Literateca
Towards a computational environment for studying literature in Portuguese

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

Alberto Simões
asimões@ipca.pt

DH Budapest 2019, 25-27 September 2019

Presentation

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 corpus infrastructure on top of Open CWB, containing several annotation layers.
Introduction

- 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.

Contents

Works

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:

**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
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

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Count</th>
<th>Emotion</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>amor</td>
<td>107,203</td>
<td>desexo</td>
<td>72,242</td>
</tr>
<tr>
<td>feliz</td>
<td>61,541</td>
<td>infeliz</td>
<td>61,488</td>
</tr>
<tr>
<td>medo</td>
<td>38,473</td>
<td>gen</td>
<td>30,237</td>
</tr>
<tr>
<td>vergonha</td>
<td>28,135</td>
<td>orgulho</td>
<td>25,439</td>
</tr>
<tr>
<td>feliz &amp; satisfeito</td>
<td>22,889</td>
<td>coragem</td>
<td>21,117</td>
</tr>
<tr>
<td>surpresa</td>
<td>20,637</td>
<td>humildade</td>
<td>20,282</td>
</tr>
<tr>
<td>odio</td>
<td>19,502</td>
<td>esperanca</td>
<td>19,259</td>
</tr>
<tr>
<td>furia</td>
<td>14,943</td>
<td>satisfeito</td>
<td>14,651</td>
</tr>
<tr>
<td>desespero</td>
<td>14,601</td>
<td>saudade</td>
<td>13,037</td>
</tr>
</tbody>
</table>
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
Health and pain

The medical professions: Doctors and nurses
Thank you for your attention!