http://jerteh.rs/

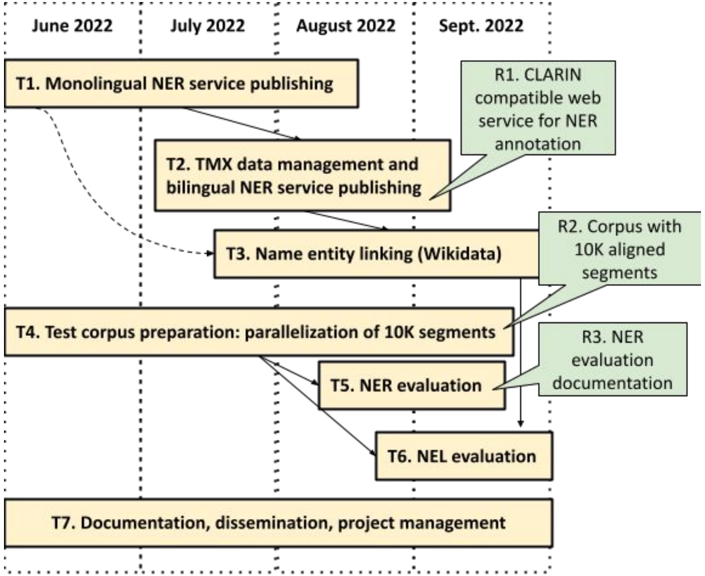
It-Sr-NER: CLARIN compatible NER and geoparsing web services for parallel texts: case study Italian and Serbian

Olja Perišić & JeRTeh team

20.10.2022, Jerteh seminar

Project intro

- Project title: *It-Sr-NER: CLARIN compatible NER and geoparsing web services for parallel texts: case study Italian and Serbian*
- grantNo: CE-2022-2070, funding type: EU - CLARIN Bridging Gaps project.
- Project URL: https://github.com/rankastankovic/It-Sr-NER/
- Demo data URL: https://github.com/rankastankovic/It-Sr-NER/tree/main/corpus
- Service URL: http://ners.jerteh.rs/
- Project tasks T1-T7, Project outcomes: R1-R3
- **Project leader: *Olja Perišić,*** Università degli Studi di Torino, Dipartimento di Lingue e Letterature Straniere e Culture Moderne, Italy
- **Project team** Duško Vitas, Ranka Stanković, Milica Ikonić, Cvetana Krstev, Saša Moderc, Mihailo Škorić.


June 2022
July 2022
August 2022
Sept. 2022
T1. Monolingual NER service publishing
R1. CLARIN compatible web service for NER annotation
T2. TMX data management and bilingual NER service publishing
R2. Corpus with 10K aligned segments
T3. Name entity linking (Wikidata)
T4. Test corpus preparation: parallelization of 10K segments
R3. NER evaluation documentation
T5. NER evaluation
T6. NEL evaluation
T7. Documentation, dissemination, project management

T1. Monolingual NER service publishing

- It-Sr-NER-ws web service is a CLARIN-compatible web service accessible via a convenient web interface, with options to either upload monolingual, text files with output containing NE annotations. Main web service site is http://ners.jerteh.rs/

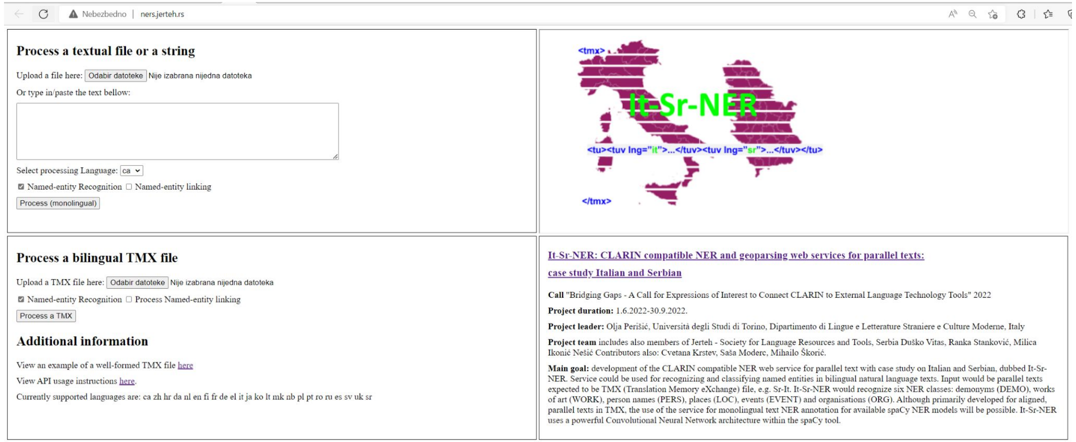
- The user can input text or upload the text file in one of the offered languages.

- If the short text is pasted into the corresponding text area, the application will show the input text containing named entity annotations. In the case of a text file, the same text file but containing NE-annotated text will be returned.

- The annotation is possible for languages with spaCy NER models on the backhand.

- For web application and service development, Python programming language was used. The guidelines published on https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Developer_Manual will be followed.

- It-Sr-NER-ws is hosted at JERTEH infrastructure and integrated into the CLARIN ERIC Switchboard, with consultation with the central CLARIN developer's office.


Nebezbedno | ners.jerteh.rs
<tmx>
It-Sr-NER
<tu><tuv lng="it">...</tuv><tuv lng="sr">...</tuv></tu>
</tmx>

Process a textual file or a string

Upload a file here: Odabir datoteke Nije izabrana nijedna datoteka

Or type in/paste the text bellow:

Select processing Language: ca

☑ Named-entity Recognition ☐ Named-entity linking

Process (monolingual)

Process a bilingual TMX file

Upload a TMX file here: Odabir datoteke Nije izabrana nijedna datoteka

☑ Named-entity Recognition ☐ Process Named-entity linking

Process a TMX

Additional information

View an example of a well-formed TMX file here

View API usage instructions here.

Currently supported languages are: ca zh hr da nl en fi fr de el it ja ko lt mk nb pl pt ro ru es sv uk sr

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

Call "Bridging Gaps - A Call for Expressions of Interest to Connect CLARIN to External Language Technology Tools" 2022

Project duration: 1.6.2022-30.9.2022.

Project leader: Olja Perišić, Università degli Studi di Torino, Dipartimento di Lingue e Letterature Straniere e Culture Moderne, Italy

Project team includes also members of Jerteh - Society for Language Resources and Tools, Serbia Duško Vitas, Ranka Stanković, Milica Ikonić Nešić Contributors also: Cvetana Krstev, Saša Moderc, Mihailo Škorić.

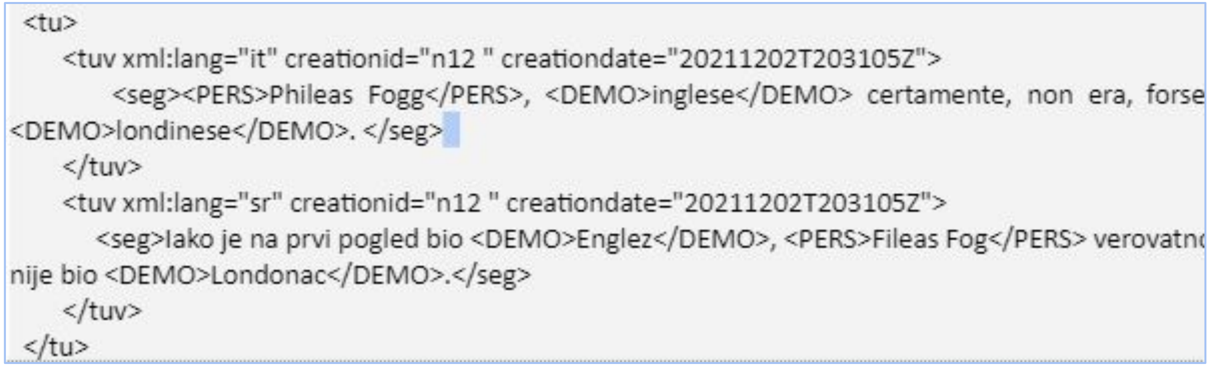
Main goal: development of the CLARIN compatible NER web service for parallel text with case study on Italian and Serbian, dubbed It-Sr-NER. Service could be used for recognizing and classifying named entities in bilingual natural language texts. Input would be parallel texts expected to be TMX (Translation Memory eXchange) file, e.g. Sr-It. It-Sr-NER would recognize six NER classes: demonyms (DEMO), works of art (WORK), person names (PERS), places (LOC), events (EVENT) and organisations (ORG). Although primarily developed for aligned, parallel texts in TMX, the use of the service for monolingual text NER annotation for available spaCy NER models will be possible. It-Sr-NER uses a powerful Convolutional Neural Network architecture within the spaCy tool.

T2. TMX management and bilingual NER services publishing

It-Sr-NER-ws function for processing TMX files as input, then providing appropriate named entity annotations, retaining the same TMX format.

Here is an example translation unit (<tu>) with annotated translation unit variants (<tuv>), where each aligned segment (<seg>) is annotated using the appropriate NER model

(e.g. Italian or Serbian)



```
 <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>
```

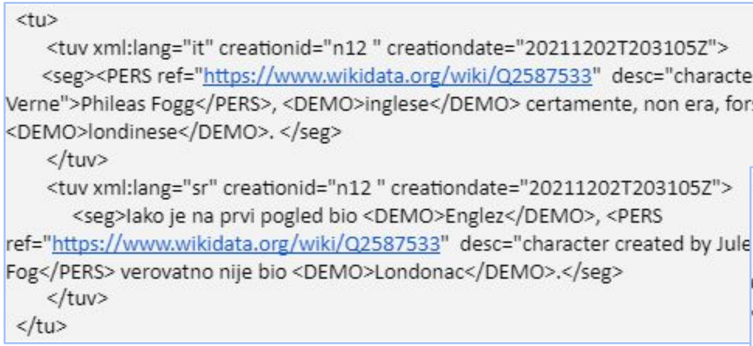
T3. Named entity disambiguation and linking (NEL) (geolocation and other entity types) using Wikidata

NEL is based on spacyopentapioca

For example the person name <PERS>Phileas Fogg</PERS> found in the Italian sentence, and <PERS>Fileas Fog</PERS> in the Serbian sentence, would be both mapped to Wikidata Q2587533 that can be useful for many applications.

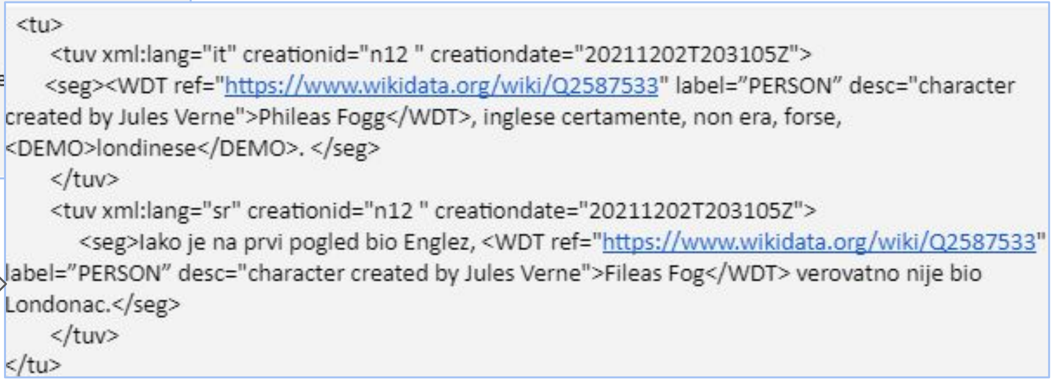
Namely, these NEs are co-indexed to refer to the same person, providing named-entity identification by referring to the ID in a common knowledge base Wikidata and using common QID= Q2587533 for the entry for Phileas Fogg on Wikidata https://www.wikidata.org/wiki/Q2587533.

Service provide two options:



```
<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 Jule
Fog</PERS> verovatno nije bio <DEMO>Londonac</DEMO>.</seg>
    </tuv>
</tu>
```

1. NEL relies on entities that are previously annotated with NER tags, and linking is performed just on annotated segments



```
<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>
```

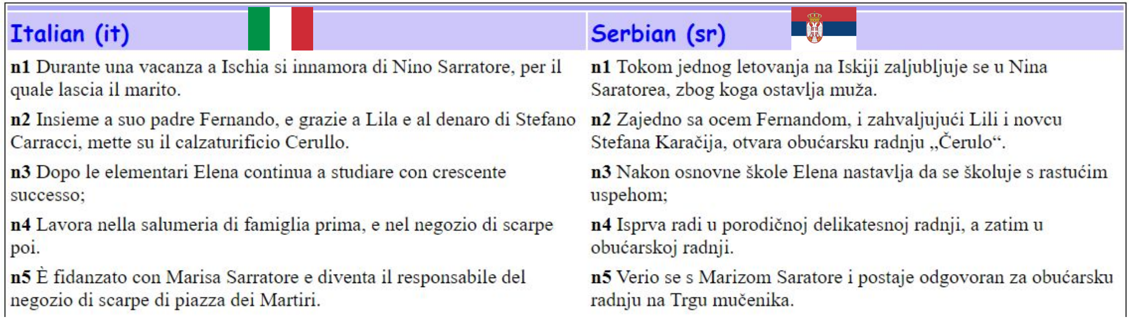
2. NEL is applied on input text without NER annotation, using only opentapioca annotation

T4. Test corpus preparation: parallelization of 10K segments

- Corpus consists of 10000 aligned segments (sentences) from several novels.
- Novels are represented with samples in which segments are shuffled in order to avoid copyright problems.
- The novels are aligned and prepared following the previously used pipeline for parallelization (Utvić et al. 2008; Krstev and Vitas 2011).
- The corpus is published at the ILC4CLARIN B Centre, so that it will be visible via the VLO (Virtual Language Observatory) https://vlo.clarin.eu/.
- Corpus is also accessible from working github location https://github.com/rankastankovic/It-Sr-NER/tree/main/corpus

Novels:

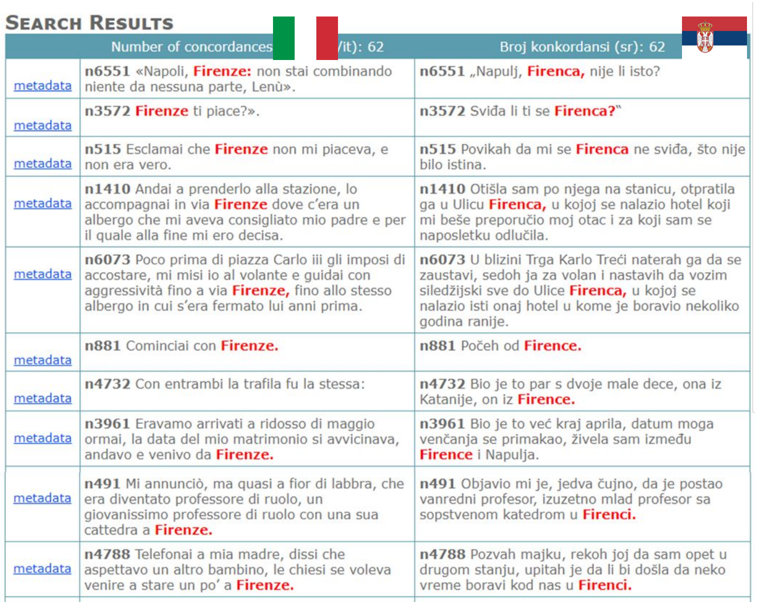
❖ Umberto Eco: The Name of the Rose;
❖ Carlo Collodi: The Adventures of Pinocchio;
❖ Elena Ferrante: Those Who Leave and Those Who Stay,
❖ Luigi Pirandello: One, No one and One Hundred Thousand,
❖ Jules Verne: Around the World in Eighty Days;
❖ Ivo Andrić: Legends of Anika and The Bridge on the Drina,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



Italian (it)	Serbian (sr)
n1 Durante una vacanza a Ischia si innamora di Nino Sarratore, per il quale lascia il marito.	**n1** Tokom jednog letovanja na Iskiji zaljubljuje se u Nina Saratorea, zbog koga ostavlja muža.
n2 Insieme a suo padre Fernando, e grazie a Lila e al denaro di Stefano Carracci, mette su il calzaturificio Cerullo.	**n2** Zajedno sa ocem Fernandom, i zahvaljujući Lili i novcu Stefana Karačija, otvara obućarsku radnju „Čerulo“.
n3 Dopo le elementari Elena continua a studiare con crescente successo;	**n3** Nakon osnovne škole Elena nastavlja da se školuje s rastućim uspehom;
n4 Lavora nella salumeria di famiglia prima, e nel negozio di scarpe poi.	**n4** Isprva radi u porodičnoj delikatesnoj radnji, a zatim u obućarskoj radnji.
n5 È fidanzato con Marisa Sarratore e diventa il responsabile del negozio di scarpe di piazza dei Martiri.	**n5** Verio se s Marizom Saratore i postaje odgovoran za obućarsku radnju na Trgu mučenika.

T4. Test corpus preparation: parallelization of 10K segments

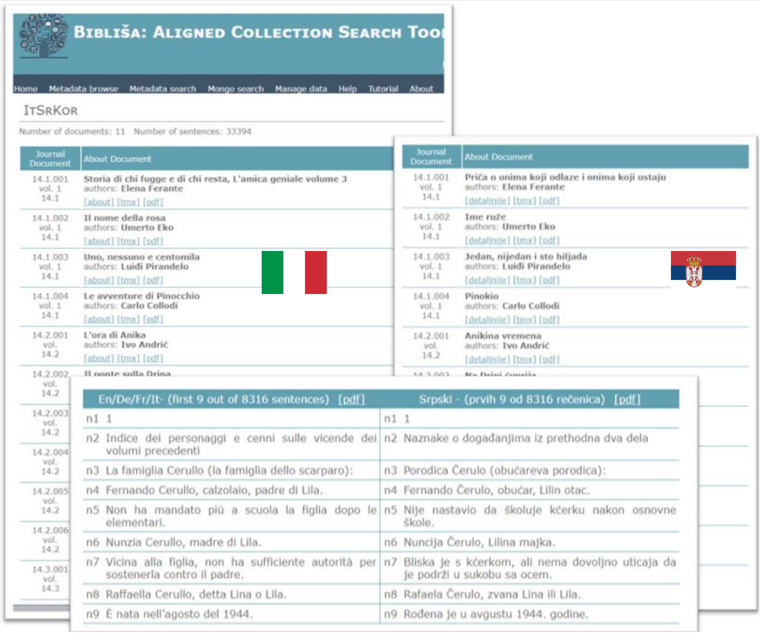
Apart for download, corpus is available also for search on the digital library Bibliša for search with possibilities of morphological and semantic query expansion for Serbian.



SEARCH RESULTS

	Number of concordances (it): 62	Broj konkordansi (sr): 62
metadata	n6551 «Napoli, Firenze: non stai combinando niente da nessuna parte, Lenù».	n6551 „Napulj, Firenca, nije li isto?
metadata	n3572 Firenze ti piace?».	n3572 Sviđa li ti se Firenca?"
metadata	n515 Esclamai che Firenze non mi piaceva, e non era vero.	n515 Povikah da mi se Firenca ne sviđa, što nije bilo istina.
metadata	n1410 Andai a prenderlo alla stazione, lo accompagnai in via Firenze dove c'era un albergo che mi aveva consigliato mio padre e per il quale alla fine mi ero decisa.	n1410 Otišla sam po njega na stanicu, otpratila ga u Ulicu Firenca, u kojoj se nalazio hotel koji mi beše preporučio moj otac i za koji sam se naposletku odlučila.
metadata	n6073 Poco prima di piazza Carlo iii gli imposi di accostare, mi misi io al volante e guidai con aggressività fino a via Firenze, fino allo stesso albergo in cui s'era fermato lui anni prima.	n6073 U blizini Trga Karlo Treći naterah ga da se zaustavi, sedoh ja za volan i nastavih da vozim siledžijski sve do Ulice Firenca, u kojoj se nalazio isti onaj hotel u kome je boravio nekoliko godina ranije.
metadata	n881 Cominciai con Firenze.	n881 Počeh od Firence.
metadata	n4732 Con entrambi la trafila fu la stessa:	n4732 Bio je to par s dvoje male dece, ona iz Katanije, on iz Firence.
metadata	n3961 Eravamo arrivati a ridosso di maggio ormai, la data del mio matrimonio si avvicinava, andavo e venivo da Firenze.	n3961 Bio je to već kraj aprila, datum moga venčanja se primakao, živela sam između Firence i Napulja.
metadata	n491 Mi annunciò, ma quasi a fior di labbra, che era diventato professore di ruolo, un giovanissimo professore di ruolo con una sua cattedra a Firenze.	n491 Objavio mi je, jedva čujno, da je postao vanredni profesor, izuzetno mlad profesor sa sopstvenom katedrom u Firenci.
metadata	n4788 Telefonai a mia madre, dissi che aspettavo un altro bambino, le chiesi se voleva venire a stare un po' a Firenze.	n4788 Pozvah majku, rekoh joj da sam opet u drugom stanju, upitah je da li bi došla da neko vreme boravi kod nas u Firenci.

The option for browsing of aligned collections is available with a limited number of segments for all and with a larger number of segments for authorized users.



ITSRKOR

Number of documents: 11 Number of sentences: 33394

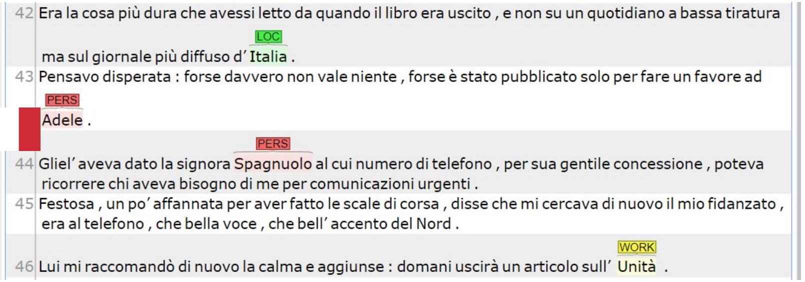
Journal Document	About Document
14.1.001 vol. 1 14.1	Storia di chi fugge e di chi resta, L'amica geniale volume 3 authors: Elena Ferante [about] [tmx] [pdf]
14.1.002 vol. 1 14.1	Il nome della rosa authors: Umerto Eko [about] [tmx] [pdf]
14.1.003 vol. 1 14.1	Uno, nessuno e centomila authors: Luiđi Pirandelo [about] [tmx] [pdf]
14.1.004 vol. 1 14.1	Le avventure di Pinocchio authors: Carlo Collodi [about] [tmx] [pdf]
14.2.001 vol. 14.2	L'ora di Anika authors: Ivo Andrić [about] [tmx] [pdf]
14.2.002 vol. 14.2	Il ponte sulla Drina
14.2.003 vol. 14.2	
14.2.004 vol. 14.2	
14.2.005 vol. 14.2	
14.2.006 vol. 14.2	
14.3.001 vol. 14.3	

Journal Document	About Document
14.1.001 vol. 1 14.1	Priča o onima koji odlaze i onima koji ostaju authors: Elena Ferante [detaljnije] [tmx] [pdf]
14.1.002 vol. 1 14.1	Ime ruže authors: Umerto Eko [detaljnije] [tmx] [pdf]
14.1.003 vol. 1 14.1	Jedan, nijedan i sto hiljada authors: Luiđi Pirandelo [detaljnije] [tmx] [pdf]
14.1.004 vol. 1 14.1	Pinokio authors: Carlo Collodi [detaljnije] [tmx] [pdf]
14.2.001 vol. 14.2	Anikina vremena authors: Ivo Andrić [detaljnije] [tmx] [pdf]
14.2.002	Na Drini ćuprija

En/De/Fr/It- (first 9 out of 8316 sentences) [pdf]	Srpski - (prvih 9 od 8316 rečenica) [pdf]
n1 1	n1 1
n2 Indice dei personaggi e cenni sulle vicende dei volumi precedenti	n2 Naznake o događanjima iz prethodna dva dela
n3 La famiglia Cerullo (la famiglia dello scarparo):	n3 Porodica Čerulo (obućareva porodica):
n4 Fernando Cerullo, calzolaio, padre di Lila.	n4 Fernando Čerulo, obućar, Lilin otac.
n5 Non ha mandato più a scuola la figlia dopo le elementari.	n5 Nije nastavio da školuje kćerku nakon osnovne škole.
n6 Nunzia Cerullo, madre di Lila.	n6 Nuncija Čerulo, Lilina majka.
n7 Vicina alla figlia, non ha sufficiente autorità per sostenerla contro il padre.	n7 Bliska je s kćerkom, ali nema dovoljno uticaja da je podrži u sukobu sa ocem.
n8 Raffaella Cerullo, detta Lina o Lila.	n8 Rafaela Čerulo, zvana Lina ili Lila.
n9 È nata nell'agosto del 1944.	n9 Rođena je u avgustu 1944. godine.

T5. NER evaluation

Automatically annotated 1000 sentence pairs imported to the INCEPTION tool http://inception.jerteh.rs for manual correction

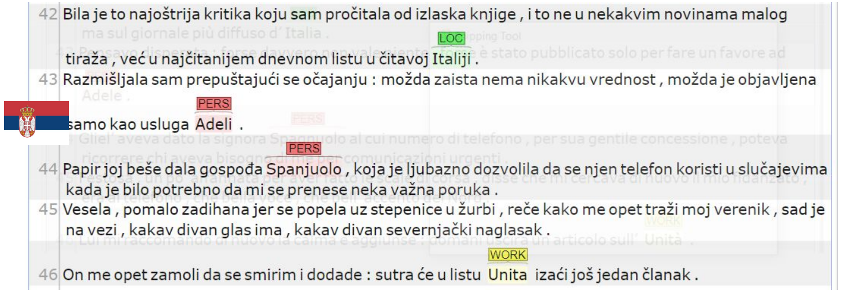

42 Era la cosa più dura che avessi letto da quando il libro era uscito , e non su un quotidiano a bassa tiratura ma sul giornale più diffuso d' Italia .

43 Pensavo disperata : forse davvero non vale niente , forse è stato pubblicato solo per fare un favore ad Adele .

44 Gliel' aveva dato la signora Spagnuolo al cui numero di telefono , per sua gentile concessione , poteva ricorrere chi aveva bisogno di me per comunicazioni urgenti .

45 Festosa , un po' affannata per aver fatto le scale di corsa , disse che mi cercava di nuovo il mio fidanzato , era al telefono , che bella voce , che bell' accento del Nord .

46 Lui mi raccomandò di nuovo la calma e aggiunse : domani uscirà un articolo sull' Unità .


42 Bila je to najoštrija kritika koju sam pročitala od izlaska knjige , i to ne u nekakvim novinama malog tiraža , već u najčitanijem dnevnom listu u čitavoj Italiji .

43 Razmišljala sam prepuštajući se očajanju : možda zaista nema nikakvu vrednost , možda je objavljena samo kao usluga Adeli .

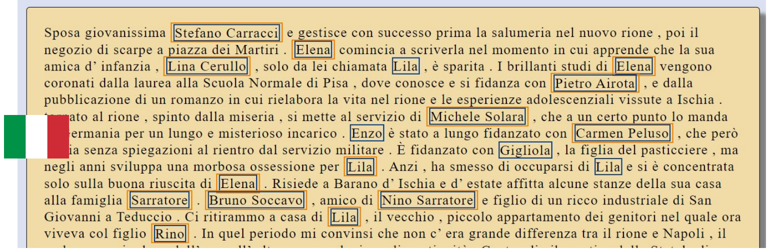
44 Papir joj beše dala gospođa Spanjuolo , koja je ljubazno dozvolila da se njen telefon koristi u slučajevima kada je bilo potrebno da mi se prenese neka važna poruka .

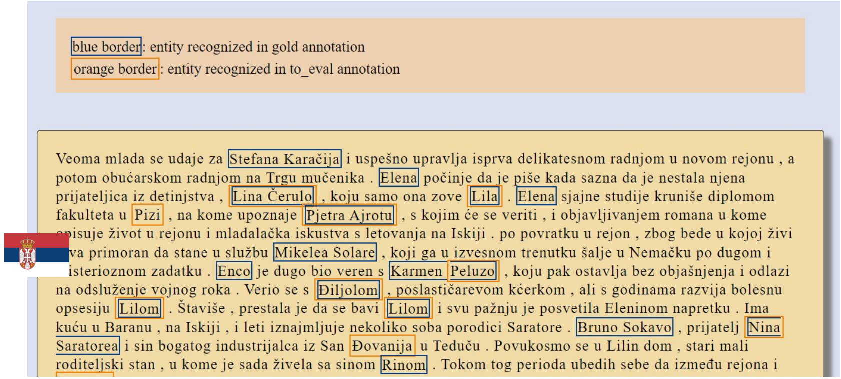
45 Vesela , pomalo zadihana jer se popela uz stepenice u žurbi , reče kako me opet traži moj verenik , sad je na vezi , kakav divan glas ima , kakav divan severnjački naglasak .

46 On me opet zamoli da se smirim i dodade : sutra će u listu Unita izaći još jedan članak .

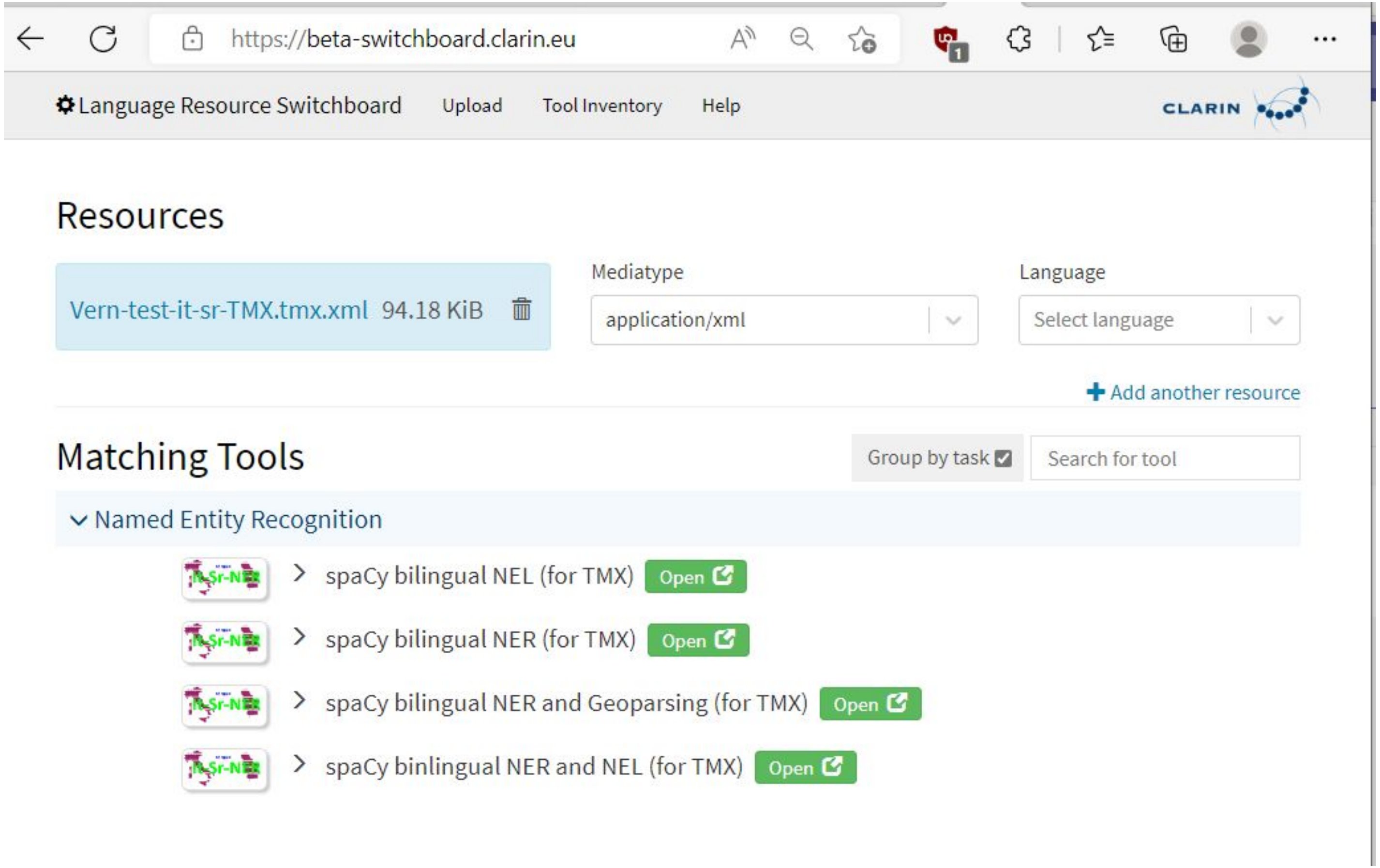
Visual comparison of automatic and manual annotation

Manual annotation in blue and automatic in orange
For each name entity type one HTML file per language is generated and published as Italian set and Serbian set .


Sposa giovanissima Stefano Carracci e gestisce con successo prima la salumeria nel nuovo rione , poi il negozio di scarpe a piazza dei Martiri . Elena comincia a scriverla nel momento in cui apprende che la sua amica d' infanzia , Lina Cerullo , solo da lei chiamata Lila , è sparita . I brillanti studi di Elena vengono coronati dalla laurea alla Scuola Normale di Pisa , dove conosce e si fidanza con Pietro Airota , e dalla pubblicazione di un romanzo in cui rielabora la vita nel rione e le esperienze adolescenziali vissute a Ischia . t ato al rione , spinto dalla miseria , si mette al servizio di Michele Solara , che a un certo punto lo manda ermania per un lungo e misterioso incarico . Enzo è stato a lungo fidanzato con Carmen Peluso , che però ia senza spiegazioni al rientro dal servizio militare . È fidanzato con Gigliola , la figlia del pasticciere , ma negli anni sviluppa una morbosa ossessione per Lila . Anzi , ha smesso di occuparsi di Lila 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 Sarratore . Bruno Soccavo , amico di Nino Sarratore e figlio di un ricco industriale di San Giovanni a Teduccio . Ci ritirammo a casa di Lila , 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 , il


blue border: entity recognized in gold annotation
orange border: entity recognized in to_eval annotation

Veoma mlada se udaje za Stefana Karačija i uspešno upravlja isprva delikatesnom radnjom u novom rejonu , a potom obućarskom radnjom na Trgu mučenika . Elena počinje da je piše kada sazna da je nestala njena prijateljica iz detinjstva , Lina Čerulo , koju samo ona zove Lila . Elena sjajne studije kruniše diplomom fakulteta u Pizi , na kome upoznaje Pjetra Ajrotu , s kojim će se veriti , i objavljivanjem romana u kome opisuje život u rejonu i mladalačka iskustva s letovanja na Iskiji . po povratku u rejon , zbog bede u kojoj živi va primoran da stane u službu Mikelea Solare , koji ga u izvesnom trenutku šalje u Nemačku po dugom i isterioznom zadatku . Enco je dugo bio veren s Karmen Peluzo , koju pak ostavlja bez objašnjenja i odlazi na odsluženje vojnog roka . Verio se s Điljolom , poslastičarevom kćerkom , ali s godinama razvija bolesnu opsesiju Lilom . Štaviše , prestala je da se bavi Lilom i svu pažnju je posvetila Eleninom napretku . Ima kuću u Baranu , na Iskiji , i leti iznajmljuje nekoliko soba porodici Saratore . Bruno Sokavo , prijatelj Nina Saratorea i sin bogatog industrijalca iz San Đovanija u Teduču . Povukosmo se u Lilin dom , stari mali roditeljski stan , u kome je sada živela sa sinom Rinom . Tokom tog perioda ubedih sebe da između rejona i


https://beta-switchboard.clarin.eu
Language Resource Switchboard
Upload
Tool Inventory
Help
CLARIN

Resources

Vern-test-it-sr-TMX.tmx.xml 94.18 KiB

Mediatype: application/xml

Language: Select language

Matching Tools

Group by task ☑

Named Entity Recognition

- spaCy bilingual NEL (for TMX) Open
- spaCy bilingual NER (for TMX) Open
- spaCy bilingual NER and Geoparsing (for TMX) Open
- spaCy binlingual NER and NEL (for TMX) Open

Summary of the project results

R1. CLARIN compatible web service for NER annotation of 1) bilingual texts tested on it-sr with 6 NE classes 2) monolingual texts with supporting documentation (open source), tested on Serbian and Italian.

R2. Corpus with 10000 aligned segments to be used for testing (open data)

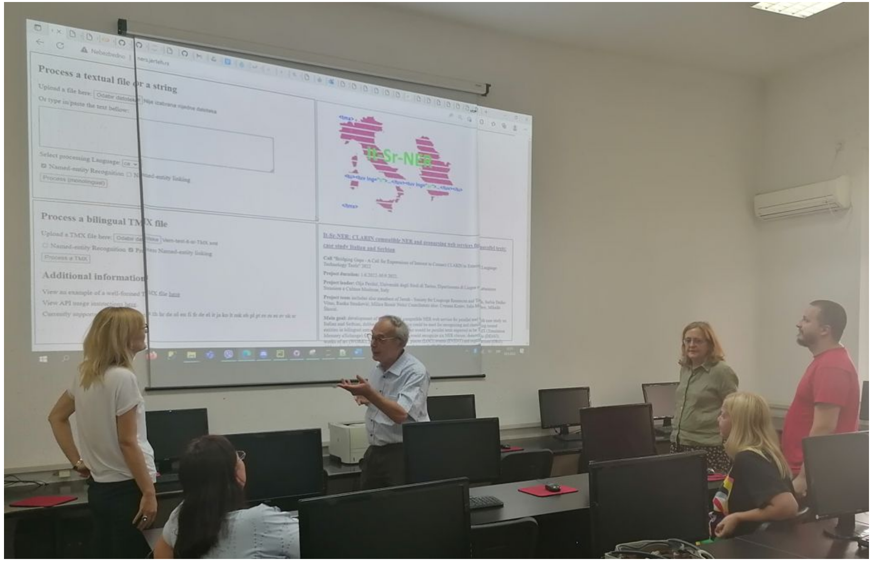
Automatically linked LOC-related entities to the corresponding pages on Wikidata, followed by an analysis of the NEL success on a validation subset of parallel sentences.

R3. NER evaluation documentation that will include manually corrected 1000 segments with metrics and visualization of differences.

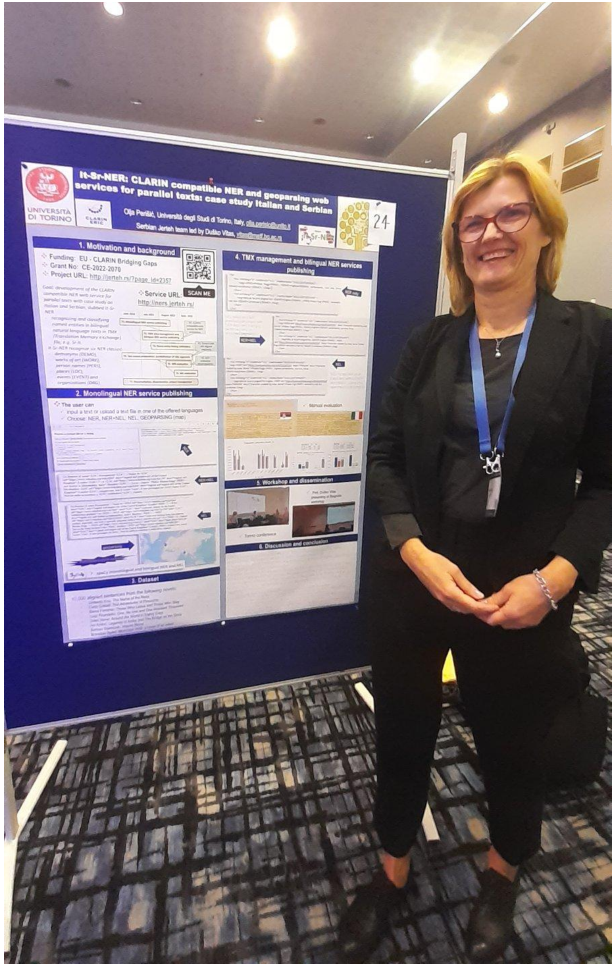
We will conduct an exhaustive evaluation on a subset of the parallel sentences. These results will be published and would enable users to gain insight into the model's performance and to understand on what kind of input these models yield the best results.

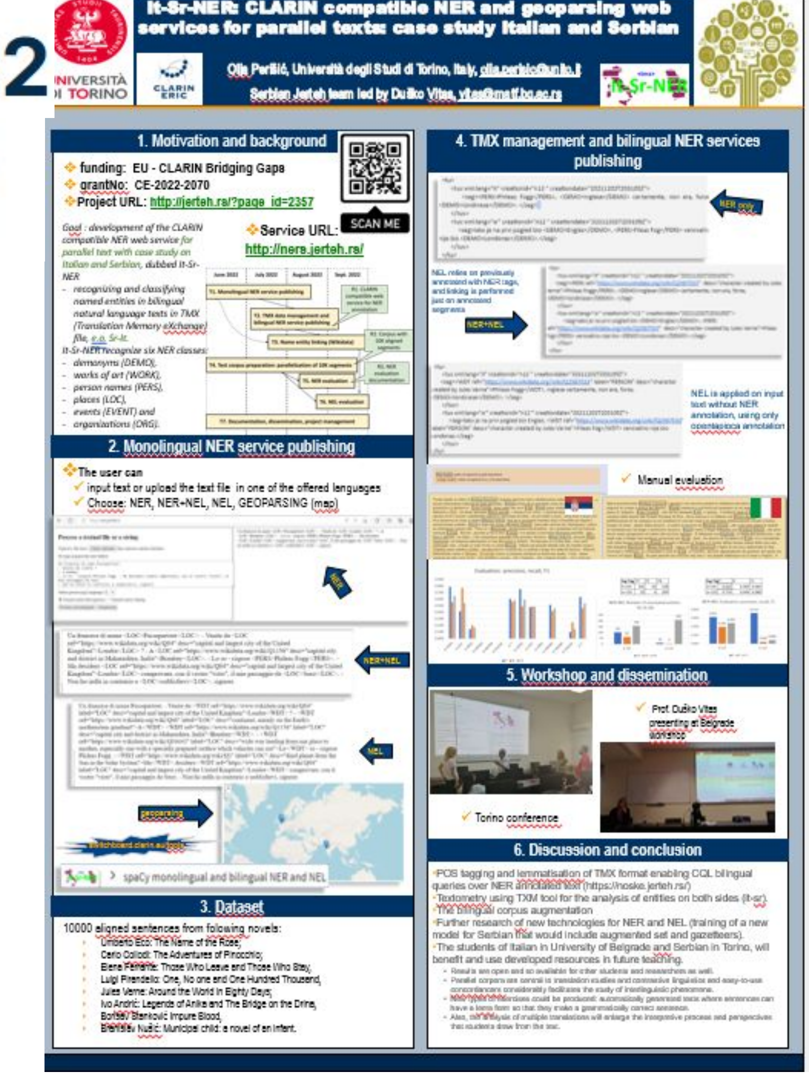
Belgrade workshop

The meeting in Belgrade was held in September 2022 for final discussions, web service testing and evaluation of results. Apart from quantitative analysis, qualitative analysis was discussed and possibilities for improvement of the result of automatic annotation.


Process a textual file or a string
Process a bilingual TMX file
Additional information
It-Sr-NER

The CLARIN Bazaar 2022





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

grantNo: 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 text 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).*

June 2022
July 2022
August 2022
Sept. 2022
T1. Monolingual NER service publishing
T2. TMX data management and bilingual NER service publishing
T3. Name entity linking (Wikidata)
T4. Text corpus preparation: parallelization of 10K segments
T5. NER evaluation
T6. NEL evaluation
T7. Documentation, dissemination, project management
R1. CLARIN compatible web service for NER annotation
R2. Corpus with 10K aligned segments
R3. NER evaluation documentation

2. Monolingual NER service publishing

The user can

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

spaCy monolingual and bilingual NER and NEL

3. Dataset

10000 aligned sentences from folowing 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

NEL relies on previously annotated with NER tags, and linking is performed just on annotated segments

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

✓ Manual evaluation

5. Workshop and dissemination

✓ Torino conference

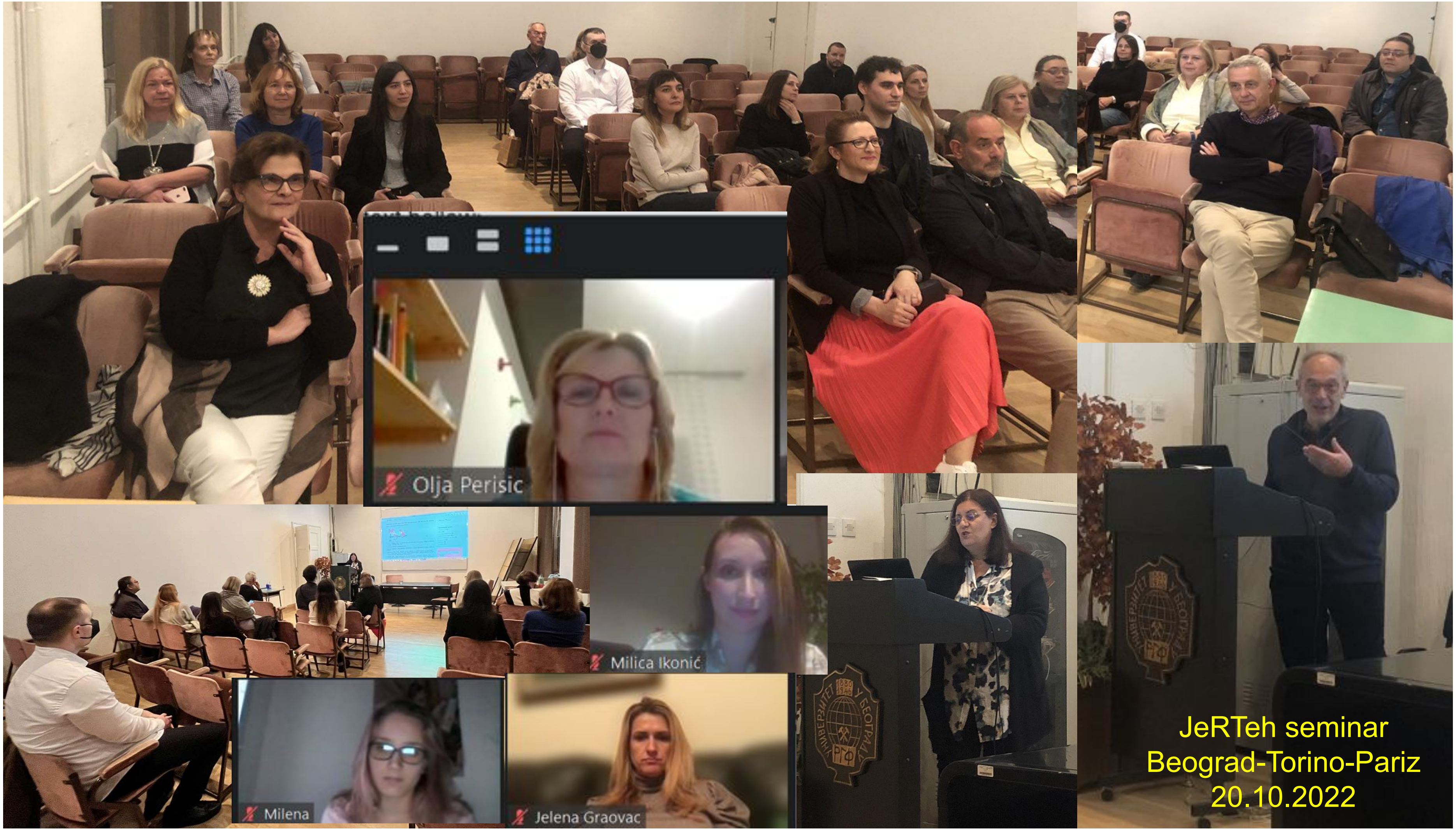
✓ Prof. Duško Vitas presenting at Belgrade workshop

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 would include augmented set and gazetteers).
- The students of Italian in University of Belgrade and Serbian in Torino, 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 in translation studies and contrastive linguistics and easy-to-use concordancers considerably facilitates the study of interlinguistic phenomena.
  - New types of exercises could be produced: automatically generated texts where sentences can have a lemma form so that they make a grammatically correct sentence.
  - Also, the analysis of multiple translations will enlarge the interpretive process and perspectives that students draw from the text.

Discussion and future plans

- POS tagging and lemmatisation of TMX format enabling CQL bilingual queries over NER annotated text: the corpus will be published on https://noske.jerteh.rs/ .
- Textometry using TXM tool will be used for the analysis of entities on both sides (it-sr).
- The bilingual corpus augmentation (set of novels are already collected for preparation and alignment).
- Further research will explore new technologies for NER and NEL: training of a new model for Serbian that would include training set augmentation and use of gazettiers.
- The students of Italian in University of Belgrade as well as students of Serbian in Torino, 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 facilitates the study of interlinguistic phenomena.
    - New types of exercises could be produced: automatically generated tests where sentences can have a lema form so that they make a grammatically correct sentence.
    - Also, the analysis of multiple translations will enlarge the interpretive process and perspectives that students draw from the text. The parallel corpus will
- The team will look for possibilities for the new projects and frameworks to cooperate.



JeRTeh seminar
Beograd-Torino-Pariz
20.10.2022