Leitura Distante: o que podemos com corpora anotados?

Cláudia Freitas

Linguateca

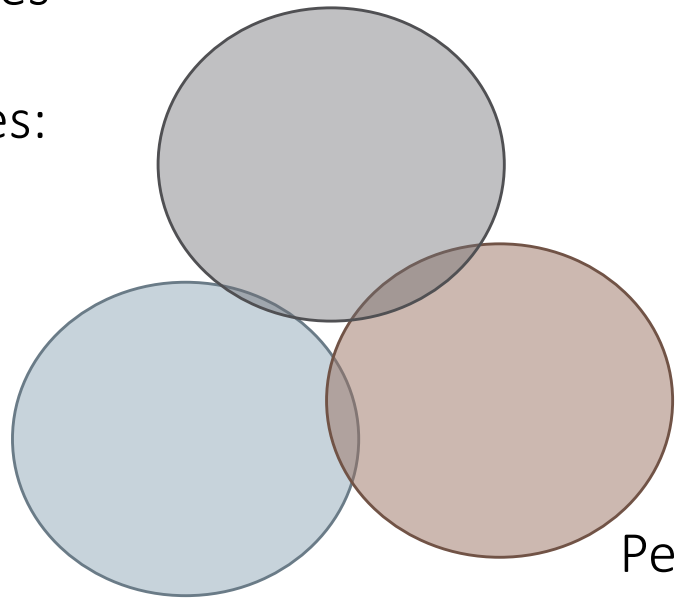
PUC RIO

CAPES

Preâmbulo

Leitura Distante (Moretti, 2003)
A distância como aliada → padrões

Dados que dão forma aos padrões:
metadados (em geral)
Estão *fora* do texto



Corpus, anotação e ferramentas
(Santos, 2009)
Um corpus é um trio: o texto, a anotação, a ferramenta de busca

Leitura não-linear (Freitas 2017)
Padrões, mas também *singularidades*

Rearranjos tornados possíveis apenas com o texto descorporificado (Paixão de Souza, 2013)

Permite alternar lentes entre a leitura distante e a leitura aproximada

Estratégias e abordagens

Analisar dados

Atribuir sentido aos dados
(para analisá-los)

Um exemplo - gênero de personagens como operador analítico

- Como são caracterizadas personagens masculinas e femininas em obras literárias brasileiras dos séculos XIX e XX ?

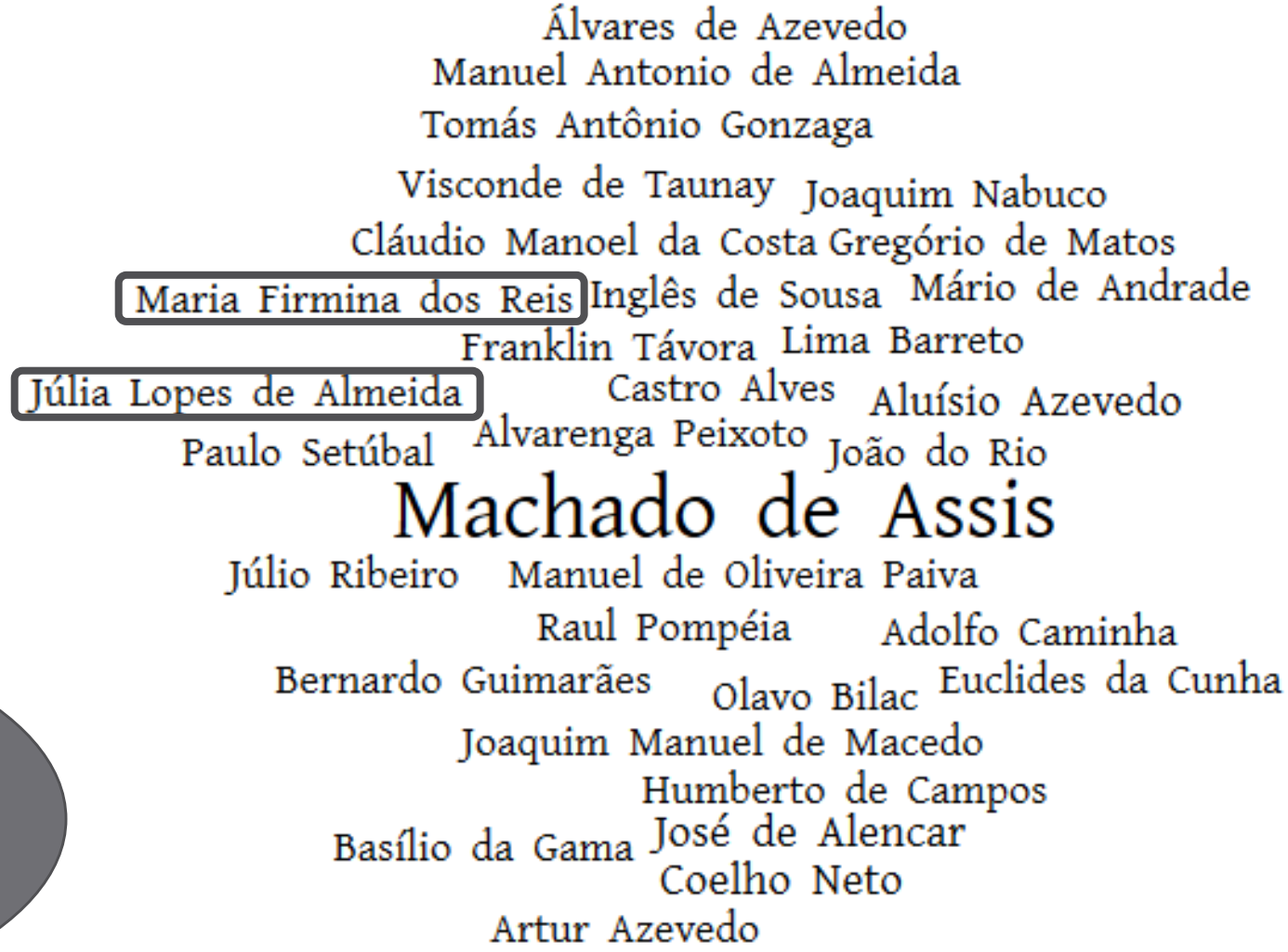
Corpus OBras

248 obras da literatura brasileira – v 5.3

32 autores séc XIX-XX

3 obras de autoria feminina

Obras escritas por quem e para quem?
Censo de 1872: 16% de população brasileira era alfabetizada....



Um corpus é um trio….

Projeto AC/DC: corpo OBras

AC/DC : Linguateca

O corpo **OBras** (Obras Brasileiras) é um corpo de textos brasileiros que já alcançaram o domínio público, criado numa colaboração entre a Linguateca, a Universidade de Oslo, a PUC-Rio, a Universidade Estadual do Maranhão (UEMA) e Anya Campos. Para mais informações veja-se a página do projeto.

Procurar: OK

Resultado:

- ○ Concordância
- ○ Distribuição das formas (word)
- ◉ Distribuição dos lemas (lema)
- ○ Distribuição da categoria gramatical (PoS) (pos)
- ○ Distribuição do tempo verbal e/ou do caso pronominal (temcagr)
- ○ Distribuição de pessoa e/ou número (pessnum)
- ○ Distribuição do género morfológico (gen)
- ○ Distribuição da função sintáctica (func)
- ○ Distribuição pelas obras (obra)
- ○ Distribuição por autores (autor)
- ○ Distribuição por género de texto (classe)
- ○ Distribuição pela corrente literária (escola)
- ○ Distribuição pelo sexo do entrevistado, do biografado ou do autor (sexo)
- ○ Distribuição por original ou traduzido (trad)
- ○ Distribuição por campo semântico (sema)
- ○ Distribuição por grupo (de cor, roupa, etc.) (grupo)
- ○ Distribuição das dependências (dependencias)

Tipo	Literário
Variante(s)	BR
Tamanho (unidades)	8.9 milhões
Tamanho (palavras)	6.3 milhões

Carateres úteis: | { } [ ]
Página principal
Procure noutros corpos:
AmostRA-NILC ANCIB Avante! Corpus Brasileiro CD HAREM CETEMPúblico CHAVE Ciência Viva Colonia CONDIVport CONDIVport2 CoNE C-Oral-Brasil DHBB DiaCLAV Diáspora TL-PT ECI-EBR ECI-EE ENPCPUB (parte em português) Floresta FrasesPB FrasesPP Mariano Gago Literateca Marielle, presente! Moçambula Museu da Pessoa Natura/Minho NOBRE **OBras** P'lo

Metodologia

1. Mineração de textos: Busca por estruturas predicadoras

função
APOSTO
ADJUNTO ADNOMINAL
PREDICATIVO DO SUJEITO

+

NOMES COMUNS
NOMES PRÓPRIOS
PRONOMES PESSOAIS

Concretamente....

id="Os_irmãos_Leme Prosa:romance PS 1933 histórico masc ":
Rodrigo César, iradíssimo, deliberou invadir as atribuições do Ouvidor.

id="O_Coruja Prosa:romance AA 1889 realismo masc ":
é feio, mas enfim, sempre há **homens sérios**, cujo nome o público não ignora;

id="O_Cabeleira Prosa:romance FT 1876 realismo_regionalismo_romantismo masc ":
Mas **José César era ativo**, enérgico, esforçado e de grandes espíritos.

id="Uma_lágrima_de_mulher prosa:romance AA 1880":
E **Rosalina**, **meiga**, encarava com chorosa ternura o olhar sombrio de Miguel.

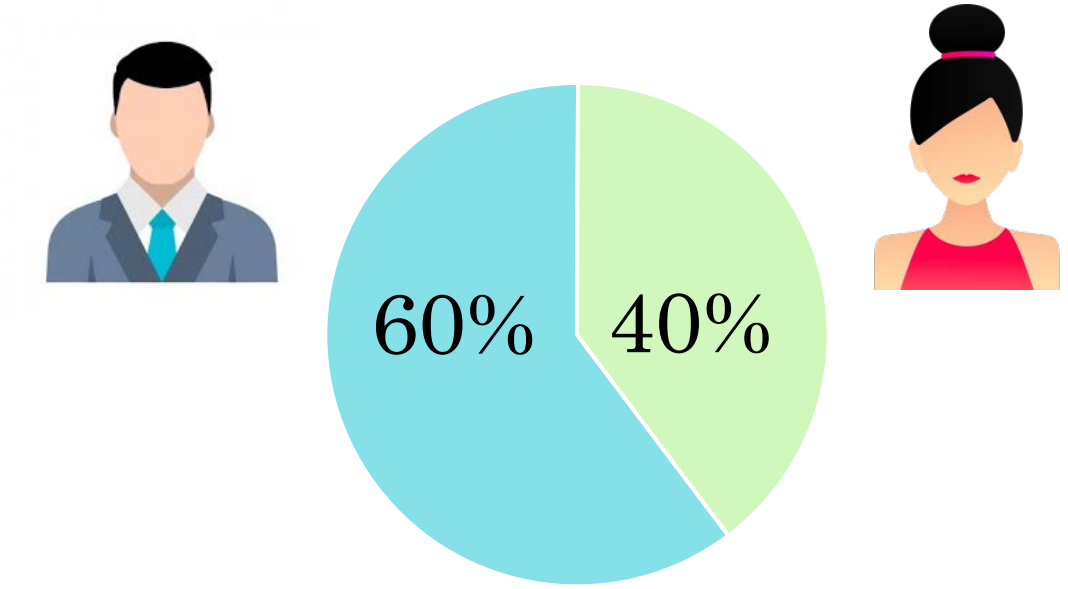
id="A_serpente_de_bronze prosa:conto HC 1921":
Criemos as meninas com decoro, vestindo-as com discrição, e teremos **moças discretas**, **pudicas**, **decorosas**, **ciosas** do seu corpo e dos seus encantos.

id="O_seminarista prosa:romance BG 1872":
A **mãe** de Eugênio era **fanática** e **supersticiosa**.

Mais concretamente…

```
[pos="PROP.*" & func="SUBJ>"] [lema="ser|estar"] [pos="ADV.*"]* @[temcagr!=".*PASS.*" & pos="ADJ|N|V" & gen="M" & func="<SC"]
[pos="PROP.*"] "era" [pos="ADV.*"]* [pos="ADJ.*" & gen="M" & func!=">N"] [word="e"] @[pos="ADJ.*" & gen="M" & func!=">N"]
[lema="ele" & func="SUBJ>"] [lema="ser|estar"] [pos="ADV.*"]* @[temcagr!=".*PASS.*" & pos="ADJ|N|V" & gen="M" & func="<SC"]
[pos="PROP.*" & func!="P<"] "," [pos="ADV.*"]* @[func="N<PRED|.*APP.*" & gen="M" & pos="ADJ"]
[pos="PROP.*" & func!="P<"] "," [pos="ADV.*"]* [func="N<PRED|.*APP.*" & gen="M" & pos="ADJ"] "e" @[gen="M" & pos="ADJ"]
[lema="homem|moço|rapaz|marido"] @[pos="N|ADJ|V" & func="<PRED|<OC|N<"]
[lema="homem|moço|rapaz|marido"] [pos="N|ADJ|V" & func="<PRED|<OC|N<"] "," @[pos="ADJ" & gen="M"]
[lema="homem|moço|rapaz|marido"] [pos="N|ADJ|V" & func="<PRED|<OC|N<"] "e" @[pos="ADJ" & gen="M"]
```

Primeiros resultados: o que vemos?



3.862 Predicações

O que vemos?

sentada perdida linda alta feia doente solteira casada boa bela pobre idosa velha livre moça rica capaz bonita filha loura amada alegre pálida feliz gorda viúva morena honesta formosa encantadora

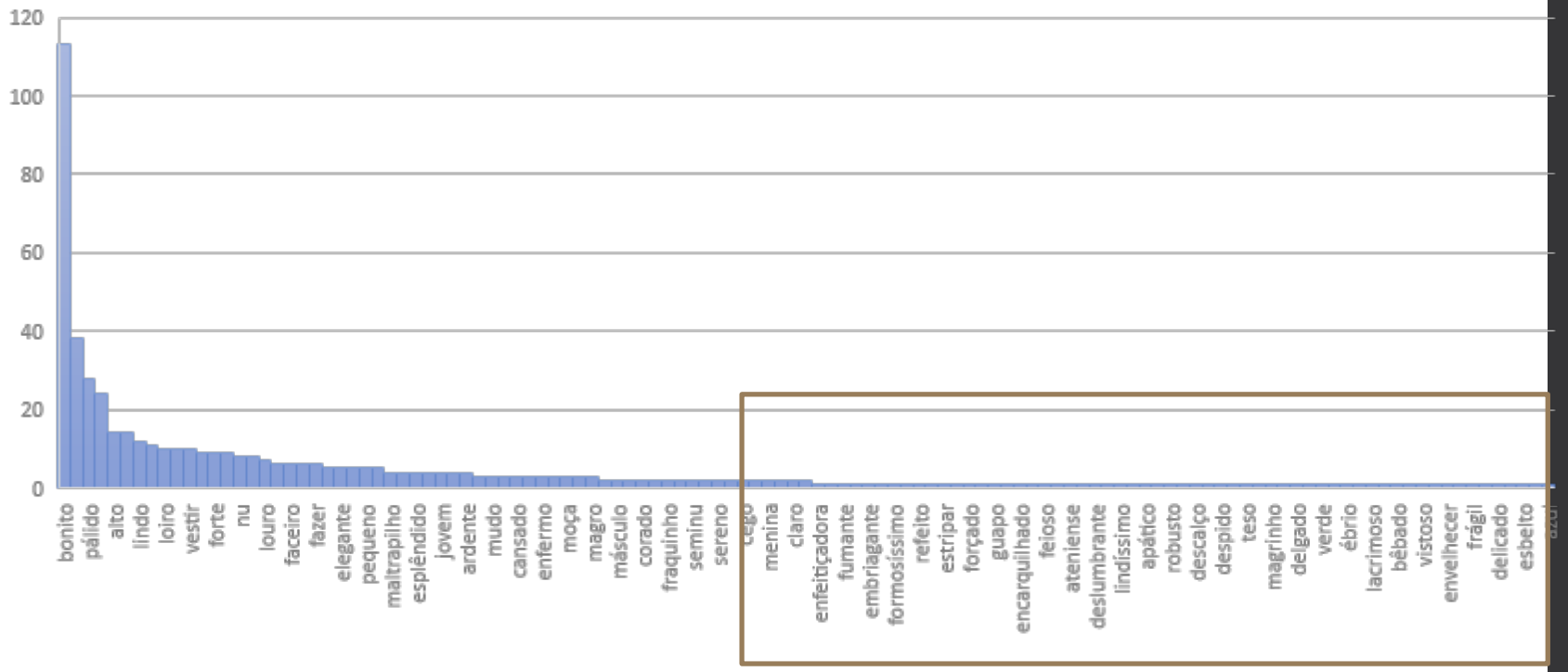
<500 predicações diferentes

alto inteligente feliz gordo pobre solteiro capaz moço cândido mau bom doente digno forte sério honrado robusto doido feito magro pálido público honesto alegre grave rico velho baixo bonito branco

< 800 predicações diferentes

Primeiros resultados

Cerca de 60% dos predicadores masculinos e femininos aparece uma única vez



Como generalizar?

Metodologia

2. Categorização: Criação de eixos para *dar sentido* à cauda longa.

Análise manual

APARÊNCIA / CORPO
CARÁTER
EMOÇÃO / AFETO
LUGAR SOCIAL

Classificação múltipla

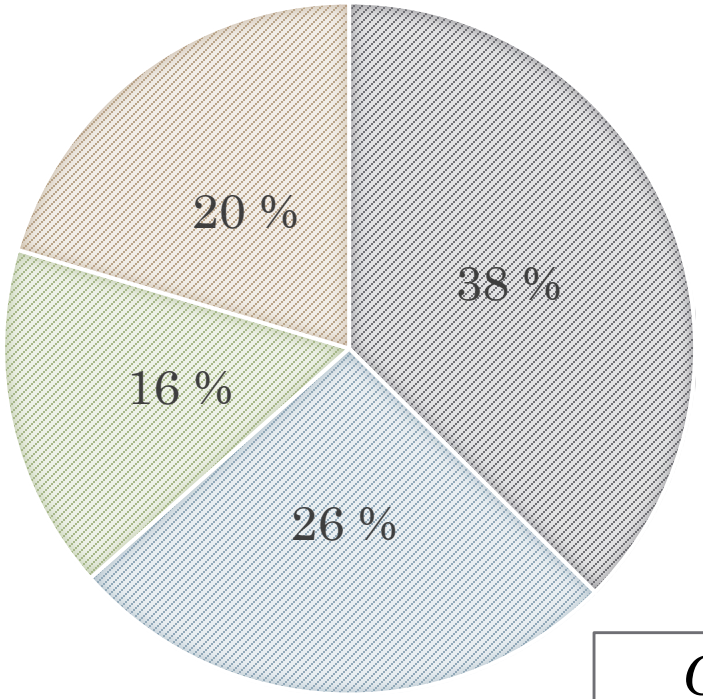
- "Nem os moços **valentes**[emo & caráter], nem os senhores respeitáveis, nem os jornalistas vão sequer à delegacia. " (*A Alma Encantadora das Ruas*)

- "Eugênio, **trêmulo**[físico & emo], confuso e de olhos no chão deixou cair sobre sua cabeça toda esta tremenda trovoada. " (*O Seminarista*)

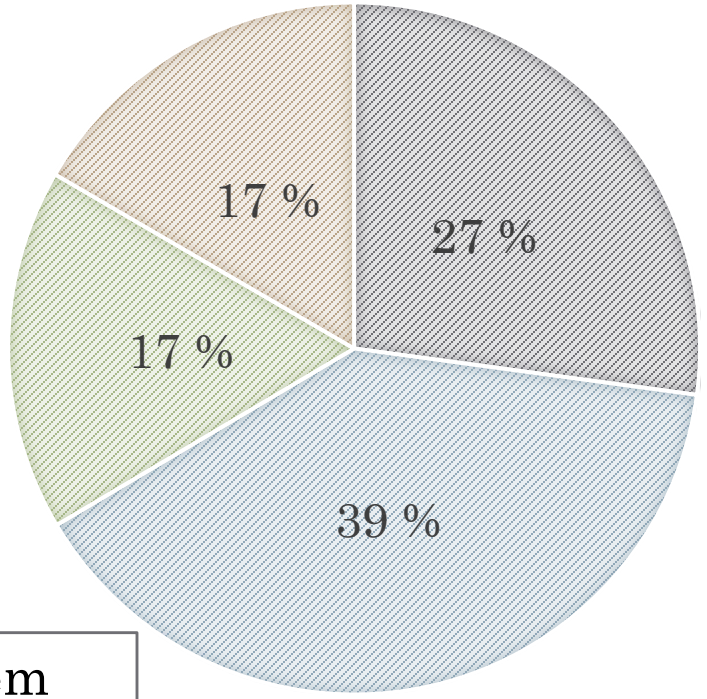
Metodologia

predicador	qtd	eixo
bonito	146	corpo
belo	66	corpo
amado	43	emo
casado	43	social
formoso	38	corpo
pálido	37	corpo
honesto	36	caráter
velho	32	corpo
alto	30	corpo
doente	25	corpo
rico	24	social
bom	24	caráter
idoso	22	corpo
capaz	22	caráter
solteiro	21	social
moreno	21	corpo
moço	21	corpo
pobre	21	social
sentado	20	erro
feliz	19	emo
lindo	18	corpo
encantador	17	corpo

O que vemos?



caráter
corpo
emo
social



caráter
corpo
emo
social

Caráter e *corpo* tem distribuição invertida

Emo e *social* têm distribuição parecida

O que vemos?

	Predicações masculinas	Predicações femininas
Emoção/afeto	Feliz, amigo, contente, alegre, valente*, atônito, amado, triste, espantado, pensativo, comovido, apaixonado, respeitado*, radiante, ansioso, trêmulo, entusiasmado*, aborrecido, indignado, orgulhoso, furioso*	Amada, feliz, alegre, amante*, triste, faceira*, trêmula, impaciente, amiga, assustada, agitada*, chorosa*, radiante, espantada, tranquila, nervosa, admirada, pensativa, risonha, coitada, ardente*
Caráter	Bom, sério, grande, honesto, digno, capaz, inteligente, mau, honrado*, doido, grave*, valente*, generoso, frio, vulgar, direito, calmo, ilustre*, livre, simples, hábil	Honesta, boa, perdida*, capaz, livre, vulgar, de verdade*, digna, impaciente*, fácil, indiferente, discreta, fresca, infame, inteligente, ingênua*, inocente, fria, pura, santa*, namoradeira*
Aparência/corpo	Alto, magro, bonito, pálido, gordo, forte, velho, (homem)feito, moço, branco, doente, baixo, elegante, robusto, vestido, morto, fraco, novo, maduro*, moreno	Bonita, bela, pálida, formosa*, alta, feia, linda*, velha, loira, morena, vestida, encantadora*, forte, moça, nua, doente, loura, gorda, faceira*, trêmula, (mulher)feita
Lugar social	Rico, pobre, solteiro, público*, amigo, casado, filho, político*, capitão*, feiticeiro, notável*, português*, sertanejo, primitivo*, importante*, nobre, ativo, desconhecido*, sozinho, abastado*, civilizado	Rica, casada, pobre, perdida, filha, viúva, solteira, mãe*, sozinha, amiga, distinta, empregada, guerreira*, feiticeira, virgem, suprema*, íntima*, irmã, cozinheira*, brasileira

Quadro 3 – Predicações masculinas e femininas, listadas por frequência. As palavras com * referem-se a predicações usadas

O que vemos?

	Predicações exclusivamente masculinas	Predicações exclusivamente femininas
emoção/afeto	Valente, respeitável, entusiasmado, furioso, terrível, rude, indiferente, deslumbrado, feroz, cruel, apressado, hediondo, corajoso, destemido, triunfante, abatido, maravilhado, ruim, desconfiado, pacato	Amante; faceira, agitada, chorosa, ardente, ciumenta, adorável, zangadíssima, corada, sôfrega, espavorida, afortunada, desgraçada, adorada, soberbíssima, brabinha, tiriricas, desaventurosa, agastada
caráter/personalidade	Mau, honrado, grave, valente, ilustre, extraordinário, ébrio, bobo, rude, sábio, perdido, seguro, franco, religioso, singular, polido, cirscunspeto, teimoso, leal, desgraçado	Perdida, de verdade, impaciente, ingênua, santa, namoradeira, obediente, ímpia, travessa, má, carinhosa, bondosa, insensível, angélica, morfética, frívola, trabalhadeira, rebelde, curiosa
aparência/corpo	Maduro, baixinho, barbado, sadio, limpo, galhardo, calvo, miúdo, rijo, quadragenário, encolhido, vesgo, míope, idoso, suado, tisnado, espadaúdo, bonitinho, macilento, preto	Formosa, linda, encantadora, faceira, maltrapilha, tísica, esplêndida, ardente, muda, grávida, cansada, moça, coxa, máscula, cardíaca, corada, fraquinha, sã, seminua, deliciosa
lugar social	Público, político, capitão, notável, português, primitivo, importante, desconhecido, abastado, célebre, ilustrado, útil, formado, oficial, negociante, poderoso, pa engenheiro	Mãe, livre, viúva, guerreira, suprema, íntima, cozinheira, noiva, esposa, donzela, núbil, troiana

Lugares públicos

Parecidos em quantidade; diferentes em qualidade

Lugares domésticos

O que ainda poderíamos ver?

CRUZAMENTO COM OUTRAS VARIÁVEIS

- Diferenças conforme período/tempo
- Diferenças conforme a escola literária
- Diferenças conforme os autores
- Diferenças conforme o gênero (M/F) dos autores
- Diferenças conforme o país dos autores

?

- Diferenças conforme *quem predica*
  - Personagens...
  - Narradores...

Outro exemplo – incompleto...
Discurso relatado (*reported speech*) como operador analítico

- Quem fala (mais)?
- O quanto fala?
- Fala sobre o quê?

([sema="dizer_relato.*"] [func="<SUBJ" & gen="M" & pos="PROP.* | PERS.*"] | [func="SUBJ.*" & gen="M" & pos="PROP.* | PERS.*"] [pos!="V"] {0,6} @[sema="dizer_relato.*"]) within s

O AC/DC já tem anotação semântica dos verbos de elocução :)

26 de outubro de 2019

Procura: **([sema="dizer_relato.*"] [func="<SUBJ" & gen="F" & pos="PROP.*|PERS.*"]|[func="SUBJ.*" & gen="F" & pos="PROP.*|PERS.*"] [pos!="V"] {0,6} @[sema="dizer_relato.*"]) within s**
Pedido de uma concordância em contexto
Corpo: OBras v. 7.12

814 ocorrências.

Número de ocorrências excessivo! Tente restringir a sua procura a menos de 500 casos.

Concordância

Procura: **([sema="dizer_relato.*"] [func="<SUBJ" & gen="F" & pos="PROP.*|PERS.*"]|[func="SUBJ.*" & gen="F" & pos="PROP.*|PERS.*"] [pos!="V"] {0,6} @[sema="dizer_relato.*"]) within s**.

Apresenta-se uma amostra aleatória de 500 das **814** ocorrências encontradas.

id="Histórias_sem_Data Prosa:conto MdA 1884 masc ": -- Nunca fui janeleira, **dizia ela**, e acho muito feio que uma moça viva com o sentido na rua .

id="Dona_Guidinha_do_Poço Prosa:romance MdOP 1891 naturalismo_regionalismo masc ": **Guida noticiara-lhe** que logo que o Silveira estivesse de volta mandava-o tomar conta das bestas e outros animais .

id="Casa,_não_casa Prosa:conto MdA 1975 masc ": -- Bem, **disse ela**, já sei que me despreza .

id="Iaiá_Garcia Prosa:romance MdA 1878 romantismo masc ": **Estela supunha** que o amor de Jorge, ao fim de tão longo período, estaria acabado de todo, como produto da primeira estação .

id="Memorial_de_Aires Prosa:romance MdA 1908 realismo masc ": como se falasse da morte do Barão de Santa-Pia e da situação da filha, **D. Cesária perguntou** se ela realmente não casava .

id="O_Sacrifício Prosa:romance FT 1879 romantismo_realismo masc ": **perguntou Virgínia**, tanto que por entre árvores e sombras reconheceu Maurícia .

id="Clara_dos_anjos Prosa:romance LB 1948 naturalismo_realismo masc ": -- Penso que sim -- **disse Clara**, e acrescentou: -- olhe, papai, não pude passar a limpo a música .

id="A_Escrava_Isaura Prosa:romance BG 1875 romantismo masc ": Leôncio! ... onde vais! -- **exclamou Malvina** precipitando-se para ele; mal, porém, havia ela chegado à porta, ouviu-se a explosão atroadora de um tiro .

Quem fala mais? Fala como? Sobre o quê?

814 casos

dizer	98
responder	58
falar	17
perguntar	13
dizer+ele	13
declarar	11
supor	10
acrescentar	9
insistir	9
responder+ele	6
suspeitar	5
explicar	5
perguntar+ele	5
prometer	5
negar	5
asseverar	4
gritar	4
dizer+eu	4
afirmar	4
contar	4
advertir	3
declarar+ele	3
teimar	3
admitir	3
confirmar	3
referir	2
aconselhar	2
propor	2
anunciar	2
concluir	2
repetir	2
noticiar+ele	1
alegar	1
acrescer	1
avisar	1
explicar+ele	1
recomendar	1
perguntar+eu	1
contar+ele	1
recordar	1
indicar	1
defender	1
gritar+ele	1
indagar	1
murmurar	1
destacar	1
replicar	1
jurar	1
recusar	1
falar+ele	1
chamar	1

Como "dar sentido" a essas listas?

Como depreender padrões?

dizer	201
responder	69
falar	43
declarar	38
dizer+ele	27
supor	20
afirmar	16
perguntar	16
jurar	16
insistir	14
contar	14
negar	9
concluir	8
acrescentar	8
gritar	7
suspeitar	6
anunciar	6
explicar	6
advertir	5
teimar	5
responder+ele	5
chamar	5
contestar	4
explicar+ele	4
indagar	4
perguntar+ele	4
considerar	4
recusar	4
opinar	3
recomendar	3
dizer+eles	3
aconselhar	3
propor	3
admitir	3
revelar	3
repetir	3
replicar	3
prometer	3
definir	3
afirmar+ele	2
alegar	2
declarar+se	2
referir	2
berrar	2
comunicar	2
informar	2
citar	2
dizer+eu	2
observar	2
exclamar	2
avisar+eles	1
avisar	1
asseverar	1
assegurar	1
advogar	1
cumprimentar	1
relatar	1
argumentar	1
concluir+se	1
justificar	1
apurar	1
noticiar	1
declarar+ele	1
perguntar+eu	1
proclamar	1
controverter+se	1
indicar	1
adiantar	1
destacar	1
completar	1
jurar+ele	1
confirmar	1
comentar	1
falar+ele	1

1733 casos

QUEM fala? *para além do gênero...*

- Caracterização dos personagens de Dickens por meio dos verbos de elocução

Ruano San Segundo, Pablo. 2016. "**A Corpus-Stylistic Approach to Dickens' Use of Speech Verbs: Beyond Mere Reporting**." Language and Literature 25 (2): 113-129.

Utilização de corpus para recuperar verbso de elocução e revelar como Dickens usa certos verbos para reportar a fala de certas personagens, contribuindo para a sua caracterização. A análise de 14 obras completes mostra que a prática não é um fenômeno isolado, mas um recurso estilístico importante em suas obras.

Cuidados - porque nossa língua é o Português...

Tabela 1. Distribuição de sujeitos ocultos por corpus

Corpus	Frases com sujeito oculto
Bosque-UD (v.2.4)	15.8%
DHBB	39.5%
Machado de Assis	28.42%

Desafios (para o português também): *Automatic Attribution of Quoted Speech in Literary Narrative* (Elson & McKeown, 2010)

Columbia Quoted Speech Attribution Corpus (2017)

AC/DC: QUEMDISSE (Freitas et al., 2016)

Quantos desses casos estão envolvidos em predicações?

Quantos desses casos referem-se a verbos de elocução?

Cláudia Freitas, Elvis de Souza e Luisa Rocha. *Quantificando e qualificando o sujeito oculto em Português*. Jornada de descrição do Português, Salvador, BA, 16-17 Outubro 2019.

Considerações finais (?)

MUITOS caminhos para explorar!

- Uma gramática das predicações humanas?
- Mas quem são os humanos (para além dos PROP)?
  - Humanos são personagens...
- Que categorias de análise são relevantes para enriquecer a leitura distante?
- Faz sentido pensar em categorias novas?
- O que já temos (para o português)?
- O que o AC/DC nos dá hoje?
- O que podemos fazer para deixar ainda melhor?
- O quão boa precisa ser a análise automática (morfológica, sintática, semântica) para nos oferecer resultados confiáveis?

Desafios do pré-processamento (ainda):
Do pdf ao txt (bem tratado), o caminho é longo...
: (

Referências

ELSON, D. & MCKEOWN K. 2010. Automatic attribution of quoted speech in literary narrative. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI'10) , p. 1013-1019.

FREITAS, C. Estudos linguísticos e Humanidades Digitais: corpus e descorporificação. *Gragoatá*, vol. 22, n. 44, p.1207-1227.

FREITAS, B., FREITAS, C. & SANTOS, D. "QUEMDISSE?: Reported speech in Portuguese". In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)

FREITAS, C., DE SOUZA, E. & ROCHA, L. *Quantificando e qualificando o sujeito oculto em Português*. Jornada de descrição do Português, Salvador, BA, 16-17 Outubro 2019.

SABURI COSTA, B. & FREITAS, C. "Verbos de Elocução em Português: um estudo descritivo com base em grandes corpora e motivado pela Linguística Computacional". Revista Fórum Linguístico, vol. 14, n. 3, p. 2266-2285, 2017.

PAIXÃO de SOUZA, M. C. P. A Filologia Digital em Língua Portuguesa: alguns caminhos. In: GONÇALVES e BANZA, Ana Paula Banza (Eds.). Património Textual e Humanidades Digitais: da antiga à nova Filologia. Évora: CIDEHUS, 2013. Disponível em <http://dspace.uevora.pt/rdpc/bitstream/10174/10468/1/e-book.pdf>. Acessado em: 30 out. 2017.

RUANO SAN SEGUNDO, Pablo. 2016. "A Corpus-Stylistic Approach to Dickens' Use of Speech Verbs: Beyond Mere Reporting." Language and Literature 25 (2): 113-129.

SANTOS, D., FREITAS, C. & BICK, E. OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in the public domain, OpenCor, Canela, RGS, Brasil, 24 de setembro de 2018.

SANTOS, D.; BICK, E. Providing Internet access to Portuguese corpora: the AC/DC project. In Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhauer (eds.), Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000) (Atenas, Grécia, 31 de Maio a 2 de Junho de 2000), pp. 205-210.

SANTOS, D. "Pesquisando corpos: Foi você que pediu um Corpo Ferreira?". Escola de Verão Belinda Maia (Edv 2009; Edv 2009) (FLUP, Porto, Portugal, 29 de Junho - 3 de Julho 2009)

Obrigada!

claudiafreitas@puc-rio.br