

Linguateca & University of Oslo, Norway  
Linguateca & 2Ai Lab, IPCA, Portugal

## Literateca

Towards a computational environment for studying  
literature **in Portuguese**

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

Alberto Simões  
asimoes@ipca.pt

DH Budapest 2019, 25-27 September 2019



## Literateca

Introduction

## Contents

Works

Annotation

Size

## Example studies

Emotions across time

Speech description profiles

Reference to clothing

Health, pain, and the medical professions



- ▶ A literary and linguistic environment to study texts in Portuguese.



- ▶ A literary and linguistic environment to study texts in Portuguese.
- ▶ A corpus infrastructure on top of Open CWB, containing several annotation layers.



- ▶ A literary and linguistic environment to study texts in Portuguese.
- ▶ A corpus infrastructure on top of Open CWB, containing several annotation layers.
- ▶ With an interface to dedicated R programs for statistical computations and their visualization.



Literateca contains the following kinds of works:

- ▶ Classical works from 1300 onwards
- ▶ Canonical literary texts
- ▶ Non-canonical literary texts
- ▶ Excerpts of literary texts that have been translated into other languages



All texts have been annotated by a broad-coverage parser for Portuguese, PALAVRAS (Bick, 2000). In addition:



All texts have been annotated by a broad-coverage parser for Portuguese, PALAVRAS (Bick, 2000). In addition:

## Semantic domains

- ▶ Colours, clothing, body, family, emotions and health have been annotated
- ▶ and (partially) human-revised

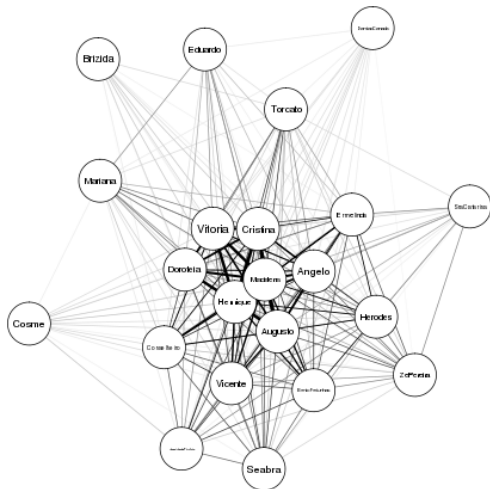
## "Literary" concerns

- ▶ Metadata like gender, genre and literary school has been added
- ▶ NE categories like person, place and works have been revised
- ▶ Characters have been added



# Contents

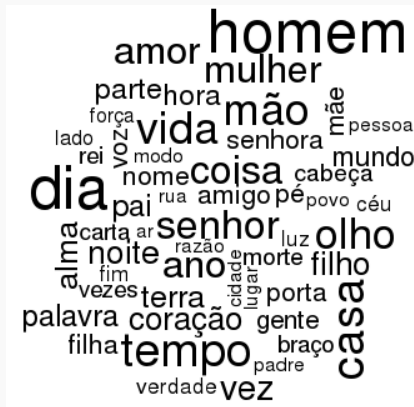
## Character network example





Current (15 august 2019) numbers

- ▶ 28 million words
- ▶ 784 works from 212 authors
- ▶ 187 novels from the "COST period"
- ▶ 6 novels annotated with characters





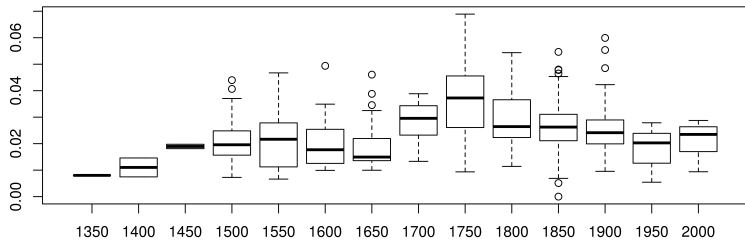
## A lexicon of emotion words

More than 4,000 lemmas divided in 24 categories. Not yet revised.

## Presence in the texts

amor	107,203	desejo	72,242
feliz	61,541	infeliz	61,488
medo	38,473	gen	30,237
vergonha	28,135	orgulho	25,439
feliz & satisfeito	22,889	coragem	21,117
surpresa	20,637	humildade	20,282
odio	19,502	esperanca	19,259
furia	14,943	satisfeito	14,651
desespero	14,601	saudade	13,037

# Emotions across time

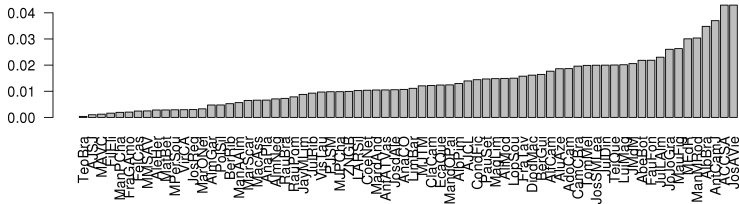




## Speech verbs

- ▶ Three kinds of report (direct speech, indirect speech, mixed), and simple mention
- ▶ use of - for direct speech
- ▶ often expressing attitude or feeling

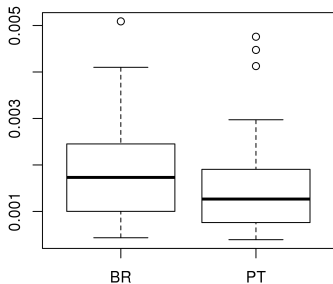
Speech reporting by 66 authors of novels



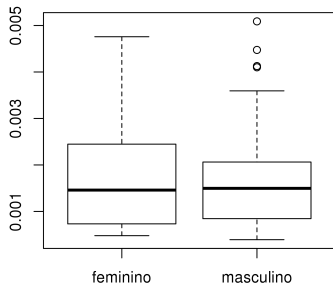
# Reference to clothing in novels



### Brazil vs. Portugal

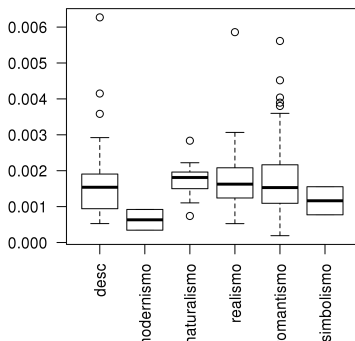


### Women vs. men

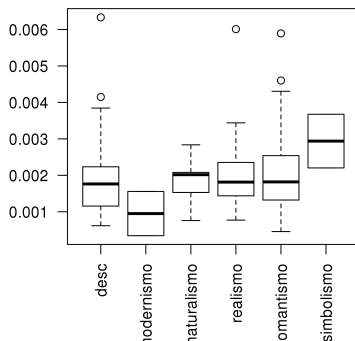




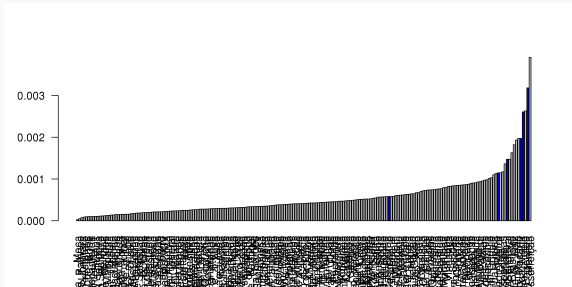
### Reference to health



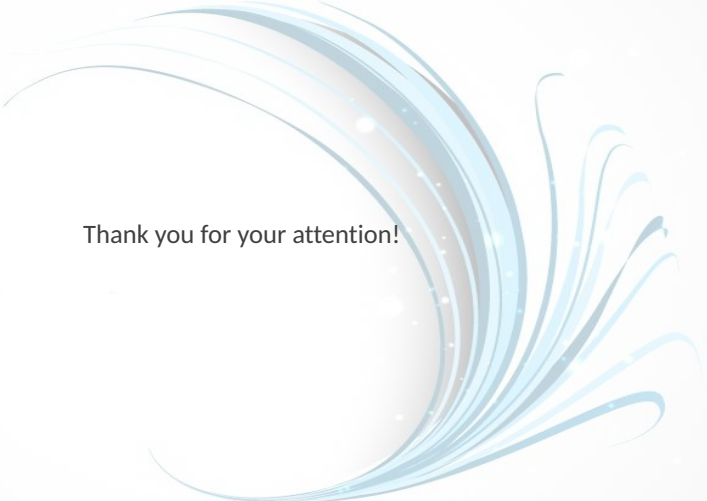
### Reference to health and pain



# The medical professions: Doctors and nurses







Thank you for your attention!