OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in the public domain*

Diana Santos[1,2][0000−0002−3108−7706], Cláudia Freitas[1,3][0000−0001−6807−85580], and Eckhard Bick[1,4][0000−0002−5505−4861]

[1] Linguateca https://www.linguateca.pt
[2] University of Oslo, HF, ILOS, Pb 1013 Blindern, Oslo, Norway
d.s.m.santos@ilos.uio.no
[3] PUC-Rio, Pontifícia Universidade Católica do Rio de Janeiro, Rua Marqus de São Vicente, 225, Gávea - Rio de Janeiro, RJ - Brasil - 22451-900
maclaudia.freitas@gmail.com
[4] Institute of Language and Communication, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark eckhard.bick@mail.dk

Abstract. OBras is an open corpus which can be downloaded from the Linguateca site and which was created in order to be the Brazilian counterpart of corpo Vercial, a corpus of public domain literary works from Portugal. Being a part of the AC/DC project, it is searchable from the Web, as well as regularly annotated with syntactic and semantic information, and featuring a new "edition" every two months.
Syntactical annotation is provided by the PALAVRAS parser [1], while colour, clothing, body, emotions and speech are annotated with specific rules. We present in this text several ways of quantifying its content.

Keywords: Portuguese · Brazilian literature · Web-searchable corpora

1 The AC/DC project

Ever since 1999 Linguateca has developed the AC/DC project so that people could interrogate annotated corpora in Portuguese, with ever increasing quality of annotation, a wider genre palette, and more kinds of information, compare [2] with [3].

Even though all material is fully available for querying, not all corpora included in AC/DC can be distributed in their entirety, due to copyright limitations. Literature is one of the genres included in AC/DC since its creation, but it is especially prone to availability restrictions. This is why most literary corpora only include old texts which are already in the public domain, or have restrictive conditions, like COMPARA.

* Thanks to all who have contributed with text preparation and annotation, under the scope of Linguateca.

This is why, at least in a first phase, we invested in literature which was already in the public domain, which basically spans, in what Brazil is concerned, one century. We named the corpus "Obras Brasileiras" with acronym OBras, and we are still in the process of adding more texts. http://www.linguateca.pt/OBras shows the works included, together with their metadata and size in tokens, as well as how to downnload it.

2 Corpus description

Through the several annotations one can describe a corpus on many levels. We start by the description in terms of part-of-speech, as well as the size and variety of proper names. Then we present the quantities of all different semantic data, for version 5.3 of 18 June 2018, see [4] for the annotation.

Table 1. Quantification of OBras according to annotation fields. NB! We have not included the multiword cases of colour, body and clothing for type counting

Annotation	Tokens	Types (lemmas)
Size	ca. 5 millions	151,676
Verbs	842,736	17,134
Nouns	965,805	26,426
Adjectives	289,507	11,087
Proper names	132,210	21,320
Colours	11,932	258
Clothing	10,395	208
Body	54,762	242
Saying verbs	78,219	825
Emotions	132,336	2,185

References

1. Bick, Eckhard: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Dr.phil. thesis. Aarhus University. Aarhus University Press, Aarhus, Denmark (2000)
2. Santos, Diana, Bick, Eckhard: Providing Internet access to Portuguese corpora: the AC/DC project. In: Gavrilidou, Maria et al. (eds.), Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), pp. 205-210. ELRA (2010)
3. Santos, Diana: Corpora at Linguateca: Vision and Roads Taken. In: Berber Sardinha, Tony, Ferreira, Telma de Lurdes São Bento (Eds.), Working with Portuguese Corpora, pp. 219-236. Bloomsbury (2014)
4. Anotaceḑilão. http://www.linguateca.pt/acesso/anotacao.html. Last accessed 29 July 2018