ComCorHd - Linguística Computacional,
Corpus e Humanidades Digitais

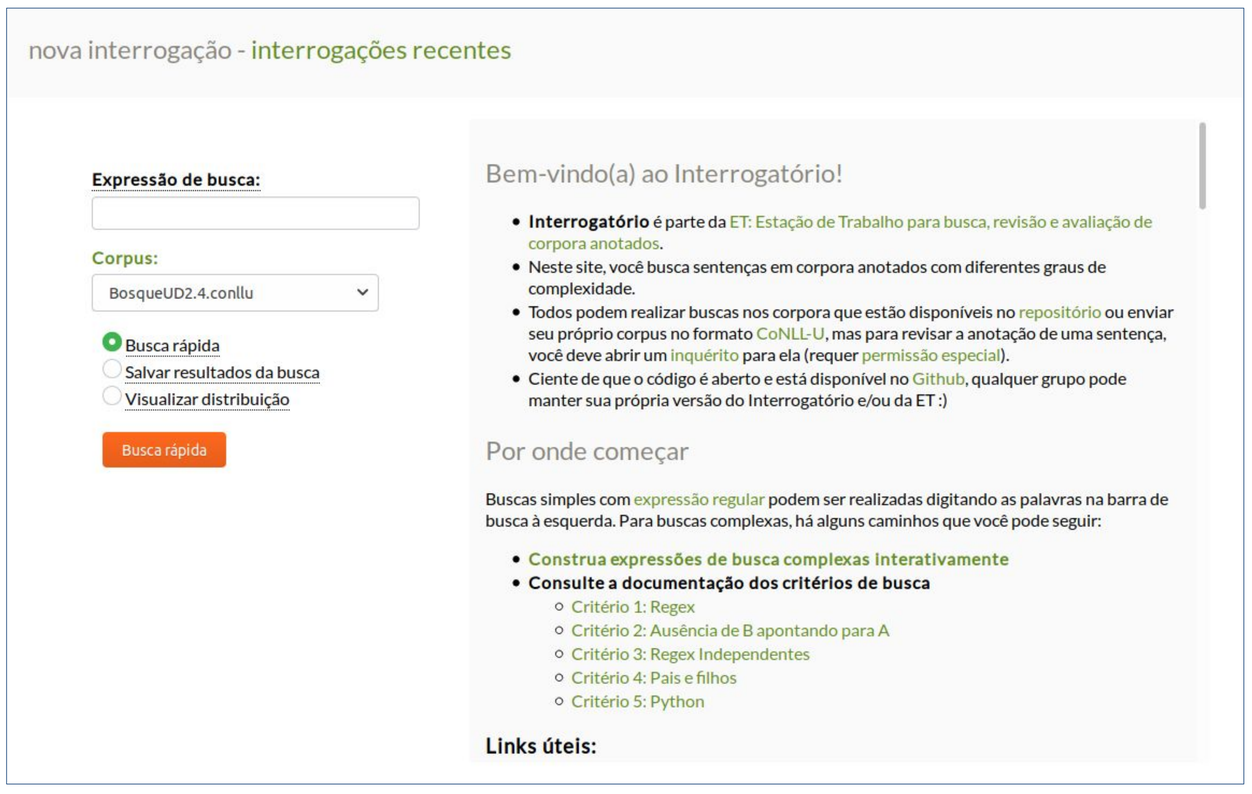
ET: uma Estação de trabalho para busca, revisão e avaliação de corpora anotados morfossintaticamente

Elvis de Souza
Departamento de Letras
PUC-Rio, Brasil

Cláudia Freitas
Departamento de Letras
PUC-Rio, Brasil

Nossa ET

- Desenhada a partir do ponto de vista linguístico
- Objetivo de facilitar a busca (1), a revisão (2) e a avaliação (3) de corpora anotados
- Permite o alinhamento do trabalho feito pelos especialistas em língua e o desempenho de sistemas de PLN
- Acesso público pelo Interrogatório, ambiente de busca e revisão em corpora anotados no formato CoNLL-U: http://comcorhd.letras.puc-rio.br/ET



Busca (1) e revisão (2)

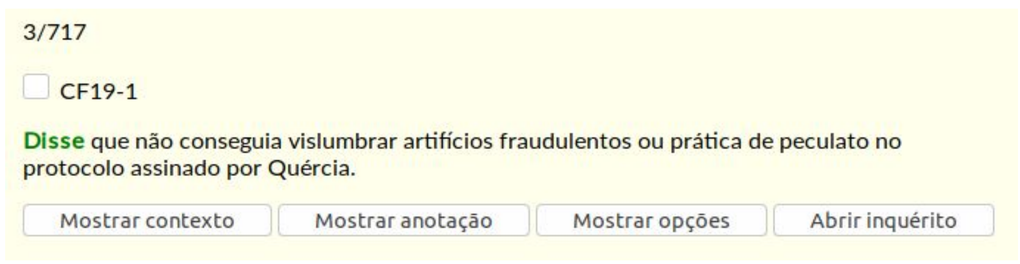


Figura 1: Frase encontrada como resultado da primeira busca por sujeitos ocultos no Interrogatório

# text = Os três cortadores de cana eram de Alagoas e estavam na cidade havia 15 dias.									
# source = CETENFolha n=269 cad=Cotidiano sec=soc sem=94a									
# sent_id = CF269-2									
# id = 1131									
1	Os	o	DET	_	Definite=Def|Gender=Masc|Number=Plur|PronType=Art	3	det	_	_
2	três	três	NUM	_	NumType=Card	3	nummod	_	_
3	cortadores	cortador	NOUN	_	Gender=Masc|Number=Plur	8	nsubj	_	_
4	de	de	ADP	_	_	5	case	_	_
5	cana	cana	NOUN	_	Gender=Fem|Number=Sing	3	nmod	_	_
6	eram	ser	AUX	_	Mood=Ind|Number=Plur|Person=3|Tense=Imp|VerbForm=Fin	8	cop	_	_
7	de	de	ADP	_	_	8	case	_	_
8	Alagoas	Alagoas	PROPN	_	Gender=Masc|Number=Sing	0	root	_	_
9	e	e	CCONJ	_	_	8	cc	_	_
10	estavam	estar	VERB	_	Mood=Ind|Number=Plur|Person=3|Tense=Imp|VerbForm=Fin	8	conj	_	_
11-12	na	_	_	_	_	_	_	_	_
11	em	em	ADP	_	_	13	case	_	_
12	a	o	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	13	det	_	_
13	cidade	cidade	NOUN	_	Gender=Fem|Number=Sing	10	iobj	_	_
14	havia	haver	VERB	_	_	10	advcl	_	_
15	15	15	NUM	_	NumType=Card	16	nummod	_	_
16	dias	dia	NOUN	_	Gender=Masc|Number=Plur	14	obj	_	SpaceAfter=No
17	.	.	PUNCT	_	_	8	punct	_	_

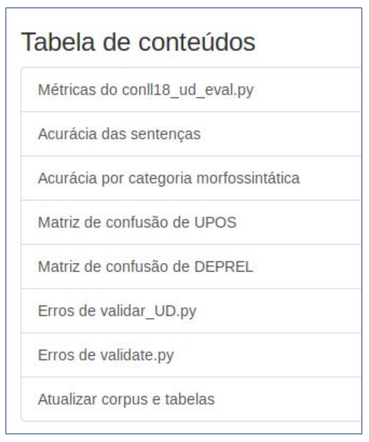
Figura 2: Inquérito - Interface de edição da anotação de uma sentença

Como buscar sujeitos ocultos

Critério	Expressão de busca
2	root#8#nsubj|csubj|nsubj:pass#8
5	lemma = "haver" and deprel = "root" and feats = "Number=Sing" and feats = "Person=3"
4	!\tAUX\t.*cop :: ^(?!.*VERB).*root
5	lemma = "(chover|ventar|anoitecer|amanhecer|entardecer|relampejar|trovejar|escurecer|clarear" and deprel = "root"

Tabela 1: Expressões de busca utilizadas na busca por sujeitos ocultos

Avaliação (3)



UD[2] / UD[1]	ADJ	ADP	ADV	AUX	CCONJ	DET	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X	_	All
ADJ	386	1	2	0	0	0	27	0	0	0	5	0	0	0	23	0	0	444
ADP	0	1616	1	0	0	4	0	0	0	0	2	0	0	0	0	0	0	1623
ADV	2	3	350	0	0	2	3	0	0	1	1	0	1	0	1	0	0	364
AUX	0	0	0	272	0	0	1	0	0	0	0	0	0	0	10	0	0	283
CCONJ	0	1	3	0	203	0	2	0	0	0	0	0	1	0	0	0	0	210
DET	1	1	3	0	0	1543	0	1	0	6	6	0	0	0	0	0	0	1561
NOUN	29	2	3	2	0	3	1837	0	0	0	41	0	0	0	8	0	0	1925
NUM	2	2	0	0	0	1	3	230	0	0	10	0	0	0	0	0	0	248
PART	0	3	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	17
PRON	0	0	6	0	0	7	2	1	0	295	2	0	5	0	0	0	0	318
PROPN	4	0	1	0	0	0	29	2	0	1	812	0	0	0	2	0	0	851
PUNCT	0	0	0	0	0	0	0	0	0	0	0	1343	0	0	0	0	0	1343
SCONJ	0	1	3	0	0	0	0	0	0	12	0	0	84	0	0	0	0	100
SYM	0	0	0	0	0	0	0	0	0	0	1	0	0	29	0	0	0	30
VERB	13	1	1	7	0	0	5	0	0	0	5	0	0	0	821	0	0	853
X	1	2	0	0	0	0	7	0	0	0	17	0	0	0	0	3	0	30
_	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	742	742
All	438	1633	373	281	203	1560	1916	234	14	315	902	1343	91	29	865	3	742	10942

Figura 3: Julgamento - diferentes formas de avaliação de corpora no formato CoNLL-U, e uma matriz de confusão gerada dentro do ambiente

Agradecimentos

Elvis de Souza é bolsista de Iniciação Científica do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) no projeto *Construção de datasets para o PLN de língua portuguesa*. Número do processo da bolsa: 128693/2019-3.

Referências

C. D. Manning, "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?" in International conference on intelligent text processing and computational linguistics. Springer, 2011, pp. 171– 189.

A. Rademaker, F. Chalub, L. Real, C. Freitas, E. Bick, and V. de Paiva, "Universal dependencies for portuguese," in Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), 2017, pp. 197–206.

R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, T. Oscar et al., "Universal dependency annotation for multilingual parsing," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2013, pp. 92–97.

M. Straka, J. Hajic, and J. Strakova, "Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing," in Proceedings of the tenth international conference on language resources and evaluation (LREC 2016), 2016, pp. 4290–4297.

D. Santos, "Podemos contar com as contas?" In Sandra Maria Alu´ısio; Stella E O Tagnin (ed) New Language Technologies and Linguistic Research: A Two-Way Road Cambridge Scholars Publ 2014; 2014, 2014. [6] F. M. Tyers, M. Sheyanova, and J. N. Washington, "Ud annotatrix: an annotation tool for universal dependencies," 2017.

C. Freitas, E. de Souza, and L. Rocha, "Quantificando (e qualificando) o sujeito oculto em português," in ˆ VI Jornada de Descric¸ao do Português, ˆ STIL 2019, 2019.

L. Rocha, I. Soares-Bastos, C. Freitas, and A. Rademaker, "Scavenger hunt: what do we find when look for confusions," in International Conference on the Computational Processing of Portuguese, PROPOR 2018, 2018.

C. Freitas, L. F. Trugo, F. Chalub, G. Paulino-Passos, and A. Rademaker, "Tagsets and datasets: Some experiments based on portuguese language," in International Conference on Computational Processing of the Portuguese Language. Springer, 2018, pp. 459–469.