# TNA ACTIVITY REPORT

## Compiling a literary corpus with minimal resources

Author: Andressa Rodrigues Gomide

Current position: Researcher

Affiliation: Universidade de Coimbra

Host institution: Universität Trier

Mentor(s): Prof. Dr. Christof Schöch

Period of stay: 26/09/2022 – 02/12/2022

## Introduction

I have applied for the CLS INFRA Fellowship Programme to have the opportunity to learn from and interact with researchers at the Trier Centre for Digital Humanities (TCDH) in the context of the compilation of a corpus of literary production in Portuguese.

The current project I work with at the Centro de Estudos de Linguística Geral e Aplicada da Universidade de Coimbra (CELGA-ILTEC), at Coimbra University (Portugal) is centred on the development of linguistic and computational resources for Portuguese as a pluricentric language. The main expected outcomes are the creation of a pluricentric dictionary and pedagogical material for the teaching of Portuguese language. To meet this goal, the Corpus Pluricêntrico da Língua Portuguesa (CPLP), a large reference corpus composed of different varieties of Portuguese, is currently being compiled. Mostly due to copyright issues, the literary subcorpus has proven to be the most difficult subset to compile.

As TCDH is a reference in literary corpus compilation, my main goal during my visit was to create a framework that allows for the creation of decent literary corpus with financial and time constraints.

### Research Questions

To accomplish the goal above, the project approached the following research questions:

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

A. How and to which extent do literary texts published more than 70 years ago (public domain) linguistically differ from recently published ones?
B. Do automatic corpus analysis techniques (e.g. keyword analysis) yield significantly different results when comparing two literary corpora, one made with complete pieces and another one featuring less than 10% (fair use) of the full content of each piece?
C. When using only small part of the entire literary piece, what is the best extraction technique (e.g.: keeping a continuous chunk of sentences; gathering full sentences randomly selected)?

## Methodological plan and the visit

### Literature review

Little literature has been found on how to deal with copyright issues when compiling literary corpora. I dedicated the first three weeks of the residence to explore new approaches and readings suggested by the members of TCDH. I explored other approaches to address the main goal with this project and fine-tune the methodology.
At this initial stage I contributed to the writing of short survey papers on key methodological concerns for Computational Literary Studies under the CLS INFRA project. I've also became familiar with Pydistinto, a "Python implementation of different measures of distinctiveness for contrastive text analysis"[1], and added this tool to my methodological plan.
In those first weeks I also had the opportunity to share my office in Trier with another visiting scholar, from Leibniz Institute for the German Language, who introduced me to very useful tools such as the KorAP[2].

### Preparing new and existing corpora

This second step consisted in preparing the dataset to pilot the analysis. To pilot this framework, I used only literary texts written in European Portuguese. This is because the procedures applied for the text preparation are not greatly affected by minor differences observed among Portuguese varieties; and because some data for the European variety was already available.
The dataset was divided in two main groups: copyrighted books and public domain books The collection and preparation process varied according to the datasets.  Each text was minimally marked with XML; segmented and annotated (for POS and lemma); and the texts metadata were stored externally to the text, in tabular format.

### Copyrighted

This dataset comprises the texts that are copyrighted and cannot be freely acquired or distributed. Given the characteristics of these texts (frequently read; real representation of current standard language in use, etc.), this data is highly valuable for the creation of dictionaries and teaching material. During my visit I develop some strategies to circumvent some of the many obstacles found when preparing this dataset. An article with the outcomes is in preparation and to be submitted soon.

[1] https://tcdh.uni-trier.de/en/projekt/pydistinto
[2] https://korap.ids-mannheim.de/

### Public Domain

For the pilot study I used the Portuguese section of the European Literary Text Collection (ELTeC)[3] composed of 100 Portuguese literary pieces that are already in public domain. To make access to the data easier I made the text collection available via CQPweb[4], a corpus analysis web application. As some members of the TCDH were working with the French component of ELTeC, I also convert the French texts to CQPweb format and made them available[5].

### Creative Commons

I also collected recently written and published books and short stories published under the Creative Commons license, which will be made available online for research purposes.

### Contrasting different datasets

Once all data was collected, I prepared the corpora using Spacy[6] and contrast different combination of subsets of my dataset using pydistinto. An article with the results is in preparation.

## Outcomes

The visit to the TCDH proved to be more fruitful then expected. I not only managed to work on my pre-stablished research questions, but I also expand my knowledge from traditional corpus linguistic approaches to the always evolving world of digital humanities.
In addition to the framework developed the dataset collected and the articles in progress, my stay in Trier made it even clearer for me how curating existing datasets for optimal usage and adapting existing tools for our own research questions can make research more efficient. On a personal level, I also benefit from a very welcoming team, who gave me all the necessary support during my visit.

[3] https://www.distant-reading.net/eltec/

[4] https://ola.unito.it/CQPweb32/eltec_pt/

[5] https://ola.unito.it/CQPweb32/eltec_fr

[6] https://spacy.io/