



Annotating your texts, manually and automatically

Distant  *Reading*

COST Action “Distant Reading” (COST Action CA16204) training school
March 22, 2021

Programme

- ❑ Introduction to the COST action annotation campaign and datasets 12.30-13.05 (Francesca Frontini, Ranka Stanković)
 - ❑ Motivation and annotation guidelines
 - ❑ Presentation of tools used for manual and automated annotation and testing results
- ❑ Break 10 min (13:05-13:15)
- ❑ Practical exercise (13:15-13:50) (Ranka Stanković)
 - ❑ Brat + NerBeyond (20 min)
 - ❑ Practical exercise (15 min)
- ❑ Break 10 min (13:50-14:00)
- ❑ CLARIN PL
 - ❑ CLARIN tools (1h) : 14h - 15h
 - ❑ Practical exercise

Francesca Frontini
ILC CNR & CLARIN ERIC

The ELTEc NER annotation campaign

- Requirements
 - Why ad hoc annotations
- Annotation guidelines
 - What to annotate (categories)
 - How to annotate it (nested annotations, include determiners, ...)
- The process
 - Alignment of annotation for the various languages
 - Choice and configuration of annotation tool

The ELTEc NER annotation campaign

WG3 - “Literary Theory and History” expressed a list of desiderata

<https://www.distant-reading.net/wg-3/>

WG2 - “Methods and tools” adapted the guidelines and carried out the annotation

<https://www.distant-reading.net/wg-2/>

Objective: test existing NER tools in various languages and see how adapted they are.

Desiderata

- A first set of desiderata emanates from the idea that a novel is an epic set in the private space of a bourgeois home, something which demands researchers to be able to **detect indicators of social structure and roles, such as honorifics, names of professions, etc.**
- Another set of research topics touches upon questions about identity, otherness, but also the distinction between urban and rural spaces, which require the annotation of **demonyms**, as well as a **higher granularity in the annotation of toponyms**, to facilitate detecting different types of locations (cities vs villages and countryside).
- Finally, questions about cultural references, role models and cosmopolitanism can only be answered if references to **works of art, authors, folklore and periodical publications are detected**.

Desiderata

- A first set of desiderata emanates from the idea that a novel is an epic set in the private space of a bourgeois home, something which demands researchers to be able to **detect indicators of social structure and roles, such as honorifics, names of professions, etc.** > **ROLE and DEMONYM**
- Another set of research topics touches upon questions about identity, otherness, but also the distinction between urban and rural spaces, which require the annotation of **demonyms**, as well as a **higher granularity in the annotation of toponyms**, to facilitate detecting different types of locations (cities vs villages and countryside).
- Finally, questions about cultural references, role models and cosmopolitanism can only be answered if references to **works of art, authors, folklore and periodical publications are detected**.

Desiderata

- A first set of desiderata emanates from the idea that a novel is an epic set in the private space of a bourgeois home, something which demands researchers to be able to **detect indicators of social structure and roles, such as honorifics, names of professions, etc.** > **ROLE and DEMONYM**
- Another set of research topics touches upon questions about identity, otherness, but also the distinction between urban and rural spaces, which require the annotation of **demonyms**, as well as a **higher granularity in the annotation of toponyms**, to facilitate detecting different types of locations (cities vs villages and countryside) > **PLACE attributes**
- Finally, questions about cultural references, role models and cosmopolitanism can only be answered if references to **works of art, authors, folklore and periodical publications are detected.**

Desiderata

- A first set of desiderata emanates from the idea that a novel is an epic set in the private space of a bourgeois home, something which demands researchers to be able to **detect indicators of social structure and roles, such as honorifics, names of professions, etc.** > **ROLE and DEMONYM**
- Another set of research topics touches upon questions about identity, otherness, but also the distinction between urban and rural spaces, which require the annotation of **demonyms**, as well as a **higher granularity in the annotation of toponyms**, to facilitate detecting different types of locations (cities vs villages and countryside) > **PLACE attributes**
- Finally, questions about cultural references, role models and cosmopolitanism can only be answered if references to **works of art, authors, folklore and periodical publications are detected.** > **WORK and its attributes**

The Tagset

The mandatory tag set used here combines the 8 categories

PERS, ROLE, DEMO, ORG, PLACE, FAC, WORK and EVENT, while MISC is optional.

Each category tag can be annotated with attributes, but this annotation is not mandatory.

The guidelines

Tag	Brief Description	Attributes (optional)	Comment	Examples (separated by ;)
PERS	Person names	real/fiction male/female/ collective	Proper names of people including first names, last names, individual or family names, fictional names and unique nicknames. This applies also to gods. Generational markers (Jr., VIII), and royal titles (Queen, Sir are included). Honorific titles (Mr., Mrs., Miss, Ms, Dr, Prof) are included when they occur followed by the proper name. Category to be used for named animals too.	Anna Eleanor Roosevelt Jr.; St. Nicolas; King George V; Princess Diana; Harry Potter; Psammead; Thanatos; God; Christ; Lord Eugen de Vere;

The guidelines

Tag	Brief Description	Attributes (optional)	Comment	Examples (separated by ;)
ROLE	Names of posts and job titles	profession nobility office military	profession (doktor, profesor, sailor, nurse) nobility (Lord, Viscount, Baron, Earl) office (President, Governor,...) military (Military rank such as Colonel, marechal)	Archbishop of Canterbury; Sa Majesté Britannique: marquez de Cannaes; Emperor; procurador de Celorico;

The guidelines

Tag	Brief Description	Attributes (optional)	Comment	Examples (separated by ;)
ORG		committee government workplace administrative political religious education	Names of companies, political parties, educational institutions, sport teams, hospitals, museums, libraries etc. Especially, hotels, museums, hospitals, libraries, churches and temples, commercial facilities,	White House; Green Party; Opera;

The guidelines

Tag	Brief Description	Attributes (optional)	Comment	Examples (separated by ;)
PLACE	Location	continent country region city village mountain waterbody astronym	Names of locations including celestial bodies, stars, continents, mountains, oceans, coasts, rivers, lakes, borders,	West Balkan; Europe; South Region

The guidelines

Tag	Brief Description	Attributes (optional)	Comment	Examples (separated by ;)
WORK	titles of	book song play newspaper paintings sculptures	Titles of books, songs, plays and other creations	
EVENT		natural disaster political war sport culture business	Named events and phenomena including natural disasters, hurricanes, revolutions, battles, wars, demonstrations, concerts, sports events, etc. Both cyclic one-off events	Christmas; Napoleonic Wars

The guidelines

Tag	Brief Description	Attributes (optional)	Comment	Examples (separated by ;)
FAC	Facility (human-made)		buildings, airports, roads, stations, infrastructures (bridges and streets), monuments, hospitals, named ships, trains	Opera; Theater, Prison; café; Main street
MISC optional	Temporal Quantitative Other		Date, time, ordinals, measures,...	
DEMO	Names of kinds of people	national regional political		Frenchwoman; German; Parisiens; Midi Whigs, Reds, Christian

Annotation in context

The same word or group of words can have different meanings, as we see with the Opera example:

I work for the Opera. Opera=ORG

I met you at the Opera. Opera=FAC

Nested annotations:

<PERS><ROLE>directeur</ROLE>de l'Opéra</PERS>

Choice of annotation tool

- **Brat** - <https://brat.nlplab.org/>

- Multiple users
- Customizable categories and annotation feature
- Export format compatible with other tools
- Parallel annotation and comparison features (inter annotator agreement)

Other tools similar tools are available, such as **Inception**
(<https://inception-project.github.io/>)

Annotation interface and examples

The full annotation

http://brat2.jerteh.rs/index.xhtml#/eltec-manual/eng/ENG18450_Disraeli_sample

A simplified version for testing purposes:

http://brat2.jerteh.rs/index.xhtml#/eltec-simplified/eng/ENG18450_Disraeli_sample

http://brat2.jerteh.rs/index.xhtml#/eltec-simplified/fra/rd0019_sample

Ranka Stanković
University of Belgrade

Outline

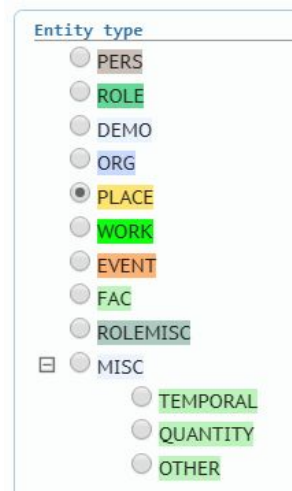
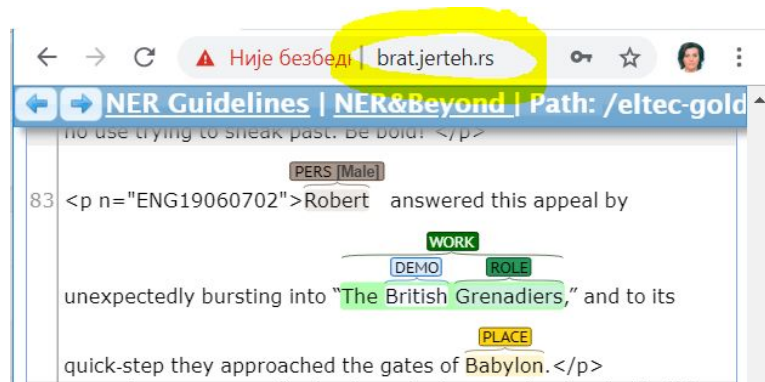
- Software infrastructure - presentation on tools related to NER
 - BRAT - manual annotation
 - Formats and transformations NER & Beyond
- Annotation campaigns
 - dataset preparation (all languages)
 - dataset publishing (txt+ann)
 - manual annotation or correction of automatic annotations
 - detailed and simplified annotation
- Inter-annotator agreement
 - why?
 - how?
- Manual annotation vs automatic annotations
 - problems, testing and comparing issues

Tools used for manual annotation and transformation

- Brat <https://brat.nlplab.org/> rapid annotation tool online, flexible environment for collaborative text annotation
- Source code: <https://github.com/nlplab/brat>
 - Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012, April). BRAT: a web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 102-107).
- Manual annotation for ELTEC
 - available on <http://brat.jerteh.rs/> collection eltec-gold,
 - according to [guidelines](#), where overview of manual annotation is given.
- Guidelines were inspired by WG3 [Research questions-NER](#) -Antwerpen.
- Set of tools is combined in the portal <http://nerbeyond.jerteh.rs/> with [tutorial](#)

Annotation text samples

- 9 languages:
 - Czech, English, French, German, Hungarian, Norwegian, Portuguese, Serbian and Slovene
- "Lightweight semantic annotation" defined marking:
 - persons, demonyms, professions and other roles, works, places, facilities and organizations



Annotation of text samples

- 1st: detailed annotation with properties:
<http://brat.jerteh.rs/index.xhtml#/eltec-manual/>
- 2nd: simplification (automatic), adapted to tools for automatic NER:
<http://brat.jerteh.rs/index.xhtml#/eltec-simplified/>

avec leur goût dans un panier, et toujours en bout de cheminée qui passait / p

<p n="FR00201652">C'étaient nos élèves. Mme Ouly leur apprenait des cantiques ; moi, je les initiais aux mystères de l'alphabet. J

<p n="FR00201653">Quelquefois aussi, quand « bonne amie » avait sa goutte, c'était moi qui balayais la classe, besogne bien peu digne d'

servi et la mère Jacques qui m'attendait... Après dîner, quelques tours de as, puis la veillée au coin du feu... Voilà t

l'oncle Baptiste ; M. Eyssette voyageait toujours pour la Compagnie vin

<p n="FR00201654">Les affaires n'allaient pas trop mal. Les dettes de Lyon étaient aux trois quarts payées. Dans un an ou deux, tout sera

<p n="FR00201655">Moi, j'étais d'avis, en attendant, de faire venir Mme Eyssette à l'hôtel Pilois avec nous, mais Jacques n

Entity attributes

PLACE: ?

PLACE: ?

PLACE: Mountain

PLACE: City

PLACE: Country

PLACE: Region

PLACE: Waterbody

PLACE: Village

PLACE: Astronym

PLACE: Continent

Entity attributes

ORG: ?

ORG: ?

ORG: Government

ORG: Administrative

ORG: Committy

ORG: Political

ORG: Workplace

ORG: Education

ORG: Religious

Entity attributes

EVENT: ?

EVENT: ?

EVENT: Bussiness

EVENT: Natural_disaster

EVENT: Political

EVENT: Culture

EVENT: Sport

EVENT: Religious

EVENT: War

The corpus

- <https://github.com/COST-ELTeC/WG2-Sample>
- 5 random passages of 400 white space-delimited tokens taken from 20 novels from each language, manually annotated using brat
- Thank Lou Burnard for preparing all samples
- Samples encoded as TEI documents level 1
(<https://distantreading.github.io/Schema/eltec-1.html>)

Page break <pb/>

Title page <div type="titlepage"> within <front>

Authorial preface, foreword, appendix, etc <div type="liminal"> within <front> or <back> as appropriate
volume, chapter etc. <div> nested as necessary within <body>

Heading or title <head> at start of <div>; <trailer> at end

Running title/page footer Omitted

Prose paragraph or list item <p>

Verse line <l>

deu
eng
fra
hun
ita
nor
por-1
por
slv
slv_deduplicated
srp

Towards level 2

<https://distantreading.github.io/Schema/eltec-2.html>

... all existing elements are retained and two new elements [<s>](#) and [<w>](#) are introduced to support segmentation of running text into sentence-like and word-like sequences respectively. Individual tokens are marked using the [<w>](#) element, and decorated with one or more of the TEI-defined linguistic attributes pos, lemma, and join. Both words and punctuation marks are considered to be ‘tokens’ in this sense, although the TEI suggests distinguishing the two cases using [<w>](#) and [<pc>](#) respectively.

The [<s>](#) (segment) element is used to provide an end-to-end tessellating segmentation of the whole sequence of [<w>](#) elements, based on orthographic form. This provides a convenient extension of the existing text-body-div hierarchy within which tokens are located.

Each [<s>](#) element can contain a sequence of [<w>](#) elements, ...

Element [<rs>](#) (referring string) for the encoding of any form of entity name, such as a Named Entity Recognition procedure might produce.

Preparation of text collection

- Files *.txt + *.ann
- http://brat.ierteh.rs/index.xhtml#/eltec-simplified/eng/ENG18450_Disraeli_sample

```
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
mischiefs, 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 so much more!
But I am surprised that old families can be so dead to our national art; so full of our
```

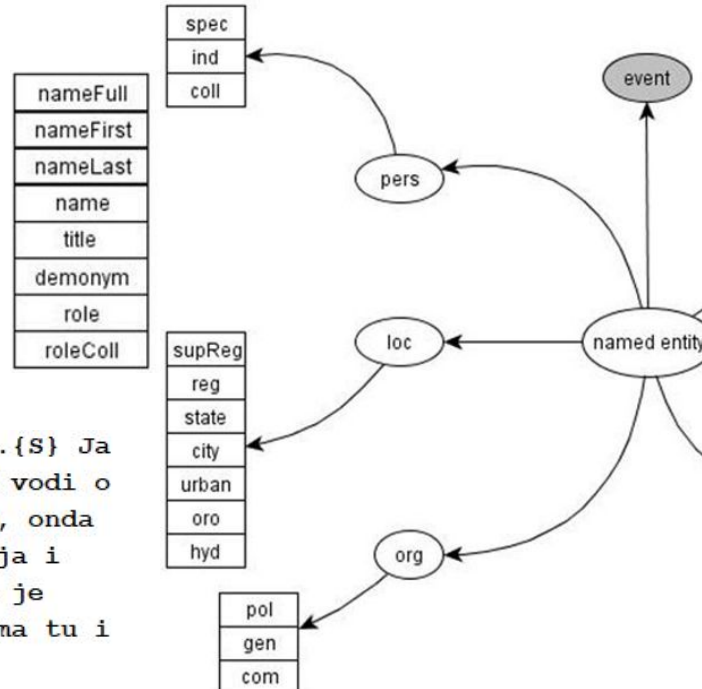
```
ENG18450_Disraeli_sample.txt ENG18450_Disraeli_sample.ann
1 T904 → LOC 135 141 → Marney
2 T907 → PERS 254 262 → Egremont
3 T908 → PERS 266 275 → Lady Maud
4 T909 → LOC 362 374 → Marney Abbey
5 T9014 → WORK 861 883 → Christian architecture
6 T9015 → PERS 891 900 → Lady Maud
7 T9018 → PERS 1255 1263 → Egremont
8 T9022 → PERS 1566 1575 → Lady Maud
9 T9026 → PERS 1920 1928 → Egremont
10 T9028 → PERS 2254 2263 → Lady Maud
11 T9033 → PERS 2468 2480 → Lady Mowbray
12 T9034 → PERS 2482 2490 → Egremont
13 T9036 → PERS 2576 2587 → Hippocrates
14 T9038 → DEMO 2659 2666 → English
15 T9040 → DEMO 2719 2725 → Norman
16 T9041 → DEMO 2742 2747 → Saxon
17 T9042 → LOC 3027 3034 → England
18 T9048 → PERS 3349 3354 → Dandy
19 T9049 → EVENT 3357 3377 → the National Holiday
```

```
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
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
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 so much more! But I am s
```

SerbianNER example of xml embedded annotation

`<p>Već sam dva meseca u <top.gr>Beogradu</top.gr>, a ništa nisam pisala.{S} Ja mnogo volim <top.gr>Beograd</top.gr>.{S} Tu te niko ne poznaje, niko ne vodi o tebi računa i kad si ti pošten čovek, a nemaš ni na koga da se obazireš, onda imaš punu slobodu, da živiš kako voliš i da radiš šta hoćeš.{S} Gimnazija i <org>Velika škola</org> – kakva razlika.{S} Volim moju školu.{S} Ona mi je najdraža kuća u <top.dr>Srbiji</top.dr>.{S} Svega znanja i sve lepote ima tu i`

`<p>... jer je stiže sudbina materina – dotle nam <pers><role>gdica</role><persName.last>Janković</persName.last></pers> iznosi istoriju svih onih nemira, kidanja i buđenja neodređenih i potajnih sila kod <role>devojke</role> koja počinje da zre.{S} U to prvo doba kod <pers><role>gospodice</role><persName.last>Jele</persName.last></pers> sve je nekako drhtavo, nervozno, nejasno, jer „ona još ni sama ne zna šta hoće”.{S} Kad se tome još doda njezina nesrećna ljubav prema <persName.first>Nikoli</persName.first>, onda je to dovoljno da u takoj osetljivoj, „bolesnoj” ženskoj duši započne vrlo zanimiva drama.</p>`



Stanford format

- Manual annotation from row, unlabeled text or pre-annotated
- Stanford and spaCy used for pre-annotation of some samples
- Manual correction followed
- <PERSON> or <PERS> or <Pers.Full> or <Pers.Name> or <Pers.FirstName> or <Pers.LastName>...
- <LOCATION> or <LOC> or GPE or <Top.City> or ...

```
<?xml version="1.0" encoding="UTF-8"?>
<samples n="ENG18450">
  <sample><p n="ENG18450657">"I shall be delighted; I hope he
will come to <PERSON>Marney</PERSON> in October. I keep the
blue ribbon cover for him."</p>
  <p n="ENG18450658">"What you suggest is very just," said
Egremont to <PERSON>Lady</PERSON> <PERSON>Maud</PERSON>. "If
only in our own spheres made the exertion, the general effect
would be great. <PERSON>Marney</PERSON> <PERSON>Abbey</PERSON>
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
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>
  <p n="ENG18450659">"If this movement in the church had only
revived a taste for Christian architecture," said <PERSON>Lady</PERSON>
<PERSON>Maud</PERSON>, "it would not have been
barren, and it has done so much more! But I am surprised that
old families can be so dead to our national art; so full of
ancestors, their exploits, their mind. Indeed you and I have
excuse for such indifference Mr Egremont."</p>
  <p n="ENG18450660">"And I do not think I shall ever again be
justly accused of it," replied <LOCATION>Egremont</LOCATION>,
"you plead its cause so effectively. But to tell you the truth
I have been thinking of late about these things; monasteries
and so on; the influence of the old church system on the
```

Results of annotation

- How many NE annotated
 - per sample,
 - per language, or per collection, or per annotator,...
 - per NER type?
 - ...

NER stats on .ann files

Upload a .zip file with following structure:
f.zip

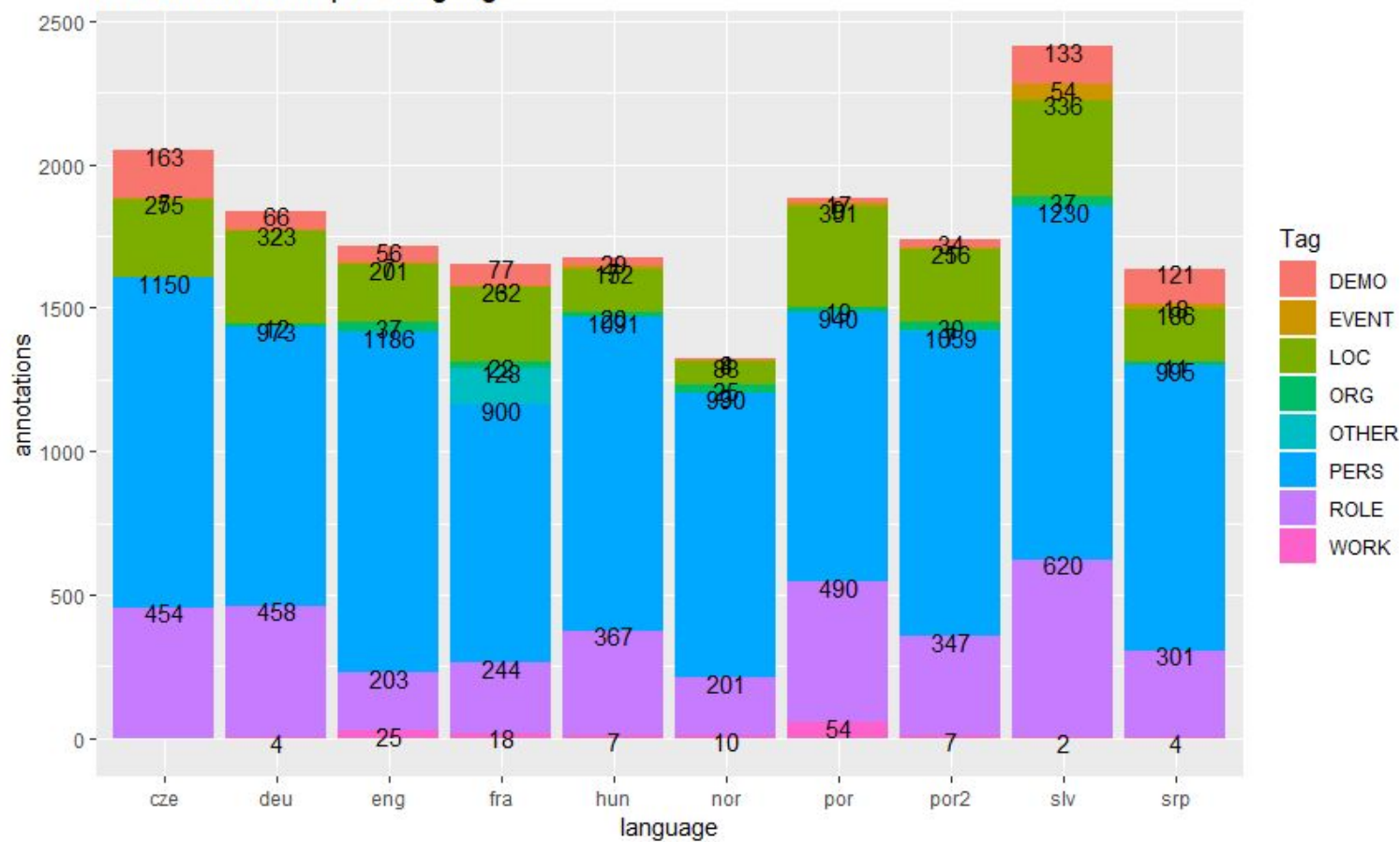
- eng/
 - f1.ann
 - f2.ann
- slv/
 - f1.ann
 - f2.ann

[Download Sample Archive](#)

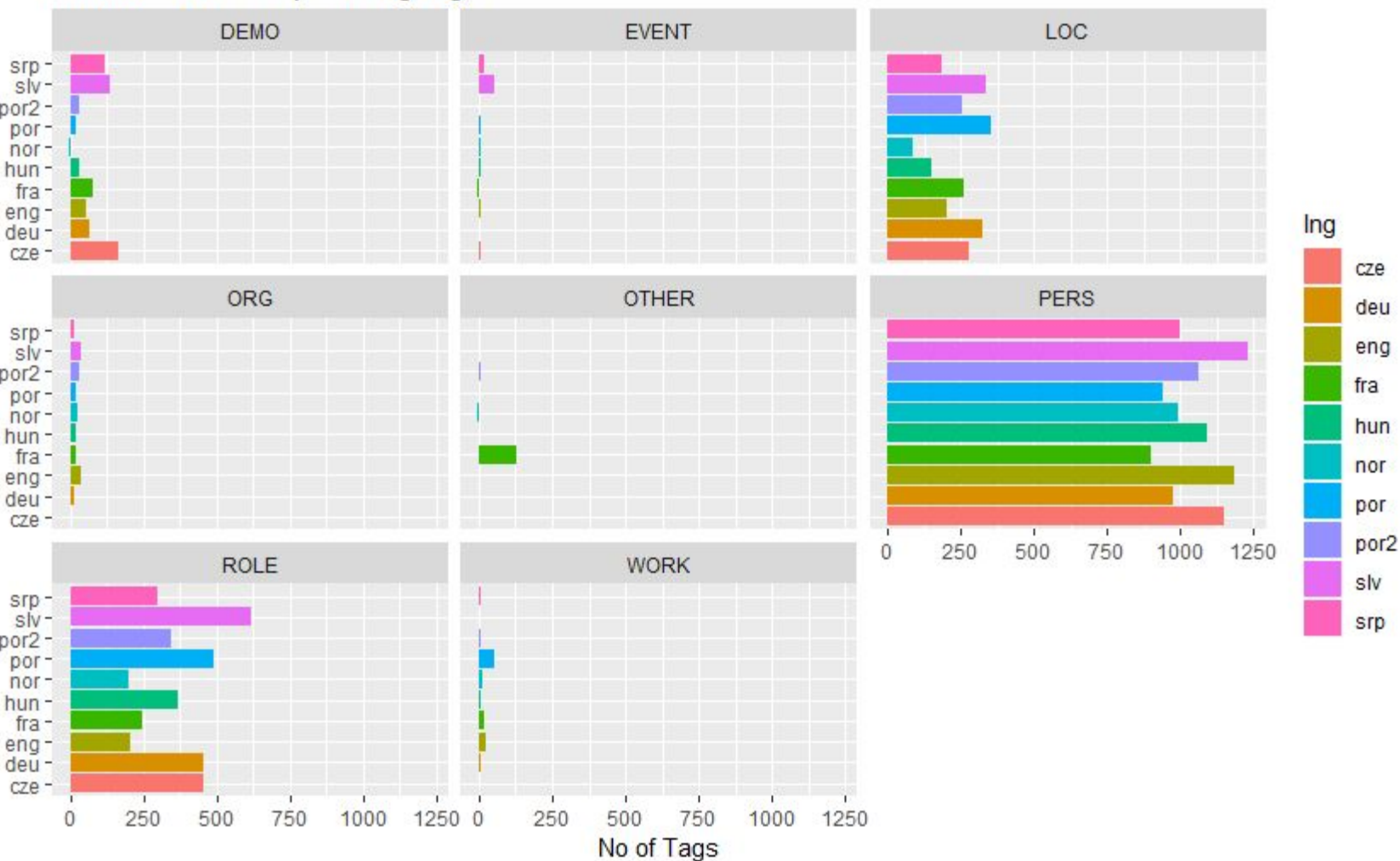
Select an archive: ner_stats (1).zip

Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, et al..
Named Entity Recognition for Distant Reading in ELTeC. CLARIN Annual Conference 2020, Oct
2020, Virtual Event, France. <https://hal.archives-ouvertes.fr/hal-03160438/document>

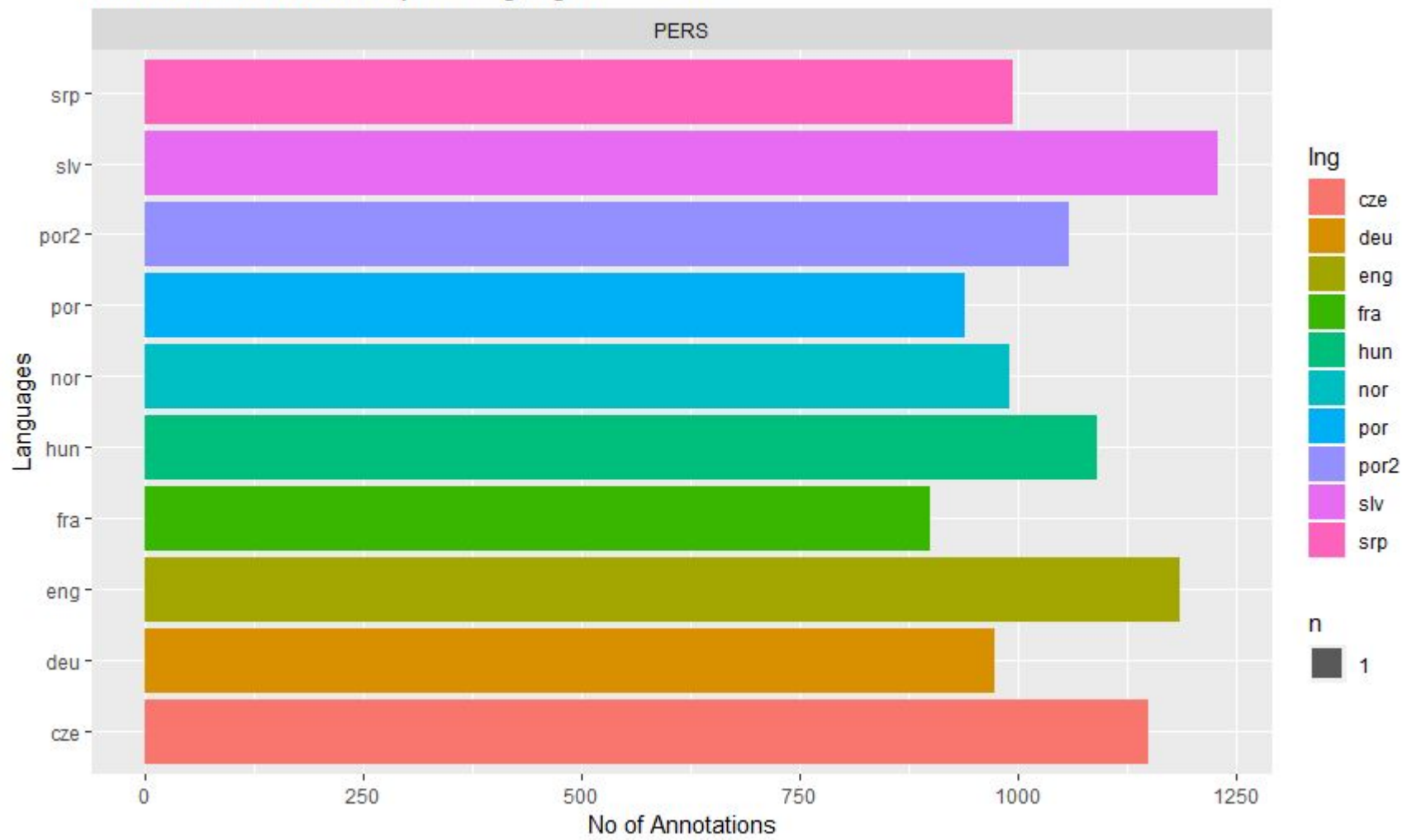
NER Annotations per Language



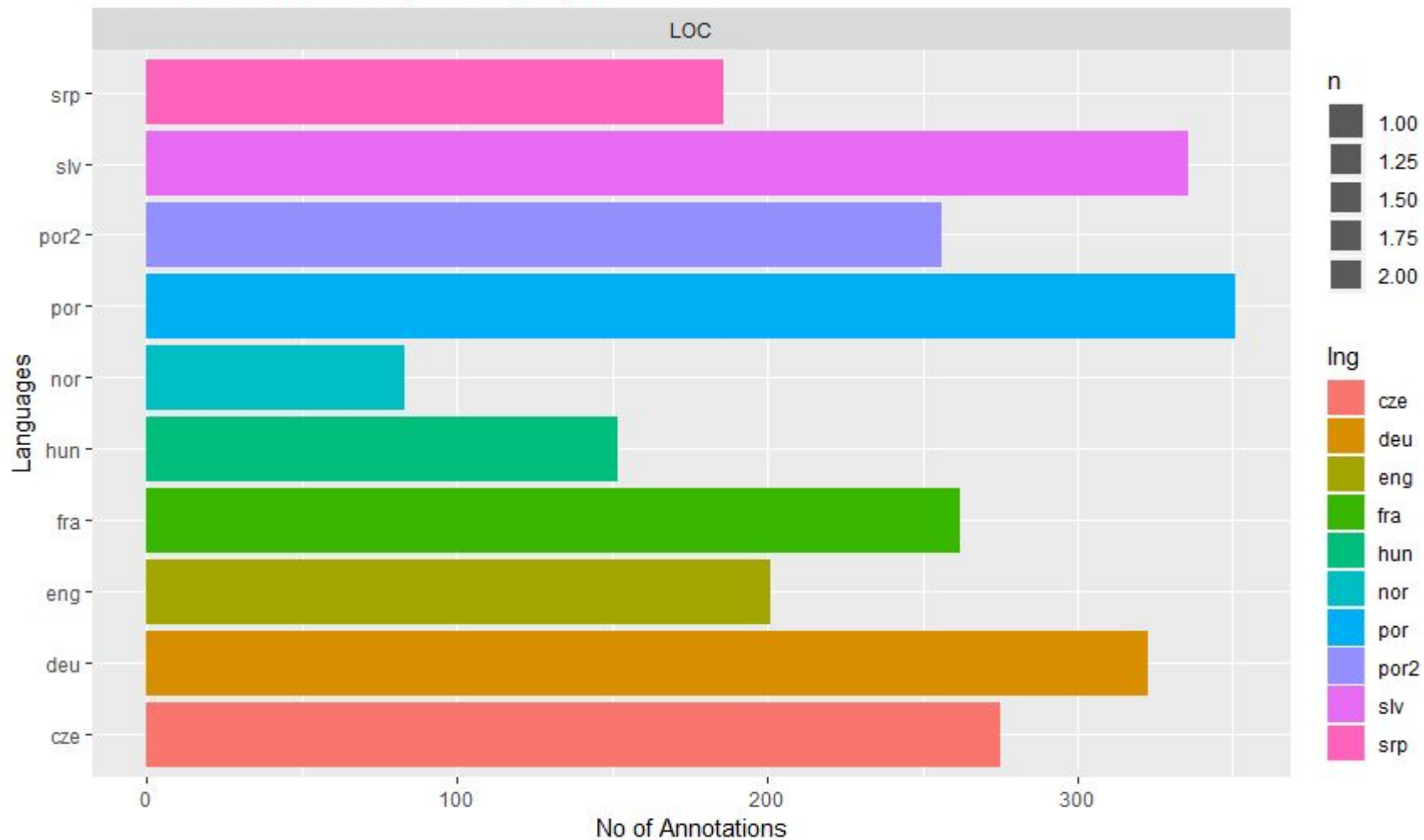
NER Annotations per Language



Number of Annotations per Language for PERS



Number of Annotations per Language for LOC



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

NER Tools that have been tested

English (ranka@rgf.rs, branislava.sandrih@fil.bg.ac.rs):

spaCy:

- /eltec-automatic/eng-spaCy-orig/ annotated with 4 class model: CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME, WORK_OF_ART
- /eltec-automatic/eng-spaCy-mapped/ annotation mapped to ELTeC tagset: EVENT->EVENT; GPE,LOC->PLACE; NORP->DEMO; ORG->ORG; PERSON->PERS; QUANTITY->QUANTITY; WORK_OF_ART->WORK;
- [Visual comparison](#) and example for [PERS](#) (spaCy: problem with < and >)

Stanford :

- /eltec-automatic/eng-Stanf-orig/ annotated with 4 class model: Location, Person, Organization, Misc
- /eltec-automatic/eng-Stanf-mapped/ annotation mapped to ELTeC tagset: Location->PLACE; Person->PERS; Organization->ORG; Misc->OTHER

NER Tools that have been tested

Portuguese (d.s.m.santos@ilos.uio.no):

NER-PALAVRAS on 5 full eltec-por novels amounting to 450,000 words together, with categories corresponding to PERSON; PLACE; ORG; WORK; DEMONYM; PROFESSION; DATE and OTHER, achieving a precision of 0.663 and 0.657 recall.

The process was as follows:

- run PALAVRAS-NER on the files
- use XML to BRAT program from <http://nerbeyond.jerteh.rs/>
- revise manually the output with BRAT, providing thus a golden set
- compare the two annotated files in BRAT format, using a simple Perl program

We are now studying the performance in order to improve the NER system, and also to check whether some of the results are due to different theoretical approaches about what a named entity is.

NER Tools that have been tested

French (carmen.brand@gmail.com ; francescafrontini@gmail.com ; ioana.galleron@gmail.com): Test planned with SEM:

- Tests done with SEM: <http://apps.lattice.cnrs.fr/sem/> on rd0019, still need to be confronted with BRAT annotations.
- Evaluation of french NER Spacy model "fr_core_news_md", below the results in terms of P, R, F.
- Simplifications: remove nested annotations (no more than 10 annotations out of the total), categories to match Spacy categories (PER, LOC, ORG, MISC).
- Next steps: work on the golden french corpus, perform specific training using tools such as Spacy or Google-Bert.

	P	R	F
PER	28,84	53,57	37,49
LOC	7,39	56,81	13,08
ORG	0,08	25,00	0,16
MISC	0,11	2,00	0,22

Testing NER Tools

Hungarian (palko.gabor@btk.elte.hu):

[NER of the Institute of Literary Studies](#) les than 50%

Serbian (ranka@rgf.rs, cvetana@matf.bg.ac.rs):

- Rule- and Lexicon- Based NER for Serbian [Srp-NER](#) has very detailed annotation
- SrpNER, spaCy and Stanford (models only PERS evaluated) :

Branislava Šandrih, Cvetana Krstev, Ranka Stanković, “Development and Evaluation of Three Named Entity Recognition Systems for Serbian - the Case of Personal Names”, in *Proceedings of the International Conference Recent Advances in Natural Language Processing - RANLP 2019*, 2-4 September 2019, Varna, Bulgaria, eds. G. Angelova et als., pp. 1061-1068, 2019. DOI: 10.26615/978-954-452-056-4_122 <https://www.aclweb.org/anthology/R19-1122.pdf>

- PLACE, ORG,... - further model training in progress

Other languages to come ...

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

<sample><p n="ENG1845036">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.</p>

<p n="ENG1845037">"Ah, Egremont! come and sit here," exclaimed more than one banqueter.</p>

<p n="ENG1845038">"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."</p>

<p n="ENG1845039">"How shall we all feel this time to-morrow?" said Egremont, smiling.</p>

<p n="ENG1845040">"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."</p>

<p n="ENG1845041">"Poor Cockie," said Mr Berners; "he has asked me to dine with him at the Clarendon on Saturday."</p>

<p n="ENG1845042">"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

<p n="ENG1845043">"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

Problems, solutions, points which should be discussed

- Main problems:
 - Compare (evaluate) results from different tools with our gold datasets, alignment of manual and automatic annotation non trivial (annotation - token alignment)
 - Align with morphosyntactic tagging and lemmatization (most tools have plain text as input, output standoff or inline)
 - Output from WG2 TEI, Conll, and/or ???
- Smaller issues:
 - Testing more tools for automatic annotation and models
Automatic annotation for other languages
 - Satisfy fine-grained wish-list tagset with additional attributes
(use of additional lexicons, knowledge base and ontologies?)
 - How to handle nested annotation
 - Accuracy will be needed per NER class

Practical work

<http://brat.jerteh.rs/>

Brat guest account for all course participants during training:

user: brat_guest
pswd: guest_brat_2021

collections:

1. http://brat.jerteh.rs/index.xhtml#/DR_NER_TrainingSchool-simplified/
2. http://brat.jerteh.rs/index.xhtml#/DR_NER_TrainingSchool-full/

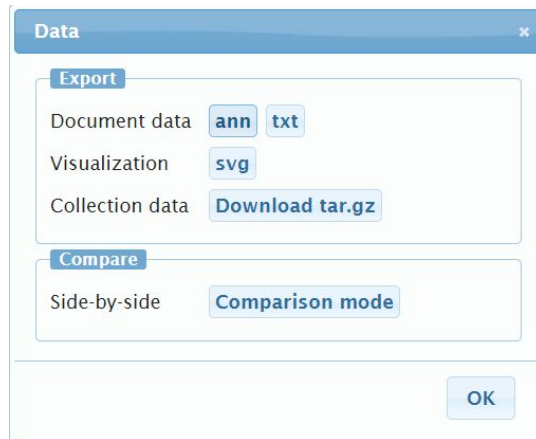
See later:

<https://zenodo.org/record/4274954>

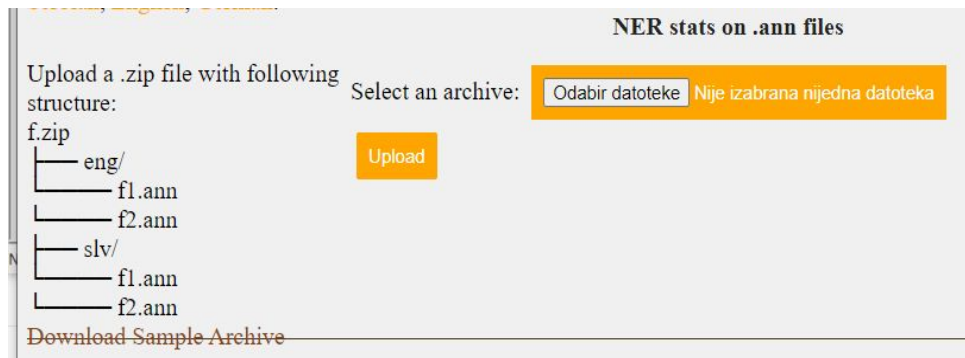
Lou, Burnard	Sample1	English, French, Italian, German
Natalija, Tomić	Sample2	Serbian, English, Russian, Ukrainian
Vojtěch, Malínek	Sample3, POL0092_Przedwiosnie_sample	Czech, English, German, Polish
Anders Skare, Malvik	Sample4	English, Scandinavian
Lovro, Škopljanač	Sample5	Croatian, English, German, Japanese
Katie, Mishler	Sample6	English, French, Spanish, Irish
Lauren, Cassidy	Sample7	English
Katrin, Horn	Sample8	English, German, Italian, Latin
Andrejka, Žejn	Sample9	Slovene, English, German, Croatian
Luiza, Marinescu	Sample10	Romanian, English, Bulgarian
Fotini, Koidaki	Sample11	Greek, English, Spanish
Jana-Katharina, Mende	Sample12, POL0016_Kariera_sample	German, English, French, Polish, Turkish, Russian, Hungarian
Susana, Sotelo Docío	Sample13	Spanish, Portuguese, English, German, French
Marek, Debnár	Sample14, POL0021_Paziowie_sample	Slovak, Czech, English, Polish, Russian, French
Ioana Alexandra, Lionte	Sample15	Romanian, French, English, German, Italian, Spanish
Alexandra, Olteanu	Sample16	Romanian, French, English
Lucreția - Elena, Pascariu	Sample17	Romanian, English, French
Cezary, Rosiński	Sample18, POL0070_Ogniem_sample	Polish, English
Lucija, Mandić	Sample19	Slovene, English, Croatian, Czech, Slovak, Polish, French

Practical work workflow

- Manually annotate sample
 - first simplified (15 min)
 - Download collection
DR_NER_TrainingSchool-simplified/
(brat - Data - Download tar.gz)
 - Statistical analysis from
<http://nerbeyond.jerteh.rs/>
Section NER stats on .ann files
 - optional full
 - Discussion (last 10 min)



The 'Data' window in the brat interface has a blue header with a close button. It contains two sections: 'Export' and 'Compare'. The 'Export' section has three rows: 'Document data' with buttons for 'ann' and 'txt', 'Visualization' with a button for 'svg', and 'Collection data' with a button for 'Download tar.gz'. The 'Compare' section has one row: 'Side-by-side' with a button for 'Comparison mode'. An 'OK' button is located at the bottom right of the window.



The 'NER stats on .ann files' section has a title bar. Below it, there is a text prompt 'Upload a .zip file with following structure:' followed by a tree diagram showing a folder 'f.zip' containing two subfolders 'eng/' and 'slv/', each with two files 'f1.ann' and 'f2.ann'. To the right, there is a label 'Select an archive:' followed by a text input field containing 'Odabir datoteke' and a message 'Nije izabrana nijedna datoteka'. Below the input field is an orange 'Upload' button. At the bottom, there is a link 'Download Sample Archive'.