It-Sr-NER: CLARIN Compatible NER and Geoparsing Web Services for Italian and Serbian Parallel Text

Olja Perišić
University of Turin, Italy
olja.perisic@unito.it

Ranka Stanković
University of Belgrade, Serbia
ranka.stankovic@rgf.bg.ac.rs

Milica Ikonić Nešić
University of Belgrade, Serbia
milica.ikonic.nesic@fil.bg.ac.rs

Mihailo Škorić
University of Belgrade, Serbia
mihailo.skoric@rgf.rs

Abstract

The paper will showcase the outcomes of the "It-Sr-NER: Web services for named entities recognition, linking and mapping" project for Serbian and Italian languages. The project was a collaboration between the University of Turin and the Society for Language Resources and Technologies JeRTeh, with the goal of creating the It-Sr-NER web service. This service is designed to annotate named entities such as people, places, organizations, ethnicities, events, and works of art in text, and display them on a map.

1 Introduction

The main motivation for starting "It-Sr-NER: Web services for named entities recognition, linking and mapping" project was the lack of tools and resources for annotating, researching, and analyzing bilingually aligned Italian-Serbian texts. At the same time there is a significant absence of corpus tools in the teaching of foreign languages in Serbia, as noted in recent research (Vitaz and Poletanović, 2020). However, on an individual level and through personal initiatives in the teaching of Serbian as a foreign language in Italy and the Italian language in Serbia, it has been shown that corpora can be highly beneficial in many ways and that students are eager to use them in collaborative work and independent research (Moderc, 2015b; Perišić, 2021). One of the challenges in teaching the Serbian language to foreign students is the rich and sophisticated morphology, which includes declensions of toponyms and other named entities that can be difficult for students to recognize and reduce to their basic form. This is due to factors such as similar endings for the masculine and neuter gender in most grammatical cases, the presence of certain toponyms only in the plural form, the so-called *pluralia tantum* (Berane, Udine, etc.), phonetic transcriptions of foreign names, and some orthographic inconsistencies (Vitas and Lažetić-Pavlović, 2008).

A team of experts from the University of Turin and the Society for Language Resources and Technologies JeRTeh have partnered as part of CLARIN's call "Bridging Gaps" to develop web services for annotating named entities in text. These services ensure the linking of named entities with Wikidata and provide geoparsing, which includes geolocating recognized locations and displaying them on a map. These services specifically target names of persons, places, organizations, ethnicities, events, and works of art as named entities.

The primary objective of the project was to create and publish web applications and services for monolingual and bilingual parallel texts within the CLARIN infrastructure as well as on the platform of the Society for Language Resources and Technologies JeRTeh. The project also aimed to create and publish an Italian-Serbian corpus of 10,000 segments of extracted and aligned sentences, selected from classics of Italian and Serbian literature. The outcomes of the project are not restricted to the Italian-Serbian language combination, the developed services can be used to process texts in twenty-four different languages.

The project was initiated and led by Olja Perišić, a professor at the University of Turin, where she teaches the Serbian and Croatian language. On behalf of JeRTeh, the development of the services was

Olja Perišić, Ranka Stanković, Milica Ikonić Nešić and Mihailo Škorić 2023. It-Sr-NER: CLARIN Compatible NER and Geoparsing Web Services for Italian and Serbian Parallel Text. *Selected papers from the CLARIN Annual Conference 2022. Ed. by Tomaž Erjavec and Maria Eskevich.* Linköping Electronic Conference Proceedings 198, pp. 99–110. DOI: https://doi.org/10.3384/ecp198010

led by Professor Ranka Stanković in collaboration with Professor Duško Vitas. Further information about the project can be found on the website of the Society for Language Resources and Technologies JeRTeh, which includes the It-Sr-NER CLARIN compatible named entity recognition (NER), named entity linking (NEL) and geoparsing web services for parallel texts.[1]

2 Italian-Serbian Parallel Corpus

Parallel corpora have been found to be a crucial tool in foreign language teaching as they help, at the beginner level, to acquire morphosyntax and lexis. Observing two or more languages in parallel facilitates contrastive analysis, which allows for the examination of similarities and differences in language structures by providing a large number of sentence examples in context. As Sinclair pointed out at an early stage of development of corpus linguistics: "The language looks rather different when you look at a lot of it at once" (Sinclair, 1991). At the intermediate and advanced level, parallel corpora are an effective tool in teaching translation as they facilitate disambiguation of word senses and definition of polysemic vocabulary, which are often underrepresented in bilingual dictionaries especially for this language combination (Moderc, 2015a; Perišić Arsić, 2018). A translation is always linked to the context of the target language, to the individual style of each translator and his/her interpretation of the original text. The possibility to compare different translations of a single text can highlight any ambiguities or inconsistencies already present in the source text. These ambiguities, due to several linguistic reasons, concern the register and are attributable to various cultural factors, but they are hardly noticed in the monolingual analysis of a text (Perišić, 2023). At the same time, it has been noted that there is a lack of representative parallel corpora even for major world languages (Granger, 2018).

To overcome this problem, as a first step in the project, it was necessary to create an Italian-Serbian corpus of 10,000 aligned segments (sentences) taken from ten different novels. The novels by Italian writers represented in the corpus are: Umberto Eco's "The Name of the Rose", Carlo Collodi's "The Adventures of Pinocchio", Elena Ferrante's "Those Who Leave and Those Who Stay", and Luigi Pirandello's "One, None and a Hundred Thousand". The corpus also includes five novels by Serbian writers: Ivo Andrić's "Legends of Anika" and "The Bridge on the Drina", Borisav Stanković's "Impure Blood", Branislav Nušić's "Municipal child: the novel of an infant", Danilo Kiš's "Garden, Ashes". Additionally, the corpus also includes Italian and Serbian translations of Jules Verne's "Around the World in Eighty Days" in order to support the main task of the project which is annotating named entities.

The novels were aligned and converted to TMX (Translation Memory eXchange) format using the ACIDE program, which is designed for creating parallel corpora (Obradović et al., 2008; Krstev and Vitas, 2011). Figure 1 on the left presents the samples of translation units, which contain the translation equivalents in tag `<tuv>`. The segments in Italian and Serbian are paired and numbered, with each segment indicating the language through the attribute "xml:lang". The ACIDE program not only creates the TMX document, but also generates an HTML representation, as shown in Figure 1 on the right.

The It-Sr-NER corpus[2] is available on the ILC4CLARIN B Center and can be accessed through the VLO (Virtual Language Observatory) [3]. The corpus, in a compressed format, includes the aligned bilingual version, as well as individual monolingual versions, and named entities that have been automatically tagged (as detailed in Section 3). The corpus and additional information can also be found in the Github[4].

The corpus, which includes the complete novels from which the published version of 10,000 segments were extracted, is not only downloadable but is searchable on the Bibliša[5] digital library. The left side of the Figure 2 presents browsing of documents (novels) with additional possibilities for authorised users for editing of metadata end aligned sentences (on the right).

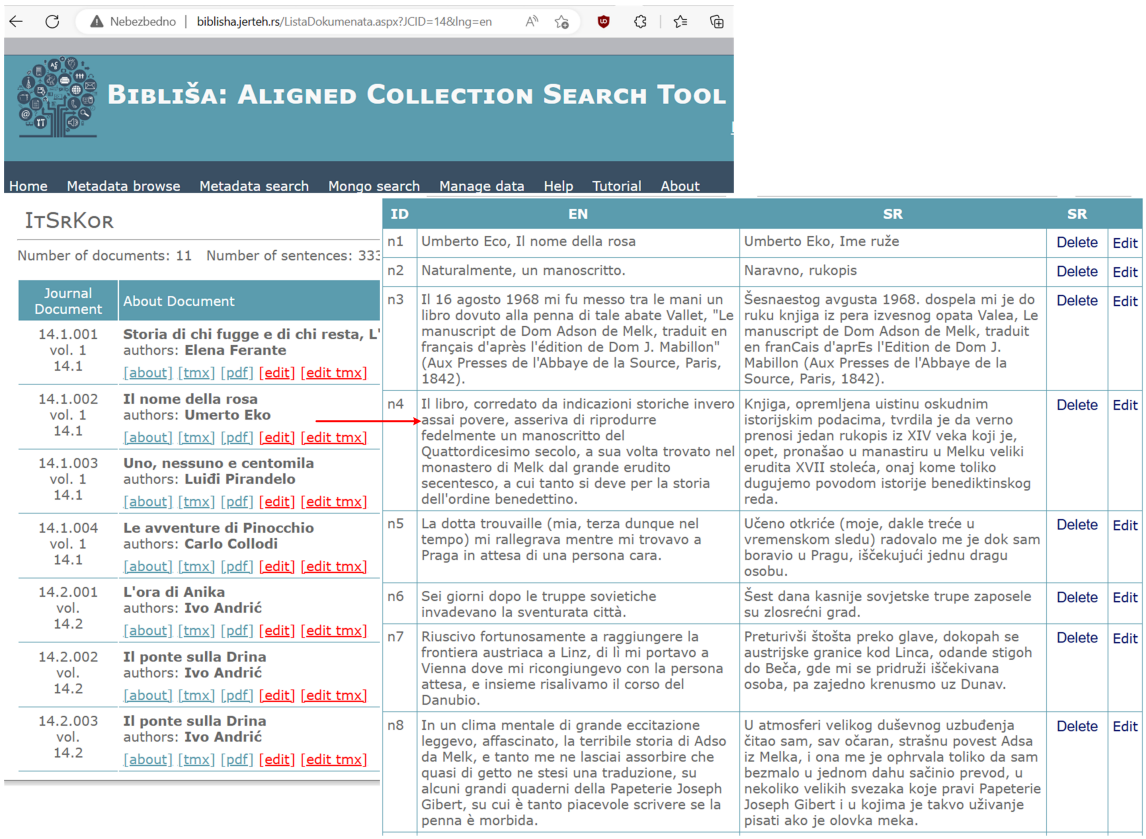
Parallel corpora are useful for translation research, and the use of concordances in contrastive linguistics can improve the study of cross-linguistic phenomena. The resources developed in this project can

[1] https://jerteh.rs/index.php/it-sr-ner-3/
[2] http://hdl.handle.net/20.500.11752/OPEN–980
[3] https://vlo.clarin.eu/
[4] https://github.com/jerteh/It–Sr–NER/tree/main/corpus
[5] http://biblisha.jerteh.rs

Italian (it)	Serbian (sr)
n2 Lasciai il tavolo frastornata, stentavo a prendere atto che Nino era davvero lì, a Milano, in quella saletta.	**n2** Ustadoh od stola pometena, bilo mi je teško da ubedim samu sebe da je Nino zaista tu, u Milanu, u toj maloj sali.
n3 Eppure eccolo, già mi veniva incontro sorridendo ma con passo controllato, senza fretta.	**n3** Pa ipak, eto ga kako mi ide u susret, sa osmehom ali odmerena koraka, bez žurbe.

```
<tu>
    <prop type="Domain"/>
    <tuv xml:lang="it" creationid="n2" creationdate="20211014T224355Z">
        <seg>Lasciai il tavolo frastornata, stentavo a prendere atto che Nino era davvero
            lì, a Milano, in quella saletta.</seg>
    </tuv>
    <tuv xml:lang="sr" creationid="n2" creationdate="20211014T224355Z">
        <seg>Ustadoh od stola pometena, bilo mi je teško da ubedim samu sebe da je Nino
            zaista tu, u Milanu, u toj maloj sali.</seg>
    </tuv>
</tu>
<tu>
    <prop type="Domain"/>
    <tuv xml:lang="it" creationid="n3" creationdate="20211014T224355Z">
        <seg> Eppure eccolo, già mi veniva incontro sorridendo ma con passo controllato,
            senza fretta.</seg>
    </tuv>
    <tuv xml:lang="sr" creationid="n3" creationdate="20211014T224355Z">
        <seg> Pa ipak, eto ga kako mi ide u susret, sa osmehom ali odmerena koraka, bez
            žurbe.</seg>
    </tuv>
</tu>
```

Figure 1: Two aligned segments of translation equivalents, Italian and Serbian (HTML left, TMX right).



ID	EN	SR	SR	
n1	Umberto Eco, Il nome della rosa	Umberto Eko, Ime ruže	Delete	Edit
n2	Naturalmente, un manoscritto.	Naravno, rukopis	Delete	Edit
n3	Il 16 agosto 1968 mi fu messo tra le mani un libro dovuto alla penna di tale abate Vallet, "Le manuscript de Dom Adson de Melk, traduit en français d'après l'édition de Dom J. Mabillon" (Aux Presses de l'Abbaye de la Source, Paris, 1842).	Šesnaestog avgusta 1968. dospela mi je do ruku knjiga iz pera izvesnog opata Valea, Le manuscript de Dom Adson de Melk, traduit en franCais d'aprEs l'Edition de Dom J. Mabillon (Aux Presses de l'Abbaye de la Source, Paris, 1842).	Delete	Edit
n4	Il libro, corredato da indicazioni storiche invero assai povere, asseriva di riprodurre fedelmente un manoscritto del Quattordicesimo secolo, a sua volta trovato nel monastero di Melk dal grande erudito secentesco, a cui tanto si deve per la storia dell'ordine benedettino.	Knjiga, opremljena uistinu oskudnim istorijskim podacima, tvrdila je da verno prenosi jedan rukopis iz XIV veka koji je, opet, pronašao u manastiru u Melku veliki erudita XVII stoleća, onaj kome toliko dugujemo povodom istorije benediktinskog reda.	Delete	Edit
n5	La dotta trouvaille (mia, terza dunque nel tempo) mi rallegrava mentre mi trovavo a Praga in attesa di una persona cara.	Učeno otkriće (moje, dakle treće u vremenskom sledu) radovalo me je dok sam boravio u Pragu, iščekujući jednu dragu osobu.	Delete	Edit
n6	Sei giorni dopo le truppe sovietiche invadevano la sventurata città.	Šest dana kasnije sovjetske trupe zaposele su zlosrećni grad.	Delete	Edit
n7	Riuscivo fortunosamente a raggiungere la frontiera austriaca a Linz, di lì mi portavo a Vienna dove mi ricongiungevo con la persona attesa, e insieme risalivamo il corso del Danubio.	Preturivši štošta preko glave, dokopah se austrijske granice kod Linca, odande stigoh do Beča, gde mi se pridruži iščekivana osoba, pa zajedno krenusmo uz Dunav.	Delete	Edit
n8	In un clima mentale di grande eccitazione leggevo, affascinato, la terribile storia di Adso da Melk, e tanto me ne lasciai assorbire che quasi di getto ne stesi una traduzione, su alcuni grandi quaderni della Papeterie Joseph Gibert, su cui è tanto piacevole scrivere se la penna è morbida.	U atmosferi velikog duševnog uzbuđenja čitao sam, sav očaran, strašnu povest Adsa iz Melka, i ona me je ophrvala toliko da sam bezmalo u jednom dahu sačinio prevod, u nekoliko velikih svezaka koje pravi Papeterie Joseph Gibert i u kojima je takvo uživanje pisati ako je olovka meka.	Delete	Edit

Figure 2: Aligned text from ItSrKor in Bibliša digital library.

be utilized by students of Italian language in Serbia and Serbian language in Italy as they are open and accessible to other students and researchers in tertiary and pre-tertiary education.

3 Web Services for Named Entity Recognition and Linking

The It-Sr-NER services [6], which are stored in the CLARIN repository, can process not only monolingual texts in 24 languages, but also bilingual texts (represented in the TMX format), and successfully annotate them.

[6] http://hdl.handle.net/20.500.11752/OPEN–981

The ultimate goal of the project was to integrate the developed web services into the European infrastructure for language resources and technologies, specifically the Language Resource Switchboard platform[7]. The initial aim to annotate named entities for Italian and Serbian was later extended to other languages for which models were available. These models, which were trained using the spaCy[8] library, were downloaded for each language from the corresponding repository [9] in order to incorporate them.

For Italian, the *it_core_news_sm3.4.0* model, which was trained on the automatically created corpus, *WikiNER*[10], based on Wikipedia text and structure (Nothman et al., 2013), was used. For Serbian, a model trained on the *SrpCNNER* corpus of old Serbian novels (Šandrih Todorović et al., 2021) available on the European Language Grid (ELG) platform was used[11].

The University Library of Mannheim developed an open source system OpenTapioca[12], which links named entities to concepts in Wikidata (Delpeuch, 2019). By using the spaCy wrapper spaCyOpenTapioca[13], the application can not only recognize and annotate named entities, but also link them with items in Wikidata. The final outcome is a web service that can display recognized named entities on a map.

Four types of web services have been developed: NER, NER+NEL, NEL and geoparsing. Further on, for each service type two services were developed: one for monolingual and one for bilingual resources.

- The NER (Named Entity Recognition) process uses trained language models from the spaCy library to recognize named entities based on the classes listed in Table 1.
- NER+NEL is an extension of the NER process. In addition to recognizing named entities, it also links the annotated entities with Wikidata when possible. This is achieved by using the functions of the spacyOpenTapioca service, and is applied only to the recognized named entities, which are the text inside the XML tag.
- NEL (Named Entity Linking) is the process of recognizing and linking named entities with Wikidata, using the recognition capabilities of the spacyOpenTapioca system. The recognized named entities are annotated with the tag `<WDT>` and the class of the named entity is identified using the label attribute.
- Geoparsing - using the geopylibrary[14] for geolocating named entities of the *LOC* class that are present in wikidata, and then displaying them on a map using the folium[15] library.

To standardize the labels used for different classes of named entities across language-specific models, unification of tagset is prepared as presented in Table 1. The *PERS* class label, which marks persons, was set as the default label to which corresponding labels from other models were mapped. Furthermore, the labels denoting locations and geopolitical entities have been unified to the *LOC* label, regardless of any other labels that may have been used (such as *GPE*, *LC*, *placeName*, or *geogName*).

The label *NORP* (nationalities or religious or political groups) for nationalities, political and religious groups from Japanese and Finnish models and *NAT_REL_POL* from Romanian model were mapped to the *DEMO* label, which denotes demonyms and ethnic relations (Stanković et al., 2021). This mapping is done consistently for all classes and can be found in the configuration file[16]. Since some language models have a more extensive set of named entity classes, for example, English has 18 classes and Romanian has 16, a column for ignored labels is defined in the configuration file.

[7] https://switchboard.clarin.eu/tools
[8] https://spacy.io/
[9] https://spacy.io/models
[10] https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500
[11] https://live.european–language–grid.eu/catalogue/ld/9484
[12] https://opentapioca.org
[13] https://pypi.org/project/spacyopentapioca/
[14] https://pypi.org/project/geopy/
[15] https://python–visualization.github.io/folium/
[16] https://github.com/jerteh/It–Sr–NER/blob/main/config/lng_config.csv

NER class tag	Description of the entity class	Mapped also
PERS	Names, surnames, nicknames and their combinations (of real people and fictional characters, including gods and saints).	PER (French, German, Italian, Portuguese, Spanish...), PRS (Swedish), PERSON (English, Finnish, Greek, Macedonian...), persNAME (Polish), PS (Korean)
LOC	Continents, countries, regions, settlements, oronyms, bodies of water, names of celestial bodies, city locations.	LOC+GPE (Chinese, English, Finnish, Greek, Macedonian, Romanian...), LC (Korean), placeName and geogName (Polish)
ORG	Names of companies, political parties, educational institutions, sports teams, hospitals, museums, libraries, hotels, cafes, churches and shrines.	ORGANIZATION (Romanian), orgName (Polish), ORG+GPE_ORG (Norwegian Bokmål), OG (Korean)
DEMO	Residents of countries, cities, regions or ethnic groups; derived adjectives from the name of the location.	NORP (Chinese Dutch, English, Finnish, Japanese, Macedonian), NAT_REL_POL (Romanian)
EVENT	Names of events that recur regularly or happened once but they have their own name: natural disasters, revolutions, battles, wars.	EVN (Swedish), EVT (Norwegian Bokmål)
WORK	Titles of books, plays, poems, paintings, sculptures, newspapers.	WORK_OF_ART (Romanian, Dutch, English, Japanese, Macedonian, Finnish), WRK (Swedish)

Table 1: Named entity classes.

Web services that use named entity linking (NER+NEL and NEL) provide additional information about named entities as *xml* attributes: the entity type (*label*), description (*desc*), and a link to the Wikidata knowledge base (*ref*), in addition to the classes already associated to the entities.

It was previously stated that the input can be either monolingual or bilingual text. For bilingual resources, the input must be in the form of a valid *TMX* document. The output of three services for bilingual resources is shown in Figure 3, where the first possibility is NER, the second is NER+NEL, and the third is NEL service, with spacyOpenTapioca-based services linking recognized named entities to items in Wikidata for both languages.

The program code, web services, web application, and parallel corpora from the project have all been released[17] under open licenses, allowing for free use in research and commercial activities.

The primary development of the project took place from June to September 2022, with further adjustments made during the subsequent fine-tuning phase. In order to achieve all the results, the core team of four researchers received support from an additional three researchers. Adapting the web service to function with the CLARIN infrastructure presented a challenge, but thanks to the assistance of the CLARIN team, the verification and publication of the service were successful. The evaluation was conducted on a limited dataset for Serbian and Italian. Results revealed that PERS and ORG were better identified in Italian than in Serbian, while Serbian LOC performed better. The Italian model did not include DEMO, WORK, and EVENT. Overall, the evaluation indicates the need for further model improvement.

[17] https://github.com/jerteh/It–Sr–NER

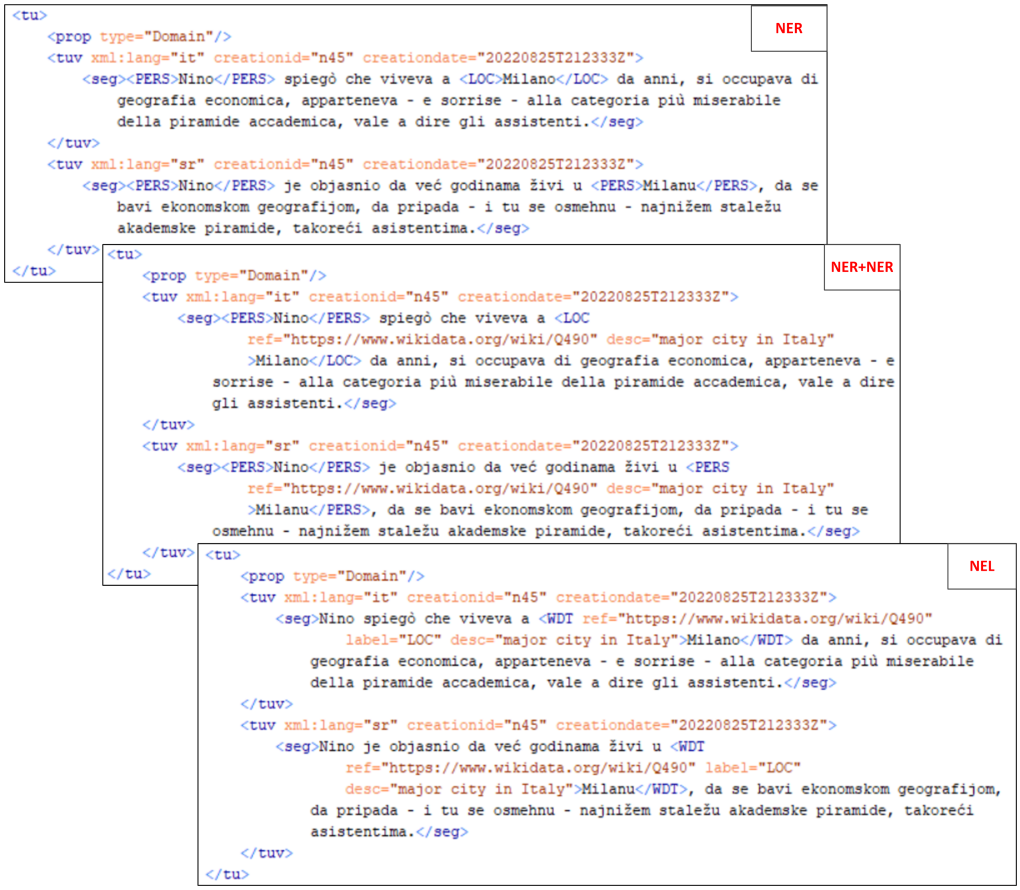


Figure 3: NER, NER+NEL and NEL output for bilingual resources in TMX format.

4 Use Cases

The web services discussed in Section 3 can be accessed and utilised in a variety of ways. Figure 4 illustrates the web services integrated on the Language Resource Switchboard platform. Bilingual resources must be inputted as an XML file (in TMX format), while monolingual resources can be submitted as a text file or entered directly into the provided field on the web application form. The integration of the web service in the CLARIN infrastructure allows for greater visibility and accessibility for researchers and educators, and the ability to easily share resources and collaborate on projects. (de Jong et al., 2022; Draxler et al., 2022)

Figure 5 illustrates an example of the results of processing a bilingual text (submitted as a TMX document) on the CLARIN platform Language Resource Switchboard using the NER+NEL service. The output shows the processing results for both languages presented simultaneously, displaying the named entities recognized and linking them to knowledge base. Each named entity category is color-coded, in order to better visualize the results to the end user.

The figure also illustrates the capability of displaying the link to wikidata and description of an item (determiner), in this case, Florence (Q2044). It is shown that the recognized named entity Florence (*Firenze* in Italian) is associated with an underlined style, which is a feature of the web services. Users can hover over the underlined text to see the description of the item, this is an additional feature provided to help users understand the context and meaning of the named entities.

The web services described are also accessible via the web application at https://ners.jerteh.rs. This is an implementation of the previously mentioned web application (developed for the It-Sr-Ner project), with embedded API endpoints, that are targeted by the app instance and the Switchboard alike. Since

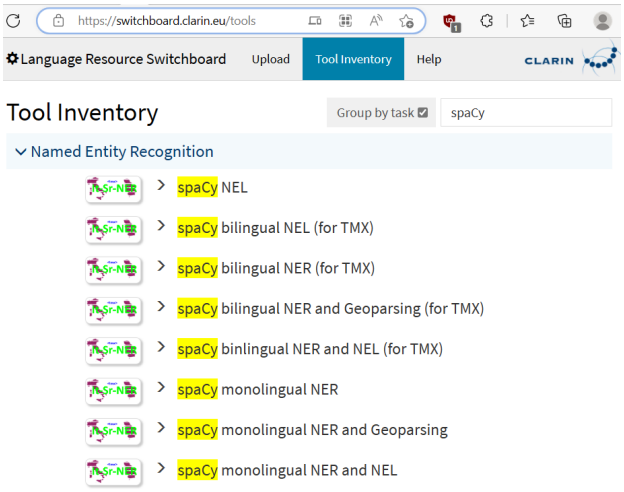


Figure 4: Presentation of the integrated web service on the CLARIN infrastructure.

Figure 5: Display of bilingual text processing using the NER service.

the API-s are opened to the web, they can also be integrated into other applications (e.g. for Python applications by using the requests module). Also, since the complete application is available in open access, other instances of it can be run on user-computers locally (which requires certain packages to be preinstalled) or run as another instance (with the same capabilities) on the web. With each of these methods providing access to the services with the same functionality, users can choose the method that

best suits their needs to access the web services. Here's an example how to send requests using the Python requests module to access the web services:

```
# Define the endpoint and parameters
endpoint = "https://ners.jerteh.rs/ner"
params = {"text": "example text", "language": "en"}
# Send the request
response = requests.post(endpoint, json=params)
# Print the response
print(response.json())
```

In this example, the endpoint is set to the Named Entity Recognition service and the parameters include the text to be processed and the language of the text. The requests module is used to send a POST request with the parameters as JSON. The response is then printed in JSON format. This example can be adapted to use other services and parameters as needed.

```
import requests
# choose language - lang
# @param ['ca','zh','hr','da','nl','en','fi','fr','de','el','it',
'ja','ko','lt','mk','nb','pl','pt','ro','ru','es','sv','uk','sr']
lang = "it"
# choose service option - feat
# @param ['ner', 'nel', 'nernel','geo']
feat = "nel"
# use api
API_KEY = ["file", "data", "lng", "feat"]
url = 'https://ners.jerteh.rs/api'
params = dict(key=API_KEY, data=data, lng=lang, feat=feat)
res = requests.get(url, params=params)
```

All of the mentioned services offer two different formats for displaying the processing results: HTML and XML. This is demonstrated in the following example. Figure 6 shows the processing of text entered directly into the text box for Italian. A selection of language and the NER service option were also selected, and as a result, HTML was generated. The frame with the resulting HTML contains javascript-powered button that, when clicked, downloads the mentioned XML result to a local computer. Additionally, using the NEL and NER+NEL services, the output is similar, but it includes links to annotated wikidata items. Furthermore, this service also provides a description of the entity by mouse-over event on the entity, by using the description of the corresponding item in wikidata. Current implementation is using description of tagged named entities in English, provided by embedded library, regardless of the text language. Descriptions in English are the most common, because wikidata is the most developed for that language, so it is implemented in this version. In the next versions of the service, the approach will be modified so that the description language corresponds to the text's language.

As with the previous web services, geoparsing is available for both bilingual and monolingual resources. Only recognized named entities of the *LOC* class by *NER+NEL* are displayed on the (HTML-based) map. Figure 7 illustrates geoparsing for a small monolingual text in Italian. It shows the location of the named entities recognized in the text on a map, providing a visual representation for each location mentioned in the text, which can be useful for various research and educational purposes.

It should be noted that for different languages, there may be variations in the recognition of named entities and differences in geoparsing due to several reasons:

- For the given language, there may not be a corresponding item (headword) in the knowledge base for the annotated entity.
- The named entity in Serbian (or any other language with rich inflections) may not be recognized because the system does not recognize inflected forms for the language (such as cases different from the nominative singular: *Srbije*, *Beogradu*, etc.).
- The translation equivalents (in this case study Serbian and Italian) may not be literal, so the named entity may not appear in one of the equivalents (see segment number 1994 in Figure 5 and entity *Mar Nero*).

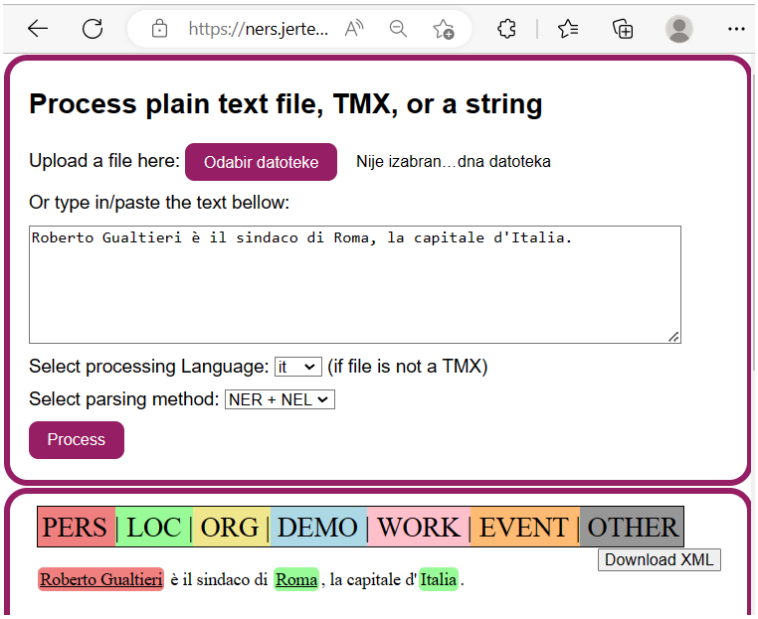


Figure 6: Display of directly entered text processing.

These variations and differences occur due to the specific characteristics of the languages and the resources used. However, the services were developed to handle these variations and differences as much as possible and to provide accurate and useful results for the end users.

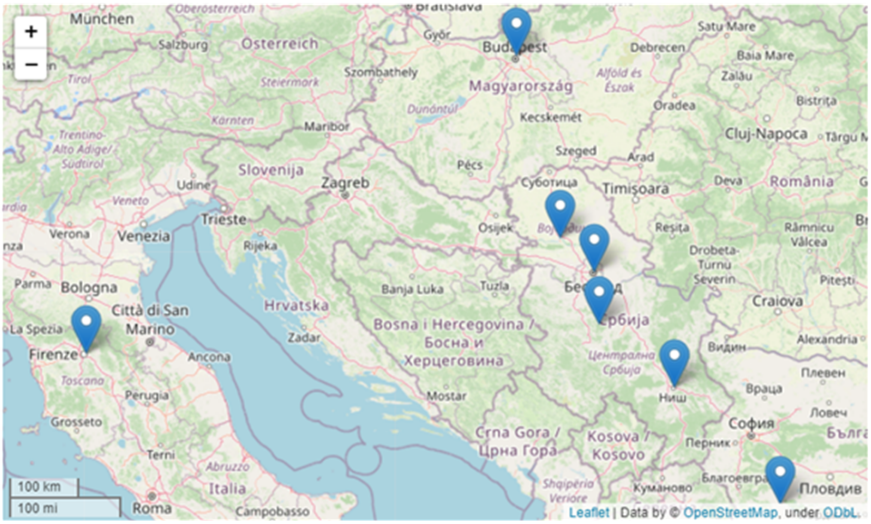


Figure 7: Geoparsing of text and presentation of locations on OpenStreetMap.

5 CQL Corpus Search

Sketch Engine[18] is a widely used tool for exploring the workings of language, based on the analysis of corpora compiled from authentic texts of billions of words. The Sketch Engine allows users to search for a word, phrase or pattern, and results can be presented in various forms such as word sketches, concordances, word lists, frequency graphs, sketch differences, etc. This tool is widely used by researchers,

[18]https://www.sketchengine.eu/

educators and linguists to analyze and understand language usage, patterns and trends. It was developed by Adam Kilgarriff and his team at the Institute of Formal and Applied Linguistics at the Charles University in Prague and it is a popular tool in the field of corpus linguistics. With its sophisticated features, it allows users to easily extract information from large corpora and analyze it in different ways. (Kilgarriff et al., 2004; Kilgarriff et al., 2014)

NoSketch Engine is an open source edition of the Sketch Engine, which offers core corpus processing and search features, but does not include advanced features such as word sketches and preinstalled corpora. A NoSketch Engine node[19] is installed and maintained by the Society for Language Resources and Technologies JeRTeh, and provides access to several monolingual and bilingual corpora, some of which are available to authorized users only.

The *ItSrNER* corpus in NoSketch is part of speech annotated and lemmatized using TreeTagger[20] (Schmid, 1999). The Italian part of corpus is tagged using TreeTagger parameter file prepared by Prof. Achim Stein, University of Stuttgart (Schmid et al., 2007) with 38 tabs in the tagset [21]. The Serbian parametric language parameter file is trained on the harmonized resources, which have been manually annotated within different projects (Stanković et al., 2020), consulting the system of morphological electronic dictionaries of the Serbian language (Krstev, 2008; Vitas and Krstev, 2012). The POS tagset for Serbian part is Universal Dependencies Tagset [22]

The ItSrNER corpus can be freely accessed and searched using CQL (Corpus Query Language). Figure 8 presents parallel concordances of ItSrNER corpus for query: *family* (Italian: *famiglia*, Serbian: *porodica*) with an option that NER tags are visible.

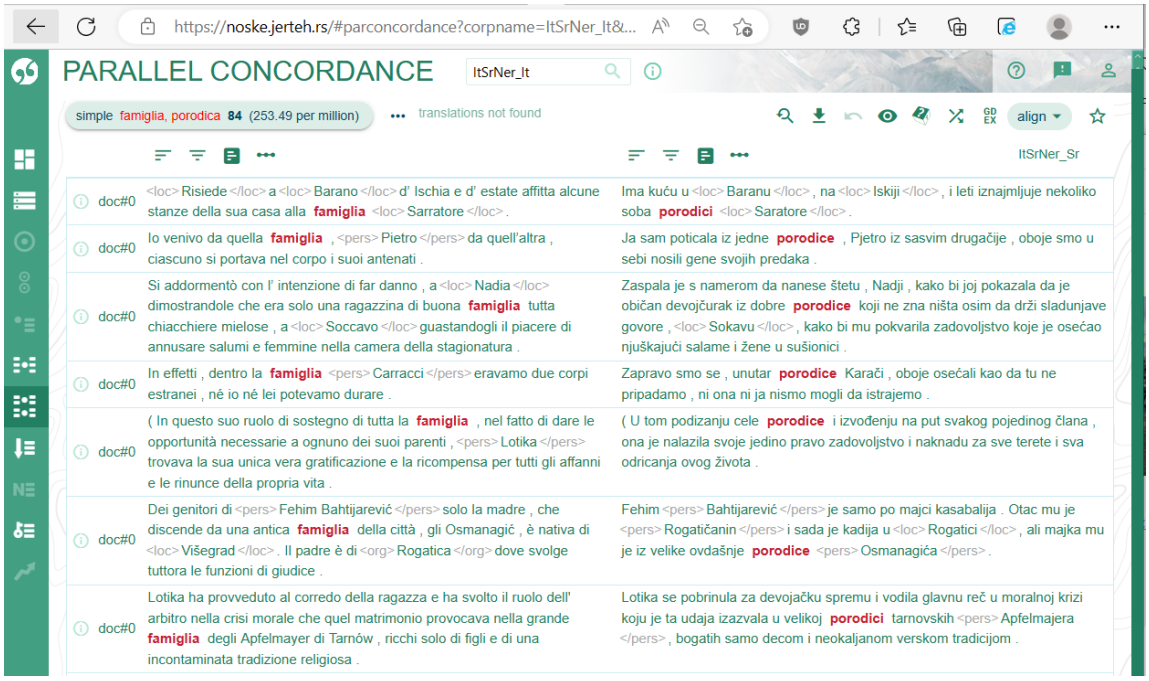


Figure 8: Annotated aligned ItSrNER corpus with NER tags on on NoSketch engine platform.

When researching the translation of location names, including those with adjectives, advanced CQL queries can be used to limit the search to the context of annotated location named entities. Figure 9 presents a page with a simple CQL query `[tag="ADJ"]+[tag="NOM"] within <loc/>`, which retrieves concordances with the names of locations in the following form: noun preceded by one or more adjectives. The accuracy of the labeling of named entities, especially the class association of the named

[19] https://noske.jerteh.rs/#dashboard?corpname=srpELTeC
[20] https://www.cis.lmu.de/ schmid/tools/TreeTagger/
[21] https://www.cis.lmu.de/ schmid/tools/TreeTagger/data/italian-tagset.txt
[22] https://universaldependencies.org/u/pos/index.html

entity, is not always correct, as can be seen in the first example.

Figure 9: Advanced CQL query within `<loc>` tag.

6 Conclusion

In the paper, we discussed the outcomes of the It-Sr-NER project, which is a web service for annotating named entities for 24 languages and displaying them on a map, with the case study on Italian and Serbian parallel texts. The project was supported by the Common Language Resources and Technology Infrastructure, CLARIN ERIC, and involved a collaboration between the University of Turin and the Society for language resources and technologies JeRTeh. The goal of the project was to improve the teaching of Italian and Serbian languages and to support translation studies. The lack of specific language technologies for the Serbian language has for years been an obstacle in the introduction of the new methodologies in teaching like corpus-based and Data Driven learning. Isolated efforts to incorporate corpora into teaching, although efficient, do not provide enough incentive for researchers and educators in the field of teaching Serbian as a foreign language. At the same time if the students are introduced to corpora and other linguistic tools through proper training, they may gradually develop researcher attitude which allow them to be more creative and to participate actively in the construction of their own learning process.

The primary outcome of the project was the release of a suite of web services for monolingual and bilingual parallel texts available on the CLARIN platform Language Resource Switchboard. Additionally, the project accomplished several secondary objectives that were equally important, such as the creation and publication of a parallel Italian-Serbian corpus, and the development of a web application and service on the JeRTeh platform for language resources and technologies. In total, eight services were created, four for monolingual and four for bilingual resources. These services can process text through direct input at the sentence level, or by processing user-uploaded files. The services also include linking of named entities with wikidata and geoparsing. While the project focused on Serbian and Italian resources, the developed services are capable of processing texts in 24 languages.

Additional research will be conducted to promote the use of the web services and integrate them into teaching. A key objective is to expand the corpus and enhance the model for annotating named entities and linking them to knowledge bases.

Acknowledgements

The authors are thankful to CLARIN ERIC, Common Language Resources and Technology Infrastructure, for supporting our project within the ”Bridging Gaps Call 2022”. The authors are also grateful to prof. Cvetana Krstev, prof. Duško Vitas, prof. Saša Moderc and Nikola Janković for providing the parallelization.

References

[de Jong et al.2022] Franciska de Jong, Dieter Van Uytvanck, Francesca Frontini, Antal van den Bosch, Darja Fišer, and Andreas Witt. 2022. Language matters. the european research infrastructure clarin, today and tomorrow. In *CLARIN. The infrastructure for language resources*, pages 31–57. de Gruyter.

[Delpeuch2019] Antonin Delpeuch. 2019. Opentapioca: Lightweight entity linking for wikidata. *CoRR*, abs/1904.09131.

[Draxler et al.2022] Christoph Draxler, Alexander Geyken, Erhard Hinrichs, Annette Klosa-Kückelhaus, Elke Teich, and Thorsten Trippel. 2022. How to connect language resources, infrastructures, and communities. *CLARIN: The Infrastructure for Language Resources*, 1:275.

[Granger2018] Sylviane Granger. 2018. Has lexicography reaped the full benefit of the (learner) corpus revolution? In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, page 208.

[Kilgarriff et al.2004] Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. Itri-04-08 the sketch engine. *Information Technology*, 105(116).

[Kilgarriff et al.2014] Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

[Krstev and Vitas2011] Cvetana Krstev and Duško Vitas. 2011. An aligned english-serbian corpus. *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality)*, 1:495–508.

[Krstev2008] Cvetana Krstev. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries*. Faculty of Philology of the University of Belgrade.

[Moderc2015a] Saša Moderc. 2015a. Su un modo di tradurre l'avverbio serbo "inače" in italiano: il caso dell'equivalente "altrimenti". *Università di Belgrado. In Italica Belgradensia*, 1:61–79.

[Moderc2015b] Saša G. Moderc. 2015b. Elektronski korpus srpskih književnih dela i njihovih prevoda na italijanski jezik. *Anali Filološkog fakulteta*, 27(2):301–316. 15.

[Nothman et al.2013] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

[Obradović et al.2008] Ivan Obradović, Ranka Stanković, and Miloš Utvić. 2008. Integrisano okruženje za pripremu paralelizovanog korpusa. *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*, pages 563–578.

[Perišić Arsić2018] Olja Perišić Arsić. 2018. L'uso dei corpora nella didattica della traduzione: l'esempio del verbo serbo "prijati" e i suoi traducenti italiani. *Italica Belgradensia*, 2018(1):49–64. 3.

[Perišić2021] Olja Perišić. 2021. Corpora in the classroom-the case of the serbian language for italian speakers. *New Trends in Slavic Studies*, pages 126–137.

[Perišić2023] Olja Perišić. 2023. *Il corpus per imparare il serbo. Il futuro dell'apprendimento linguistico*. Edizioni dell'Orso.

[Šandrih Todorović et al.2021] Branislava Šandrih Todorović, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. Serbian ner&beyond: The archaic and the modern intertwinned. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1252–1260.

[Schmid et al.2007] Helmut Schmid, Marco Baroni, Erika Zanchetta, and Achim Stein. 2007. Il sistema 'treetagger arricchito'–the enriched treetagger system. *IA Contributi Scientifici*, 4(2):22–23.

[Schmid1999] Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

[Sinclair1991] J. McH. Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

[Stanković et al.2020] Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian. In *Proc. of The 12th LREC*, pages 3947–3955, Marseille, France. European Language Resources Association.

[Stanković et al.2021] Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, and Mihailo Škorić. 2021. Annotation of the serbian eltec collection. *Infotheca*, 21(2):43–59. 3.

[Vitas and Krstev2012] Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, XVIII:279–292.

[Vitas and Lažetić-Pavlović2008] Duško Vitas and Gordana Lažetić-Pavlović. 2008. Resources and Methods for Named Entity Recognition in Serbian. *Infotheca*, 9(1–2):35a–42a, May.

[Vitaz and Poletanović2020] Milica Vitaz and Milica Poletanović. 2020. Data-driven learning the serbian case. *EL.LE*, pages 409–422, april.