Gramateca: corpus-based grammar of Portuguese

DIANA SANTOS

http://www.linguateca.pt

RATIONALE

We believe that grammar emerges from consideration of (large amounts of) real text. Most grammars do not allow checking the actual data on which they were created – even if they purport to be corpus based. Gramateca

- gives access to the data
- allows querying the data
- allows for alternative questions
- allows linguists to share their interpretation of the material
- allows the revision and creation of alternative datasets

FOUNDATIONS

Underlying Gramateca, there is a set of services on the Web called the AC/DC services

- serving 27 different corpora, 1.5 thousand million words
- marked up and/or annotated with several kinds of information

AC/DC has been serving the community of the computational processing of Portuguese – and the Portuguese linguistics comunity – for 15 years, thanks to Open CWB and PALAVRAS.

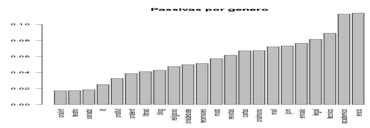
RESEARCH THEMES

Depending on the team members

- Conditional constructions

Construction	cases
se	CIRC 47 ENUN 9 CONT 4 OUTRO 22 unknown 2
a	OUTR 7 CIRC 2 ENUN 1
caso	CIRC 4 OUTR 2

- Oral vs. written Portuguese



- Body language

	Total before	Total after	BR	PT
corpo	4547	1289	764	525
corpo:outros		799	486	318
corpo:sentimento		78	59	19
corpo:animal		42	23	4
corpo:posição		34	24	10
corpo:partedeobjeto		19	9	10
corpo:faculdade		18	12	6
corpo:vegetal		16	14	2
corpo:opinião		13	11	2
corpo:lugar		12	10	2
corpo:doença		2	1	1

- Emotions
- Causation

SERVICES

Procura http://www.linguateca.pt/ACDC/

Distribuidor http://www.linguateca.pt/Distribuidor/

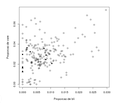
Comparador http://www.linguateca.pt/Comparador/

Ordenador http://www.linguateca.pt/Ordenador/

VARRA http://www.linguateca.pt/VARRA/

Ensinador http://www.linguateca.pt/Ensinador/

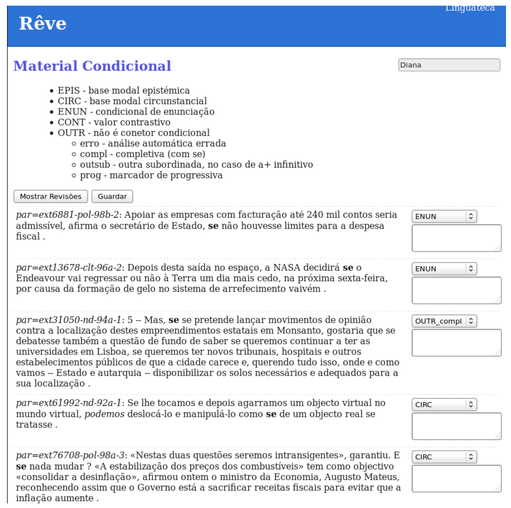
Rêve http://www.linguateca.pt/Reve/

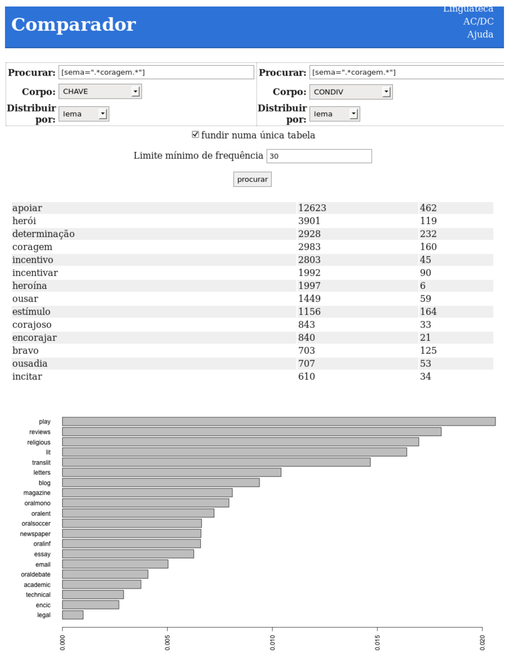


Ensinador paralelo

EXAMPLES

- inspection of annotated data
- reannotation
- comparison of corpora
- visualization capabilities





OTHER WORK

Other alleys have been started under Gramateca

- Comparison of parser schemes for Portuguese
- Parser evaluation
- Common author/text register

REFERENCES

Eckhard Bick. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Diana Santos. 2014. Corpora at Linguateca: Vision and roads taken. In Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese Corpora*, Bloomsbury, 2014, pp. 219-236.

Diana Santos. 2014. Podemos contar com as contas? In Sandra Aluísio & Stella Tagnin (eds.), *New Language Technologies and Linguistic Research: A Two-way Road*, Cambridge Scholars Publishing, 2014, pp. 194-213.

Diana Santos. 2014. Comparando corpos orais (transcritos) e escritos na Gramateca. In Camilla Bardel (ed.), *Proceedings from the conference Parler les langues romanes/Parlare le lingue romanze/Hablar las lenguas romances/Falando línguas românicas GSCP 2014*, University Press Università di Napoli L'Orientale.

Alberto Simões e Diana Santos. 2014. Nos bastidores da Gramateca: uma série de serviços. In *TorPorEsp 2014*, São Carlos

Diana Santos, Rui Pedro Ribeiro Marques, Cláudia Freitas, Cristina Mota & Alberto Simões. 2014. Comparando anotações na Gramateca. In *ELC 2014*, Uberlândia

ACKNOWLEDGEMENTS

Linguateca was supported by MCTES, UMIC and FCT for 10 years. From 201 it has been financed by the research institutions to which the members belong For the work presented here, I am grateful to the University of Oslo and to th Research Computing Group, and especially to all Gramateca participants.