[1]University of Belgrade, Serbia
[2]Linguateca & University of Oslo, Norway
[3]Praxiling UMR 5267 CNRS - UPVM3, Montpellier, France
[4]Jožef Stefan Institute, Ljubljana, Slovenia
[5]EHESS Paris, France

Named Entity Recognition for Distant Reading in Several Languages

Ranka Stanković[1], Diana Santos[2], Francesca Frontini[3], Tomaž Erjavec[4], Carmen Brando[5]

DH_Budapest
25–27 September 2019

Overview of the talk

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."

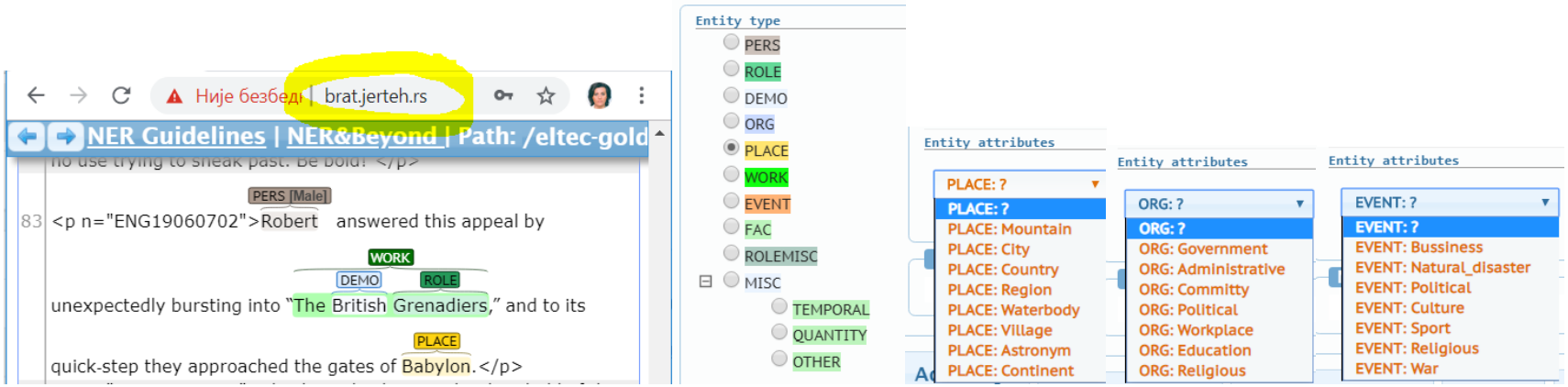
https://www.distant-reading.net

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, 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

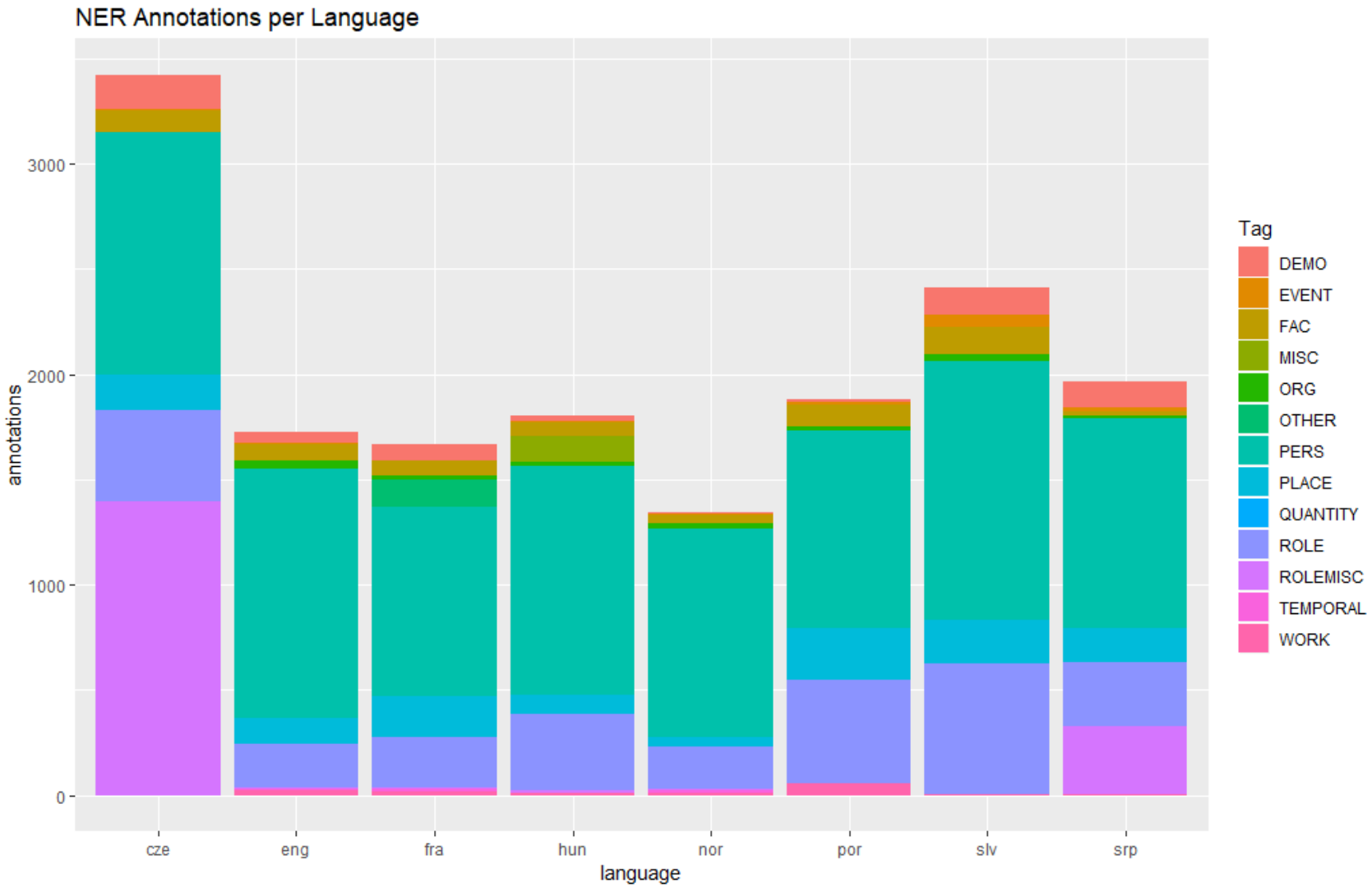


Quantitative data

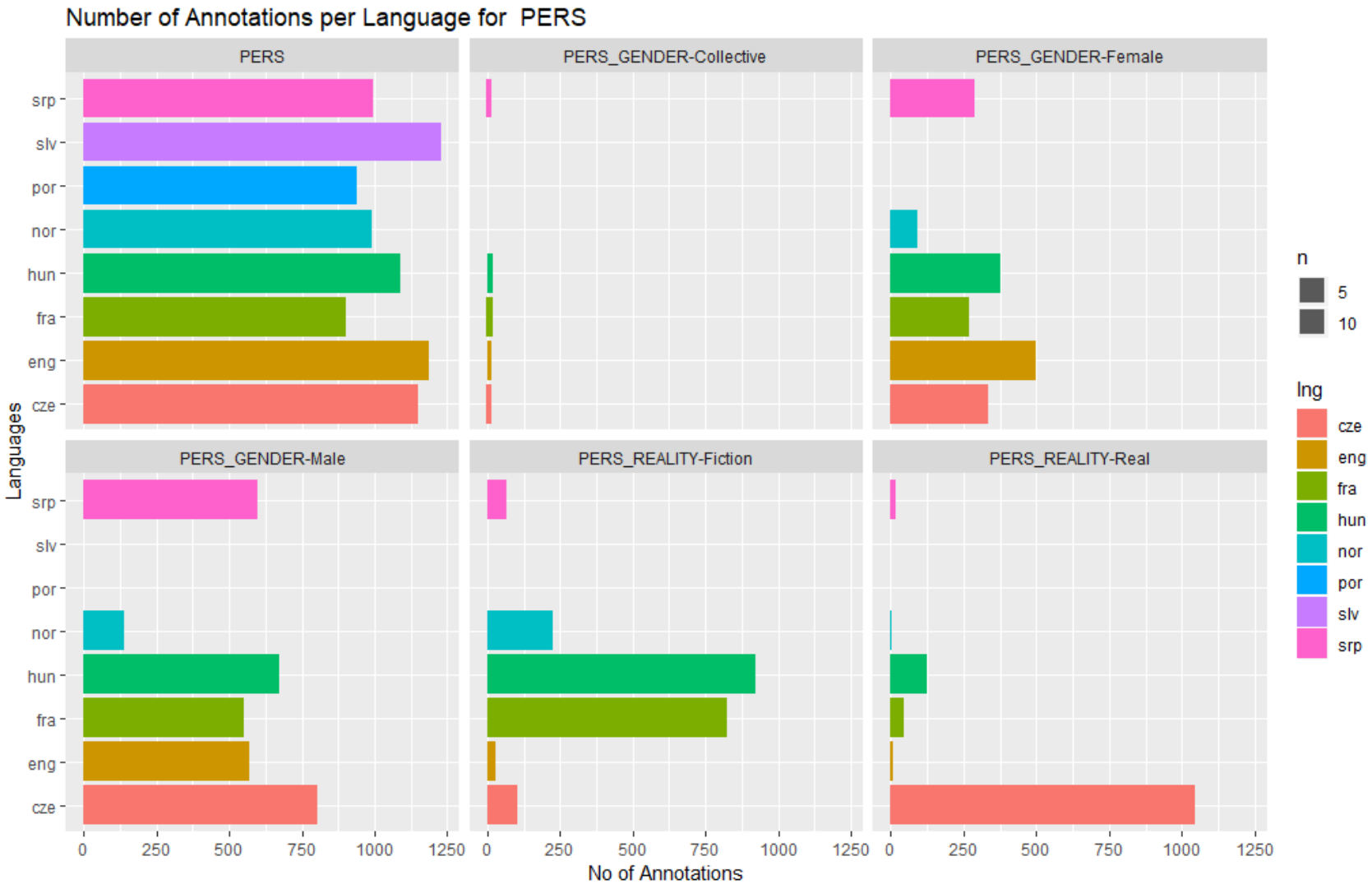
Table: Preliminary data on main classes of NE (as of May 21st, 2019, 20:44)

	eng	fra	por	hun	nor	slv	srp	cze
Words	53093	52031	53958	58825	52900	53348	67232	52650
NE	1744	1669	1883	1805	1344	2412	1962	3424
DEMO	56	77	17	29	4	133	121	163
EVENT	7	3	9	7	8	54	18	5
FAC	75	70	103	65	41	128	23	108
ORG	37	22	19	20	25	37	11	0
PERS	1186	900	940	1091	990	1230	995	1150
PLACE	126	192	248	87	42	208	163	167
ROLE	203	244	490	367	201	620	301	454
WORK	25	18	54	7	10	2	4	0

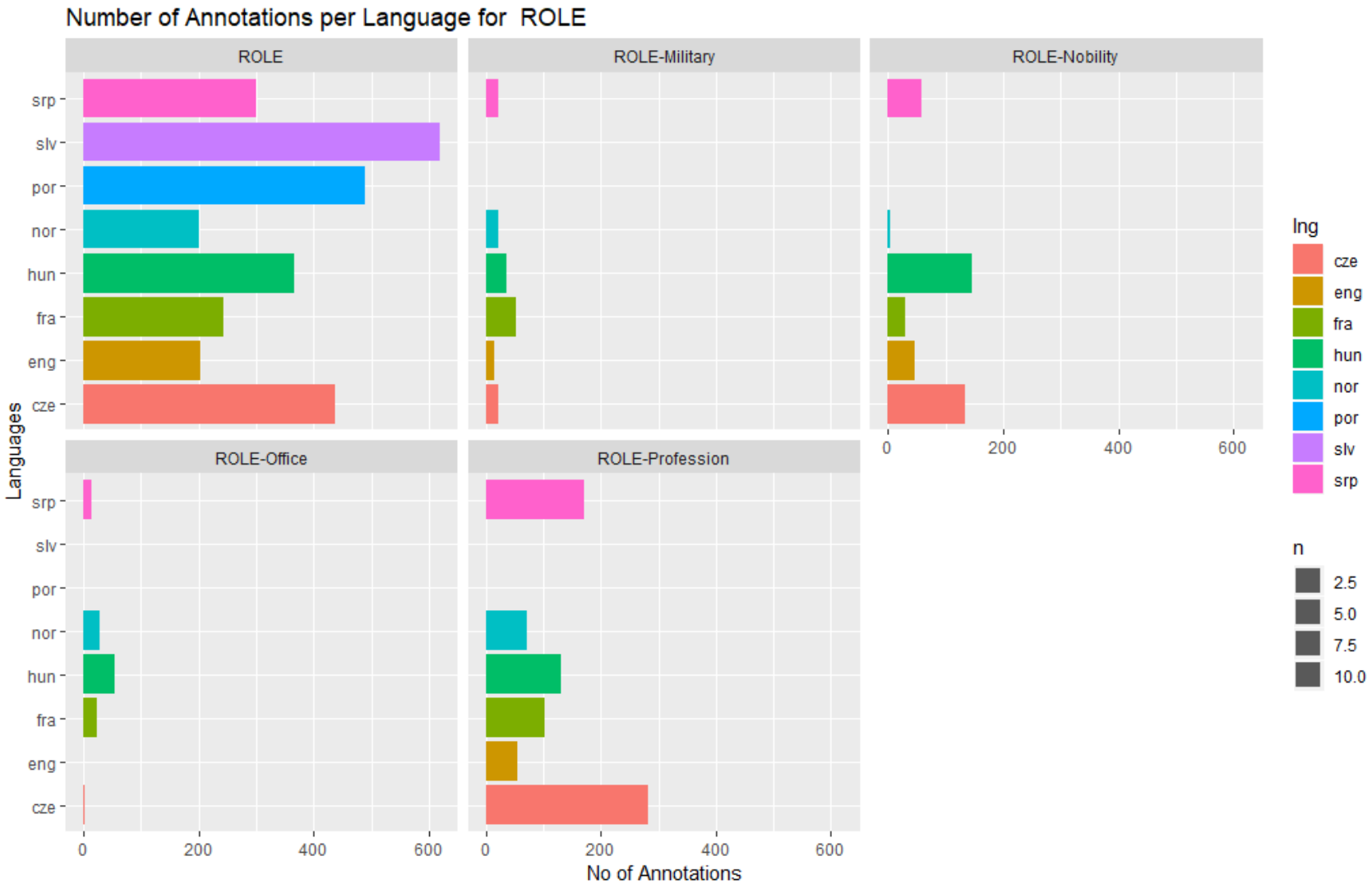
Quantitative data - all classes



Quantitative data - PERS class



Quantitative data - ROLE class



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).

Interesting issues

Interesting issues that emerged and which our guidelines have tried to address:

- ROLE is used to identify professions and job titles (Archbishop of Canterbury; Sa Majesté Britannique), and ROLEMISC for family relations and epitaphs
- Mentions of gods and other supernatural entities, also named animals, were annotated as PERS (persons)
- Swear words or exclamations containing proper names were not considered as named entities ("Oh my God!")
- Fixed expressions or idioms which included proper names, like *Éva leánya* (Eva's daughter) in Hungarian to describe "woman" were tagged as MISC

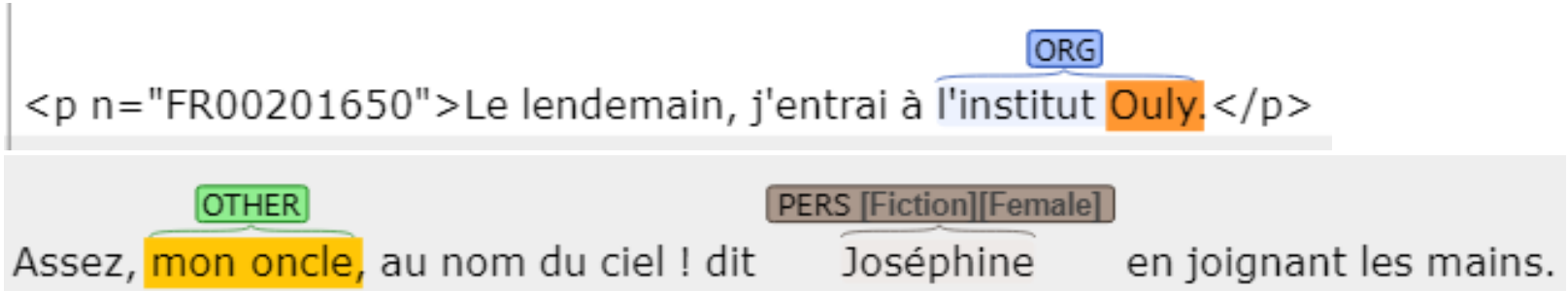
Problems detected

Some issues remain open such as:

- It is sometimes difficult to distinguish types of places, e.g. distinguish between a region and a country, or a village and a city)
- We created a separate class for FACILITIES covering "localised artefacts", e.g. Opera, Theater, Prison, etc.
- So far no distinction has been made between forms of addressing and naming the addressee, both have been marked PERS

Further problems

- Identification of fictional places and characters, which was initially thought as a requirement for literary analysis, was problematic
- Difficulty for deciding whether determinants make part of the named entity (e.g. **l'institut Ouly** or **institut Ouly**) or possessive pronouns in family epithets (e.g. **mon oncle**), a known issue in NER and recurrent in literary texts; this sometimes leads to an inconsistent annotation



Methodological choices

- Recursivity: We now allow nested annotations, e.g. <pers><role>General</role> Junot</pers><role>ministre de<place>Belgique</place></role>
- Capitalisation: given that not all languages follow the same capitalisation rules, the use of this feature is highly problematic in annotation guidelines. In particular, family relations or professional groups and individuals are annotated independently of capitalisation. On the other hand, in some languages capitalisation may be a clue for deciding whether an adjective is part of a NE, as in "Old John".

What about inter-annotator agreement?

Measuring inter-annotator agreement (IAA) is common practice in NLP to estimate the reliability of annotations and possible when multiple independent annotators are available, several measures exist and are implemented such as F-measure, Cohen's Kappa, ...

- unfortunately, most COST annotator teams (one per language) could not perform parallel annotation
- a small experiment was possible with two French annotators and an excerpt of the French ELTEC data set: 15 paragraphs of texts considered by some of the team members as difficult to annotate

Results on French sample

- Some results :
  - *PERS*: observed agreement: 0.69; Cohen's kappa: -0.16
  - *PLACE*: observed agreement: 0.64; Cohen's kappa: -0.13
  - k=1 is perfect agreement, k=0 means agreement is equal to chance, k=-1 means perfect disagreement
  - for the other NER categories, number of annotations is insufficient for conclusive IAA calculation
- Same segments are annotated, with only two differences in terms of NE boundaries (1. F: "Marseillaise" C: "La Marseillaise", 2. C: "représentant Charbonnel", F: "Charbonnel")
- Most differences are in the choice of different categories
- This experience helped to contribute to the NER annotation guidelines but needs to be repeated with more data and new, independent annotators

Professions mentioned

Table: Most mentioned professions

Lang	Types	Professions (word forms)
cze	195	kněz (priest) 5, rychtáři (magistrate) 5, plavec (swimmer) 6, knížete (knight) 6, rychtář (magistrate) 7, knížecí (knight) 11, kníže (knight) 12
eng	91	Sir 6, Dr. 6, Secretary 6, Colonel 7, General 8, President 10, Professor 13, Queen 13, Lord 16
fra	147	domestique (servant) 5 colonel (colonel) 6 comte (count) 6 capitaine (captain) 9 général (general) 12
hun	249	tanár (teacher, professor) 7 tiszttartó (steward, curator, bailiff) 7 gróf (count) 8 huszár (hussar, cavalry man) 10 úr (sir, lord, mr.) 11 grófnő (countess) 11 ur (sir, lord, mr.) 15

Professions mentioned 2

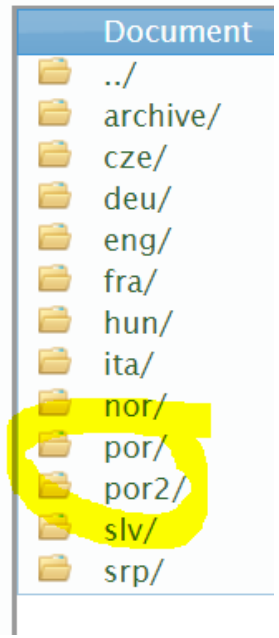
Lang	Types	Professions (word forms)
nor	116	Kapellanen (chaplain) 6, Kaptein (captain) 6, Professoren (professor) 6, Unge-Konsulen (consul) 6, Styrmanden (manager) 7, Konsulen (consul) 10, Kapteinen (captain) 11
por	221	carcereiro (jailer) 9, conselheiro (counsellor) 9, pagem (page) 9, tenente (lieutenant) 9, bispo (bishop) 10, doutor (doctor) 13, padre (priest) 17, fidalgo (nobleman) 18
slv	386	kapitan (captain) 7, sodnik (judge) 7, kmetje (peasants) 7, grajščak (nobleman) 7, grof (count) 8, vitez (knight) 13, doktor (doctor) 15, profesor (professor) 16
srp	301	лекар (physician) 5, ћата (junior clerk) 5, учитељ (teacher) 5, кмет (village leader) 6, паша (pasha) 6, биров (policeman) 6, Газда (boss) 6, дућанџија (owner of the small shop) 6, поп (priest) 7

How representative are these data?

Would our numbers be comparable if we had annotated a different set of texts? A small experiment was done with Portuguese:

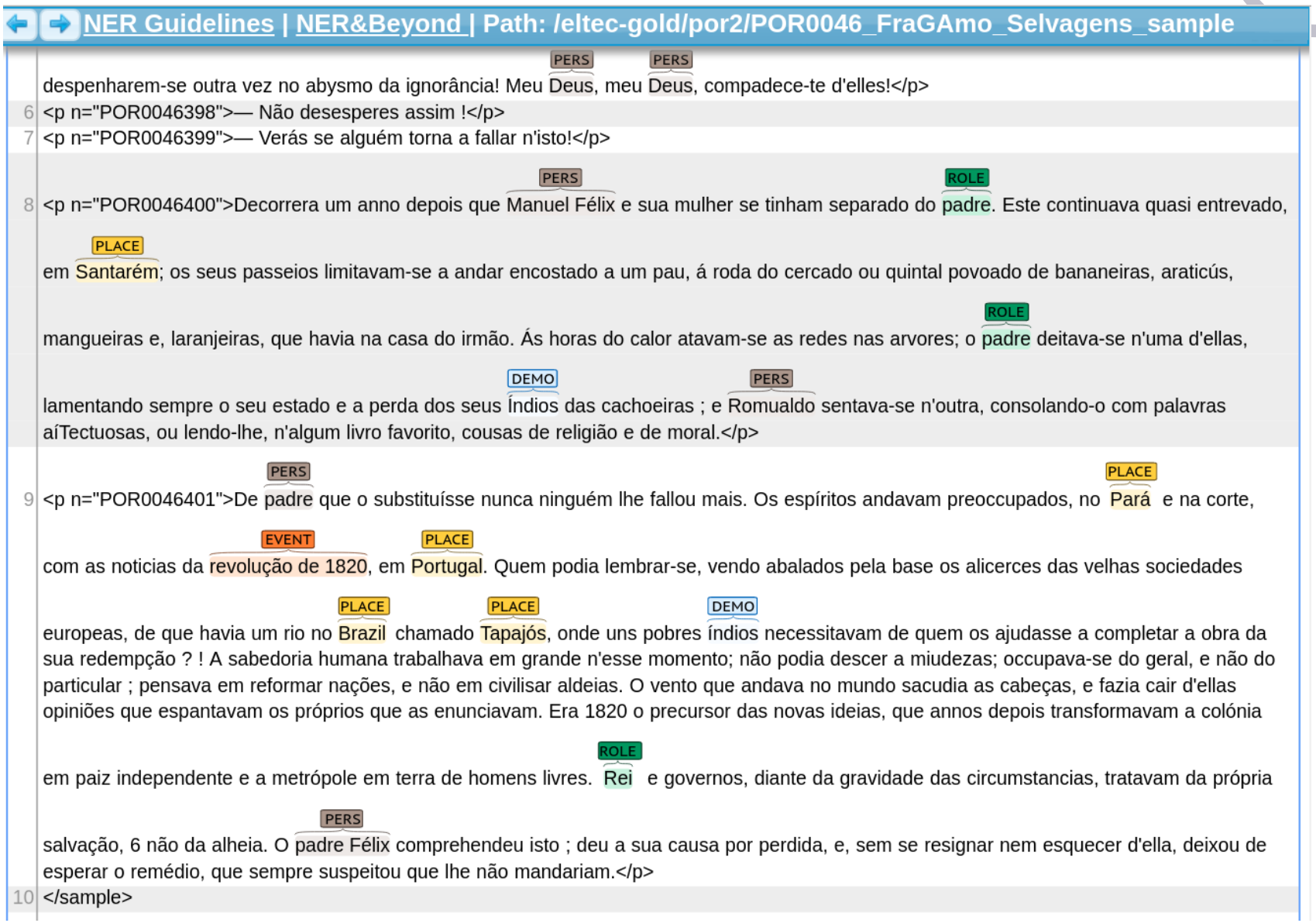
Table: Comparing two sets of texts

	por	por2
Words	53958	55429
NE	1883	1649
DEMO	17	34
EVENT	9	1
FAC	103	78
ORG	19	30
PERS	940	990
PLACE	248	169
ROLE	490	333
TEMPORAL	1	2
WORK	54	7

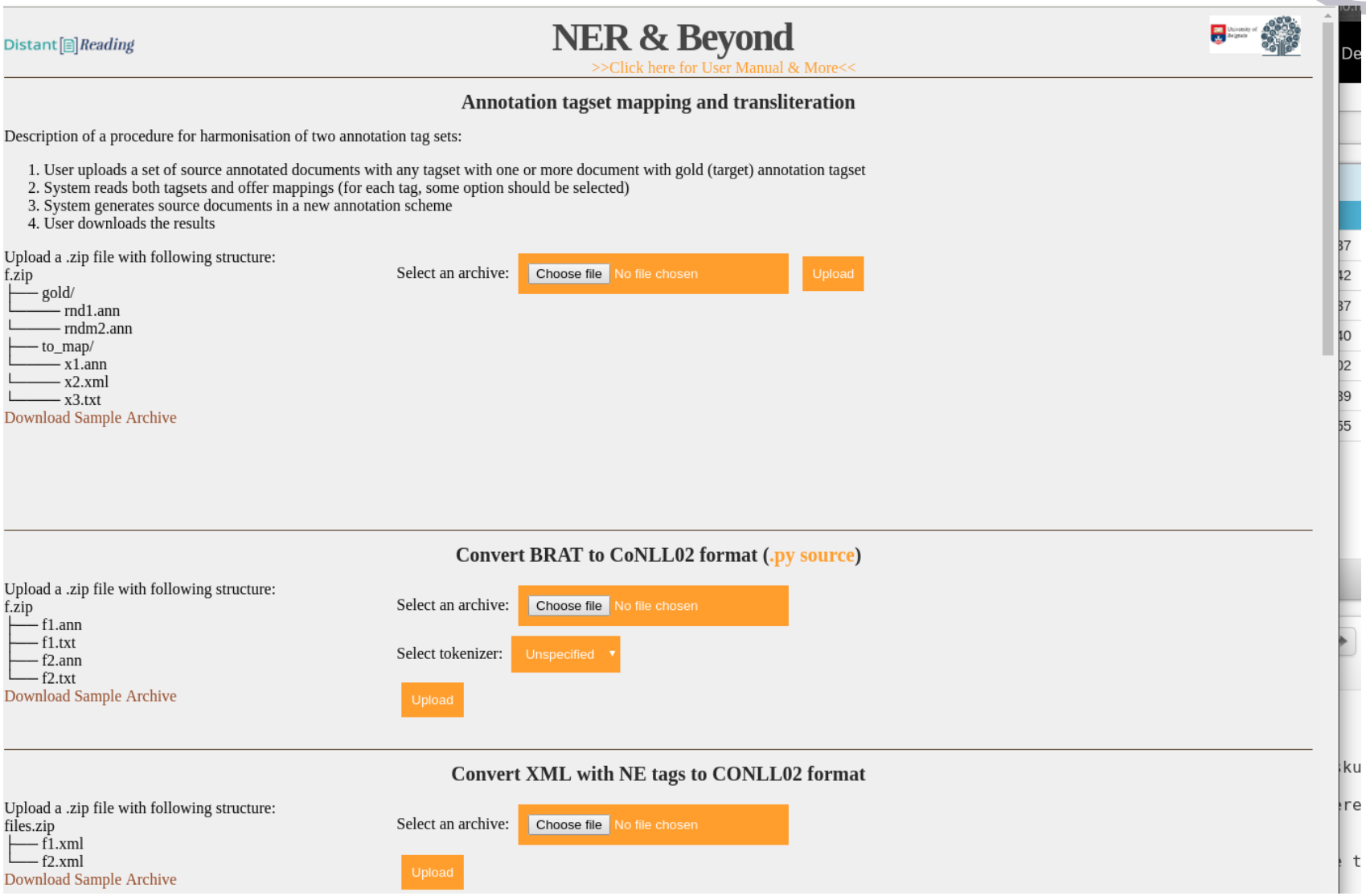


Classics vs. non-classics?

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

Further work

- Include more languages (German, Italian, Croatian,...)
- Testing more tools for automatic annotation and models
- Resolve open issues in format harmonisation (conversions restriction)
- Compare results from different tools with our gold datasets
- Entity Linking (for places)
- A way forward would be to build on a service-based digital infrastructure such as GERBIL (General Entity Annotation Benchmark Framework - http://aksw.org/Projects/GERBIL.html)

References and Bibliography

- Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii (2012). BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session* at EACL 2012.
- Honnibal, Matthew, and Ines Montani (2017). spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.
- Finkel, Jenny Rose, Trond Grenager, and Christopher Manning (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of ACL 2005*, pp. 363-370.
- Šandrih, Branislava, Cvetana Krstev and Ranka Stankovic (2019). Development and Evaluation of Three Named Entity Recognition Systems for Serbian - The Case of Personal Names, *Proceedings of RANLP 2019, NLP in a Deep Learning World*, Eds. Angelova,G.et.al., .pp. 1061-1069

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, Ioanna Galleron, 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!