It-Sr-NER: CLARIN compatible NER and geoparsing web services for parallel texts: case study Italian and Serbian

Olja Perišić, Università degli Studi di Torino, Italy, olja.perisic@unito.it

Serbian Jerteh team led by Duško Vitas, vitas@matf.bg.ac.rs

1. Motivation and background

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

SCAN ME

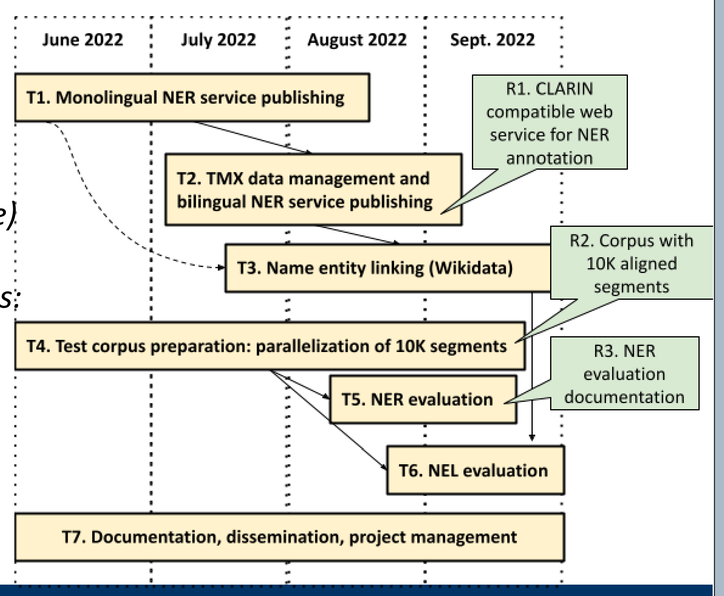
- **Service URL: http://ners.jerteh.rs/**

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;

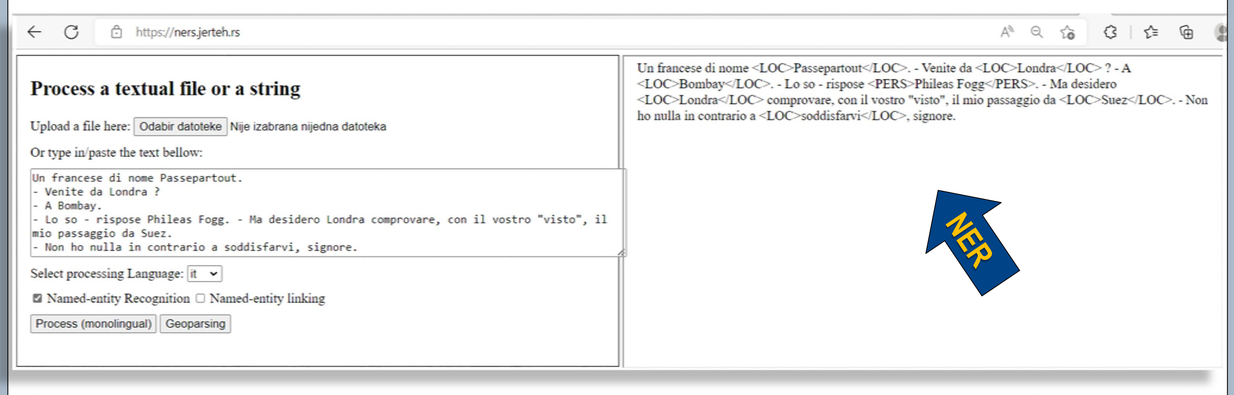
- *demonyms (DEMO),*
- *works of art (WORK),*
- *person names (PERS),*
- *places (LOC),*
- *events (EVENT) and*
- *organizations (ORG).*



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)



Un francese di nome <LOC>Passepartout</LOC>. - Venite da <LOC ref="https://www.wikidata.org/wiki/Q84" desc="capital and largest city of the United Kingdom">Londra</LOC> ? - A <LOC ref="https://www.wikidata.org/wiki/Q1156" desc="capital city and district in Maharashtra, India">Bombay</LOC>. - Lo so - rispose <PERS>Phileas Fogg</PERS>. - Ma desidero <LOC ref="https://www.wikidata.org/wiki/Q84" desc="capital and largest city of the United Kingdom">Londra</LOC> comprovare, 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> so - rispose Phileas Fogg. - <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> comprovare, con il vostro "visto", il mio passaggio da Suez. - Non ho nulla in contrario a soddisfarvi, 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
<DEMO>londinese</DEMO>. </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n12 " creationdate="20211202T203105Z">
    <seg>Iako je na prvi pogled bio <DEMO>Englez</DEMO>, <PERS>Fileas Fog</PERS> verovatn
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

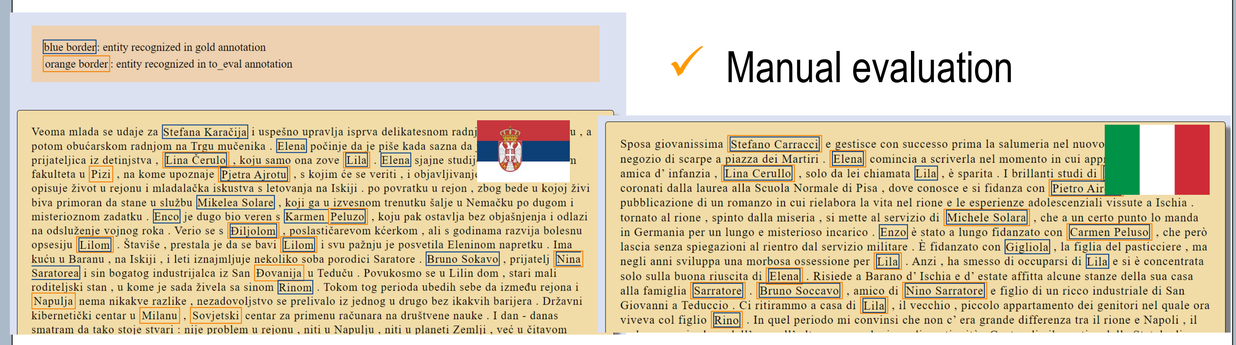
NER+NEL

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

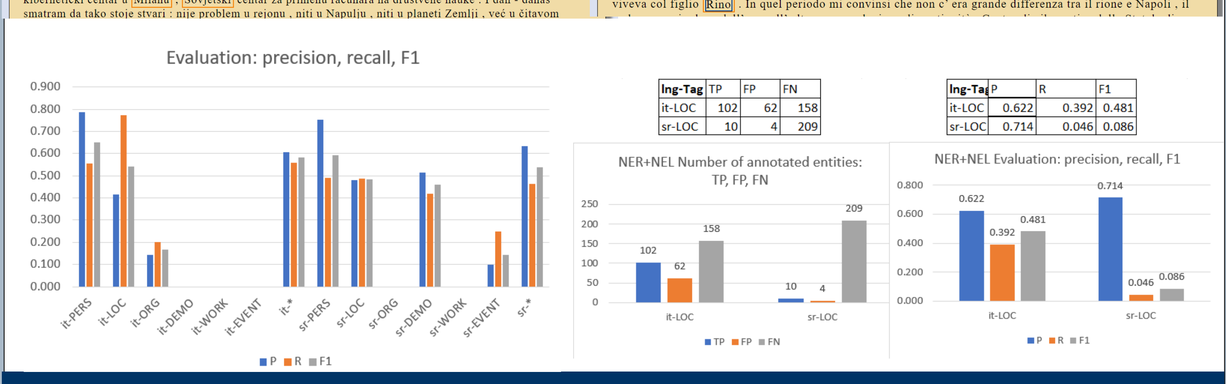
```
<tu>
  <tuv xml:lang="it" creationid="n12 " creationdate="20211202T203105Z">
  <seg><WDT ref="https://www.wikidata.org/wiki/Q2587533" label=”PERSON” desc="character
created by Jules Verne">Phileas Fogg</WDT>, inglese certamente, non era, forse,
<DEMO>londinese</DEMO>. </seg>
  </tuv>
  <tuv xml:lang="sr"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

NEL

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



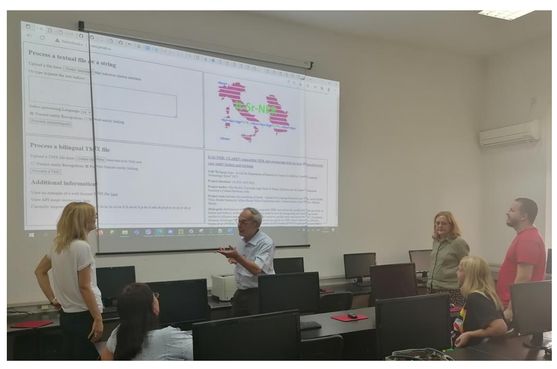
✓ Manual evaluation



Ing-Tag	TP	FP	FN
it-LOC	102	62	158
sr-LOC	10	4	209

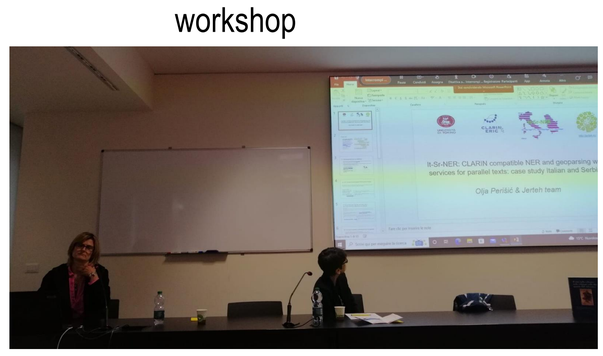
Ing-Tag	P	R	F1
it-LOC	0.622	0.392	0.481
sr-LOC	0.714	0.046	0.086

5. Workshop and dissemination



✓ Torino conference

✓ Prof. Duško Vitas presenting at Belgrade workshop



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.