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

Diana Santos

Linguateca & University of Oslo

This paper presents an on-going project, PANTERA, which aims to identify all translations ever published between the two languages Portuguese and Norwegian, and obtain a sample of each for the study of translation between the two.

PANTERA stands for Portuguese And Norwegian Texts for Education, Research and Acquisition of relevant knowledge... and started in the autumn of 2013 as a cooperation between Linguateca and the University of Oslo. Although the corpus is relatively small, it is already one of the most diversified parallel corpora in the world, if one considers source diversity – in number of authors, time stamps and genres...

PANTERA’s development was inspired by the English-Norwegian Parallel Corpus, ENPC [4] and COMPARA [2], and, just like other similar projects – such as CorTrad and PoNTE – uses DISPARA [5] as the underlying software system. PANTERA allows the study of translation and of the similarities and differences between the two languages and is publicly available for search at http://www.linguateca.pt/PANTERA/.

Building the corpus involves the following steps: we revise the digitized material in the two languages (or digitize if it is not electronically available), then apply automatic alignment using the Open CWB workbench, and perform automatic (syntactic) annotation in the two languages. PALAVRAS [1] is used for Portuguese and the Oslo-Bergen-tagger [3] for Norwegian. We also apply some semantic annotation (so far only) to the Portuguese part.

Currently 371 texts have been identified for inclusion in PANTERA: 192 translations of Norwegian texts into Portuguese, and 179 translations of texts in Portuguese into Norwegian (see PANTERA’s site for a detailed description of the texts and the sizes of the excerpts). Further information about each translation instance (163 different authors and more than 167 translators) will be stored in the STIG system, in progress (https://stig.hf.uio.no/).

1
STIG aims to be a platform for all interested in translation studies in general, and especially for the Portuguese-Norwegian pair.

Figure 1 provides a bird’s eye view in terms of the (source) Portuguese texts included, chronologically. (Gil Vicente, born in the 15th century, as well as navigation chronicles from 1500 have been omitted not to bias the picture.)

Figure 1: Publication dates of originals in Portuguese after 1800

Figure 2 provides a bird’s eye view in terms of the (source) Norwegian texts included, chronologically.

Figure 2: Publication dates of Norwegian originals
Above the axis are the translations in Portugal, below in Brazil. In the paper, we will discuss the problems involved in labelling and identifying the texts, as well as possible reasons for the choice of the actual translations.

Some linguistic studies already performed using this corpus will be presented:

- the contrast between respeito (PT) and respekt (NO) as an illustration of culture differences reflected by language

- a comparison between the use of the body parts dedo (PT) and tå and finger (NO) as an illustration of different conceptualizations of the body in the two languages

- the identification of possessive datives and null objects in Portuguese using the translations into Norwegian, as a technique to elicit complex syntactic phenomena based on contrastive patterns

Finally, the use of the corpus for the semi-automatic creation of exercises and other teaching materials with the Ensinador paralelo tool [6] will be discussed in the context of the daily work of a university teacher of Portuguese for Norwegians.

The paper will end with some considerations about the possible uses of the resource, while also warning against some dangers of non-informed uses of the corpus.

References


