

1. Motivation and background

- ❖ Funding: EU - CLARIN Bridging Gaps
- ❖ Grant No: CE-2022-2070
- ❖ Project URL: http://jerteh.rs/?page_id=2357



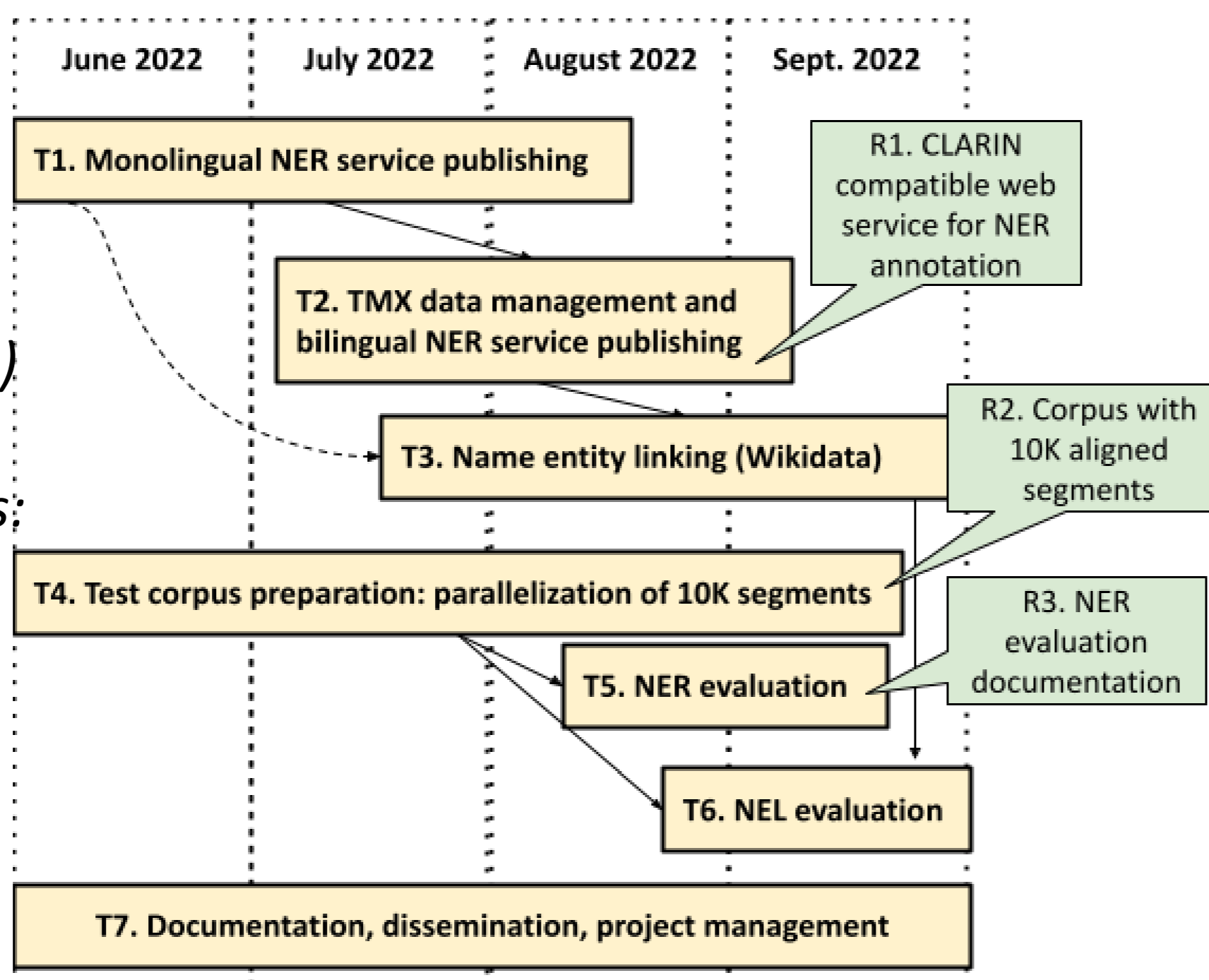
❖ Service URL: <http://ners.jerteh.rs/>

Goal: development of the CLARIN compatible NER web service for parallel texts with case study on Italian and Serbian, dubbed It-Sr-NER

- recognizing and classifying named entities in bilingual natural language texts in TMX (Translation Memory eXchange) file, e.g. Sr-It.

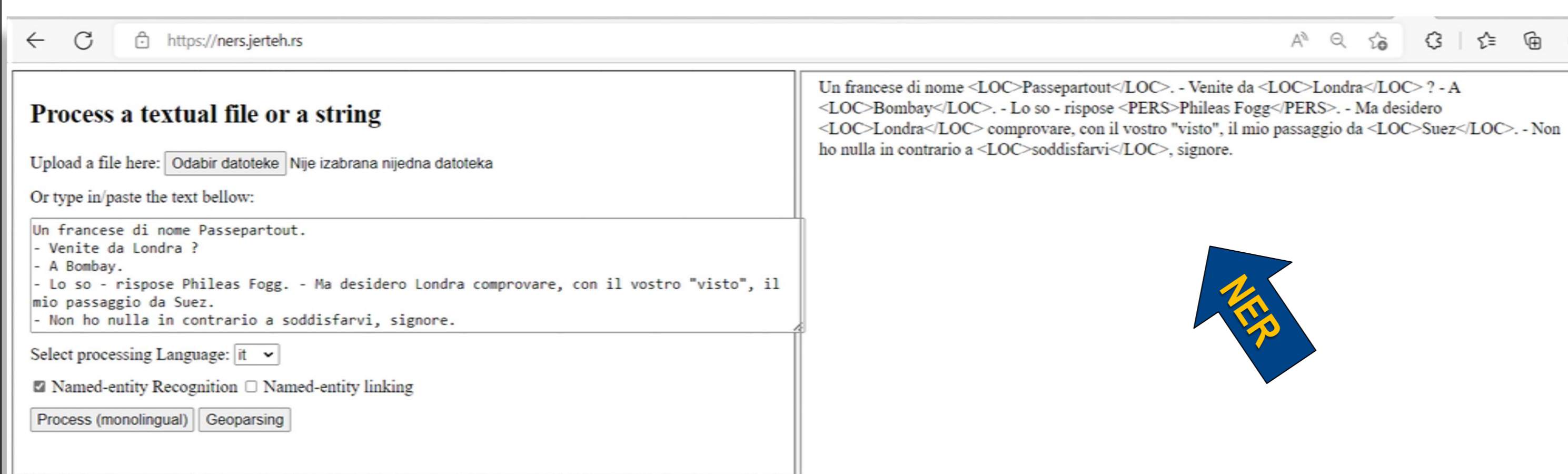
It-Sr-NER recognize six NER classes:

- demonyms (DEMO),
- works of art (WORK),
- person names (PERS),
- places (LOC),
- events (EVENT) and
- organizations (ORG).



2. Monolingual NER service publishing

- ❖ The user can
 - ✓ input a text or upload a text file in one of the offered languages
 - ✓ Choose: NER, NER+NEL, NEL, GEOPARSING (map)



NER

Un francese di nome <LOC>Passepartout</LOC>. - Venite da <LOC>Londra</LOC>? - A <LOC>Bombay</LOC>. - Lo so - rispose <PERS>Phileas Fogg</PERS>. - Ma desidero <LOC>Londra</LOC> comprare, con il vostro "visto", il mio passaggio da <LOC>Suez</LOC>. - Non ho nulla in contrario a <LOC>soddisfarvi</LOC>, signore.

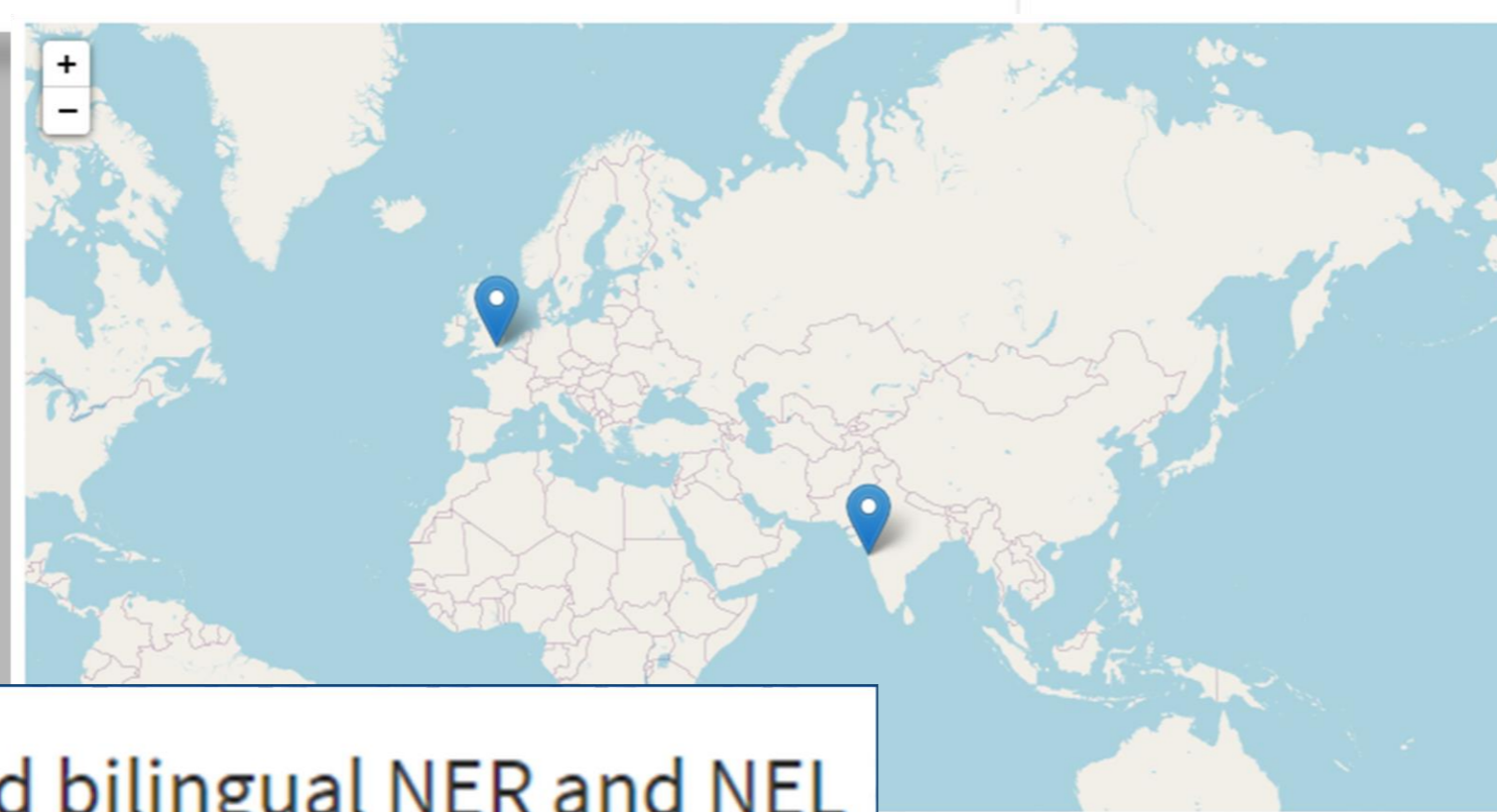
NER+NEL

Un francese di nome Passepartout. - Venite da <WDT ref="https://www.wikidata.org/wiki/Q84" label="LOC" desc="capital and largest city of the United Kingdom">Londra</WDT>? - <WDT ref="https://www.wikidata.org/wiki/Q48" label="LOC" desc="continent, mainly on the Earth's northeastern quadrant">A</WDT>? - <WDT ref="https://www.wikidata.org/wiki/Q1156" label="LOC" desc="capital city and district in Maharashtra, India">Bombay</WDT>. - <WDT ref="https://www.wikidata.org/wiki/Q34442" label="LOC" desc="wide way leading from one place to another, especially one with a specially prepared surface which vehicles can use">Lo</WDT>? - <WDT ref="https://www.wikidata.org/wiki/Q2" label="LOC" desc="third planet from the Sun in the Solar System">Ma</WDT>? - desidero <WDT ref="https://www.wikidata.org/wiki/Q84" label="LOC" desc="capital and largest city of the United Kingdom">Londra</WDT> comprare, con il vostro "visto", il mio passaggio da Suez. - Non ho nulla in contrario a <LOC>soddisfarvi</LOC>, signore.

NEL

geoparsing

switchboard.clarin.eu/tools



It-Sr-NER > spaCy monolingual and bilingual NER and NEL

3. Dataset

10,000 aligned sentences from the following novels:

- Umberto Eco: The Name of the Rose
- Carlo Collodi: The Adventures of Pinocchio
- Elena Ferrante: Those Who Leave and Those Who Stay
- Luigi Pirandello: One, No one and One Hundred Thousand
- Jules Verne: Around the World in Eighty Days
- Ivo Andrić: Legends of Anika and The Bridge on the Drina
- Borisav Stanković: Impure Blood
- Branislav Nušić: Municipal child: a novel of an infant

4. TMX management and bilingual NER services publishing

```
<tu>
  <tuv xml:lang="it" creationid="n12" creationdate="20211202T203105Z">
    <seg><PERS>Phileas Fogg</PERS>, <DEMO>inglese</DEMO> certamente, non era, forse
  </seg></DEMO> londinese</DEMO>. </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n12" creationdate="20211202T203105Z">
    <seg>lako je na prvi pogled bio <DEMO>Englez</DEMO>, <PERS>Fileas Fogg</PERS> verovatno
  nije bio <DEMO>Londonac</DEMO>. </seg>
  </tuv>
</tu>
```

NER only

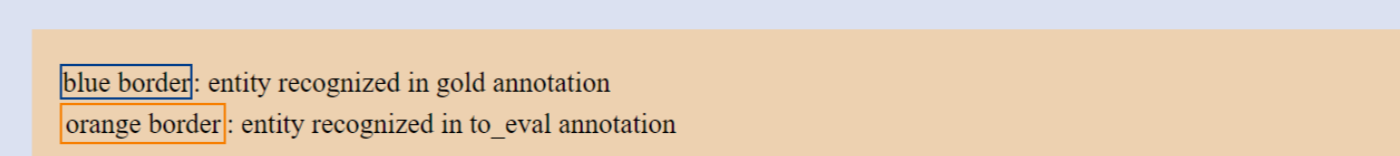
NEL relies on a text previously annotated with NER tags, and linking is performed only on annotated segments

NER+NEL

```
<tu>
  <tuv xml:lang="it" creationid="n12" creationdate="20211202T203105Z">
    <seg><PERS ref="https://www.wikidata.org/wiki/Q2587533" desc="character created by Jules Verne">Phileas Fogg</PERS>, <DEMO>inglese</DEMO> certamente, non era, forse,
  </seg></DEMO> londinese</DEMO>. </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n12" creationdate="20211202T203105Z">
    <seg>lako je na prvi pogled bio <DEMO>Englez</DEMO>, <PERS ref="https://www.wikidata.org/wiki/Q2587533" desc="character created by Jules Verne">Fileas Fogg</PERS> verovatno nije bio
  <DEMO>Londonac</DEMO>. </seg>
  </tuv>
</tu>
```

NEL

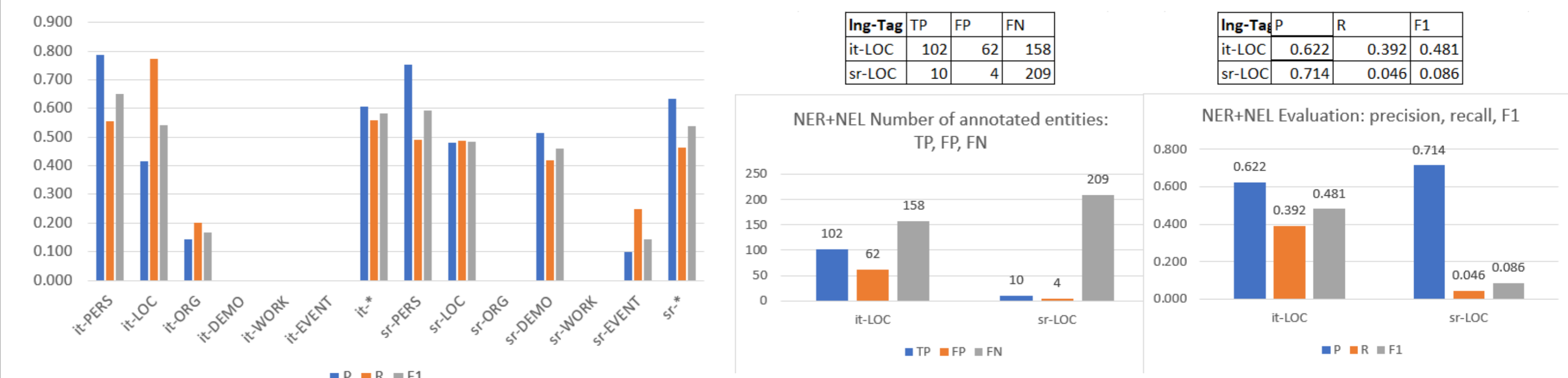
NEL is applied on an input text without NER annotation, using only opentapioca annotation



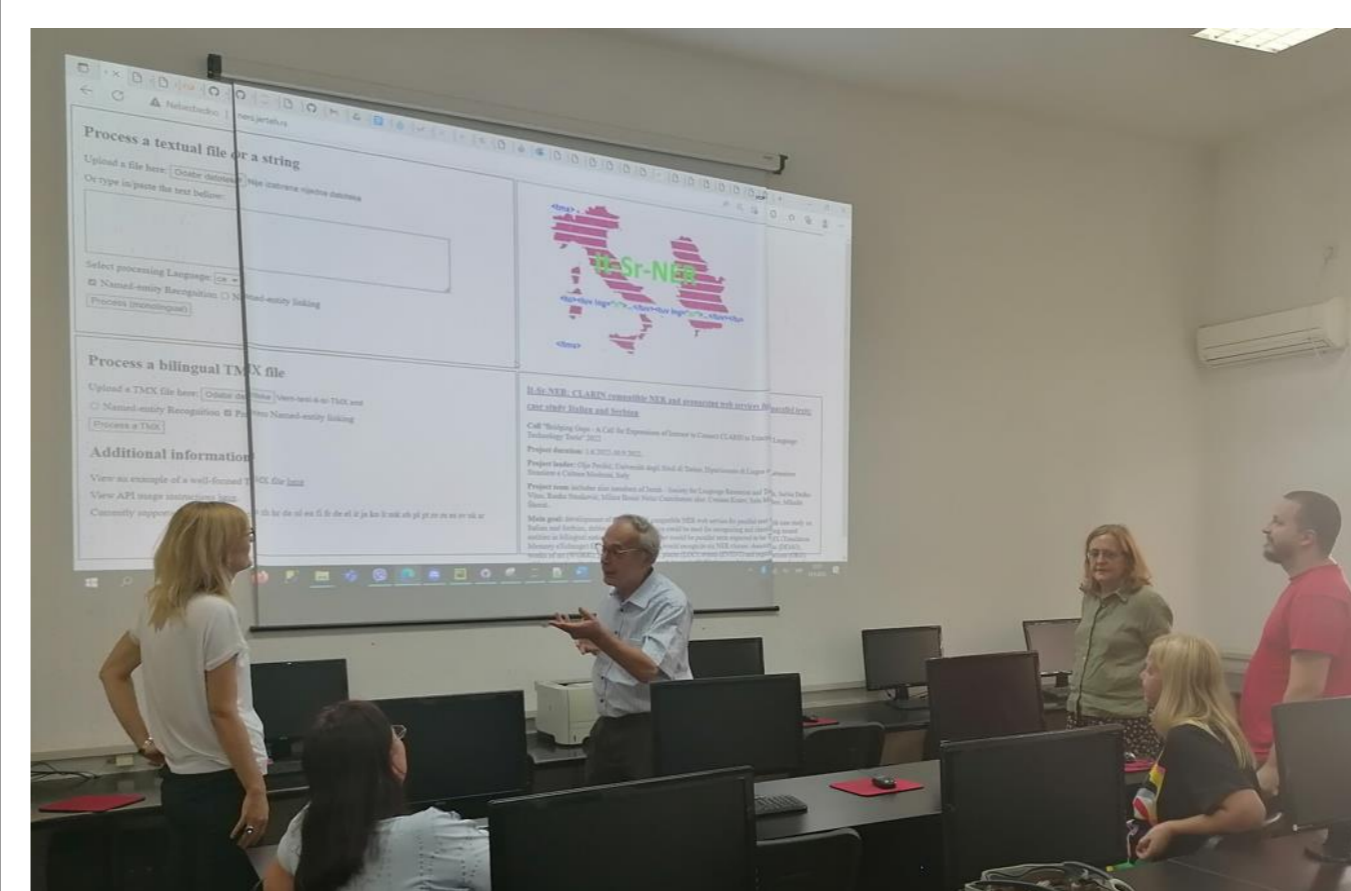
Manual evaluation

Veoma mlada se udaje za Stefana Karačića i uspešno upravlja ispravu delikatesnom radnjom... Sposa giovanissima Stefano Carracci e gestisce con successo prima la salumeria nel nuovo negozio di scappie a piazza dei Martiri. Elena comincia a scriverla nel momento in cui apprende d'infanzia una Cerullo, solo da lei chiamata Elena, è sparita. I brillanti studi di coronati dalla laurea alla Scuola Normale di Pisa, dove conosce e si fida con Pietro Arrighetti, la pubblicazione di un romanzo in cui rielabora la vita nel rione e la esperienza adolescenziale vissuta a Ischia, formato al rione, spinto dalla miseria, si mette al servizio di Michele Salari, che a un certo punto lo manda in Germania per un lungo e misterioso incarico. Enzo è stato a lungo fidanzato con Carmen Peluso, che però lascia senza spiegazioni al rientro dal servizio militare. È fidanzato con Gigliola, la figlia del pasticcere, ma negli anni sviluppa una morbosa ossessione per Elena. Ani, ha smesso di occuparsi di Tito e si è concentrata solo sulla buona riuscita di Elena. Risiede a Barano d'Ischia e d'estate affitta alcune stanze della sua casa alla famiglia Saraturo. Bruno Succavo, amico di Nino Saraturo e figlio di un ricco industriale di San Giovanni a Teduccio. Ci ritiriamo a casa di Tito, il vecchio, piccolo appartamento dei genitori nel quale ora viveva col figlio Rino. In quel periodo mi convinsi che non c'era grande differenza tra il rione e Napoli. In quel periodo mi convinsi che non c'era grande differenza tra il rione e Napoli.

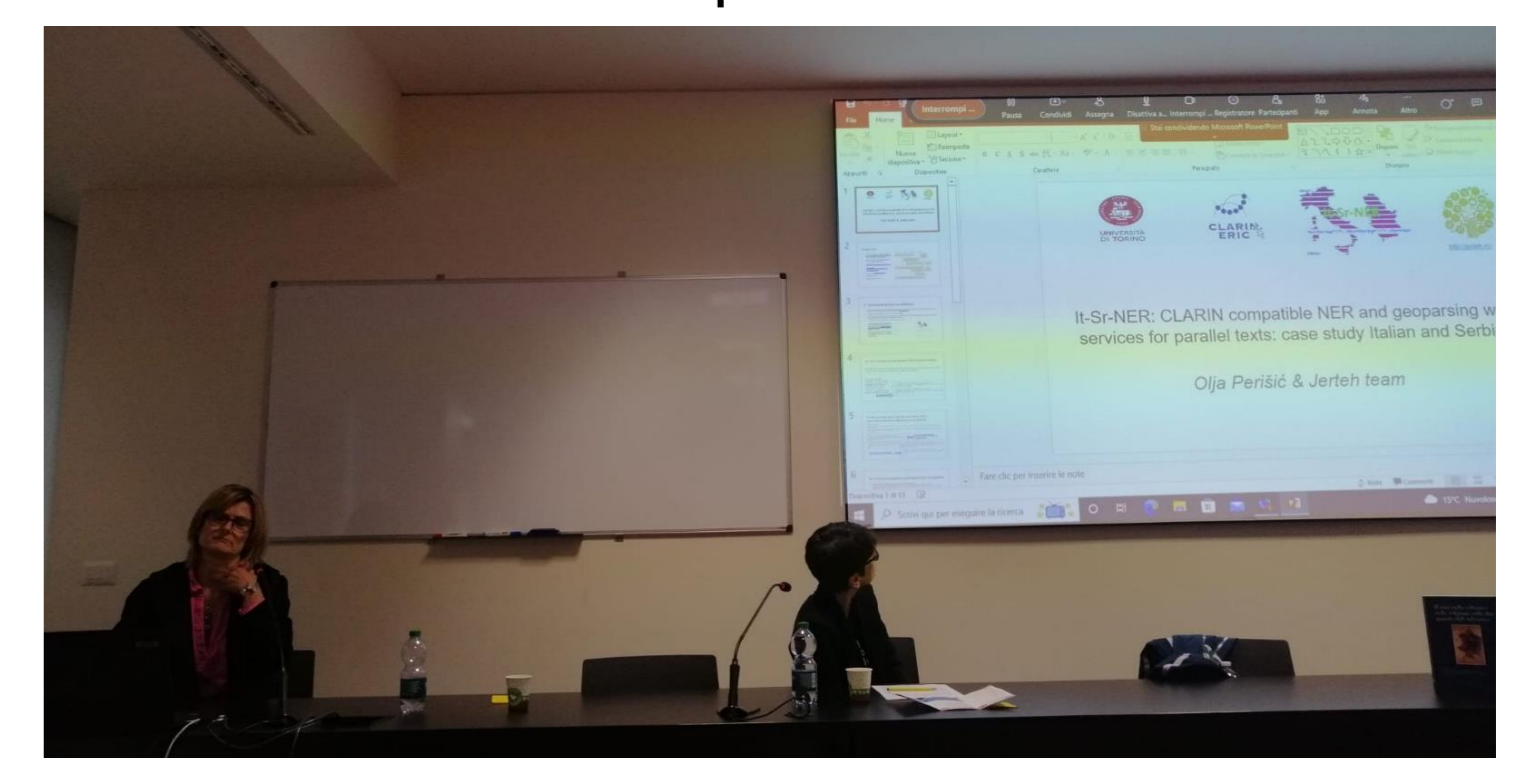
Evaluation: precision, recall, F1



5. Workshop and dissemination



Prof. Duško Vitas presenting at Belgrade workshop



Torino conference

6. Discussion and conclusion

- POS tagging and lemmatisation of TMX format enabling CQL bilingual queries over NER annotated text (<https://noske.jerteh.rs/>)
- Textometry using TXM tool for the analysis of entities on both sides (it-sr).
- The bilingual corpus augmentation
- Further research of new technologies for NER and NEL (training of a new model for Serbian that will include the augmented tag set and gazetteers).
- The students of Italian at the University of Belgrade and Serbian at the University of Turin, will benefit and use developed resources in future teaching.
 - Results are open and so available for other students and researchers as well.
 - Parallel corpora are central to translation studies and contrastive linguistics and easy-to-use concordancers considerably facilitate the study of interlinguistic phenomena.
 - New types of exercises could be produced: automatically generated tests that replace in a sentence some forms with their lemmas that students have to reproduce.
 - Also, the analysis of multiple translations will enlarge the interpretive process and perspectives that students draw from a text.