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

Diana Santos –University of Oslo & Linguateca
Cláudia Freitas – PUC-Rio & Linguateca
Eckhard Bick – University of Southern Denmark & Linguateca


UiO : University of Oslo


SDU

# OBras < AC/DC < Linguateca

- Linguateca
  - development of infrastructure for the automatic processing of the Portuguese language.
    - Resources; evaluation activities; Gramateca; Literateca
- AC/DC Project:
  - *Acesso a corpora/Disponibilização de corpora ("acess and availability of corpora")*
  - one of the activities of Linguateca, launched on 1999
  - 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
- OBras: Brazilian literature already in the public domain

# Obras – Obras Brasileiras

- Constantly increasing
  - Current version: 5.3
    - Obras: 223
    - Authors: 25
    - Tokens: 7.0 millions
    - Words: 5.0 millions

- Metadata

```
id=MdA_A_Mao_e_a_Luva
tit=A Mão e a Luva
aut=Machado de Assis
tip=Prosa
gen=romance
esc=romantismo
fonesc="Bosi 177-182"
dat=1874
fonte="http://machado.mec.gov.br/"
notas="Publicado originalmente em folhetins, a partir de 26/09/1874, em O Globo."
```

https://www.linguateca.pt/acesso/lista_autores_obras.html

| Código | Autor | Unidades | Obras |
|---|---|---|---|
| AA | Aluísio Azevedo | 310.282 | *A Mortalha de Alzira* (1894), *O Cortiço* (1890), *Aos vinte anos* (1865), *O Mulato* (1881), *Uma lágrima de mulher* (1879) |
| AdA | Álvares de Azevedo | 20.860 | *Noite na taverna* (1855) |
| ArtA | Artur Azevedo | 32.738 | *Contos fora da moda* (1894) |
| BG | Bernardo Guimarães | 166.689 | *A dança dos ossos* (1871), *A Escrava Isaura* (1875), *O ermitão do Muquém* (1868), *O seminarista* (1872) |
| CA | Castro Alves | | *O Gondoleiro do Amor* (1870) |
| CN | Coelho Neto | 158.410 | *A conquista* (1899), *Mano* (1924), *Turbilhão* (1904) |
| EdC | Euclides da Cunha | 229.324 | *Os Sertões I* (1902), *Os Sertões II* (1902), *Constrastes e Confrontos* (1866-1909) |
| FT | Franklin Távora | 208.314 | *O Cabeleira* (1876), *O Matuto* (1878), *O Sacrifício* (1879) |
| HC | Humberto de Campos | 157.300 | *A serpente de bronze* (1921), *O monstro e outros contos* (1932), *Grãos de mostarda* (1926) |
| IdS | Inglês de Sousa | 1.926 | *O Baile do Judeu* (1893) |
| JdA | José de Alencar | 302.346 | *Como e porque sou romancista* (1893), *Ubijarara* (1874), *A alma de Lázaro* (1873), *A pata da gazela* (1870), *O gaúcho* (1870), *A viuvinha* (1857), *O sertanejo* (1875) |
| JdR | João do Rio | 189.291 | *A Alma Encantadora das Ruas* (1807), *As Religiões no Rio* (1904), *Dentro da Noite* (1910) |
| JLdA | Júlia Lopes de Almeida | 129.388 | *A falência* (1901), *A viúva Simões* (1895) |
| JMdM | Joaquim Manuel de | 255.478 | *A Moreninha* (1844), *A luneta mágica* (1869), *O moço louro* (1845) |

# OBras v. 5.3

Literary manifestation

**Distribuição**

Houve **16** valores diferentes de **escola**.

| | |
|---|---|
| realismo | 682956 |
| romantismo | 647017 |
| realismo_regionalismo_romantismo | 216618 |
| naturalismo | 212297 |
| romantismo_regionalismo | 204498 |
| naturalismo_realismo | 189950 |
| histórico | 188429 |
| modernismo | 93559 |
| regionalismo | 82161 |
| impressionismo_naturalismo_realismo_simbolismo | 72184 |
| naturalismo_regionalismo | 68831 |
| naturalismo_realismo_romantismo | 57277 |
| romantismo_decadentismo | 42769 |
| indianismo_romantismo | 26221 |

Literary genres

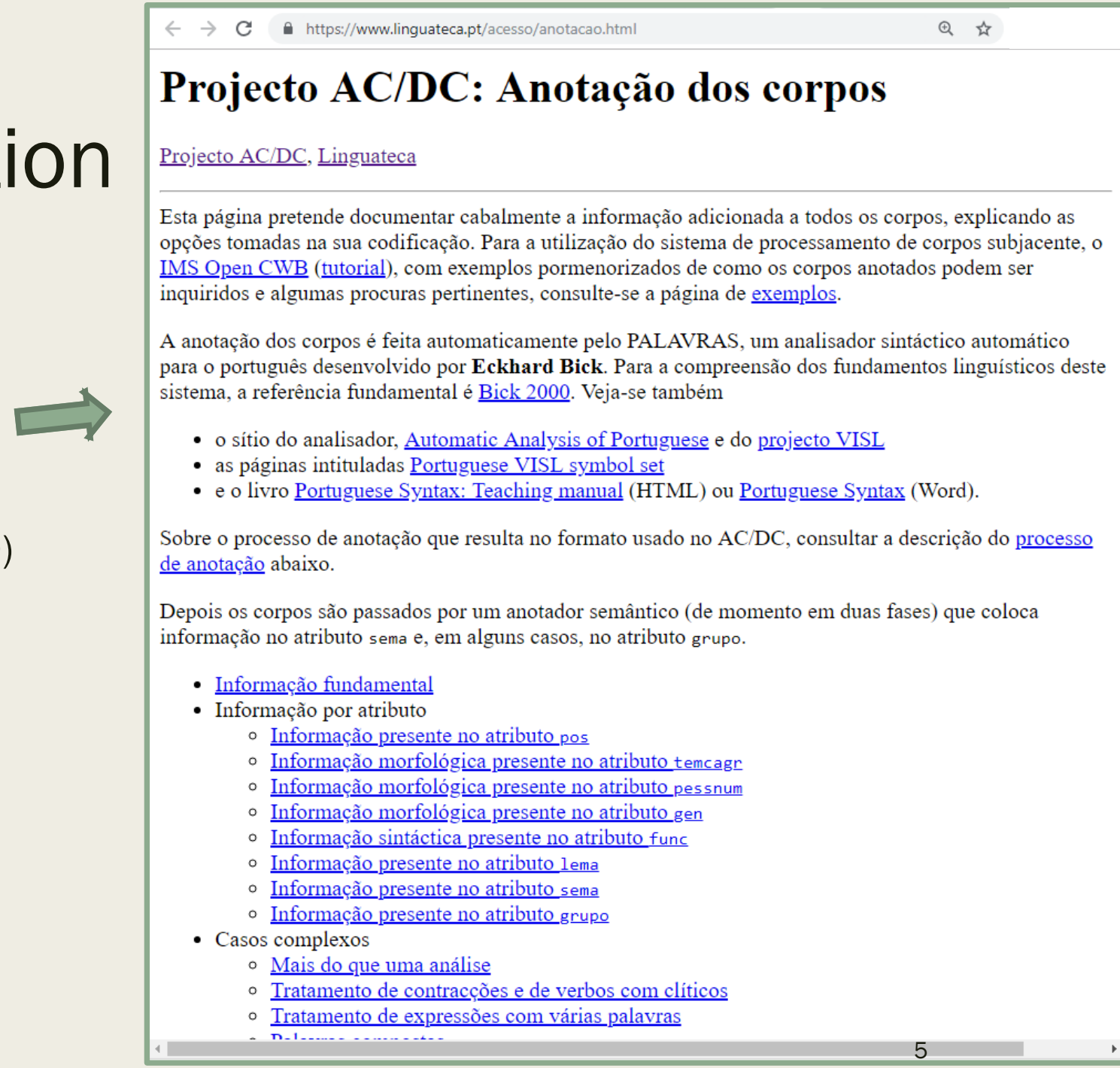
**Distribuição**

Houve **10** valores diferentes de **classe**.

| | |
|---|---|
| Prosa:romance | 3035657 |
| Prosa:conto | 1082894 |
| Prosa:crônica | 810682 |
| Prosa:prosa | 190539 |
| Prosa:novela | 186332 |
| Prosa:artigo | 46464 |
| Teatro:tragédia | 11154 |
| Prosa:autobiografia | 10875 |
| Teatro:comédia | 3272 |

# Obras - Annotation

- Morphosyntactic annotation: PALAVRAS parser (Bick, 2000)
- Semantic annotation: Corte-e-Costura (Santos & Mota, 2010)
  - *Colours*
  - *Clothing*
  - *Body*
  - *Saying Verbs*
  - *Emotions*



https://www.linguateca.pt/acesso/anotacao.html

## Projecto AC/DC: Anotação dos corpos

Projecto AC/DC, Linguateca

Esta página pretende documentar cabalmente a informação adicionada a todos os corpos, explicando as opções tomadas na sua codificação. Para a utilização do sistema de processamento de corpos subjacente, o IMS Open CWB (tutorial), com exemplos pormenorizados de como os corpos anotados podem ser inquiridos e algumas procuras pertinentes, consulte-se a página de exemplos.

A anotação dos corpos é feita automaticamente pelo PALAVRAS, um analisador sintáctico automático para o português desenvolvido por **Eckhard Bick**. Para a compreensão dos fundamentos linguísticos deste sistema, a referência fundamental é Bick 2000. Veja-se também

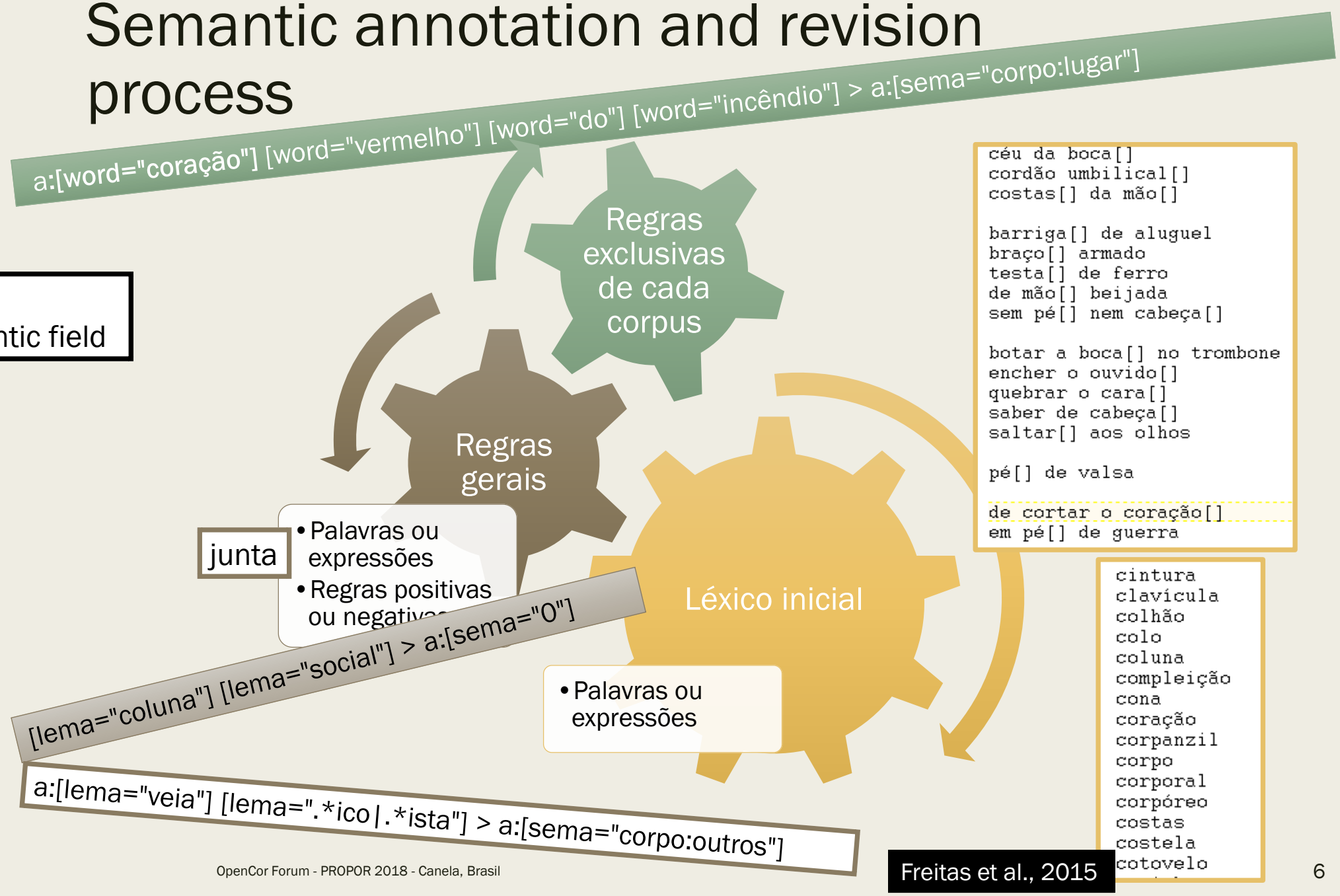
- o sítio do analisador, Automatic Analysis of Portuguese e do projecto VISL
- as páginas intituladas Portuguese VISL symbol set
- e o livro Portuguese Syntax: Teaching manual (HTML) ou Portuguese Syntax (Word).

Sobre o processo de anotação que resulta no formato usado no AC/DC, consultar a descrição do processo de anotação abaixo.

Depois os corpos são passados por um anotador semântico (de momento em duas fases) que coloca informação no atributo `sema` e, em alguns casos, no atributo `grupo`.

- Informação fundamental
- Informação por atributo
  - Informação presente no atributo `pos`
  - Informação morfológica presente no atributo `temcagr`
  - Informação morfológica presente no atributo `pessnum`
  - Informação morfológica presente no atributo `gen`
  - Informação sintáctica presente no atributo `func`
  - Informação presente no atributo `lema`
  - Informação presente no atributo `sema`
  - Informação presente no atributo `grupo`
- Casos complexos
  - Mais do que uma análise
  - Tratamento de contracções e de verbos com clíticos
  - Tratamento de expressões com várias palavras

# Semantic annotation and revision process



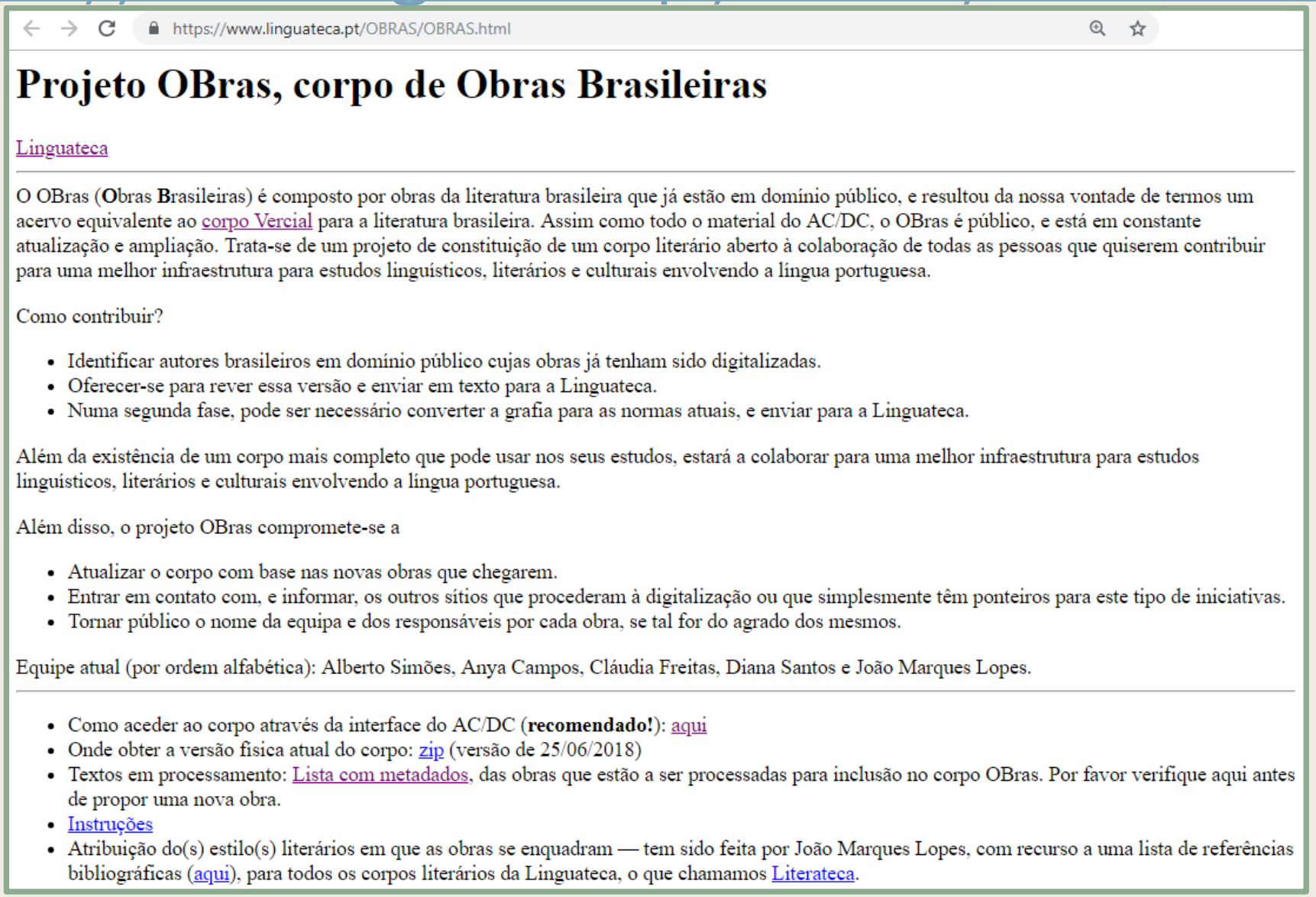

Freitas et al., 2015

# OBras: Corpus description

| 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 |

partially human-revised

https://www.linguateca.pt/OBRAS/OBRAS.html




# Projeto OBras, corpo de Obras Brasileiras

Linguateca

---

O OBras (**O**bras **B**rasileiras) é composto por obras da literatura brasileira que já estão em domínio público, e resultou da nossa vontade de termos um acervo equivalente ao corpo Vercial para a literatura brasileira. Assim como todo o material do AC/DC, o OBras é público, e está em constante atualização e ampliação. Trata-se de um projeto de constituição de um corpo literário aberto à colaboração de todas as pessoas que quiserem contribuir para uma melhor infraestrutura para estudos linguísticos, literários e culturais envolvendo a língua portuguesa.

Como contribuir?

- Identificar autores brasileiros em domínio público cujas obras já tenham sido digitalizadas.
- Oferecer-se para rever essa versão e enviar em texto para a Linguateca.
- Numa segunda fase, pode ser necessário converter a grafia para as normas atuais, e enviar para a Linguateca.

Além da existência de um corpo mais completo que pode usar nos seus estudos, estará a colaborar para uma melhor infraestrutura para estudos linguísticos, literários e culturais envolvendo a língua portuguesa.

Além disso, o projeto OBras compromete-se a

- Atualizar o corpo com base nas novas obras que chegarem.
- Entrar em contato com, e informar, os outros sítios que procederam à digitalização ou que simplesmente têm ponteiros para este tipo de iniciativas.
- Tornar público o nome da equipa e dos responsáveis por cada obra, se tal for do agrado dos mesmos.

Equipe atual (por ordem alfabética): Alberto Simões, Anya Campos, Cláudia Freitas, Diana Santos e João Marques Lopes.

---

- Como aceder ao corpo através da interface do AC/DC (**recomendado!**): aqui
- Onde obter a versão fisica atual do corpo: zip (versão de 25/06/2018)
- Textos em processamento: Lista com metadados, das obras que estão a ser processadas para inclusão no corpo OBras. Por favor verifique aqui antes de propor uma nova obra.
- Instruções
- Atribuição do(s) estilo(s) literários em que as obras se enquadram — tem sido feita por João Marques Lopes, com recurso a uma lista de referências bibliográficas (aqui), para todos os corpos literários da Linguateca, o que chamamos Literateca.

# References


- SANTOS, Diana & MOTA, Cristina. "Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora". In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), **Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010**). European Language Resources Association, pp. 1437-1444, 2010.
- BICK, Eckhard. **The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Aarhus University Press, 2000.
- FREITAS, Cláudia; SANTOS, Diana; MOTA, Cristina; CARRIÇO, Bruno; JANSEN, Heidi. O léxico do corpo e anotação de sentidos em grandes corpora: o projeto Esqueleto. **Revista de Estudos da Linguagem**, v.23, n.3, p. 641-680, 2015.
- SANTOS, Diana; MAIA, Belinda. Language, emotion, and the emotions: A computational introduction. **Language and Linguistics Compass**. 2018.