[1]Istituto di Linguistica Computazionale CNR, Pisa, Italy
[2]CRH, EHESS Paris, France
[3]Institute of Polish Language, Polish Academy of Sciences
[4]LaTTiCe CNRS, Université Sorbonne Nouvelle - Paris 3
[5]Linguateca & University of Oslo, Norway
[6]University of Belgrade, Serbia

Named Entity Recognition for Distant Reading in ELTeC

Francesca Frontini[1], Carmen Brando[2], Joanna Byszuk[3], Ioana Galleron[4], Diana Santos[5], Ranka Stanković[6]

CLARIN annual conference
5–7 October 2020

Overview of the talk

Introduction

Manual annotation
- Annotation guidelines

Software infrastructure
- BRAT
- NER & Beyond

Comparing manual and automatic annotation
- Testing Automatic NER
- Results
- Analysis

Links to CLARIN

Links

Context

Distant Reading

"Distant Reading for European Literary History (COST Action CA16204) is a project aiming to create a vibrant and diverse network of researchers jointly developing the resources and methods necessary to change the way European literary history is written. Grounded in the Distant Reading paradigm (i.e. using computational methods of analysis for large collections of literary texts), the Action will create a shared theoretical and practical framework to enable innovative, sophisticated, data-driven, computational methods of literary text analysis across at least 10 European languages."

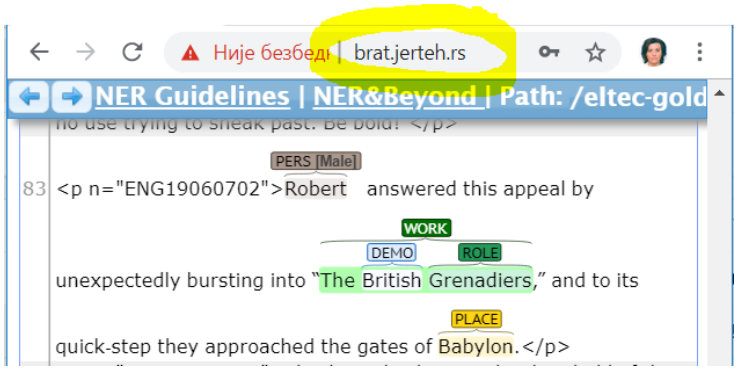
https://www.distant-reading.net

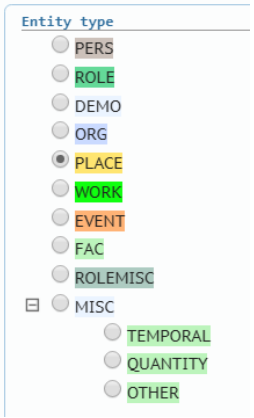
Motivation

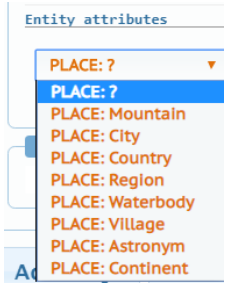
- Most NER systems and studies are
  - not specifically concerned with literary texts
  - not dealing with 19th century texts
  - not multilingual enough
- We wanted to explore the issues in COST texts
  - what kind of entities
  - what kind of problems
- Create an infrastructure to evaluate current NER systems for literary texts in the languages of COST

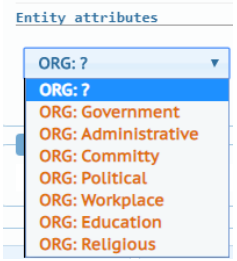
Exercise

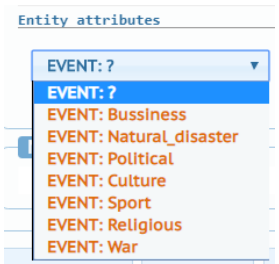
- Languages: Czech, German, English, French, Hungarian, Norwegian, Portuguese, Serbian and Slovene
- "Lightweight semantic annotation" defined: marking persons, demonyms, professions and other roles, works, places, facilities and organizations
- 5 random passages of 400 white space-delimited tokens taken from 20 novels from each language, manually annotated using brat











NE annotation

Table: Current manually NE-annotated corpus[1].

	DEMO	EVENT	LOC	ORG	OTHER	PERS	ROLE	WORK
cze	163	5	275	0	0	1150	454	0
deu	66	2	323	12	0	973	458	4
eng	56	7	198	37	0	1184	203	25
fra	77	3	262	22	128	900	244	18
hun	29	7	152	20	0	1091	367	7
nor	4	8	83	25	3	990	201	10
por1	17	9	351	19	0	940	490	54
por2	34	1	256	30	7	1059	347	7
slv	133	54	336	37	0	1230	620	2
srp	121	18	185	11	0	985	301	4
	700	114	2425	213	138	10514	3685	131

[1] Portuguese the exercise was done twice: with canonical, modernized ortography works (por1), and with non-canonical, old-ortography works (por2)

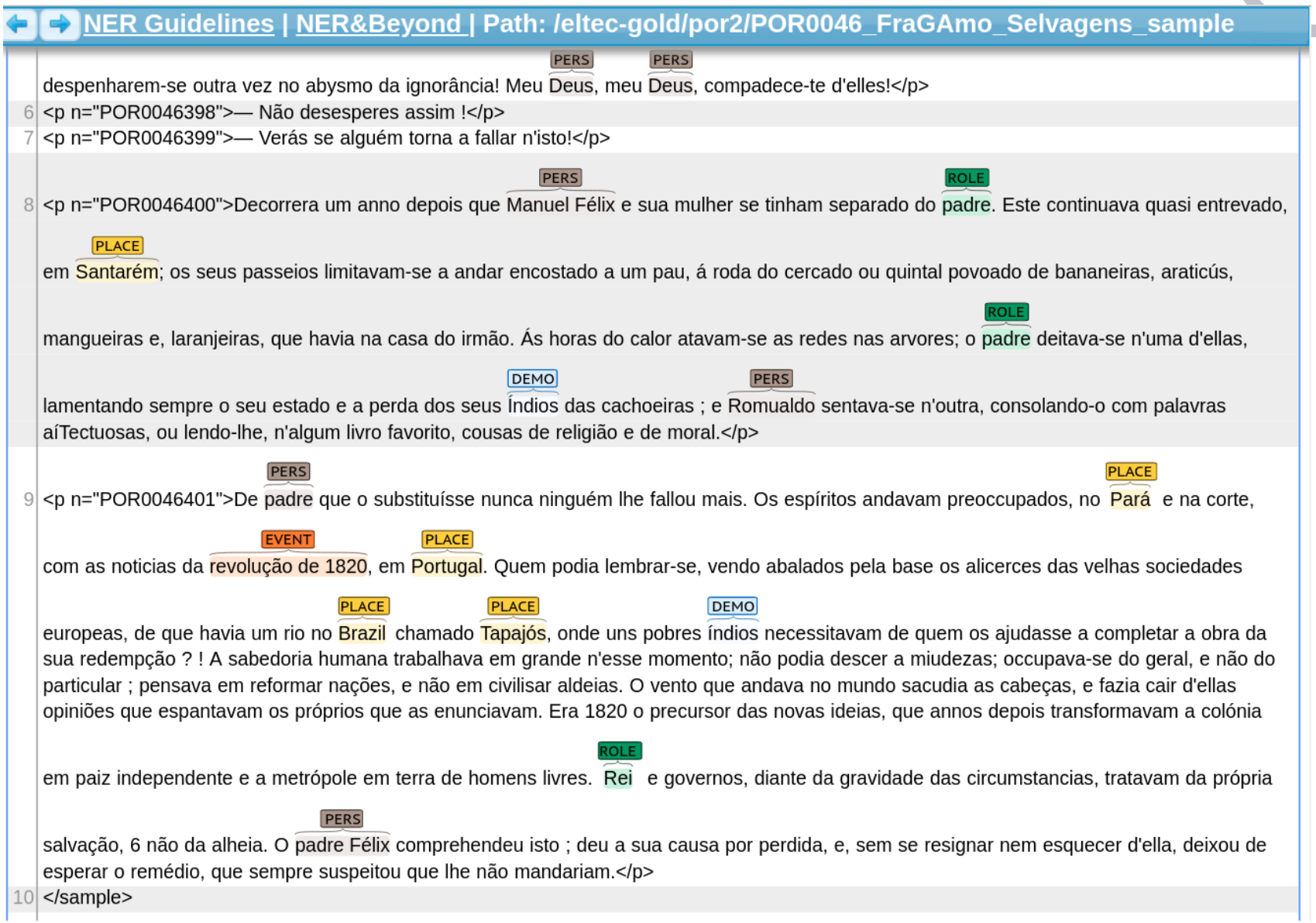
Annotation guidelines

Literary text show a broad variety NEs with respect to more standardised non fictional texts.

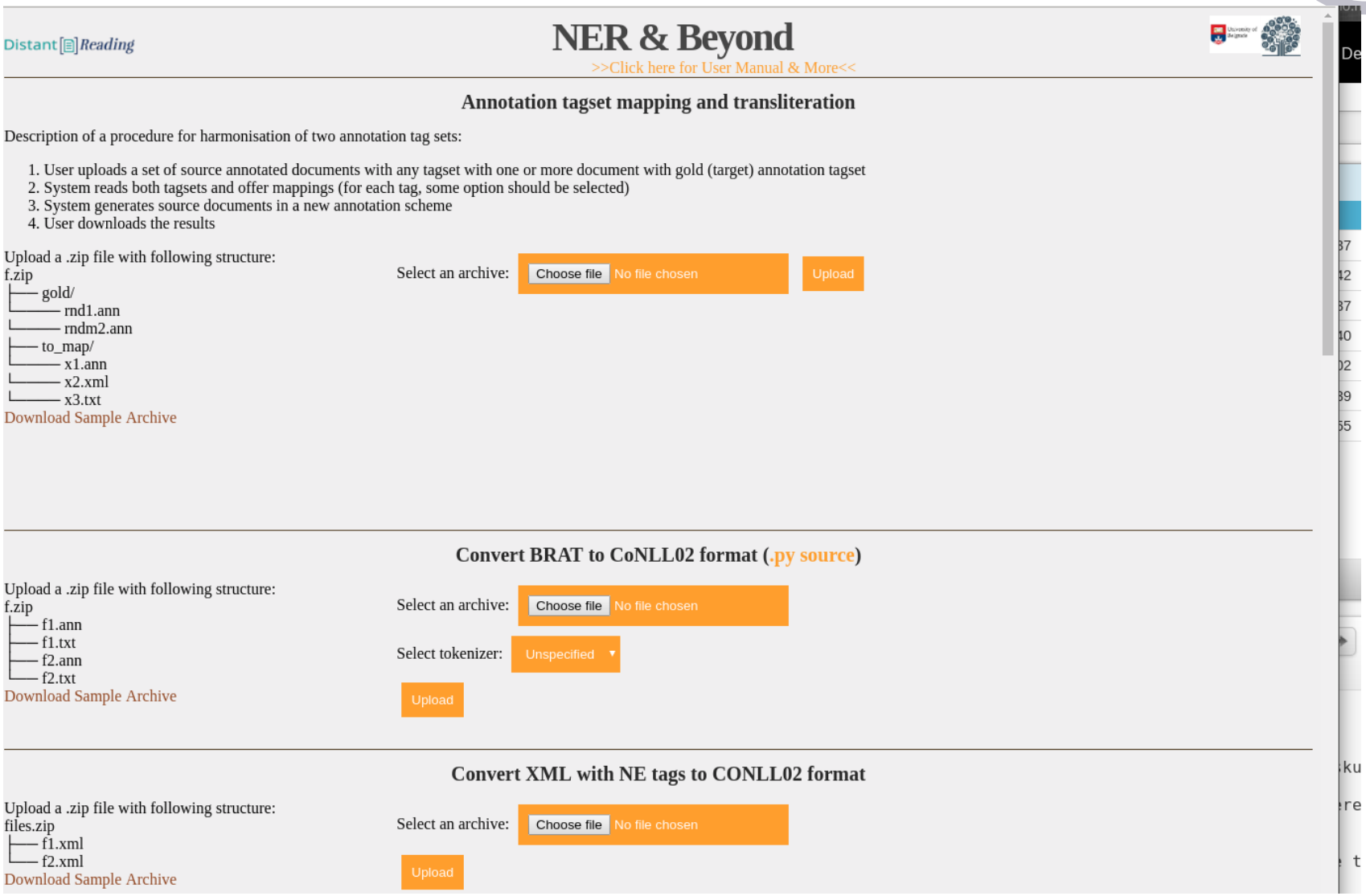
- People are often referred to by profession or their origin, as well as by family relations only ("Maman")...
- Fictional characters may be present and animals and objects may have proper names.
- Additional types of entities, which are particularly interesting for the purposes of literary and cultural analysis, require annotation, such titles of works of art, books, publications; literary movements, ...

Classification is sometimes difficult and there is a risk of proliferation of categories and sub-categories. E.g. "Vive la **Réforme**! à bas Giuzot!" (here "**Réforme**" refers to a proposal of reform of electoral law which the politician Giuzot opposed).

Manual annotation infrastructure: BRAT



Web portal NER & Beyond



Distant Reading

NER & Beyond

>>Click here for User Manual & More<<

Annotation tagset mapping and transliteration

Description of a procedure for harmonisation of two annotation tag sets:

1. User uploads a set of source annotated documents with any tagset with one or more document with gold (target) annotation tagset
2. System reads both tagsets and offer mappings (for each tag, some option should be selected)
3. System generates source documents in a new annotation scheme
4. User downloads the results

Upload a .zip file with following structure:
f.zip
├── gold/
└── rnd1.ann
└── rndm2.ann
├── to_map/
└── x1.ann
└── x2.xml
└── x3.txt
Download Sample Archive

Select an archive: Choose file No file chosen Upload

Convert BRAT to CoNLL02 format (.py source)

Upload a .zip file with following structure:
f.zip
├── f1.ann
├── f1.txt
├── f2.ann
└── f2.txt
Download Sample Archive

Select an archive: Choose file No file chosen

Select tokenizer: Unspecified

Upload

Convert XML with NE tags to CONLL02 format

Upload a .zip file with following structure:
files.zip
├── f1.xml
└── f2.xml
Download Sample Archive

Select an archive: Choose file No file chosen

Upload

Web portal NER & Beyond: tools

At University of Belgrade and Jerteh, `http://nerbeyond.jerteh.rs/`

- Conversion tools among different annotation formats
  - BRAT to CoNLL02
  - XML with NE tags to CONLL02
  - XML with NE tags to BRAT
  - BRAT (ann & txt) to XML
  - CONLL02 to BRAT
- Automatic Named entity recognition (and annotation)
  - spaCy (eng, fra, ita, nld, ger, por, srp, spa, multi)
  - Stanford (eng, ger, srp)
  - CVNER (ser)
- Computation of statistics based on BRAT `.ann` format
- NER evaluation with the Gemini tool
  - Precision and recall tables
  - Visual comparison in HTML
- Annotation tagset harmonization (mapping and transliteration)

Example of comparison of gold (manual, blue) and automatic (SpaCy, pink) annotation

<sample> And at this moment entered the room the young nobleman whom we have before mentioned, accompanied by an individual who was approaching perhaps the termination of his fifth lustre but whose general air rather betokened even a less experienced time of life. Tall, with a well-proportioned figure and a graceful carriage, his countenance touched with a sensibility that at once engages the affections. Charles Egremont was not only admired by that sex, whose approval generally secures men enemies among their fellows, but was at the same time the favourite of his own.

“Ah, Egremont! come and sit here,” exclaimed more than one banqueter.

“I saw you waltzing with the little Bertie, old fellow,” said Lord Fitzheron, “and therefore did not stay to speak to you, as I thought we should meet here. I am to call for you, mind.”

“How shall we all feel this time to-morrow?” said Egremont, smiling.

“The happiest fellow at this moment must be Cockie Graves,” said Lord Milford. “He can have no suspense. I have been looking over his book, and I defy him, whatever happens, not to lose.”

“Poor Cockie.” said Mr Berners; “he has asked me to dine with him at the Clarendon on Saturday.”

“Cockie is a very good Cockie,” said Lord Milford, “and any gentleman sportsman present wishes to give seven to two, I will take him to any ar

“My book is made up,” said Egremont; “and I stand or fall by

Annotation <PERS>

Text — File1 : ENG18450_Disraeli_sample.xml

Text — File2 : ANN FILE

Testing Automatic NER

- Case study on four collections only, in English, French, Portuguese and Serbian
- For each collection, we tested two tools: one common for all (spaCy) and another one language specific
- BRAT outputs were compared to annotations produced by these tools
- Evaluation of string detection was strict (segments must match exactly)

Testing Automatic NER - results

	Cat	Correct	Missing	Spurious	Precision	Recall	Excess
SEM-fra	LOC	73	112	84	0.465	0.395	0.535
	PERS	82	512	115	0.416	0.138	0.584
SPACY-fra	LOC	103	78	468	0.180	0.569	0.820
	PERS	329	194	297	0.526	0.629	0.474
PALAVRAS-por1	LOC	223	63	44	0.835	0.780	0.165
	PERS	816	90	86	0.905	0.901	0.095
SPACY-por1	LOC	225	84	440	0.338	0.728	0.662
	PERS	465	256	374	0.554	0.645	0.446
PALAVRAS-por2	LOC	151	67	91	0.624	0.693	0.376
	PERS	857	133	285	0.750	0.866	0.250
SPACY-por2	LOC	157	57	396	0.284	0.734	0.716
	PERS	569	236	393	0.591	0.707	0.409
Stanford-eng	LOC	98	100	126	0.438	0.495	0.563
	PERS	649	535	399	0.619	0.548	0.381
SPACY-eng	LOC	98	100	170	0.366	0.495	0.634
	PERS	536	648	240	0.691	0.453	0.309
SrpNER-srp	LOC	107	78	19	0.849	0.578	0.151
	PERS	718	267	158	0.820	0.729	0.180
SPACY-srp	LOC	57	128	104	0.354	0.308	0.646
	PERS	553	432	315	0.637	0.561	0.363

Table: Results of the strict evaluation, per language and category.

Testing Automatic NER - analysis

- A strict evaluation of detection is often penalising for PERS, because of honorifics which we chose to include in our annotation
- It is further complicated by the fact that the XML annotated input was processed as such by tools which often expect plain text
- In most cases, LOC seems to be less problematic for the pre-trained models.

Links to CLARIN

- The NE corpus, as part of the larger annotated ELTeC corpus, will have to adhere to the FAIR requirements, for long term preservation, visibility, accessibility. A deposit in a repository (possibly Textgrid) is envisaged at the end of the project.
- Accessibility could be further enhanced by making the corpus part of the CLARIN federated content search
- Finally, CLARIN collaboration is welcome for the creation of better adapted NE tools which can overcome the problems that we have identified for state of the art tools.

Useful links

- The COST Action web page: `https://www.distant-reading.net`
- The ELTeC corpus: `https://distantreading.github.io/ELTeC/index.html`
- The current NE annotated sub-corpus `http://brat.jerteh.rs/#/eltec-simplified/`

Acknowledgements

Presented research is supported by COST Action 16204 – Distant Reading for European Literary History

* * *

We thank Lou Burnard for creating the samples, Tajda Liplin, Zala Vidic, Karolina Zgaga, Michael Preminger, Tonje Vold, Emma Takács, Silvie Cinková, Klára Macháčková, and Cvetana Krstev for annotating the texts, and Branislava Šandrih for help in creating the sites at the University of Belgrade, and all our colleagues at the COST action Distant Reading for European Literary History.

Thank you!