Presentation of the NER progress
WG2

Diana Santos and the NER group

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

Distant Reading

22 January 2019

Participants so far

- Leader: Ranka Stanković
- Other initial members: Ioana Galleron, Diana Santos
- Other members: Carmen Brando (EHESS Paris), Francesca Frontini (Université Paul-Valéry)
- Annotators:
  - Norwegian: Michael Preminger, Tonje Vold, Kim Tallerås (OsloMet)
  - Hungarian: Emma Takács (Eötvös Loránd University)
  - Slovenian: Tomaž Erjavec
  - Portuguese: Diana Santos
  - French: Ioana Galleron, Carmen Brando, Francesca Frontini
  - English: Ranka Stanković, Diana Santos
- Programmers: Ranka Stanković, Branislava Šandrih (University of Belgrade), Diana Santos

Distant Reading

Work done

- setting up of initial guidelines
- installation of BRAT at the University of Belgrade for manual annotation http://147.91.183.8:12345/
- annotation of the samples selected by Lou for EN-FR-PT-HU-NO-SL
- discussion and refining of annotation guidelines
- creation of a set of difficult sentences to clarify the problem and to measure "theoretical disagrement"
- creation of several services to transform between formats and produce ConLL format (http://147.91.183.8:12346/) and invoke several "multilingual" NER systems (SpaCy, StanfordNER) at Univ. Belgrade
- scripts to provide visualization of the results

Distant Reading

Work to be continued

- Installing and/or running other NER systems
- Converting their output to ConLL
- Perform comparison to the manual annotations
- Getting feedback from WG3
- Writing up what was learned, what were the main challenges, what should be the next steps

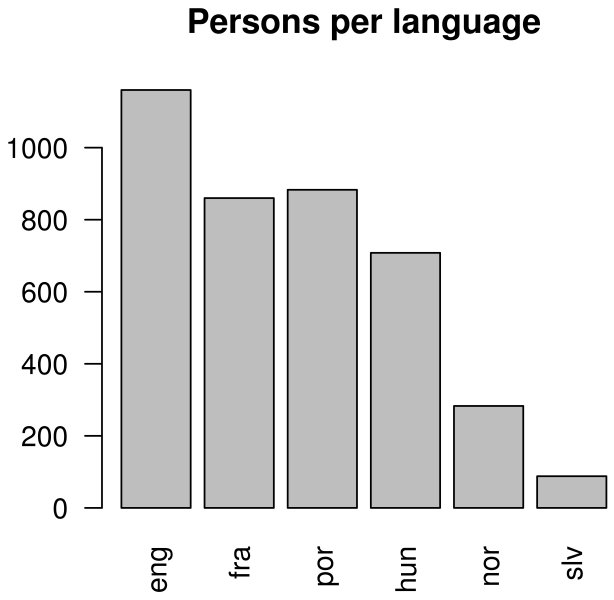
Distant Reading

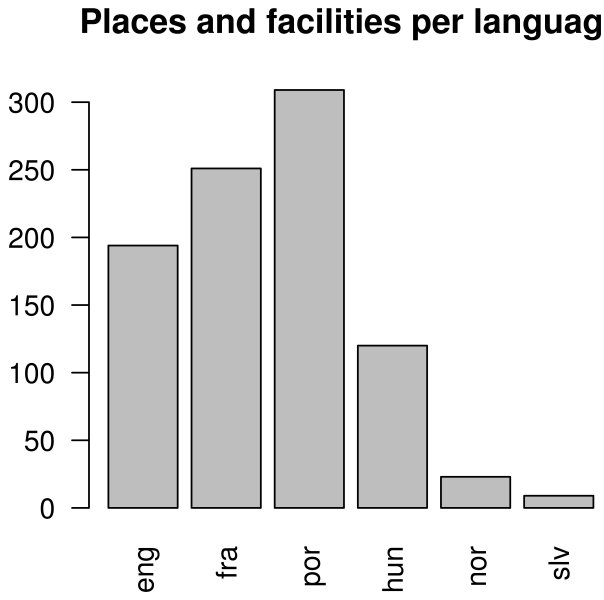
First results: quantitative

	eng	fra	por	hun	nor	slv
Words	53093	52031	53958	58825	52900	54907
NE	1562	1191	1313	1259	373	112
DEMO	54	26	6	0	0	0
EVENT	8	3	8	6	0	0
FAC	71	72	91	52	0	0
MISC	0	0	0	117	1	0
ORG	38	17	16	15	9	0
PERS	1160	860	883	708	283	88
PLACE	123	179	218	68	23	9
ROLE	85	11	36	286	46	10
WORK	23	17	51	5	2	0

Distant Reading

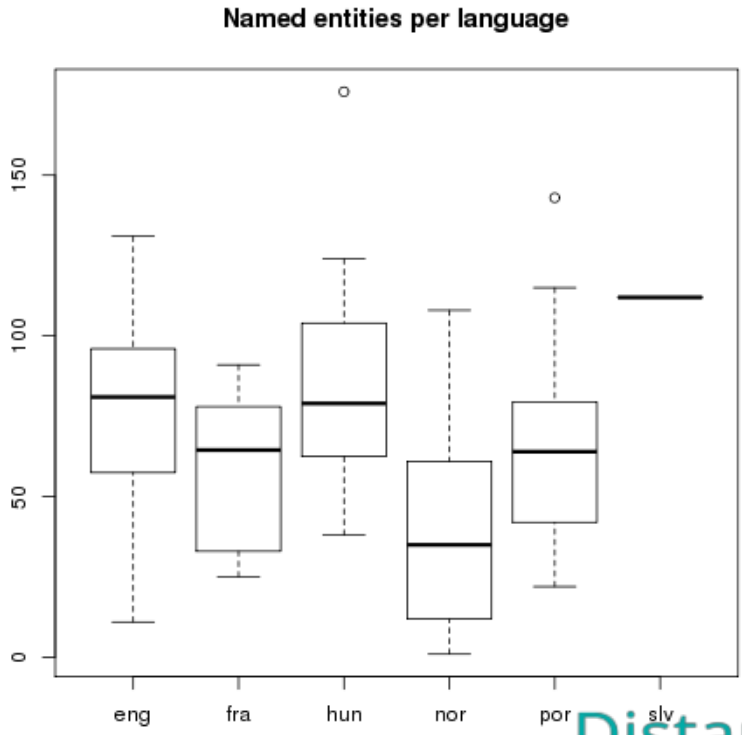
Some totals, graphically





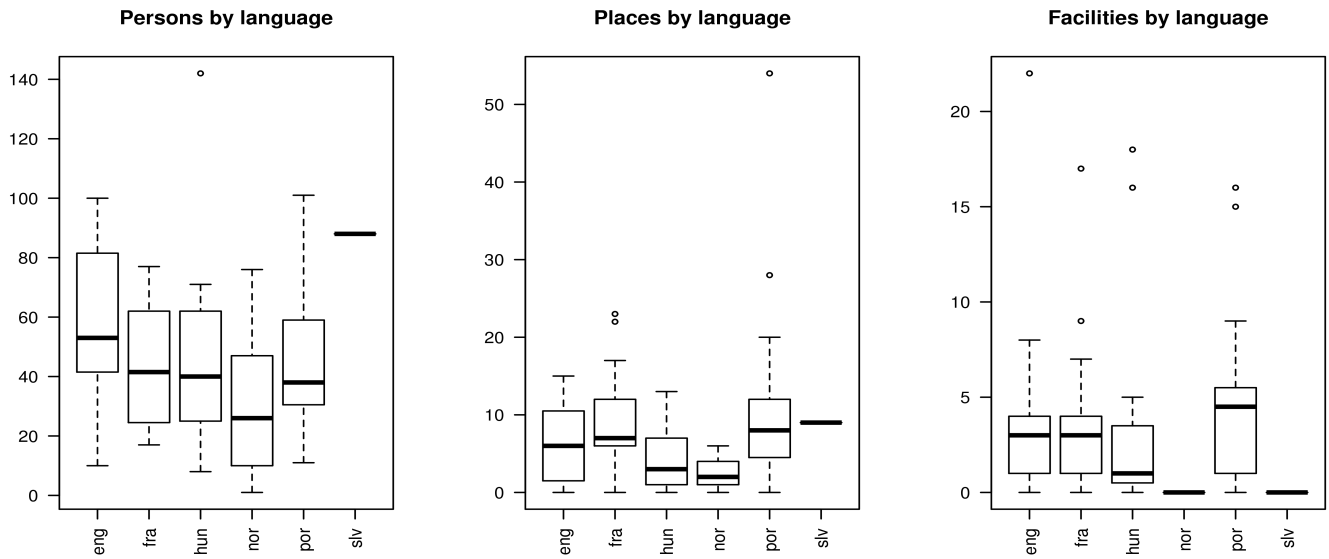
Distant Reading

Distribution of NEs



Distant Reading

Distribution of People, Places and Facilities



Distant Reading

Second results: qualitative

Different morphological and pragmatic conventions make a crosslingual definition of named entity difficult.

- What is the difference between a title and a profession? Should people who are mentioned by their titles be tagged as NE?
- Do cyclic events classify as named entities (*Christmas*)?
- Are demonyms named entities? (*Scottish*, *Nordmennane*, *Midi gascon*, *Saxon*...)
- Should abstractions based on fictional or religious origin count as named entities? (*Trojan horse*, *Éva lánya* (the daughter of Eve), *cour du roi Pétaud* (a court where nobody has authority), *como o senhor de la Palisse* (obvious statements))

Distant Reading

Second results: qualitative (contd.)

- Is it important to distinguish fictional from real people in e.g. historical novels?
- Is it correct to distinguish between French wines and German waters? (the first being a WORK, the second something from a PLACE)?
- Was it right to annotate gods and animals under PERS?

Distant Reading

Issues with tokenization

A big concern. Most NER systems we tried do their own tokenization. Two possible ways out:

- use character offsets in the file
- merge all tokens in NEs and compare afterwords

Distant Reading

What is BRAT and why it was chosen

NER Guidelines | Format2Format | Path: /eltec/eng/ENG18450_Disraeli_sa

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <samples n="ENG18450">
3 <sample><p n="ENG18450657">“I shall be delighted; I hope he will come to Marney in October. I keep the blue ribbon cover for him.”</p>
4 <p n="ENG18450658">“What you suggest is very just,” said Egremont to Lady Maud. “If we only in our own spheres made the exertion, the general effect would be great. Marney Abbey, for instance, I believe one of the finest of our monastic remains,—that indeed is not disputed—diminished yearly to repair barns; the cattle browsing in the nave; all this might be prevented, If my brother would not consent to preserve or to restore, still any member of the family, even I, without expense, only with a little zeal as you say, might prevent mischief, might stop at least demolition.”</p>
5 <p n="ENG18450659">“If this movement in the church had only revived a taste for Christian architecture,” said Lady Maud, “it would not have been barren, and it has done … much more! But I am surprised that old families can be so dead to our national art; so full of our ancestors, their exploits, their mind. Indeed
```

PLACE PERS [Male] PERS [Female] FAC [Religious][Suprastructure] WORK PERS [Female]

Distant Reading

Examples of BRAT output: file .ann

hun/01708 sample.ann

```
A1 ROLE T1 Profession
T2 PERS 225 229 Fadd
A2 PERS REALITY T2 Real
A3 PERS GENDER T2 Male
T3 ROLE 249 257 orvoshoz
A4 ROLE T3 Profession
T4 ROLE 338 342 urfi
A5 ROLE T4 Nobility
T5 ROLE 479 484 orvos
A6 ROLE T5 Profession
T6 ROLE 536 551
orvosnövendéket
```

por/PT0011.eltec sample.ann

```
T2 PERS 274 292 Jorge Mendes Nobre
T3 PERS 387 393 Cristo
T4 PERS 495 507 Maria Isabel
T5 PERS 1350 1361 George Sand
T6 PERS 1377 1388 George Sand
T7 WORK 1443 1448 Lélia
T8 WORK 1472 1485 Monte-Revêche
T9 PERS 1591 1598 Mussets
T10 PERS 1774 1779 César
T11 PERS 1802 1806 Sand
T12 PERS 2122 2126 Deus
```

Distant Reading

Examples of BRAT output converted to CoNLL format

```
<sample> O
<p n=PT0003611» O
- O
À O
saúde O
do O
doutor B-Pers
Mem I-Pers
Bugalho I-Pers
. O
</p> O
```

Distant Reading

List of difficult cases

To be parallel-annotated by several people

- To take the bull by its horns, and ask everybody to take a stance
- To illustrate the several alternatives
- To document for the rest of the COST members (and maybe the world) the non-trivial issues

Suggestion: To be analysed on paper by the (literary, all) participants during the COST meeting?

Distant Reading

Obrigada!