Processamento de corpos e literatura

Uma formação no âmbito do BILLIG

Diana Santos

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

9 e 10 de dezembro de 2020

Processamento de corpos e literatura

- O que é um corpo (linguístico)
- Linguística empírica, aplicações da linguística com corpos
- A Linguateca, o AC/DC, e a Literateca
- A ação europeia COST de leitura distante
- O BILLIG

O que é um corpo (linguístico)

Aplicado à literatura:

- o Vercial: clássicos portugueses
- o OBras: obras brasileiras
- o NOBRE: obras portuguesas não canónicas (ou não Vercial)
- peças francesas dos séculos XVII e XVIII
- literatura traduzida entre duas línguas (COMPARA, PANTERA, CorTrad...)

Razões de compilar (e usar) um corpo eletrónico

- Ninguém tem acesso / conhece a língua inteira (o que é a língua inteira?)
- Um interlocutor fantástico para perguntar repetidamente
- Permite comparar diversos registos, classes sociais, faixas etárias, épocas, regiões, ...
- Permite obter informação quantitativa sobre a língua, modelos que permitem desenvolver sistemas que façam tarefas sobre textos
- Permite sistematizar uma área de conhecimento ou aplicação

Perguntas sobre a literatura

- A literatura é escrita com a língua. Muitas (?) perguntas literárias são perguntas sobre a língua dos textos, por vezes chamada "estilo", ou sobre a língua de certos grupos de textos (géneros). Exemplos:
  - que tipo de adjetivos são usados
  - quantas personagens e como são descritas
  - tipo do narrador
  - locais e ambientes
  - intertextualidade e interculturalidade
- A história da literatura pode/deve ser feita com todos (muitos d) os textos que constituiram a literatura ao longo dos séculos (uma das ideias da leitura distante)

Técnicas de leitura distante

Eu diria que há dois tipos de técnicas

- as baseadas apenas nas palavras, como os modelos de tópicos
- as baseadas na análise linguística: além das palavras, incluem no modelo outras características provenientes da análise prévia

Em geral ambas usam técnicas de visualização para dar conta de um universo complexo. Essas técnicas não são necessariamente transparentes, e há muita gente que não se sente segura com essa nova linguagem.

A Linguateca

- Iniciada em 1998 pela mão do ministro José Mariano Gago
- Formulada como um centro de recursos para o processamento computacional da língua portuguesa
- Modelo IRA: informação, recursos e avaliação
- Em relação aos recursos: públicos e fáceis de utilizar

Recursos corpóreos

- Criação de corpos de raiz, que foram marcos no processamento da língua portuguesa, como o CETEMPúblico e o COMPARA
- Tratar (computacionalmente) a língua toda (todas as variantes)
- Disseminar e enriquecer os corpos de outrem (a maioria dos corpos servidos pelo AC/DC foram criados por outros investigadores)
- Desenvolver estudos e ensinar a usar os corpos, desenvolvendo um ambiente informático apropriado

A ação COST "Distant reading for literary history"

Distant Reading

This action will affect the way scholars in the Humanities do research, but also the way institutions like libraries will make their holdings available to researchers in the future

- `http://www.cost.eu/COST_Actions/ca/CA16204`
- `https://www.distant-reading.net/`
- `https://distantreading.github.io/`

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

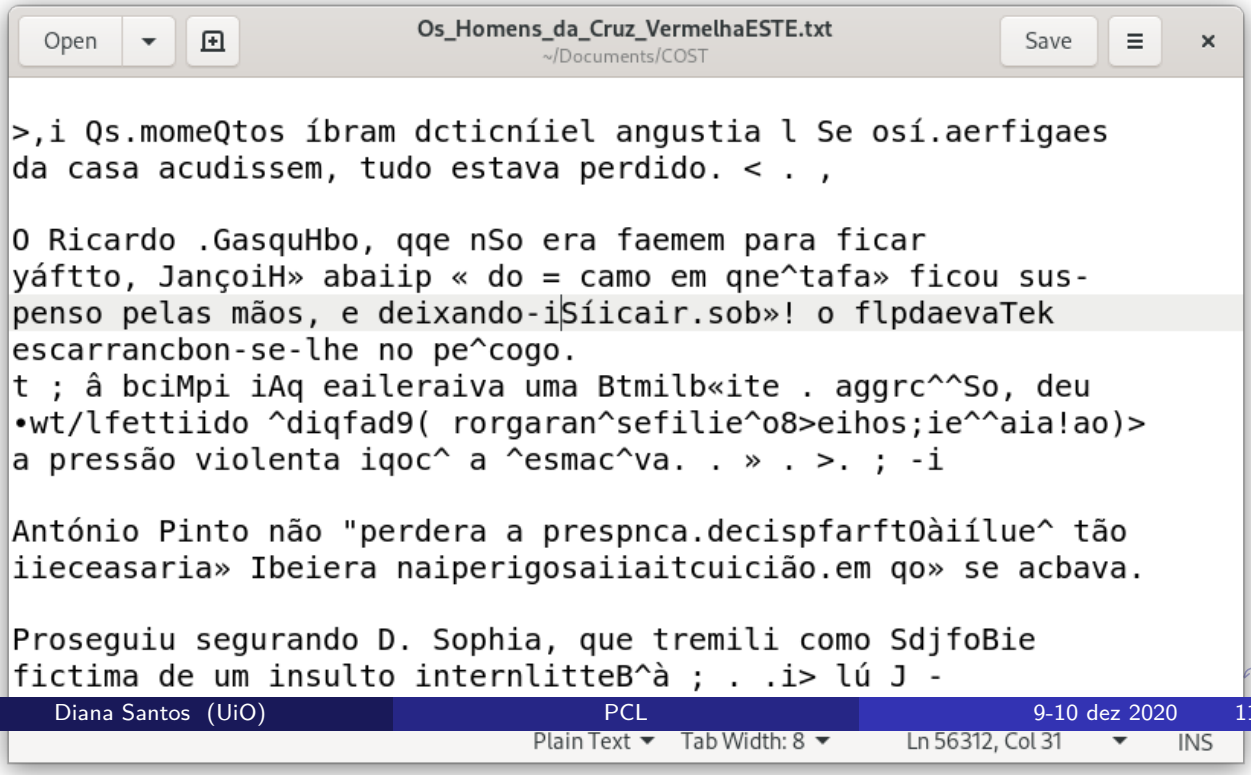
Mais sobre a ação COST

Columbano Bordalo Pinheiro (1857-1929):
O Grupo do Leão, 1885

- Duração: nov. 2017 out. 2021
- Três grupos:
  - WG1 construção da coleção ELTeC;
  - WG2 ferramentas computacionais;
  - WG3 análise literária
- A coleção ELTec-por e as outras coleções: `https://distantreading.github.io/ELTeC/`
- A coleção ELTeC-por está dentro da Literateca, e é possível obtê-la escolhendo o filtro (`costcanon="cost.*"`)

Ação COST: exemplo de problemas

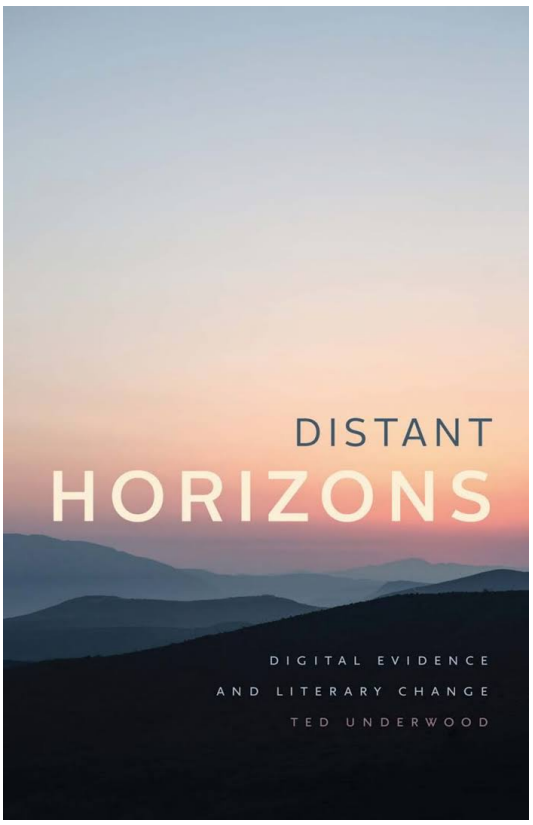
- escolha de obras: qual a população? critérios "pouco naturais", diferenças entre as línguas
- digitalização deficiente: um pesadelo!



Ação COST: resultados

- 100 obras, lançadas a 13 de novembro de 2020 (2.a publicação oficial)
- 4.a em termos de pontuação (E5C), uma medida que entra em conta com todos os critérios, depois da francesa, inglesa, e húngara.
- apenas 5 coleções têm 100 obras (as 3 anteriores e a eslovena)
- descrição quantitativa em Santos, Bick e Wlodek, `https://www.linguateca.pt/Diana/download/SantosBickWlodek.pdf`
- Um terço são romances históricos!
- Muito difícil de obter obras com mais de 100.000 palavras. Neste momento apenas temos 18, no mínimo deveriam ser 20.
- no total temos 118 obras no formato ELTeC

Leitura distante



- investigação em grande escala, quantitativa
- não necessariamente para salvar os esquecidos (moralismo de Moretti)
- outra forma de olhar: lentes que abarcam maiores períodos

  Can distant readers write quantitative literary history that is nevertheless detailed enough, streamlined enough, and lively enough to interest a wide range of readers? If we can't, then no argument will save us: what we are doing may be important, but it will belong in the social sciences. (Underwood, 2019, p. xxii)

Leitura distante segundo Underwood, cont.

- leitura distante traz novos objetos de estudo: padrões ainda sem nome
- exemplo: aumento da referência a cores durante o século XIX –> passagem de contar para mostrar ("showing to telling")? E então?
- partamos antes de temas literários, obtenhamos uma hipótese, e tentemos testá-la. Qual a diferença entre uma obra literária e uma biografia? Um contínuo entre obras imaginárias e reais. Fig. p. 23, 3832 obras

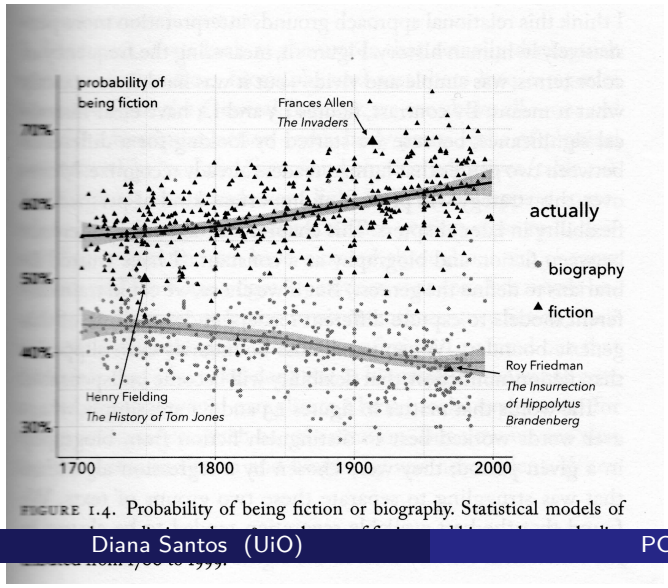


FIGURE 1.4. Probability of being fiction or biography. Statistical models of

Exemplos de estudos em português

- Atirar o barro à parede: será que alguns indicadores sintáticos e/ou semânticos podem agrupar os textos em escolas literárias?
- Será que uma análise de tópicos permite distinguir o romantismo do realismo/naturalismo?
- Existem diferenças entre a literaturas portuguesa e brasileira no que se refere a roupa?
- Como é que as personagens masculinas e femininas são descritas na literatura, considerando os eixos social, aparência e caráter?
- Como é que o Natal é descrito nos romances do período COST?
- Quais as profissões mais consideradas/mencionadas na coleção ELTeC?
- Quais são as obras "espaciáveis" (Jörn Seeman, apresentação dos atlas literários) e quais os autores mais geográficos?

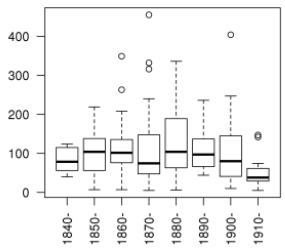
Natal na coleção COST

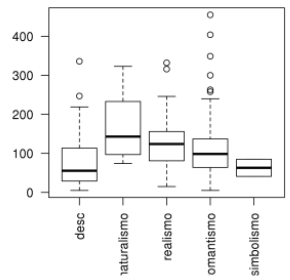
- Apenas 17 obras falam do Natal com 56 ocorrências, porque a que mais menciona Natal (29), *O agonizar de uma dinastia*, refere-se à província do Brasil.
- Apenas uma tem o Natal como tema de um dos episódios, *A Morgadinha dos Canaviais* (24 menções)
- A terceira, *Os Pobres*, usa-a para invetivar o *Natal dos pobres!*
- Mais duas obras têm episódios passados na noite de Natal: *Um duelo nas sombras, ou D. Francisco Manuel de Melo (1630)* (3) e *El-rei dinheiro* (2).
- Várias menções referem-se ao calendário e não à festa
  - oito libras que lhe emprestou pelo Natal
  - aquilo fora crescendo desde o Natal
  - teias de linho trazidas pelo Queiroz do Douro, no Natal
  - já passado o Natal

E da Páscoa? Ainda menos, 19 ocorrências em 15 obras. E do Carnaval? 38 em 10 obras (2 com cenas nesse período).

Exemplos de estudos literários na Literateca

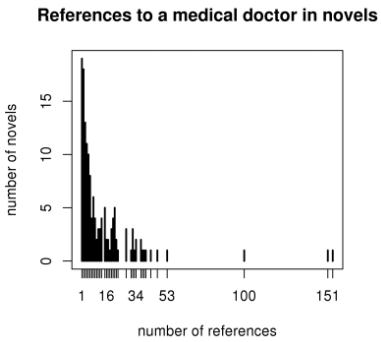
O papel do médico: https://dls.hypotheses.org/952, https://www.linguateca.pt/Diana/download/DRHealth.pdf





- generic doctor
- specific doctor
- character
- charlatan

There are 21 novels where there are more than 25 mentions to a doctor (*médico* or *facultativo*), out of the 151 where at least it is mentioned once, out of 181 total (83%).



Exemplos de estudos literários na Literateca

Personagens: https://www.linguateca.pt/Diana/download/STIL2019_SantosFreitas.pdf

REDES DE PERSONAGENS

As Figuras 2, 3 e 4 são redes de personagens obtidas após a anotação das obras, usando uma janela deslizante de 3000 palavras (sobreposição de 500).

Figura 2 – Rede associada a *Dom Casmurro*.

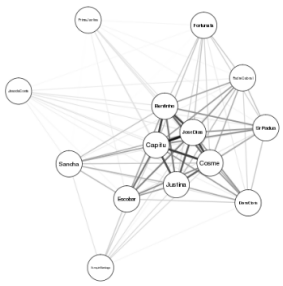


Figura 3 – Rede associada a *A Morgadinha dos Canaviais*.

Figura 4 – Rede associada a *Úrsula*.

PERSONAGENS AO LONGO DO TEMPO

As Figuras 5, 6 e 7 e são exemplos de figuras mostrando a presença dos personagens ao longo do enredo.

Figura 5 – Ao longo de *Dom Casmurro*.

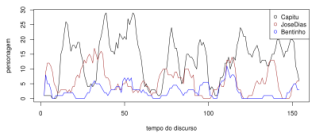


Figura 6 – Ao longo d'*A Morgadinha dos Canaviais*.

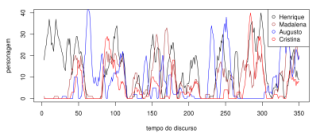
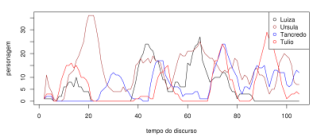
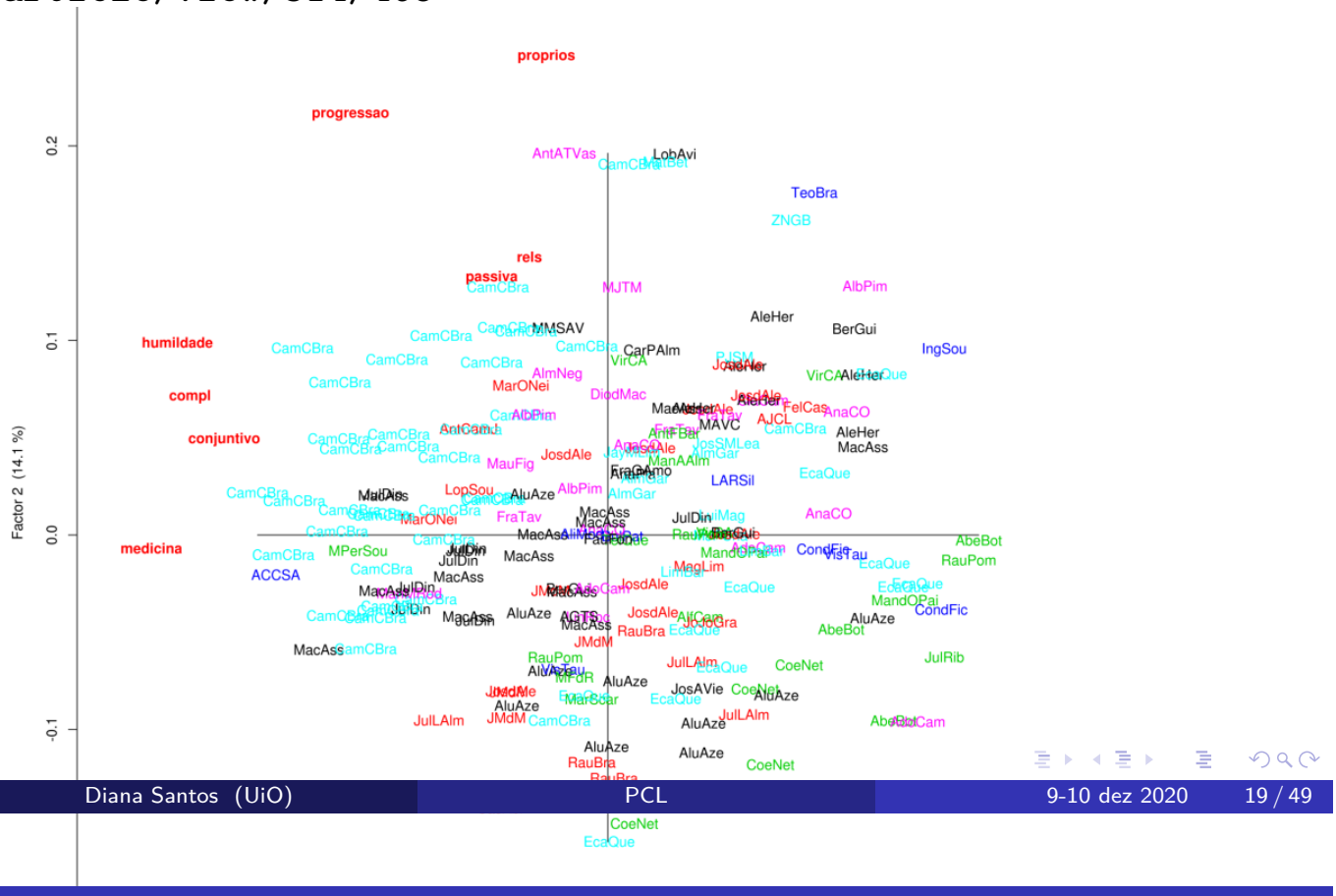


Figura 7 – Ao longo de *Úrsula*.



Exemplos de estudos literários na Literateca

Escolas: `https://linguamatica.com/index.php/linguamatica/article/view/314/465`



Exemplos de estudos literários na Literateca

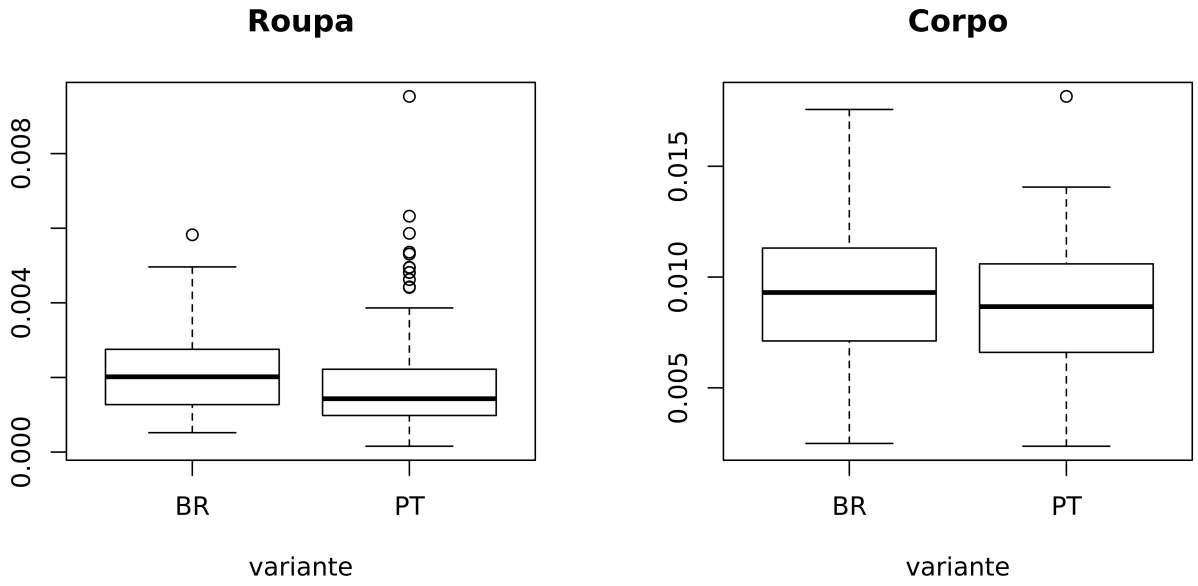
Modelos de tópicos:

Figura 12: Palavras que constituem o tópico 13

Figura 13: Palavras que constituem o tópico 90

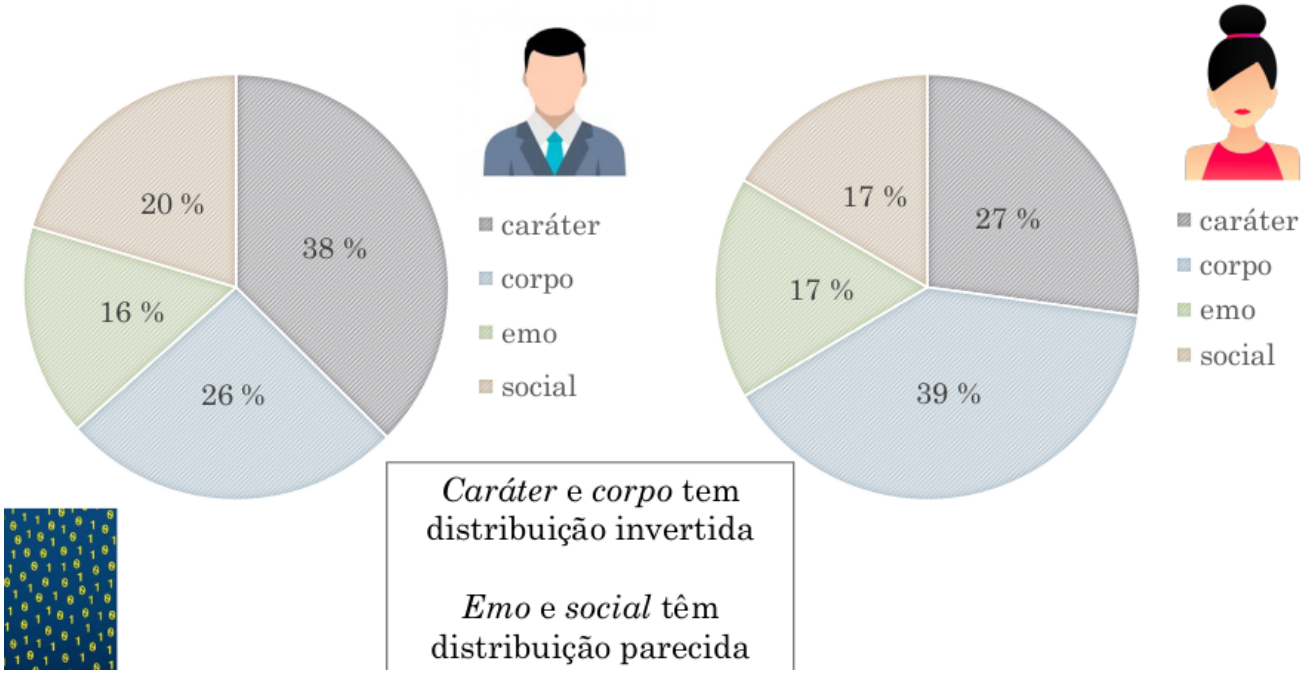
Exemplos de estudos literários na Literateca

Roupa: `https://www.linguateca.pt/Diana/download/Santos2021TradTerm.pdf`



Exemplos de estudos literários na Literateca

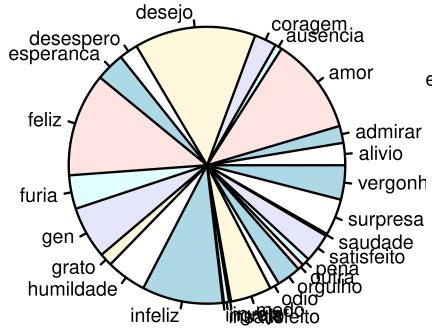
Caracterização de personagens: `http://www.linguateca.pt/ELD/aprClaudiaFreitasELD.pdf`



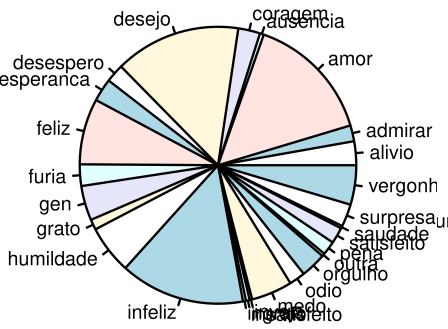
Exemplos de estudos literários na Literateca

Emoções: `https://www.linguateca.pt/Diana/download/posterSantosetal2020.pdf`

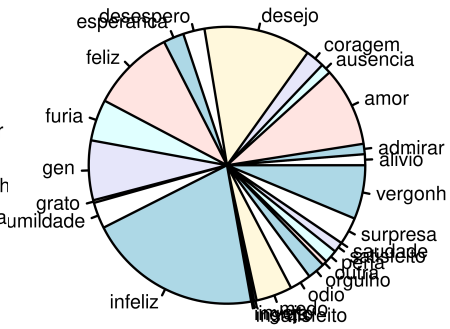
Júlio Dinis



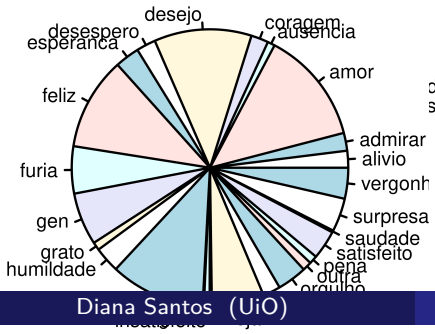
Camilo Castelo Branco



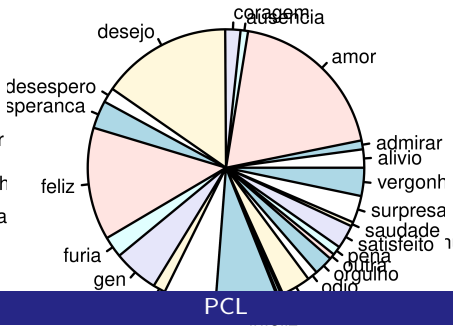
Raul Brandão



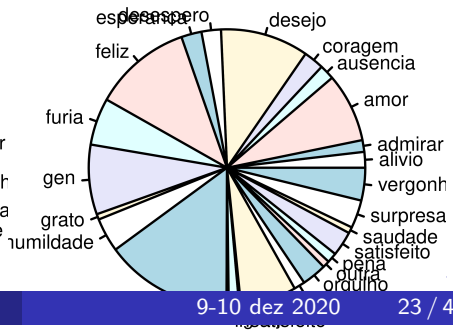
Eça de Queirós



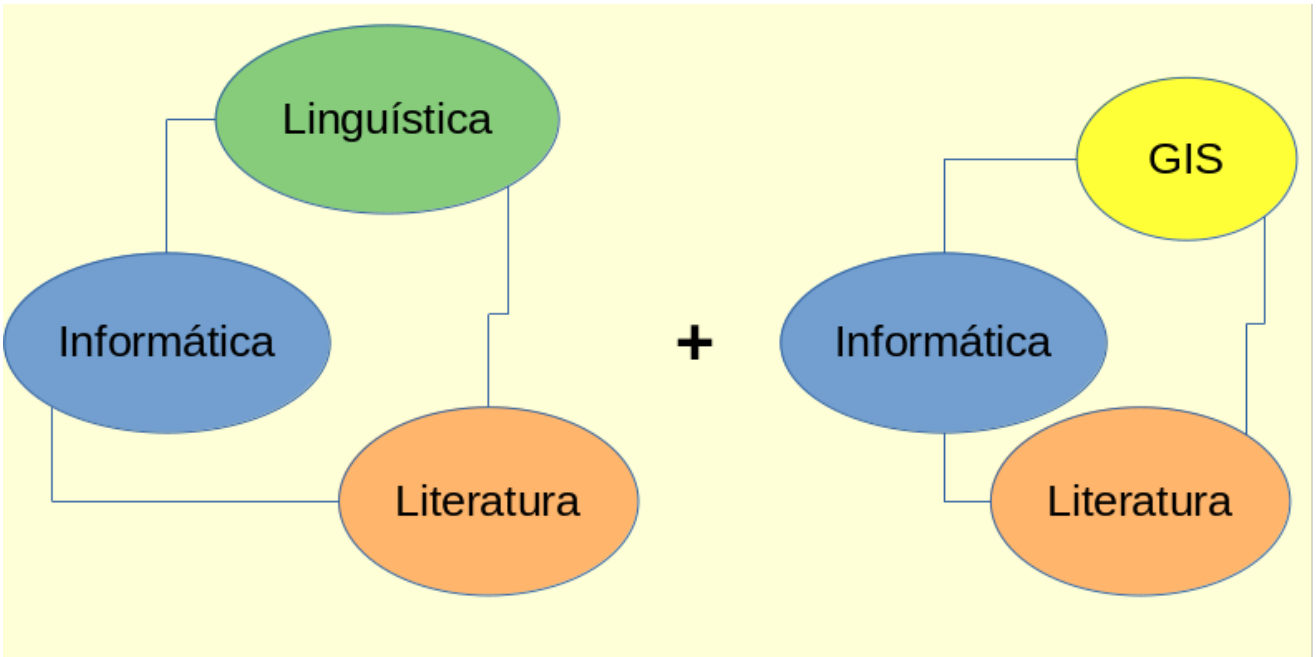
Machado de Assis



Coelho Neto



O BILLIG



Lugares na literatura

Porque é que faria sentido investigarmos a localização?

- todos os enredos acontecem em algum lugar
- a localização faz parte da caracterização narratológica de uma obra
- existe localização externa (associada à realidade) e interna (associada à ficção)
- para leitura distante, poderia ser interessante saber onde se passam as histórias
- para leitura distante, poderia ser interessante saber como as histórias caracterizam os lugares

O que é uma localização, linguisticamente?

- A resposta ingénua: os nomes de lugares
- A segunda resposta ingénua: o conteúdo dos "complementos circunstanciais de lugar"
- Olhando para a realidade linguística:
  - os nomes de lugar são usados de muitas maneiras
  - os complementos circunstanciais de lugar são muitas vezes metafóricos
  - as designações de lugar, mesmo quando se referem a lugares, são vagas, e têm vários níveis
  - os lugares não são os mesmos em tempos diferentes

Exemplos de "lugares" na Literateca: *Lisboa*

- par=CCB-A-Brasileira-de-Prazins-59: como as minhas mãos; é o vice-rei nas províncias do norte... o nosso bom padre Luís de Sousa, que pelos modos está nomeado patriarca de Lisboa...
- par=AntATVas-A-ermida-de-Castromino-1349: Nenhum dos viajantes recebera noticias de Lisboa .
- par=CCB-A-Caveira-da-Mártir-188: À herança de D. Antónia, Joaquina Xavier concorreram três famílias de Lisboa, Évora e Tavira que se apelidavam Nobres .
- par=LuiMag-O-Brasileiro-Soares-1: O abade dissertava gravemente sobre os caminhos de ferro e suas vantagens, relembrando as antigas jornadas do seu tempo, a cavalo ou de liteira, quando para ir do Porto a Lisboa era preciso fazer testamento .
- par=AntCJ-A-Ala-dos-Namorados-7954: – E como Lisboa se não entregava, estou a adivinhar que maiores foram para ela as antipatias da corte .
- par=AH-História-de-Portugal-IV-826: Livro dos Pregos, f. 3, no Cartório da Câmara Municipal de Lisboa .

Exemplos da classificação de "lugares" na Literateca

```
[lema="Espanha"] » [sema="Local:pais"]
[lema="Nova=Espanha"] » [sema="Local:territorio"]
[lema="Carlos=II=de=Espanha"] » [sema="Pessoa:hist"]
[lema="Das=Cousas=de=Espanha"] » [sema="Obra"]
[lema="El-Rei=de=Espanha"] » [sema="Pessoa"]
[lema="embaixador=de=Espanha"] » [sema="Pessoa"]
[lema="Fernando=de=Espanha"] » [sema="Pessoa:hist"]
[lema="Filipe=II=de=Espanha"] » [sema="Pessoa:hist"]
[lema="infanta=de=Espanha"] » [sema="Pessoa:hist"]
[lema="rei=de=Espanha"] » [sema="Pessoa:hist"]
[lema="Viagem=a=Espanha"] » [sema="Obra"]
[lema="Guerra=de=Espanha"] » [sema="Evento"]
```

Algo mais discutível

Espanha fica marcada com o tipo de lugar do sintagma nominal a que pertence:

```
a:[word="cadeias"] b:[word="em"] c:[lema="Espanha"] »
c:[sema="Local:organizado"]

a:[word="teatro"] b:[word="de"] c:[lema="Espanha"] »
c:[sema="Local:organizado"]
a:[word="templos"] b:[word="da"] c:[lema="Espanha"] »
c:[sema="Local:relig"]
a:[word="tribunal"] b:[word="em"] c:[lema="Espanha"] »
c:[sema="Local:organizado"]
```

Ver `https://www.linguateca.pt/Gramateca/Viagem.html`

Para ter uma visão global

Além de regras positivas, como as anteriores, também é preciso implementar regras negativas:

```
a:[lema="animado"] b:[word="pelo"] c:[lema="Colares"] » c:[sema="0

a:[lema="cálice"] b:[word="de"] c:[lema="Bordéus"] » c:[sema="0"]

a:[lema="Acre"] b:[word="beijo"] » a:[sema="0"]
```

e regras de correção da atomização (segmentação)

```
a:[lema="viscondessa"] b:[word="de"] c:[lema="Vila=Seca"] d:[lema=
» a:[lema="viscondessa=de=Vila=Seca"  pos="PROP"] b:[lema="viscond
pos="PROP"] c:[lema="viscondessa=de=Vila=Seca"] d:[lema="viscondes

a:[lema="Viso-Rei"] b:[word="da"] c:[word="Índia"] » a:[lema="Viso
pos="PROP"] b:[lema="Viso-Rei=da=Índia"  pos="PROP"] c:[lema="Viso
```

Resultados para *Espanha*

```
Local:pais 2061
Pessoa:hist 43
Pessoa 15
0 13
Obra 3
Local:territorio 2
```

e algum lixo...

Para dar uma ideia do trabalho envolvido

Neste momento (2 de dezembro de 2020, Literateca v. 5.15, 32,8 milhões de palavras),

- No Vercial há 2039 regras positivas para locais; 247 regras negativas; e 2160 regras de correção.
- No NOBRE há 1753 regras positivas para locais; 17 regras negativas; e 176 regras de correção.
- No OBras há 186 regras positivas para locais (mas sem subclassificação); 55 regras negativas, e 602 regras de correção.
- Em toda a Literateca há 78.981 casos marcados como Local, correspondendo a 1782 lemas diferentes.
- Ainda há 158.717 casos de nomes próprios anotados pelo PALAVRAS como possível Local (civ ou top), correspondentes a 20084 lemas diferentes, que é preciso rever.

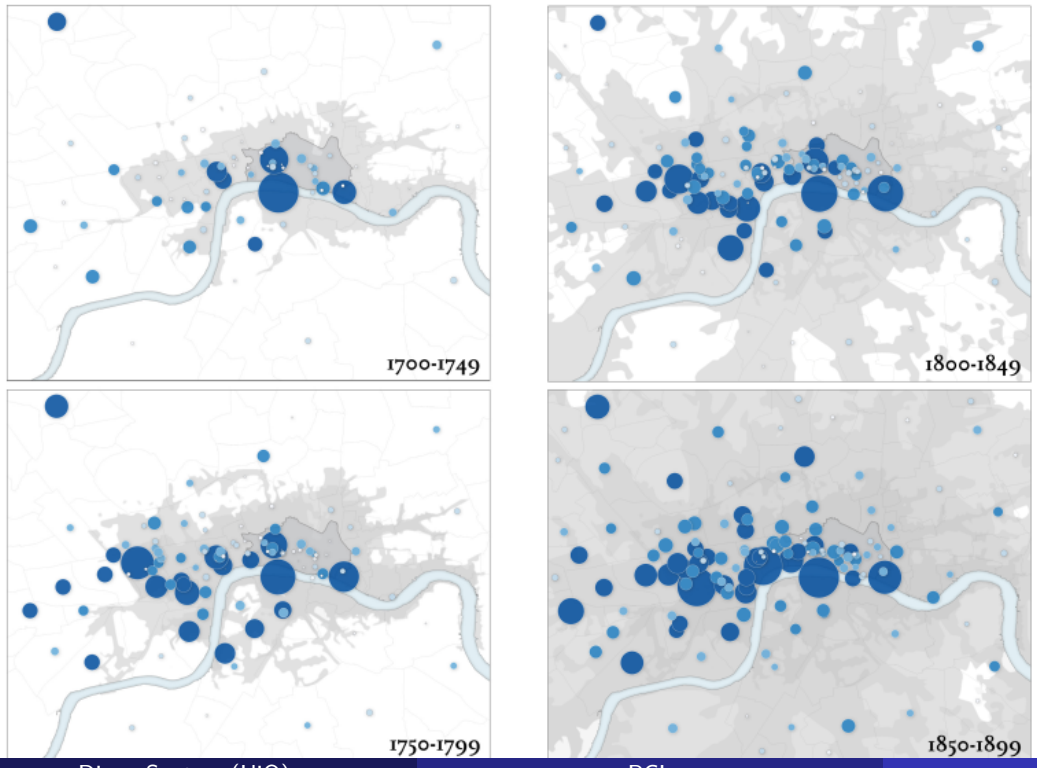
Outras formas de olhar para os lugares na literatura

- Onde é que se passam os eventos mais significativos de um romance? No campo ou na cidade?
- Quais são os locais que se mencionam mais numa aldeia? E numa cidade?
- Que locais aparecem na maior parte das obras?
- Que partes da casa são mencionadas?
- Como é que se viaja? E quanto tempo levam as viagens?
- Como é que um dado lugar é apresentado por uma literatura?

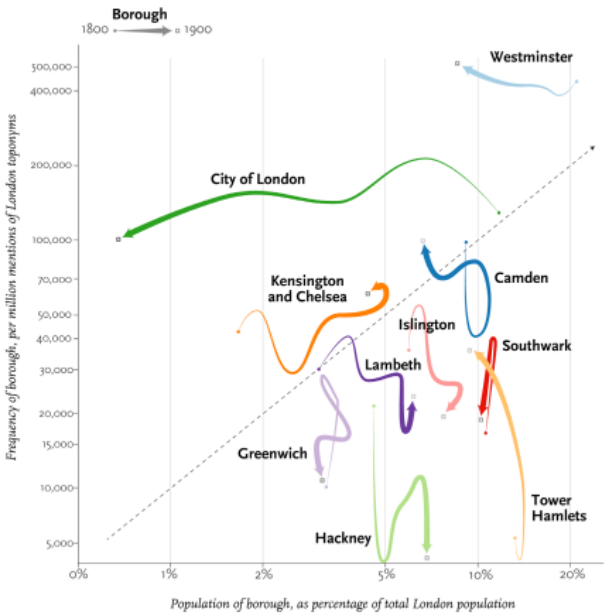
As emoções de Londres

Heuser, Ryan, Franco Moretti & Eric Steiner. "The emotions of London". *Literary Lab Pamphlet* 13, October 2016.

Figure 3.2 The stability of fictional London, 1700-1900



As emoções de Londres, cont. 2



- using a corpus of about 5,000 English novels published between 1700 and 1900: 304 for the period 1700-49, 1,079 for 1750-99, 1,290 for 1800-49, and 2,189 for 1850-99.
- comparação entre as populações reais dos bairros e as vezes que esses bairros são mencionados

As emoções de Londres, cont. 3

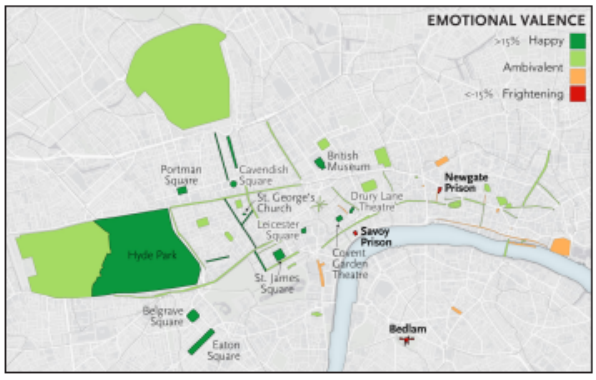


Figure 5.1 The emotions of London, 1700-1900

- Apenas duas emoções: *fear* (medo) e *happiness* (felicidade)
- "The most striking result of this map, though, was that so many passages turned out to be neither happy nor frightening."

"A map dominated by emotional neutrality. Did this mean that London novels avoided emotions? Not quite (though one does wonder what Paris novels would show)."

Inspiração do artigo anterior

- Quais as cidades mais mencionadas na literatura lusófona?
- É preciso ligar os locais/bairros/ruas à cidade...
- Qual a densidade de locais ao longo do tempo?
- Quais os romances de Lisboa, ou os romances do Rio, ou os romances do Porto, ou os romances de S. Luís do Maranhão? São categorias que façam sentido?

Outras ligações entre SIGs e literatura...

Ainda mais arbitrárias

- Muito do importante nas humanidades digitais são os métodos e a interdisciplinaridade
- porque não fazer mapas que não usam geografia?
- primeira ideia: mapa de uma casa, e quais os romances que falam mais em... quartos, cozinhas, salas, entradas, jardins... ou mapa de uma aldeia, e ver onde as cenas são passadas: na taberna, na igreja, no jornal...
- segunda ideia: corpo humano: quais os escritores que falam mais na cabeça, ou no corpo... Ramos et al. (2020)
- de facto, toda a vizualização estatística pode ser considerada uma geografização: colocar no espaço (num espaço abstrato) valores quantitativos

Mas voltando à geografia

- Muda com o tempo... vilas promovidas a cidades, províncias promovidas a países, nomes que mudam, divisões administrativas que se modificam, ...
- O mesmo nome para divisões diferentes: Lisboa cidade, Lisboa distrito, ... muitas vezes sem isso estar bem definido
- Locais diferentes com o mesmo nome. Quantas povoações chamadas *Oleiros* há em Portugal?

Representação: pontos ou polígonos?

Atlas literários

Duas formas de os conceber:

- da região para a literatura: ver que obras mencionam uma dita região
- da literatura para a região: dadas obras com menções a locais reais, o que é que elas caracterizam?

Quanto disto pode ser aplicado à literatura "toda"?

BILLIG: É uma das coisas que pretendemos investigar

- Criámos um corpo BILLIG (que por razões de direitos de autor não pode ser público)
- Anotámos automaticamente com o PALAVRAS e restante anotação da Linguateca
- Quantos dos lugares marcados pelos anotadores humanos foram identificados automaticamente?
- Quantos dos lugares identificados automaticamente foram marcados pelos anotadores humanos?
- quais as causas de diferenças?

BILLIG: protótipo com criação automática de mapas

- Melhoria da interface com adição de funcionalidades "geográficas"
- Implementado em php e leaflet por Paulo Alves. Integração com o AC/DC por Diana Santos. Exige ancoragem das expressões geográficas em termos de pares latitude-longitude.

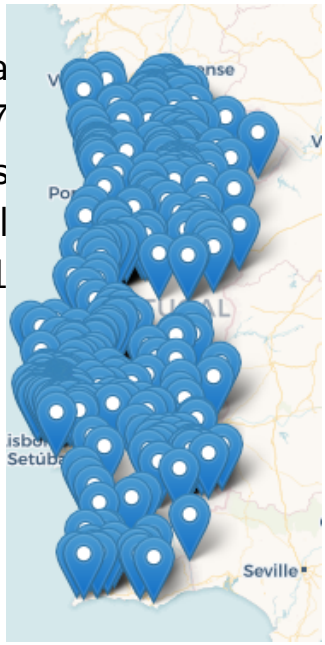
Problemas:

- Nem todas as procuras retornam locais "identificáveis"
- Nem todas as procuras retornam locais, ou locais do mesmo nível de granularidade: O que é que se marca em *Quando vou a Lisboa vou sempre aos Jerónimos!*? ou *Quando venho do Brasil vou sempre aos Jerónimos!*
- O que marcar quando se refere um rio ou uma serra ou um oceano?
- Locais mudam de sítio e de nome, desaparecem ou são construídos

BILLIG: Dados (versão da BD de 4 de dezembro de 2020)

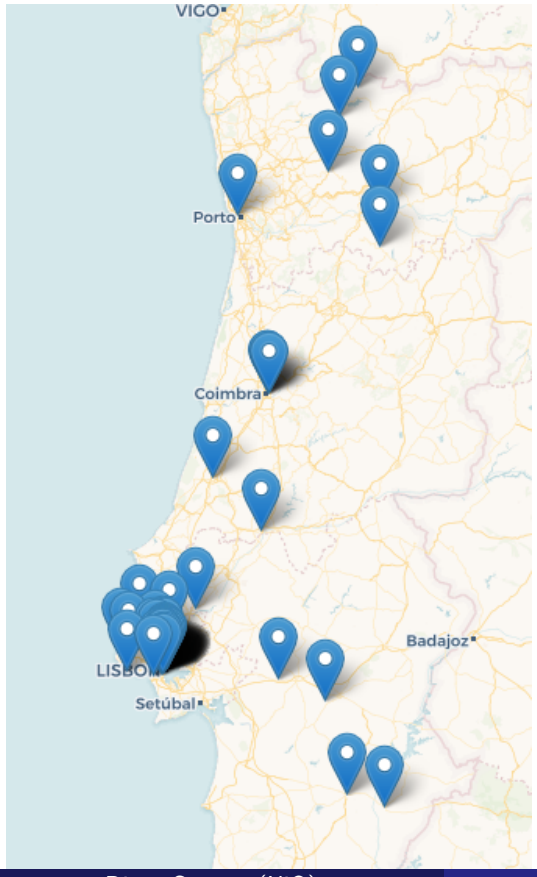
	Atlas	Literateca
Obras	363	847
Palavras	1.790.221	35,14 milhões
Locais	22.506	ca. 255 mil
Locais diferentes	3375	23.691

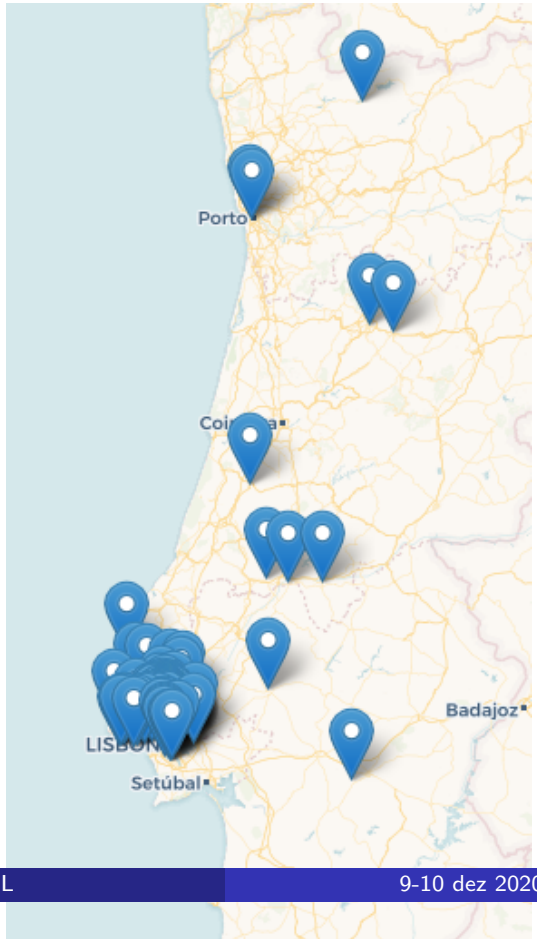
Só 21 obras comuns, 11 do Eça e 5 do Camilo. Locais georeferenciados na base de dados do Atlas: 3285/2426.



Todos os autores no Atlas.

BILLIG: Mapas por autores: Eça vs. Lobo Antunes





BILLIG: Primeiro problema

Como comparar a informação dada nos dois projetos, Atlas e Literateca?

- O Atlas anota por excerto, a Literateca por ocorrência no texto
- O Atlas anota com o nome mais alargado, a Literateca com o que aparece:
  - Varanda de Carqueijais (Serra da Estrela)
  - Monsanto (parque florestal)
  - Chiado (Espinho)
  - Varanda de Carqueijais (Serra da Estrela)
  - Discoteca "A Lontra"
  - Torre do Bugio - Forte de São Lourenço da Cabeça Secca
- Como fazer uma comparação automática?
  - Primeira tentativa: anotar automaticamente todos os casos em que a cadeia de caracteres era semelhante, o que produziu: 9337 casos, correspondentes a 774 lemas (o mais frequente: Lisboa)
  - Segundo passo: identificar os que não apareceram, e tentar "parafraseá-los" para aumentar a abrangência
  - Terceiro passo: olhar para os lugares identificados automaticamente, e que não estavam mencionados no Atlas

BILLIG: melhorias ao sistema de criação de mapas

Além das evidentes em termos de usabilidade e abrangência (juntar o Brasil, juntar o resto do mundo)

- apresentar mapas por polígonos (se estamos a falar de províncias ou países...)
- obter automaticamente a visualização mais apropriada
- usar diferentes cores e tamanhos para identificar as localizações
- ter diferentes formas de descrever os locais

BILLIG: problemas não resolvidos

- diferentes tempos, diferentes mapas
- viagens e caminhos
- sobreposição com outras indicações: emoções, eventos, personagens, etc. (possivelmente uma sobreposição de diferentes mapas)
- mapas enviesados
- mapas de outros tipos de localização (qual a aldeia/cidade/casa padrão)?
- o que fazer quando se misturam locais reais e ficcionais? Ou locais em tempos diferentes?

Concluindo

Com o BILLIG

- Aprendemos muito
- Enriquecemos os serviços da Literateca – e esperemos ter fôlego para aplicar o que aprendemos por exemplo também a outros tipos de corpos, como o DHBB
- Identificámos várias formas de prosseguir que podem ser objeto de novas propostas, ou teses – fica aqui o repto!
- Chegámos (graças ao Covid) a um número razoável de pessoas, e criámos alguns materiais que podem contribuir para a sua formação

Obrigada!

Presépio de Machado de Castro