PANTERA: a parallel corpus to study translation between Portuguese and Norwegian

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

17 August 2017

PANTERA: translation between Norwegian and Portuguese

PANTERA: *Portuguese And Norwegian Texts for Education, Research and Acquisition of relevant cultural and linguistic capabilities*

- a parallel corpus modelled on COMPARA (modelled on ENPC) but with more information
- an additional coupling to STIG (*System for Translation Information in General*) under development by DMLF at UiO
- growing every day: see quantitative description at the PANTERA site, http://www.linguateca.pt/PANTERA/

The digitization and revision of PANTERA’s texts has been financed by ILOS/UiO through several research assistants: Heidi Jansen, Fernanda Veloso, Peder Østebø.
Context: Linguateca, Gramateca, Travelling emotions, STIG, ...

Use of PANTERA for linguistics: one example

Visualization of translation between the two languages

Use of PANTERA for teaching (of language and of translation)

---

Linguateca is a network for the computational processing of Portuguese (language) started by Portuguese research ministry after a public discussion in Portugal (1988-1989) about scientific policy.

It was a distributed initiative with several nodes, the leading one at SINTEF in Oslo, which in addition to resource development had a heavy workload on evaluation venues.

From 2010 onwards its funding was severely reduced, but the repositories and a subset of the projects has continued to this day, with some new projects happening due to my work at UiO.

Gramateca, and PANTERA, are just two of them.
Context: Gramateca

An international project based on AC/DC
- Corpus-based grammar of Portuguese
- An infrastructure which gathers annotated corpora
- A meeting point for researchers all around the world

http://www.linguateca.pt/Gramateca/

Contact: Diana Santos. Participants from other institutions: Syddanske Universitet, PUC-Rio, Univ. Coimbra, Univ. Lisboa, Univ. Minho, USP-São Paulo, Yamaguchi Universitet.

AC/DC cluster

Bird’s eye view of the (highly annotated) textual resources:
An inter-departmental research network at the Arts Faculty (HF) at UiO, for the **study of emotions in language**, and the way they (their words, concepts and forms of expression) have changed/travelled in time, place – and in society and text.

Core group:

- Jens Braarvig (IKOS)
- Anne Golden (ILN)
- Diana Santos (ILOS)


I chose a sub-sub-subproject in this large network to illustrate the potentialities of mixed methods, understood here as the intelligent combination of quantitative and qualitative approaches. The study of *Respeito* (respect) in Portuguese and its (un)correlation in Norwegian.

Since all words were automatically classified as emotions, we can also study the group(s) that include(s) *Respeito* (with several lexical items), which are: **HUMILDADE** (humbleness) and **ADMIRAÇÃO** (admiration).
Respeito in Portuguese

Using the AC/DC corpora (which underlie Gramateca, 1.28 billion words), we can have a coarse picture of this and morphologically related words.

<table>
<thead>
<tr>
<th>Word</th>
<th>CP</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>respeitar</td>
<td>20037</td>
<td>70375</td>
</tr>
<tr>
<td>respeito</td>
<td>18815</td>
<td>-7336</td>
</tr>
<tr>
<td>desrespeito</td>
<td>1561</td>
<td>9394</td>
</tr>
<tr>
<td>respeitável</td>
<td>1137</td>
<td>5554</td>
</tr>
<tr>
<td>desrespeitar</td>
<td>1124</td>
<td>7585</td>
</tr>
<tr>
<td>respeitado</td>
<td>963</td>
<td>8736</td>
</tr>
<tr>
<td>respeitoso</td>
<td>231</td>
<td>2917</td>
</tr>
<tr>
<td>respeitinho</td>
<td>34</td>
<td>41</td>
</tr>
<tr>
<td>desrespeitosamente</td>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>respeitosamente</td>
<td>0</td>
<td>925</td>
</tr>
<tr>
<td></td>
<td>44,014</td>
<td>306,338</td>
</tr>
</tbody>
</table>

Using parallel corpora

One way to look at the alignment (or not) of this “emotion” is to use translations as semantic data, something I have argued ever since I started my PhD.

So, using PANTERA, one can both

- look at the general picture,

  ([sema="emo.*(humildade|admirar).*"] | [lema="respeitável|respeito|respeitar...respeitinho"])

  [data from 12 Augusto 2017] ori 33 246197 1.32e-4
  trad 50 474898 1.05e-4

- or look at specific cases. (See translation examples, http://www.linguateca.pt/Diana/download/RespeitoPANTERA.pdf)
Looking at how Portugal sees Norway using translation

One different way one can use PANTERA – actually, STIG is much better because it includes way more information on the texts that are included in PANTERA – is to appreciate the choices of translations of Norwegian texts into Portuguese.

In the *Scandinavia through Sunglasses. Spaces of Cultural Exchange between Southern/Southeastern Europe and Nordic Countries*, UiO, 28-29 september, Ana Rita Ferreira and I will discuss the translations published in Portugal:

- Which texts were selected?
- Was there mediation through other languages/cultures?

Some problems associated with macrostudies

- Initial guesses that turned out to be wrong (no Norwegian original):

- Works not published in the original language:

- (Original) works with different names in different editions

- Works that come from different originals:
Chronology, from Norwegian to Portuguese, originals

First translation in 1894
Other problems

- Literary puzzles
  - Is there a Portuguese original for Kjærlighetsbrev fra den portugisiske nonne søster Marianne til Grev de Chamilly, 1934, Fabritius & Sønners Forlag? Or was it written in French?

- Posthumous works: which dato should one use?

- Choices of poems: should each poem tell as one original?

- Different editions: a good example is Uroens bok by Fernando Pessoa

- Before modern times: which date or edition should one consider?
  - Álvaro Velho. Dagbok fra Vasco da Gamas første reise til India, 1999, Folium forlag. Should one use 1498?

Chronology, from Portuguese to Norwegian, originals
translation in 1938

Teaching with PANTERA

- Frequent use as illustration of complex grammatical and semantic differences between the two languages
- Use as assignment in advanced and master courses: let the students do their research projects aided by authentic data
- (Planned) use for automatic exercises, with ParaEnsinador (under development)
Using translation data
- to improve search (for example of phenomena not marked in one language)
  - dative possessives
  - null objects
- to discover differences in semantic domains
  - fingers and toes
  - respect
- to confirm others’ claims

Concluding remarks

- An interesting project that will be always enriched due to new translations and better information about the texts (in STIG), and which will provide students with a wealth of “problems” and research objects.
- But: although it is relevant to have a direct comparison of the two languages, it is important to understand what is being compared, what the translations imply and their history. There is a high number of variables at stake, and many individuals (authors, translators) involved.