Universal Grammatical Dependencies for Portuguese with CINTIL Data, LX Processing and CLARIN Support

António Branco, João Silva, Luís Gomes and João Rodrigues
Faculty of Sciences of the University of Lisbon

LREC 2022 Marseille

The largest UD POS corpus and UD treebank for Portuguese, and their corresponding POS tagger and dependency parser

- CINTIL Corpus
- CINTIL DependencyBank
- CINTIL DependencyBank PREMIUM
- CINTIL-UPos
- LX-UTagger
- 675k tokens
- 99.01% accuracy
- CINTIL-UDep Treebank
- LX-UDParser
- 38k sentences
- 90.87 UAS / 88.01 LAS

Tools also available as online, web, batch, and notebook services
https://portulanclarin.net/workbench/lx-udparser/

This work was supported by PORTULAN CLARIN - Research Infrastructure for the Science and Technology of Language funded by Lisboa 2020, Alentejo 2020 and FCT - Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016