Recursos da Linguateca para humanidades digitais em contexto multicultural

Oficina de Verão, Universidade Politécnica de Macau

Diana Santos

d.s.m.santos@ilos.uio.no

1-2 de julho de 2025

Audiência e objetivo

- alunos de mestrado e doutoramento, bem como aos docentes da Faculdade de Línguas e Tradução
- temas de investigação dos doutorandos: literatura portuguesa, estudos interculturais, ensino de português como língua estrangeira e estudos de tradução e interpretação entre chinês e português, etc.
- Objetivo: dar a conhecer a Linguateca e como usar os seus recursos
- Especial interesse em sessões práticas

Recursos de disseminação já existentes

Ao longo dos anos, fui dando vários cursos para várias audiências diferentes, alguns deles até foram gravados e encontram-se disponíveis, ver a minha página de publicações,
`https://www.linguateca.pt/Diana/public.html`

Dez 2024	Usando o AC/DC com o corpo DisPR (grav.)
Abr 2021	Grandes quantidades de informação: um olhar crítico (grav.)
Dez 2021	Humanidades Digitais e História: algumas observações
Nov 2021	Explorando o CorTrad para pesquisa e(m) tradução (grav.)
Dez 2020	Processamento de corpos e literatura: formação no âmbito do BILLIG
Jun 2019	Apresentando os recursos da Linguateca
Set 2013	Para que serve um corpo: O que é, o que significa a sua anotação, exemplos de uso

Também houve duas Escolas de Verão de Linguateca, ambas no Porto, em 2006 e 2009

História da Linguateca

- Os primórdios remontam ao interesse de Mariano Gago na língua, e às discussões públicas em algumas áreas da Ciência no final dos anos 90
- O pequeno projeto "processamento computacional do português" (1998-2000) preparou as bases para a discussão sobre a área, e projetou a Linguateca (na altura chamada CdRLP)
- A Linguateca foi financiada pelas autoridades portuguesas até 2011, mas teve sempre como objetivo o português em todo o mundo
- Depois disso passou a ser um projeto sem financiamento, mas com apoio institucional da FCCN/FCT, que abrigam os servidores onde estão os recursos

Principais recursos

Durante a história da Linguateca muitos recursos foram criados, e a maioria estão disponíveis ainda na internete, sobretudo os que provêm das variadas atividade de avaliação conjunta que nós organizámos.

- Santos, Diana & Cláudia Freitas. "Avaliação conjunta em português". In Helena M. Caselli & Maria das Graças Volpe Nunes (eds.), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3ed. BPLN, 2024. Disponível em https://brasileiraspln.com/livro-pln/3a-edicao. Capítulo 15

Mas eu diria que, agora, os corpos, incluindo os corpos paralelos, e a sua anotação e exploração são o mais importante que a Linguateca (ainda) oferece. Por isso vou apresentar o AC/DC e o CorTrad.

Atividades presentes

- Anotação do domínio da comida e bebida em textos literários, no âmbito da leitura distante
- Criação do CulBras (subcorpo do CorTrad para culinária brasileira)

"Só na má literatura é que as personagens não comem", afirmou Alberto Manguel (n. 1948) em Dezembro do ano passado, numa entrevista ao Ípsilon (Público). `https://www.paginaum.pt/2022/04/11/um-divertimento-literario-cheio-de-fantasia-culinaria`

Apresentação do AC/DC

com exemplos de Macau, China e macaense.
`[lema="Japão|China|Indonésia|Índia|Macau"]`
E apontando para a referência de 2022: "A sintaxe do AC/DC: apresentação do CWB e das opções tomadas"
`https://www.linguateca.pt/Diana/download/instrACDC.pdf`
Melhorias: Mais campos semânticos, donde:

- a procura `[sema="cor"]` passa a `[sema=".*(cor|cor:.*)"]`
- a procura `[sema="medo"]` passa a `[sema=".*emo:medo.*"]`
- e [lema="vermelho" & sema!="cor"] passa a `[lema="vermelho" & sema!=".*(cor|cor:.*)"]`

Roberto Carneiro

- Procurar `[lema="Roberto=Carneiro"]`
- Roberto Carneiro como objeto (linguístico): [pos="V.*"] [pos!="V.*"]* [lema="Roberto=Carneiro" & func="<ACC"] within s
- Roberto Carneiro como sujeito (linguístico): [lema="Roberto=Carneiro" & func="SUBJ>"] [pos="VAUX.*"]* @[pos="V.*"] [: pos!="V.*" :] within s [pos="V.*"] [pos!="V.*"]* [lema="Roberto=Carneiro" & func="<SUBJ"] within s

`https://www.publico.pt/2014/11/23/portugal/noticia/retrato-de-uma-familia-feliz-1676884`
`https://arquivos.rtp.pt/conteudos/roberto-carneiro-2/`

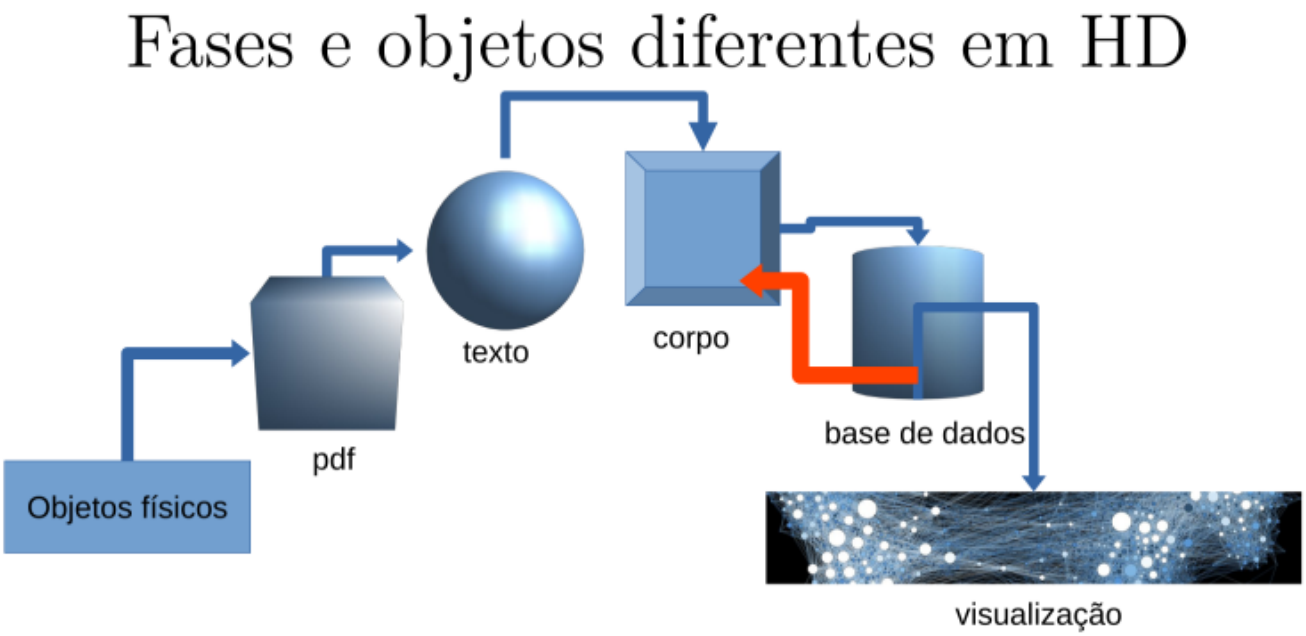
Exercícios de teste/treino, para amanhã

- Qual a região/continente mais falada na literatura? Em que contextos?
- Quais as palavras modificadas pelos adjetivos *europeu* e *asiático*?
- Quem fala mais da família no Museu da Pessoa, os homens ou as mulheres?
- Veja a distribuição de comer, beber e fumar no CONDIVport, por tema e por variante.

Apresentação da Literateca

- A ideia de leitura distante, de Moretti
- A ideia de humanidades digitais: usar métodos digitais para responder a perguntas de Letras
- A criação de corpos de obras literárias em português, para pesquisa na internete
- A necessidade de técnicas de vizualização mais complexas, usando métodos estatísticos e R – ver as lições no *Programming Historian*

A filosofia da Literateca



Apresentação da Literateca, segunda parte

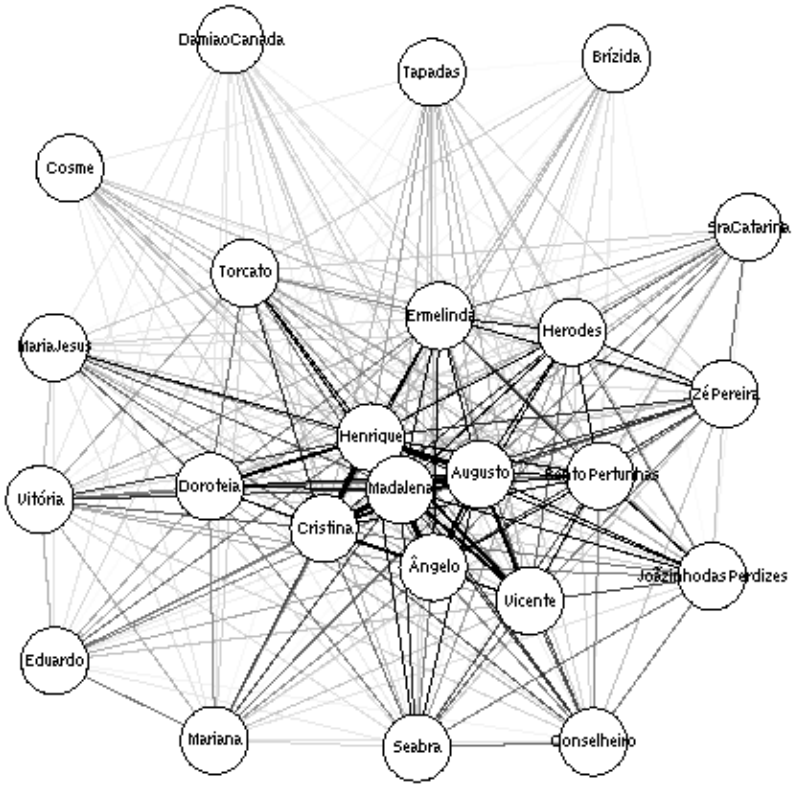
- Redes de personagens
- Escola literária
- Caracterização de pessoas
- Identificação de personagens, o DIP
- Comida e bebida na literatura

Redes de personagens

Embora exista grande conhecimento informático e matemático sobre grafos, as redes nas humanidades digitais são em geral usadas sobretudo para visualização

- Marcar nomes próprios como personagens no corpo
- Contar quantas vezes as personagens ocorrem, numa janela deslizante de 3000 unidades (com 500 de sobreposição)
- Contar quantas vezes co-ocorrem
- Usar R para desenhar a rede, usando o logaritmo do numero de co-ocorrências para a grossura dos arcos

Redes de personagens: exemplo

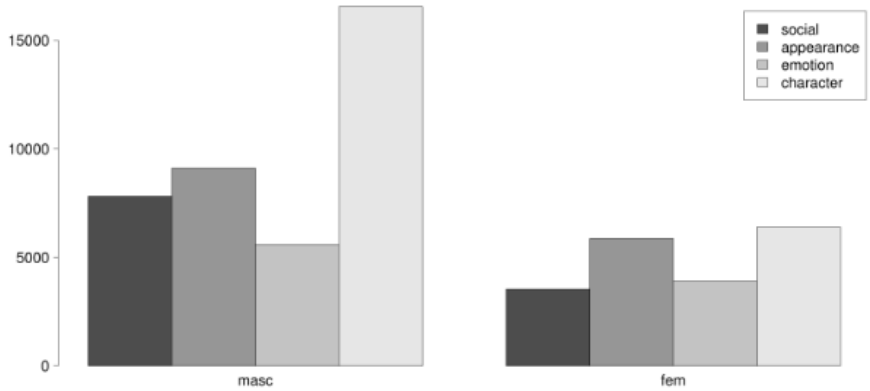


Escola literária

- Com base na "redução" de centenas de obras a características linguísticas que se podem contar (estilo?), e depois reduzindo o número de dimensões a duas, visualizam-se regiões num espaço virtual. É possível identificar regiões correspondentes a escolas literárias?
- Com base na escolha das palavras (não gramaticais) mais frequentes num conjunto de obras românticas e realistas, é possível classificar novas obras automaticamente?

Caracterização de pessoas (na literatura)

- Com base na classificação dos adjetivos (e nomes) associados a palavras referentes a pessoas em quatro grandes áreas: social, aparência, emocional, caráter. Freitas & Santos (2023)

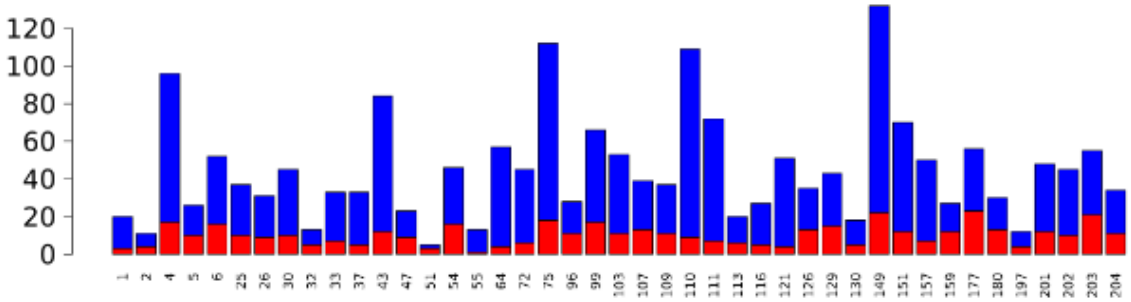


- E depois, o estudo das palavras preferidas por género...
  - *bonita, formosa, bela, linda, encantadora, feia...*
  - *velho, jovem, robusto, grande, baixo...*

Identificação de personagens, o DIP

Para um conjunto de obras literárias em português, identificar (Santos et al. 2023)

- as personagens (com os seus variados nomes)
- o seu género
- a sua profissão ou profissões
- as relações familiares com outras personagens



Comida e bebida na literatura (em apreciação)

Projeto contrastivo (literaturas portuguesa, brasileira, italiana e norueguesa) baseado em corpos

- léxico da comida e bebida nestas literaturas
- usos da expressões de comida
- questões pragmáticas associadas a comida e bebida
- descrição de atividades associadas a comida e bebida: refeições, preparações, etc.
- usos (literários) do emprego destas atividades nas obras literárias

Pressuposto/primeira fase: anotação e desambiguação de palavras de comida e bebida, e de (algumas) refeições.

Apresentação do CulBras

- Parte do CorTrad, um corpo paralelo português inglês criado em colaboração com Stella Tagnin e Elisa Teixeira, da USP. Ver Santos & Tagnin (2021)
- Culinária brasileira vs. culinária no Brasil: Este ano (2024-2025) aumentado com mais um subcorpo criado por Rozane Rebechi, sobre culinária brasileira traduzida para o inglês americano, a que chamámos CulBras.
- Corpo estruturado: título da receita, lista de ingredientes, modo de fazer, comentários, legendas de fotos
- Estudo da tradução: tradução cultural, e diferentes formas de resolver um "mesmo" desafio

`https://www.linguateca.pt/CorTrad/`

Apresentação do CulBras

Expressão de busca: **<t> []* "arroz" %c []* </t>**
Resultado escolhido: **concordância em contexto**
Corpus pesquisado: **originais** (versão 2.4)

30 ocorrências.

Voltar
Nova pesquisa

Original	Tradução publicada
Arroz de cuxá	Cuxá rice
Arroz de hauçá	Hauçá rice
Arroz com guariroba	Rice with guariroba
Arroz com pequi	Rice with pequi
Arroz de carreteiro	Carreteiro rice
Bolinho de arroz	Rice cupcakes
Arroz de forno	Rice in the oven