainer).

The person responsible for this training, Diana Santos, one of the coordinators of the Billig project, created a web environment where the exercises and presentations could be found, alongside the recordings of the classes. This facilitated access for all participants (theoretical part of the course and exercises).

All the exercises we did were with the corpus of Linguateca, focusing on the specific corpus of Literateca, where there are literary texts in Portuguese that are part of the BILLIG project. From this training, we can continue to make annotations in the database of the Atlas Project of Literary Landscapes and generate spatial analyzes from the corpus of Literateca.

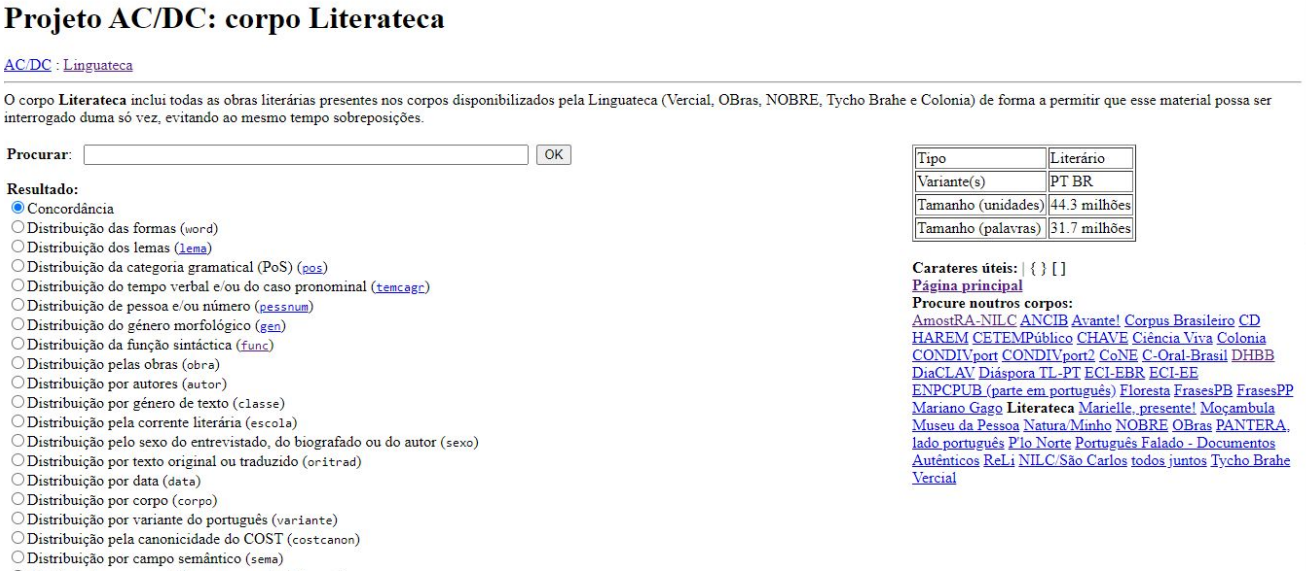

Figure 1 - Literateca page for searching the corpus

As can be seen in Figure 1, the first lessons sought to familiarize students in the Literateca web environment, showing how to search the annotated corpus. From the researches carried out we were able to extract some results that allow the textual analysis in a large volume of text.

Other studies were conducted in different corpus. The Brazilian Biographical Historical Dictionary (DHBB) was another target for searches and corpus analyzes. Figure 2 shows the results of some exercises performed in this linguistic corpus.

```
Qual a distribuição de texto por dicionário?

dhbb    312384
dhbpr   79885
dprrj   62960
__UNDEF__       12

Quantos verbetes tem cada dicionário?

__UNDEF__       12368

Quantos homens e quantas mulheres estão verbetados (ou seja, são entradas no dicionário)? e por cada dicionário?

Resposta 2
m       567
N/A     338
f       296
__UNDEF__       47

Resposta 1

m       2422341
N/A     558433
__UNDEF__       96986
f       64751

Quantos jornalistas estão mencionados no DHBB? Obtenha-os por ordem alfabética

jornalista      1802 ocorrências.
```

Figure 2 - Example of an exercise performed from DHBB

In the last training sessions, we had contact with text analysis in R software. Some information and ideas were given about research and analysis in this software, but, above all, the awakening of the possibilities of research in literary texts was covered in this training.

After all the knowledge acquired with the training carried out, I think that the applications we indicate to perform in this project, based on knowledge transfer, will be successful.