Distant reading place in Portuguese literature

Diana Santos & Eckhard Bick

University of Oslo & Linguateca & University of Southern Denmark
d.s.m.santos@ilos.uio.no, eckhard.bick@mail.dk

Picture from https://lithub.com/the-places-we-read/

NorLit 2021, 15 June 2022

In a nutshell

- This presentation is not really about mapping
- It is more about the conceptual problems brought about by the properties of natural language

Two main points

- location names are vague
  - with other (non-place) concepts
  - in their "geographical" coverage
- "location" depends on the research question

Meta-presentation

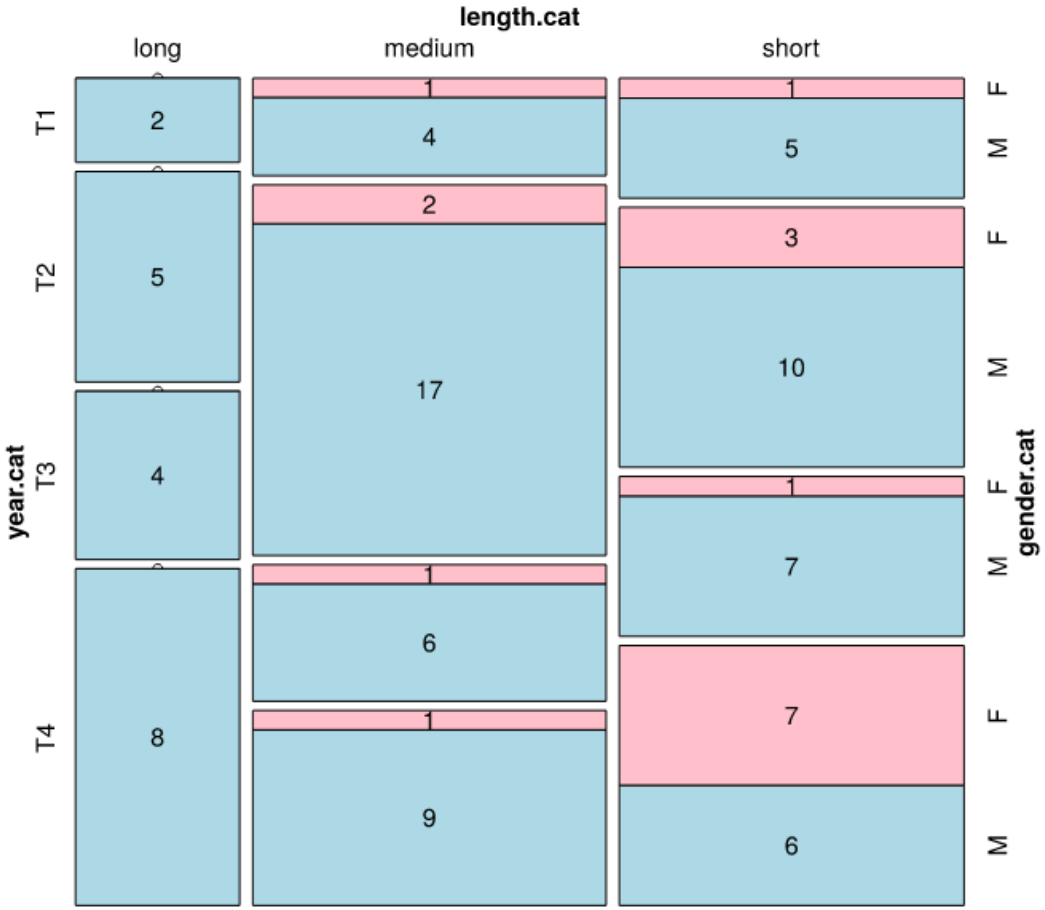
- Presenting the original presentation in a bird's eyes view
- Choosing three examples as to the conceptualization problems
- Showing some results

Outline of presentation

- Brief presentation of ELTeC-por
- *Viagem*: Place annotation in AC/DC
- Named entity recognition with PALAVRAS-NER
- Quantitative overview
- Some results

Brief presentation of ELTeC-por

ELTEC-por is a collection of 100 Portuguese novels following ELTeC's collection design, created in the scope of COST action CA16204, "Distant reading for European Literary History", see `https://www.distant-reading.net/`.



Distant Reading

Brief presentation of *Viagem*

- *Viagem* is a subproject of Literateca, where automatic annotation of proper names indicating places is manually revised, using human-computer cooperation.
- It started under the BILLIG project, therefore we started by Portuguese works, but intend to continue for all lusophone literature.
- The worflow is as follows: we use the original PALAVRAS annotation (Bick, 2014) to identify place candidates, and then we
  - replace it by our own classification
  - correct possible identification/tokenization errors
  - add geo-referencing information

`https://www.linguateca.pt/Literateca`
`https://billig.fcsh.unl.pt/`

Bick, Eckhard. "PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese." In T. Berber Sardinha T. L. S. B. Ferreira (eds.), *Working with Portuguese Corpora*, Bloomsbury, 2014, pp. 279-302.

One detailed example of work in *Viagem*

1. find which proper nouns were automatically marked as locations by PALAVRAS-NER
   `sema=".*(civ|loc|town).*" & pos="PROP.*"]`
2. look at the concordances, and create rules
3. reannotate the corpus

Let us look at *Aljubarrota*, which can be used to denote the place of a battle, the battle itself, a symbol, and the village near the battle.

Aljubarrota: in which works does it occur? (186 cases)

Distribuição

Houve **17** valores diferentes de **obra**.

A_Ala_dos_Namorados	110	271393
O_soldado_de_Aljubarrota	42	42498
O_Monge_de_Cister_I	6	69542
O_Monge_de_Cister_II	6	81895
A_Ilustre_Casa_de_Ramires	4	130722
Os_quatro_reis_impostores	4	156385
Febo_Moniz	3	81649
Arzila:_Romance_do_Século_XV	2	50783
A_Casa_dos_Fantasmas	1	108552
A_ermida_de_Castromino	1	95729
A_morte_vence	1	74311
O_Anel_Misterioso:_Cenas_da_Guerra_Peninsular	1	66351
O_Conde_de_Castel_Melhor	1	235046
O_Manuelinho_de_Évora	1	55402
Os_Maias	1	265179
Os_tripeiros:_Crónica_do_século_XIV	1	44785
Providência	1	85356

Aljubarrota: which authors mention it by name?

Distribuição

Houve **15** valores diferentes de **autor**.

AntCamJ	110	271393
MatBet	42	42498
AleHer	12	1484356
EcaQue	5	2773552
MarMes	4	181339
OliMar	3	81649
BerPPin	2	50783
AJCL	1	44785
AlbPim	1	241132
AntATVas	1	95729
AntFBar	1	198230
AugSar	1	85356
JoJoGra	1	74311
JoadCam	1	235046
LARSil	1	108552

Aljubarrota: selected concordances

- (AleHer1) *Há três anos, não longe da morada de meu velho pai, em Aljubarrota, pelejava eu na Ala dos Namorados* **PLACE|EVENT** Three years ago was I fighting not far away from my old father's abode, in Aljubarrota,
- (AleHer1) *Sim, depois de Aljubarrota, quando no seu castelo de Sintra já não podia ter voz* **EVENT** Yes, after Aljubarrota, when in his Sintra castle he was no longer heard
- (AleHer1) *o bom cavaleiro da ala de Mem Rodrigues nos campos de Aljubarrota* **EVENT** The good gentleman from M.R. battalion in the Aljubarrota battlefield
- (AleHer1) *Estando eu na tenda d' el-rei, naquela noite depois da de Aljubarrota* **EVENT|TIME** I was in the king's tent, that night after Aljubarrota's night

Aljubarrota: selected concordances

- (EcaQue1) *E os outros Ramires, o de Silves, o de Aljubarrota, os de Arzila, os da Índia !* **ABSTRACTION** And the other Ramires: the one from Silves, the one from Aljubarrota, ...
- (EcaQue2) *sua nora não tivera avós mortos em Aljubarrota !* **EVENT** His daughter-in-law could not boast of ancestors dead in Aljubarrota!
- (AntCJ1) *para os lados de Aljubarrota se estreitava consideravelmente* **PLACE - village/region?** toward vicinity of Aljubarrota it would become narrower
- (AntCJ1) *irão primeiro jantar a Aljubarrota* they will first have dinner in Aljubarrota **PLACE - village**
- (AntCJ1) *nunca mais depois daquele crepúsculo épico de Aljubarrota* never again since that epic sunset of Aljubarrota **EVENT|ABSTRACTION**

Aljubarrota: selected concordances

- (AntCJ1) *tangera as Ave-marias na igreja matriz de Aljubarrota* he had tolled Ave-maria at Aljubarrota's main church **PLACE - village**
- (AntCJ1) *resto de algum cerrado feito pelos pastores de Aljubarrota para guarida do gado* the remains of some fence built by Aljubarrota's shepherds... **PLACE - fields**
- (AntCJ1) *o rei, sempre de luto por causa de Aljubarrota*, the king, still mourning because of Aljubarrota **ABSTRACTION|EVENT**
- (AntCJ1) *Valverde foi o corolário de Aljubarrota* Valverde (battle) was the corollary of Aljubarrota (battle) **ABSTRACTION**
- (MatBet1) *perto do rio Lena que atravessa a planicie de Aljubarrota* **PLACE** near the river that crosses Aljubarrota's plain

Aljubarrota: selected concordances

- (MatBet1) *estabelecer uma padaria nas visinhanças de Aljubarrota* open a bakery near Aljubarrota **PLACE - village|battlefield**
- (MatBet1) *A pobre vida de Aljubarrota não offerecia um abrigo de caridade* The poor life in Aljubarrota did not offer a shelter **PLACE - region**
- (MatBet1) *destacamento que saía de Aljubarrota* batallion that was leaving Aljubarrota **PLACE|EVENT - village/battlefield**
- (MatBet1) *onde jazia o corpo da sobrinha da padeira de Aljubarrota* **ABSTRACTION?|BELONGING?** where was buried the body of the Aljubarrota baker(woman)
- (OliMar1) *e lhes daremos outra Aljubarrota !* and we will give them another Aljubarrota **ABSTRACTION**

Further examples, now from *Lamego*

Let me also bring other cases, now with the place name *Lamego*.

- *cónego de Lamego* - Lamego vicar
- *cortes de Lamego* - Lamego assembly
- *um presunto de Lamego* - a ham produced in Lamego, in the way they are produced there
- *uma menina de Lamego* - a young woman from Lamego
- *uma patrícia dos presuntos de Lamego* - someone from the same place as Lamego ham
- *estrada de Lamego* - road to Lamego
- *descendo de Lamego* - coming down from Lamego (southwards)

What can we (hardly) conclude?

1. One has first to distinguish place vs. non-place
2. One has to address (indirectness, vagueness) of place
3. When one anyway interprets the text as place, how to classify it?
   - what is a city? (depends on the time)
   - how to separate among *vila*, *aldeia*, *povoado*?
   - should one distinguish landmarks, individual properties, fields?
   - how often are rivers and seas named for their coast?

Often what is "place" depends on the application/study's purpose

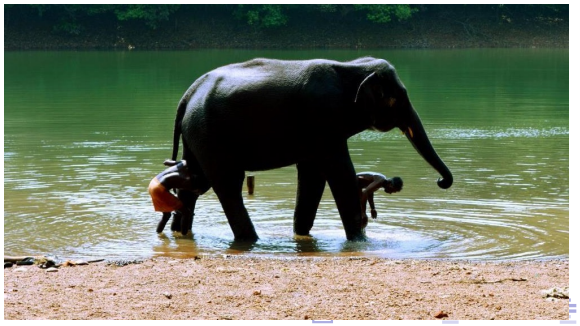
- where are characters from?
- which places in Portugal have associated products (named after the place)?
- which professions/roles are closely connected to a place?

Just one more example, now about *Índia* and *Índias* in the whole Literateca

- India has had many meanings around the centuries, and was far from a precisely delimited area, and had far different connotations and roles
- It is therefore fascinating to see the different ways it appears in Portuguese literature

Influence on the vocabulary

- materials and products: *louça da Índia, caxemira da Índia, arca da Índia, cassa da Índia, seda da Índia, colcha da Índia, terrina da Índia, jarrões da Índia, lenços da Índia, bengala da Índia, talha da Índia, taça da Índia, porcelana da Índia, foulard das Índias, madeira violeta das Índias*
- animals and plants: *elefantes da Índia, craveiro da Índia, galinha da Índia, cravo da Índia, jasmins da Índia, cana da Índia, junco da Índia, chá da Índia, cravinhos da Índia, porquinhos da Índia*
- institutions/abstractions: *carreira da Índia, naus da Índia, jornada da Índia, caminho da Índia, Casa da Índia, vice-rei da Índia*, etc.



Índia and *Índias*, as referred in the literature

- *Da actividade dos espanhóis neste período escreve Heeren, no manual histórico do Sistema político dos Estados da Europa, desde a descoberta das duas Índias* since the discovery of the two Indias
- *Trouxeram-na as gentes impressionáveis, que afluíram para a nossa terra, depois de desfeito no Oriente o sonho miraculoso da Índia .* after the wondrous dream of India was destroyed
- *Da Índia, essa avó das nações, como diz um escritor moderno,* From India, this grandmother of nations
- *O pai afinal não é nenhuma Índia !* Father, after all, is hardly India!

Aljubarrota: some rules

```
[lema="Aljubarrota"] » [sema="Evento"]
[lema="cercania|lado|..."] [lema="de"] a:[lema="Aljubarrota"]
» a:[sema="Local:campo"]
[word="chamam|deixei"] a:[lema="Aljubarrota"] »
a:[sema="Local:campo"]
[lema="jantar|regresso|..."] [lema="a"] a:[lema="Aljubarrota"]
» a:[sema="Local:campo"]
[lema="de"] a:[lema="Aljubarrota"] [word="vieram"]
» a:[sema="Local:campo"]
[lema="ir"] [lema="por"] a:[lema="Aljubarrota"]
» a:[sema="Local:campo"]
[lema="capitão"] [lema="que"] [lema="em"] a:[lema="Aljubarrota
» a:[sema="Local:campo"]
[lema="A=batalha=de=Aljubarrota"] » [sema="obra"]
```

Aljubarrota: some corrections

Before the correction of tokenization:
Distribuição

Houve **12** valores diferentes de **lema**.

Aljubarrota	186
batalha=de=Aljubarrota	30
Rimance=da=Ala=dos=Namorados=de=Aljubarrota	7
Chamorros=de=Aljubarrota	6
Namorados=de=Aljubarrota	6
A=batalha=de=Aljubarrota	4
Antes=da=de=Aljubarrota	4
Condestável=de=Aljubarrota	3
Padeira=de=Aljubarrota	3
Padeiras=de=Aljubarrota	3
Ruços=de=Aljubarrota	3
Santo=de=Aljubarrota	3

Quantitative description

Currently (25 November 2021), the ELTeC-por collection in Literateca (v. 7.5) has

Total tokens	8,162,799
Distinct tokens	213,618
Distinct lemmas	111,539
Total place tokens (revised)	27,698
Total places (revised)	24,510
Distinct places (revised)	1,066
Total place tokens (unrevised)	7,660
Total places (unrevised)	6,206
Distinct places (unrevised)	1,903

Difference between place tokens and places: *Campo Grande* has **two** place tokens, but counts as **one** place.

Quantitative description: kinds of places

Distribuição

Houve **83** valores diferentes de **sema**.

Local:cidade	9998
Local:pais	5671
Local:vila	1276
Local:continente	892
Local:territorio	804
Local:rua	734
Local:relig	567
Local:freguesia	509
Local:regiao	460
Local:municipio	458
Local:bairro	344
Local:provincia	287
Local:organizado	283
Local:aldeia	209
Local:cidade_Local:vila	194
Local:ludico	188
Local:ilha	186

Quantitative description: which places

Distribuição

Houve **1049** valores diferentes de **lema**. Apres

Lisboa	3375
Portugal	2429
Coimbra	959
Porto	909
França	714
Paris	699
Castela	476
Inglaterra	451
Europa	415
Roma	386
Hespanha	328
Santarém	315
Evora	303
Guimarães	303
Índia	289

Distribution issues

- These corpora are publicly available for interrogation.
- But ELTeC is distributed in XML.

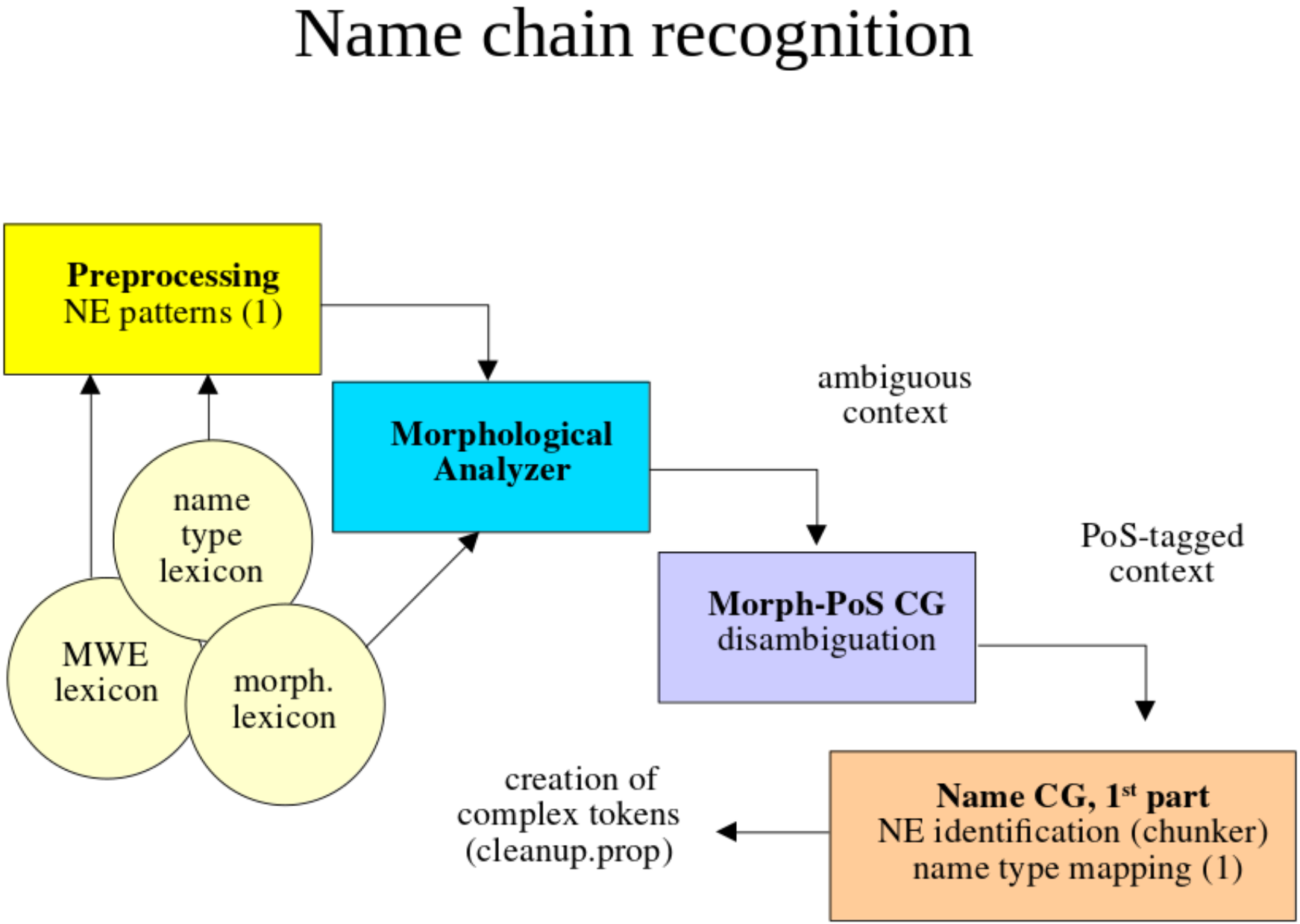
This means that the XML version of ELTeC is obtained by running PALAVRAS and PALAVRAS-NER with no revision. And this is what the researchers get, if they fetch the XML collection from github.

- We are currently investigating ways to cooperate so that the work done in AC/DC can be reused, but this is not simple – classic problem of human-machine interaction

Named entity recognition with PALAVRAS-NER

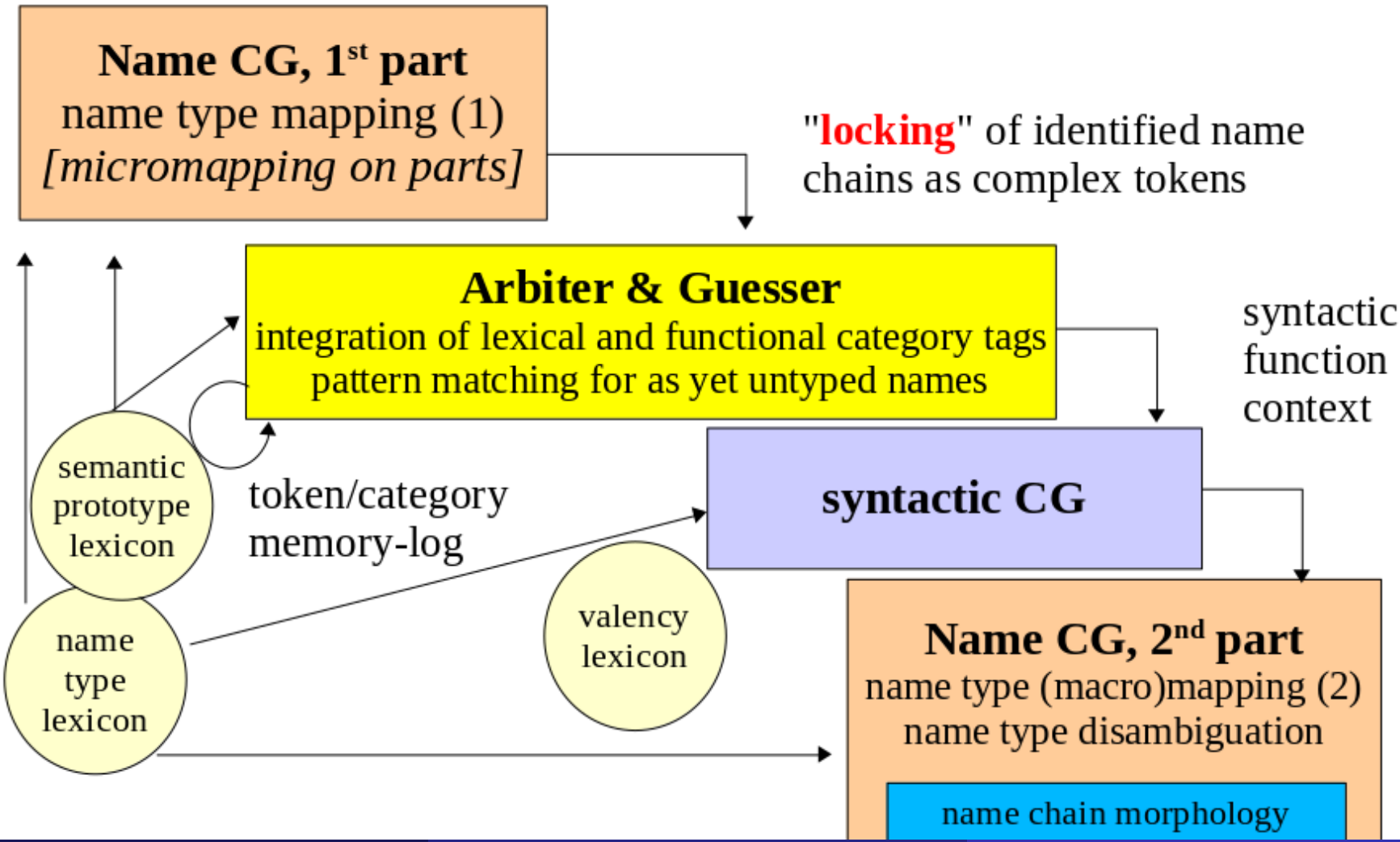
- PALAVRAS-NER is a NER system based on (and meshed with) a general parser (PALAVRAS) hand-written CG rules, both local and contextual, using ca. 40 name categories
- NE recognition: Dynamic tokenization of names as MWEs
  - pattern- and lexicon-based name chain recognition
  - rule-based chunking: fusion, expansion or splitting
- NE identification: Category assignment at 3 levels
  - Corpus-based, manually revised gazeteer lists (ca. 18,000) plus cross-language resources from other parsing projects
  - Pattern-based name type prediction
  - Context-based name type inference
- With dynamic tokenization, the grammar can "conclude" gender, number and semantics from the first (or last) sub-token
- Logging of names/types to resolve abbreviations and ambiguities elsewhere

Named entity recognition with PALAVRAS-NER (2)



Named entity recognition with PALAVRAS-NER (3)

Name type identification



PALAVRAS-NER: CG rules for NER typing

- Micromapping on 1. parts
  - NER typing of 1. parts, exploiting noun classes
  - pattern-based mapping, e.g. numerical expressions in addresses
- Macromapping on entire NE's based on syntactic and semantic context, e.g.:
  - appositions inheriting semantics from a head noun
  - matching semantic slot filler restrictions of verb arguments
  - subject predicatives inheriting sem.type from subjects
  - propagation of NER class between conjuncts

PALAVRAS-NER: Some examples

Examples of the workings of the two kinds of CG rules:

- *Câmara Municipal de Leiria*

```
  [o] <art> <dem> DET F S @y
Câmara [câmara] N F S @F @admin @prop1
Municipal [municipal] ADJ M/F S @prop2
de [de] PRP @prop2
Leiria [Leiria] <civ> <ud> PROP F S @prop2
```

- *a Expo-2010, em Yamarana*

Although *Yamarana* is not in the place lexicon, an event followed by *em* indicates a place, and even more specifically, a city (not mountain).

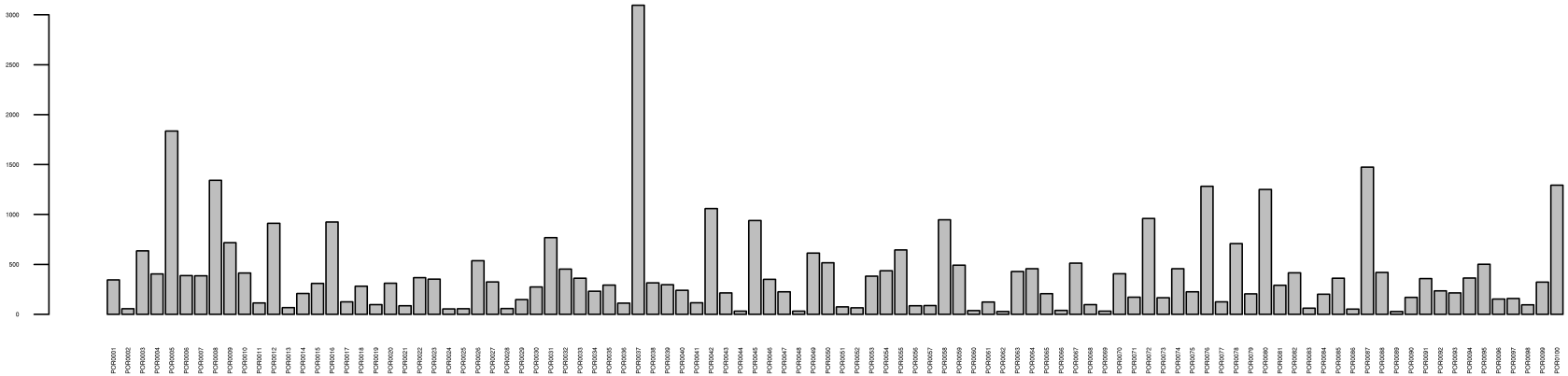
First results

If one looks at the top 10 places

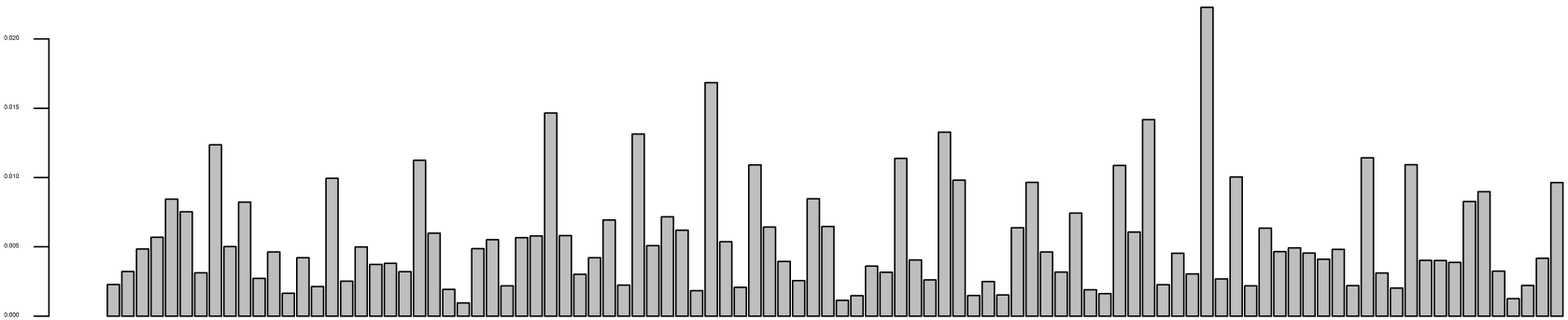
	PALAVRAS-NER	AC/DC	change
Lisboa	3267	3375	increase
Portugal	2303	2429	increase
Coimbra	1012	959	decrease
Porto	915	909	decrease
França	662	714	increase
Paris	616	699	increase
Castela	606	476	decrease
Hespanha	526	328	decrease
Europa	418	415	decrease
Inglaterra	417	451	increase
Roma	356	386	increase

Places per novel

Absolute: *A Ala dos Namorados* (most) and *Os canibais* (least)

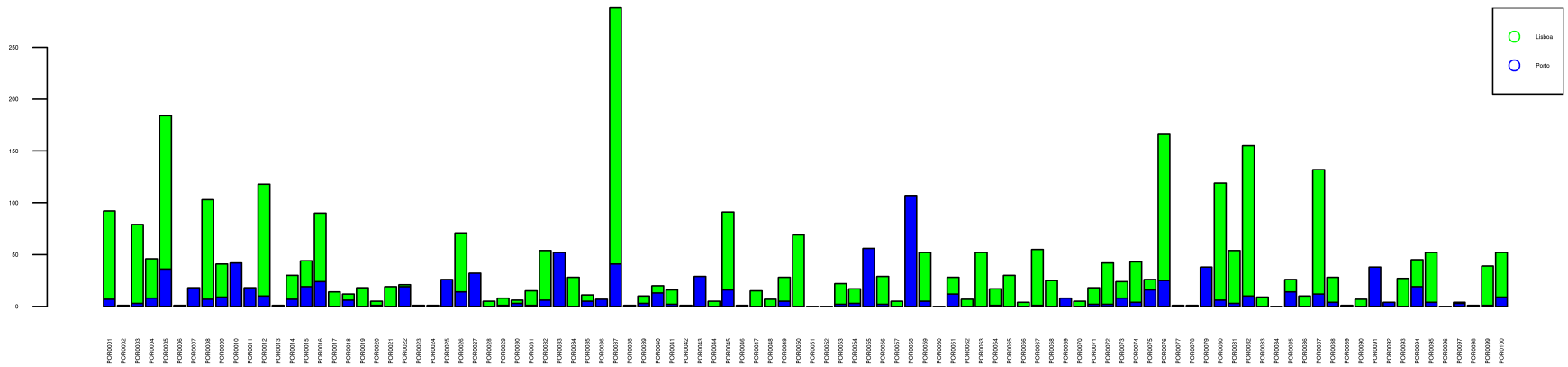


Relative: *No tempo dos franceses* (most) and *A Rosa do Adro* (least)



Presence of Lisbon vs. Porto (absolute)

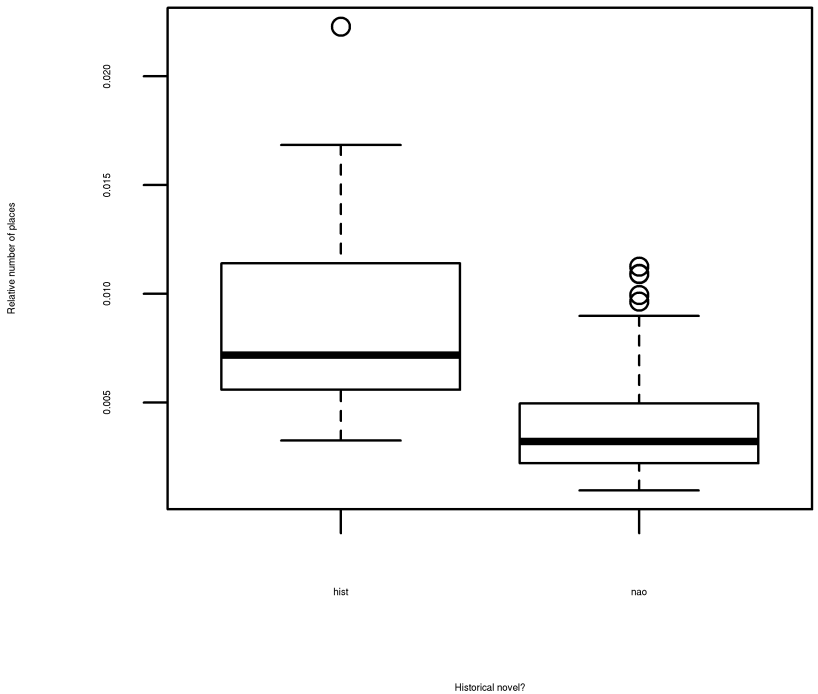
There is often some competition between the capital and the second city, although novelists seem to be well distributed between the two.

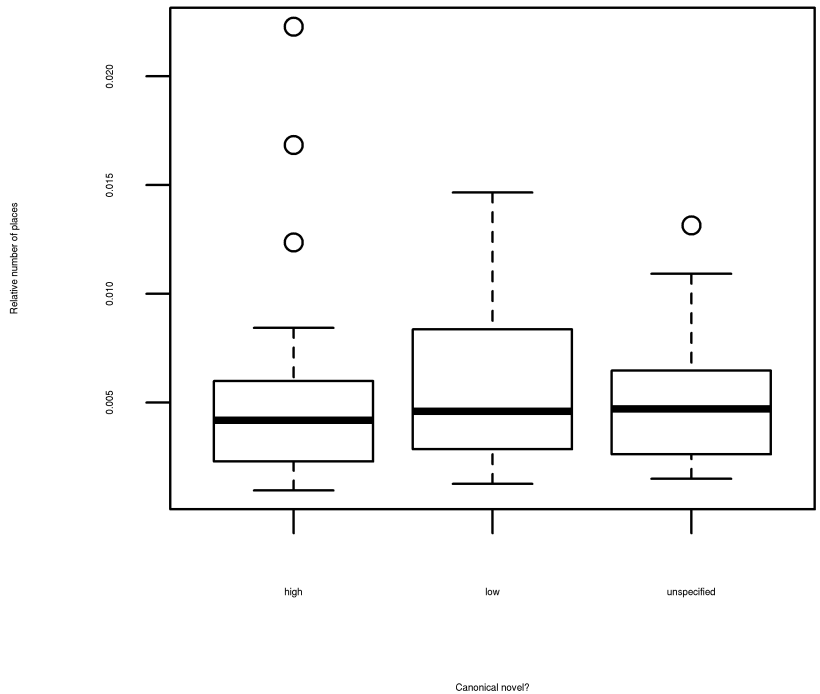


But Lisbon clearly wins in number of mentions.

Do historical novels differ from the others?

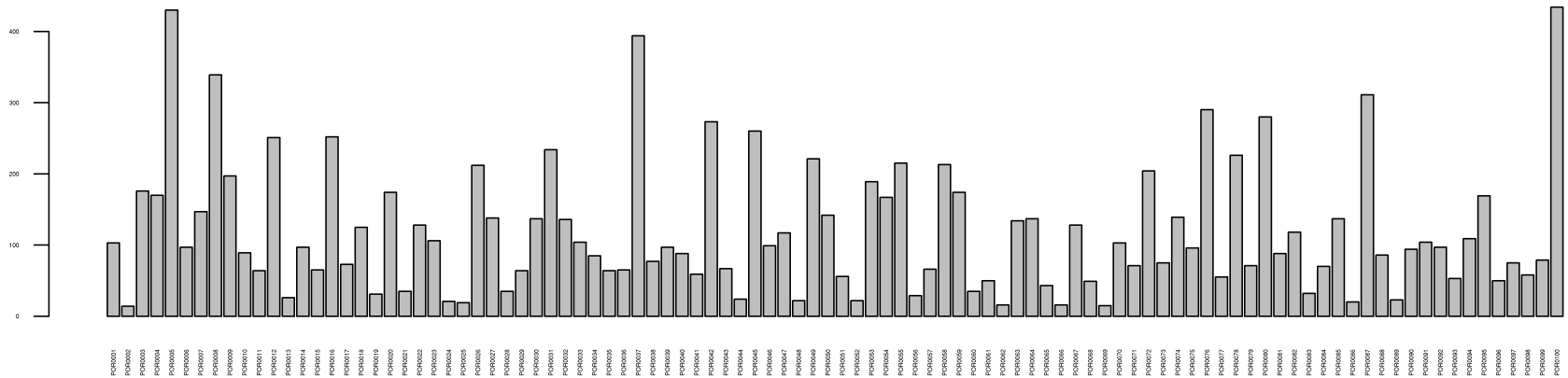
There are 32 historical novels in the ELTeC collection. And there are 26 canonical novels (operationalised by having more than one reprint in the period 1980-2010).

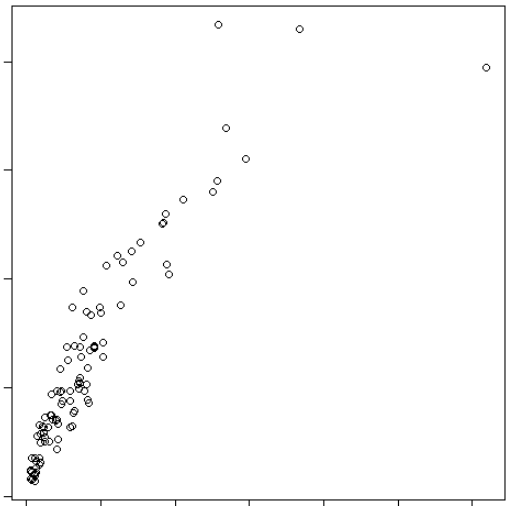




And what about place diversity?

Are there works which are more diverse placewise than others? *A senhora duquesa* (most) and *Sacrificada* (least)





Does the number of distinct places correlate with the number of places? Yes, 0.90

Last words

The goal of this presentation was

- to illustrate some choices of distant reading
- to problematize fully automatic vs. human-revised annotation: advantages and pitfalls
- to illustrate some initial analyses
- to document and publicize the available resources
  - `https://www.linguateca.pt/acesso/corpus.php?corpus=LITERATECA`
  - `https://github.com/COST-ELTeC/ELTeC-por`

Boa viagem! We wish you good reads and travels!