It-Sr-NER: Web Services for Recognizing and Linking Named Entities in Text and Displaying Them on a Web Map

Olja Perišić, Ranka Stanković, Milica Ikonić Nešić, Mihailo Škorić

It-Sr-NER: Web Services for Recognizing and Linking Named Entities in Text and Displaying Them on a Web Map | Olja Perišić, Ranka Stanković, Milica Ikonić Nešić, Mihailo Škorić | Infotheca | 2023 | |

10.18485/infotheca.2023.23.1.3

http://dr.rgf.bg.ac.rs/s/repo/item/0007790

Scientific paper

It-Sr-NER: Web Services for Recognizing and Linking Named Entities in Text and Displaying Them on a Web Map

UDC 81'322.3
DOI 10.18485/infotheca.2023.23.1.3

ABSTRACT: The paper will present the results of the project "It-Sr-NER: Web services for named entities recognition, linking and mapping," in which teams from the University of Turin and the Society for Language Resources and Technologies JeRTeh participated, and whose goal was the development of the *It-Sr-NER* web service for named entity annotations in the text and displaying them on the map. Named entities in these services are names of persons, places, organizations, demonyms (ethnicities), events and works of art.

KEYWORDS: parallel corpora, named entities, NER, NEL, geoparsing, Serbian, Italian, web services.

PAPER SUBMITTED: 29 November 2022
PAPER ACCEPTED: 31 January 2023

Olja Perišić
olja.perisic@unito.it
Università degli Studi di Torino
Dipartimento di Lingue e Letterature Straniere e Culture Moderne
Turin, Italy

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

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

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

1 Introduction

The lack of tools and resources that enable the annotation, research and analysis of bilingually aligned (parallel) Italian-Serbian texts was the main motivation and inspiration for starting this project. In the didactics of foreign languages in Serbia, there is an almost complete absence of corpus tools

in teaching (Vitaz and Poletanović 2020), while on an individual level and through personal initiatives in the teaching of the Serbian language as a foreign language in Italy and the Italian language in Serbia, it has been shown that corpora in teaching are significant in many ways and that students are glad to accept and apply them in shared work, but also in independent research (Moderc 2015a; Perišić 2021). The rich and sophisticated morphology of the Serbian language provides for the declension of toponyms and other named entities that foreign students are not always able to recognize and reduce to their basic form. Some of the reasons are the same endings for the masculine and neuter gender in most cases, the presence of certain toponyms only in the plural form, the so-called *pluralia tantum* (Pljevlja, Divčibare, etc.), but also phonetic transcription of foreign names as well as some orthographic inconsistencies (Витас and Павловић-Лажетић 2008).

As part of the "Bridging Gaps" call of the European infrastructure for language resources and technologies CLARIN[1] (Common Language Resources & Technology Infrastructure), a team of experts from the University of Turin and the Society for Language Resources and Technologies JeRTeh joined forces and developed web services for annotating named entities in text, ensuring their linking with Wikidata,[2] as well as geoparsing, i.e. geolocation of recognized locations and their display on the map. Named entities in these services are names of persons, places, organizations, demonyms (ethnicities), events and works of art.

The main goal of the project is the realization and publication of 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 envisages the creation and publication of the Italian-Serbian corpus of 10,000 segments of extracted and aligned sentences, selected from classics of Italian and Serbian literature. The results of the project are not limited to the Serbian-Italian language combination, but the developed services can be applied to the processing of texts in twenty-four different languages.

The project initiator and team leader is Olja Perišić, a professor at the University of Turin (*Università degli Studi di Torino, Dipartimento di Lingue e Letterature Straniere e Culture Moderne*) where she teaches the Serbian language. On behalf of JeRTeh, the development of the services was led by prof. Ranka Stanković in cooperation with prof. Duško Vitas. More information about the project is also available on the website of the Society for

1. CLARIN
2. Wikidata

Language Resources and Technologies JeRTeh It-Sr-NER CLARIN compatible NER and geoparsing web services for parallel texts.[3]

2 Parallel corpus

In foreign language teaching, parallel corpora have proved to be an indispensable instrument in order to acquire morphosyntax and lexis. The parallel observation of two or more languages facilitates contrastive analysis, i.e. the observation of similarities and differences of language structures thanks to the large number of examples of sentences in context, as was noticed by Sinclair at an early stage of corpus linguistics: "The language looks rather different when you look at a lot of it at once." (Sinclair 1991, 100).In teaching translation, parallel corpora enable word sense disambiguation and the definition of polysemic lexicon, which are not given enough space in bilingual dictionaries (Moderc 2015b; Perišić Arsić 2018). At the same time, it has been noticed that the number of representative, parallel corpora even for major world languages is insufficient (Granger 2018).

For all these reasons, in the first phase of the project, it was necessary to create an Italian-Serbian corpus of 10,000 aligned segments (sentences) excerpted from ten different novels. The novels are presented with random samples of segments in order to avoid copyright issues. Novels by Italian writers represented in the corpus are: 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, None and a Hundred Thousand*. Serbian writers are represented by five novels: Ivo Andrić, *Legends of Anika* and *The Bridge on the Drina*; Borisav Stanković, *Impure Blood*; Branislav Nušić, *Municipal child: the novel of an infant*; Danilo Kiš, *Garden, Ashes*. Considering that the main task of the project is to annotate the named entities, the corpus also includes translations into Italian and Serbian of Jules Verne's novel *Around the World in Eighty Days*.

The novels were aligned and prepared in TMX (Translation Memory eXchange) format using the ACIDE application, an integrated environment for the development of parallel corpora (Obradović, Stanković, and Utvić 2008; Krstev and Vitas 2011). Figure 1 presents the first two translation units, marked with the label <tu> (translation unit), within which there are translation equivalents marked with the label <tuv> (translation unit variant). Paired, i.e. aligned segments in Italian and Serbian are numbered (n1,

3. https://jerteh.rs/index.php/it-sr-ner-3/

```
<tu>
    <tuv xml:lang="it" creationid="n1" creationdate="20220825T211907Z">
        <seg>Sposa giovanissima Stefano Carracci e gestisce con
        successo prima la salumeria nel nuovo rione, poi il negozio di
        scarpe a piazza dei Martiri.</seg>
    </tuv>
    <tuv xml:lang="sr" creationid="n1" creationdate="20220825T211907Z">
        <seg>Veoma mlada se udaje za Stefana Karačija i uspešno
        upravlja isprva delikatesnom radnjom u novom rejonu, a potom
        obućarskom radnjom na Trgu mučenika.</seg>
    </tuv>
</tu>
<tu>
    <tuv xml:lang="it" creationid="n2" creationdate="20220825T211907Z">
        <seg>Elena comincia a scriverla nel momento in cui apprende
        che la sua amica d'infanzia, Lina Cerullo, solo da lei
        chiamata Lila, è sparita.</seg>
    </tuv>
    <tuv xml:lang="sr" creationid="n2" creationdate="20220825T211907Z">
        <seg>Elena počinje da je piše kada sazna da je nestala njena
        prijateljica iz detinjstva, Lina Čerulo, koju samo ona zove
        Lila.</seg>
    </tuv>
</tu>
```

Figure 1. Example of a TMX output document

n2,...) and each has an attribute indicating the language: `xml:lang="it"` or `xml:lang="sr"`. The parallelization process with the ACIDE application, in addition to generating a TMX output document, also generates an HTML representation that can be seen in Figure 2.

Italian (it)	Serbian (sr)
n1 Sposa giovanissima Stefano Carracci e gestisce con successo prima la salumeria nel nuovo rione, poi il negozio di scarpe a piazza dei Martiri.	**n1** Veoma mlada se udaje za Stefana Karačija i uspešno upravlja isprva delikatesnom radnjom u novom rejonu, a potom obućarskom radnjom na Trgu mučenika.
n2 Elena comincia a scriverla nel momento in cui apprende che la sua amica d'infanzia, Lina Cerullo, solo da lei chiamata Lila, è sparita.	**n2** Elena počinje da je piše kada sazna da je nestala njena prijateljica iz detinjstva, Lina Čerulo, koju samo ona zove Lila.

Figure 2. Example of aligned segments of translation equivalents in Italian and Serbian

The corpus is published in the ILC4CLARIN B Center, with the unique identifier http://hdl.handle.net/20.500.11752/OPEN-980, so it is visible

through the VLO (Virtual Language Observatory) https://vlo.clarin.eu/. The zipped form of the corpus contains several parts: aligned bilingual, but also individual, monolingual versions. Automatically tagged named entities (as explained in Section 3) are also part of the published corpus. The corpus can also be accessed from the working *github* site https://github.com/rankastankovic/It-Sr-NER/tree/main/corpus.

In addition to being downloadable, the corpus containing complete novels from which the published version of 10,000 segments were extracted, is also available for searching on the Bibliša[4] digital library. A feature of the Bibliša library is an advanced search with the possibility of morphological and semantic expansion of queries for the Serbian language. Figure 3 shows

metadata	**n3330** E si riunivano per lo più dietro la casa, presso le finestre della stanza grande, perchè li erano del tutto isolati dai contadini e dagli ultimi venuti e potevano gemere liberamente e inosservati, uccisi e irrigiditi dal dolore esclamare: - Ahimè, bato!	**n3330** A najviše ih je bilo iza kuće, do samih prozora velike sobe, jer tu su sasvim odvojeni od ovih varošana i od dolazećih i mogli slobodni, neopaženi, da hukću, ubijeni i zgrčeni: - Oh, bre, **bato!**
metadata	**n31** Specialmente le vedove, i cui figli appena cresciuti si davano a spendere e a sprecare, invece di pensare alla casa e a sostituire il padre, il capo della famiglia, facevano paura a questi lor figli, ricordando il "loro bato" [Bato = fratello], come tutti lo chiamavano in famiglia, e minacciavano.	**n31** Osobito udovice, čiji sinovi tek što nastali, pa mesto da preduzmu i počnu voditi brigu o kući, da zamene oca, domaćina, a oni počeli trošiti i rasipati - osobito su one te svoje sinove jednako njime, **„batom** svojim", kako su ga svi u rodbini zvali, zastrašivale i pretile im:
metadata	**n4780** Dede ha bisogno di compagnia, crescere da soli è brutto, è meglio darle un fratellino o una sorellina.	**n4780** Za Dede će biti dobro da dobije drugaricu, nije lako odrastati sam, bolje je da ima **batu** ili sestricu.
metadata	**n2565** Anche quelli di Sofia facevano a gara nel servire, stringendosi intorno a Marco, per dimostrare agli altri quanto bene volevano al loro nuovo amico, al loro bato.	**n2565** A Sofkini opet, radi njih, da bi pred njima pokazali koliko oni vole i cene svoga novog prijatelja, tog njihovog **„batu"**, svi su se utrkivali u služenju, obletanju oko Marka.
metadata	**n2523** «Nadia e suo fratello».	**n2523** „Nadja i njen **brat.**"

Figure 3. An example of expanding the search in Bibliša

a panel with the results of a semantic expansion example where the query "brat" ("brother") is automatically expanded with the synonym "bata" (little brother), and then their morphological forms are added to the query. Several segments can be seen in Figure 3.

4. Библиша

The option to view collections of parallel documents (Figure 4) is available with a limited number of segments for all users and with a larger number of segments for authorized users.

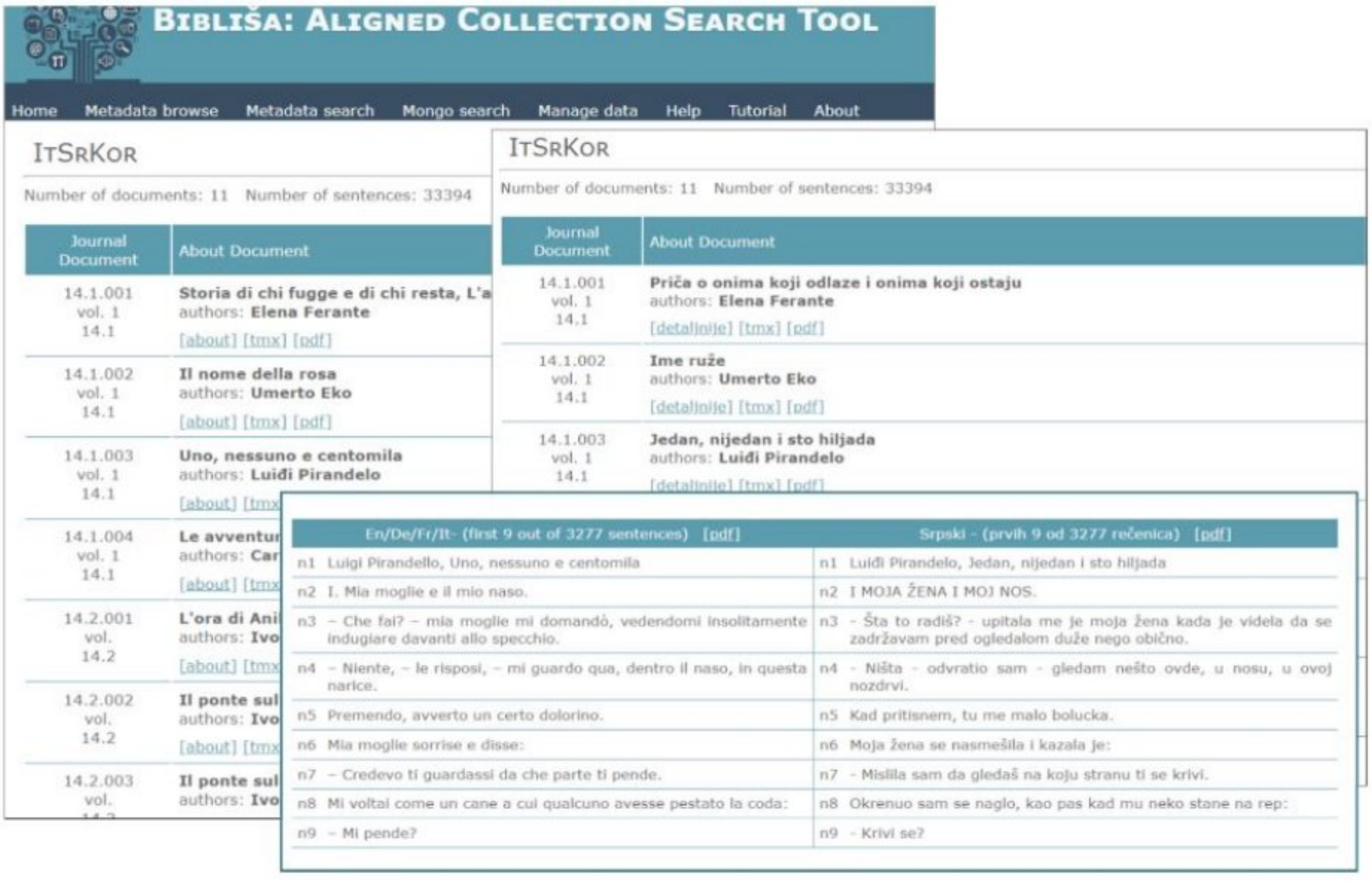

Figure 4. Browsing a collection of parallel documents ItSrKor

Parallel corpora are valuable for translation studies, while contrastive linguistics and the simple use of concordances facilitate the study of cross-linguistic phenomena (Figure 5). Students of the Italian language at the University of Belgrade, where prof. Saša Moderc teaches, as well as students of the Serbian language in Turin, where prof. Olja Perišić teaches, will be able to use the developed resources, considering that they are open and therefore available to other students and researchers.

2.1 Service implementation

It-Sr-NER services,[5] stored in the CLARIN repository with the unique identifier http://hdl.handle.net/20.500.11752/OPEN-981, not only process monolingual texts (in 24 languages), but also successfully annotate bilingual texts in the form of translation memories in TMX format.

Apart from the development of the web services, the ultimate goal was the integration into the European infrastructure for language resources and CLARIN technologies, specifically the Language Resource Switchboard platform.[6] The primary goal was to annotate named entities for Italian and Serbian languages, but it was extended to other languages for which the compatible spaCy[7] models exist. Models trained using the spaCy library, downloaded for each language from the corresponding repository https://spacy.io/models, were used to label the named entities. For the Italian language, the `it_core_news_sm-3.4.0`,[8] model was downloaded, trained on an automatically created corpus for the recognition of named entities, WikiNER,[9] based on the text and structure of Wikipedia (Nothman et al. 2013), while for the Serbian language, a model trained on the corpus was implemented of old Serbian novels SrpCNNER (Šandrih Todorović et al. 2021), downloaded from the European Language Grid (ELG) platform.[10]

In addition to recognizing named entities, the goal of the application was also to link them with items in the open knowledge base Wikidata. The University Library of Mannheim (Universitätsbibliothek Mannheim, abbreviated UB Mannheim) developed spaCyOpenTapioca[11] for the task of linking named entities to concepts (items) in Wikidata in spaCy using OpenTapioca[12] (Delpeuch 2019). The source code of the service and the application is available on the github repository. By using the spacyOpenTapioca package, It-Sr-NER services can, in addition to recognizing and annotating named entities, link entities with items in wikidata.

As a final result, a web service was created that enables geoparsing, i.e. displaying recognized named entities that exist in wikidata on a map.

5. It-Sr-NER services
6. https://switchboard.clarin.eu/tools
7. https://spacy.io/
8. it_core_news_sm-3.4.0
9. WikiNER
10. ELG
11. spacyOpenTapioca
12. OpenTapioca

Eight web services have been developed, four each for monolingual and bilingual resources, which enable:

1. NER - recognition of named entities according to the classes from Table 1 trained language models of the spaCy library;
2. NER+NEL - in addition to the recognition of named entities from the previous point, additional linking with wikidata of annotated entities, where possible by using the functions of the spacyOpenTapioca service, applied only for the recognized named entities, i.e. for the text inside the XML tag;
3. NEL - recognition and linking of named entities with wikidata relying on the recognition of named entities by the spacyOpenTapioca system, whereby the recognized named entity is annotated with the tag <WDT> and the class of the named entity with the *label* attribute.
4. Geoparsing - for those recognized named entities of LOC class that exist in wikidata, finding the latitude and longitude (geolocating) using the *geopy*,[13] library, followed by displaying them on a map using the *folium*[14] library.

NER class tag	Description of the entity class
PERS	Names, surnames, nicknames and their combinations (of real people and fictional characters, including gods and saints).
LOC	Continents, countries, regions, settlements, oronyms, bodies of water, names of celestial bodies, city locations.
ORG	Names of companies, political parties, educational institutions, sports teams, hospitals, museums, libraries, hotels, cafes, churches and shrines.
DEMO	Residents of countries, cities, regions or ethnic groups; derived adjectives from the name of the location.
EVENT	Names of events that recur regularly or happened once but they have their own name: natural disasters, revolutions, battles, wars.
WORK	Titles of books, plays, poems, paintings, sculptures, newspapers.

Table 1. Named entity classes

13. https://pypi.org/project/geopy/
14. https://python-visualization.github.io/folium/

Language-specific models often use different labels to denote classes of named entities. So, for example, LOC is usually used for the spatial location class, but one can also find GPE for geopolitical entities (English model), then LC (Korean model), as well as placeName and geogName (Polish model). In order to harmonize the labels of named entity classes between models for different languages, the labels of the classes denoting locations and geopolitical entities have been renamed to the LOC label, even if other labels (GPE, LC, placeName, geogName) were used. The mapping is done systematically for all classes and can be seen in the configuration file available at lng_config.csv.[15]

In addition, entity labels of the PERS class, which mark persons, were set as the basic tag into which corresponding labels from other models were mapped: PER (Italian), PRS (Swedish), PERSON (Macedonian), persNAME (Polish). The label NORP (nationalities or religious or political groups) of nationalities, political and religious groups from the Japanese and Finnish models, then NAT_REL_POL from the Romanian one, are mapped into the label of the class DEMO, which are marked with demonyms, ethnic relations (Stanković et al. 2021). Since some language models have a much richer set of named entity classes, for example English has 18 classes, Romanian 16, a column with a list of ignored labels is defined in the configuration file.

In addition to the mentioned classes with which the named entities are associated, the NER+NEL and NEL web services provide information about the entity type (attribute *label*), description (attribute *desc*) and a link in the Wikidata knowledge base (attribute *ref*).

It has already been mentioned that the input can be monolingual or bilingual text. In the case of processing bilingual resources, the input file must be a valid TMX document. Figure 5 shows the output of the NER+NEL service for bilingual resources, while Figure 6 shows the output of the NEL service, also for bilingual resources, but this link recognized named entities to items in wikidata using the spacyOpenTapioca library for both languages.

All project results: program code, web services, web application, as well as parallel corpora, have been published with open licenses, and as such can be freely used for research and commercial purposes.

15. lng_config.csv

```
<tu>
    <prop type="Domain"/>
    <tuv xml:lang="it" creationid="n54" creationdate="20220825T211907Z">
        <seg>Progettava di raggiungere <LOC ref="https://www.wikidata.org/wiki/Q90" desc=
        "capital and largest city of France">Parigi</LOC> insieme ad altri suoi compagni,
        mi invitò ad andare con lei in automobile.</seg>
    </tuv>
    <tuv xml:lang="sr" creationid="n54" creationdate="20220825T211907Z">
        <seg>Plan joj je bio da stigne u <LOC ref="https://www.wikidata.org/wiki/Q90" desc
        -"capital and largest city of France">Pariz</LOC> zajedno sa drugim svojim
        kolegama, pozvala me je da joj se pridružim, išle bismo automobilom.</seg>
    </tuv>
</tu>
```

Figure 5. NER+NEL output for bilingual resources in TMX format

```
<tu>
    <prop type="Domain">
    <tuv xml:lang="it" creationid="n934" creationdate="20220825T211907Z">
        <seg>Ranko Mihailović è un giovane silenzioso e bravo, anch'egli frequenta la facoltà
        di legge a <WDT ref="https://www.wikidata.org/wiki/Q1435" label="LOC" desc="capital
        city of Croatia">Zagabria</WDT>, si vede già dinanzi una carriera nell'amministrazione
        e prende raramente e tiepidamente parte alle discussioni e ai dibattiti dei suoi amici
        sull'amore, la politica, le diverse concezioni della vita e l'ordine sociale.</seg>
    </tuv>
    <tuv xml:lang="sr" creationid="n934" creationdate="20220825T211907Z">
        <seg>Ranko Mihailović, ćutljiv i dobroćudan mladić, koji studira pravo u <WDT ref=
        "https://www.wikidata.org/wiki/Q1435" label="LOC" desc="capital city of Croatia">
        Zagrebu</WDT>, pomišlja već sada na činovničku karijeru i slabo i mlako učestvuje u
        drugarskim prepirkama i razgovorima o ljubavi, politici, i pogledima na život i
        društveno uređenje.</seg>
    </tuv>
</prop></tu>
<tu>
```

Figure 6. NEL output for bilingual resources in TMX format

3 Methods of use

The described web services are available via the web application at the address https://ners.jerteh.rs/, as well as by using the requests module within the Python application[16] as follows:

```
import requests
# choose language - lang, and feat
lang = "it"  #@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']
feat = "nel" #@param ['ner', 'nel', 'nernel','geo']
# use api
API_KEY = ["file", "data", "lng", "feat"]
```

16. https://colab.research.google.com

```
url = 'https://ners.jerteh.rs/api'
params = dict(key=API_KEY, data=data, lng=lang, feat=feat)
res = requests.get(url, params=params)
```

Figure 7 shows the integrated web services on the Language Resource Switchboard platform. The input data can be submitted as an XML file in case of bilingual resources, and for monolingual resources a text file can be passed or the text can be entered directly into the provided field of the web application form.

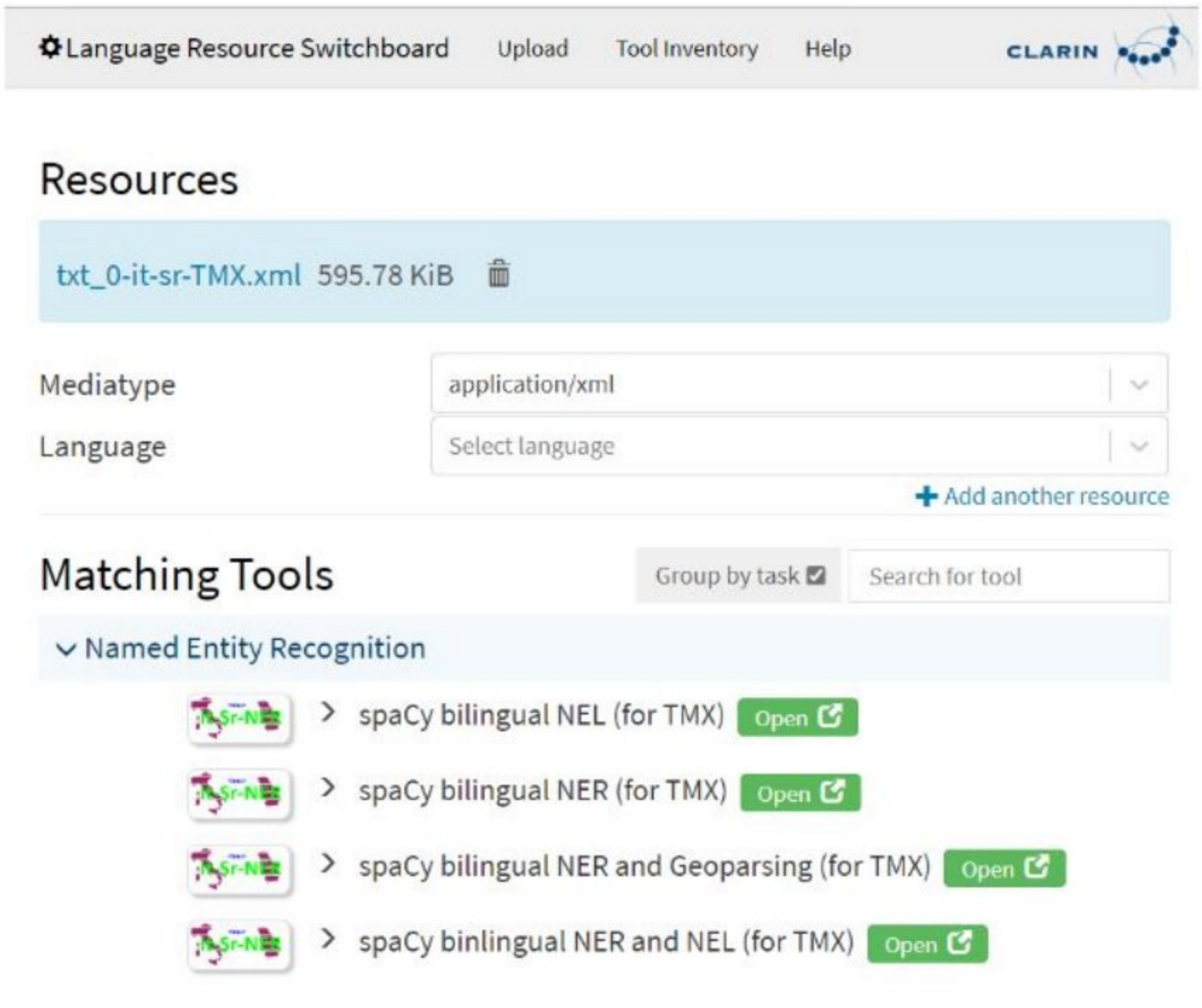

Figure 7. Presentation of the integrated web service on the CLARIN infrastructure

The developed services allow two different formats of display of processing results, HTML and XML documents as illustrated in the following example. Figure 8 shows the processing of text directly written in the textbox for

Serbian (left) and Italian (right). For both languages, a selection of NER services (number 1 for Serbian and number 5 for Italian) was presented, and as a result of the work, HTML and XML documents were presented (number 2 for Serbian and number 6 for Italian). Using the NEL service (number 3) for the Serbian language, the output is presented in the form of HTML and XML documents (number 4), where it can be seen that this service also provides a description of the entity by mouseover event on the entity.

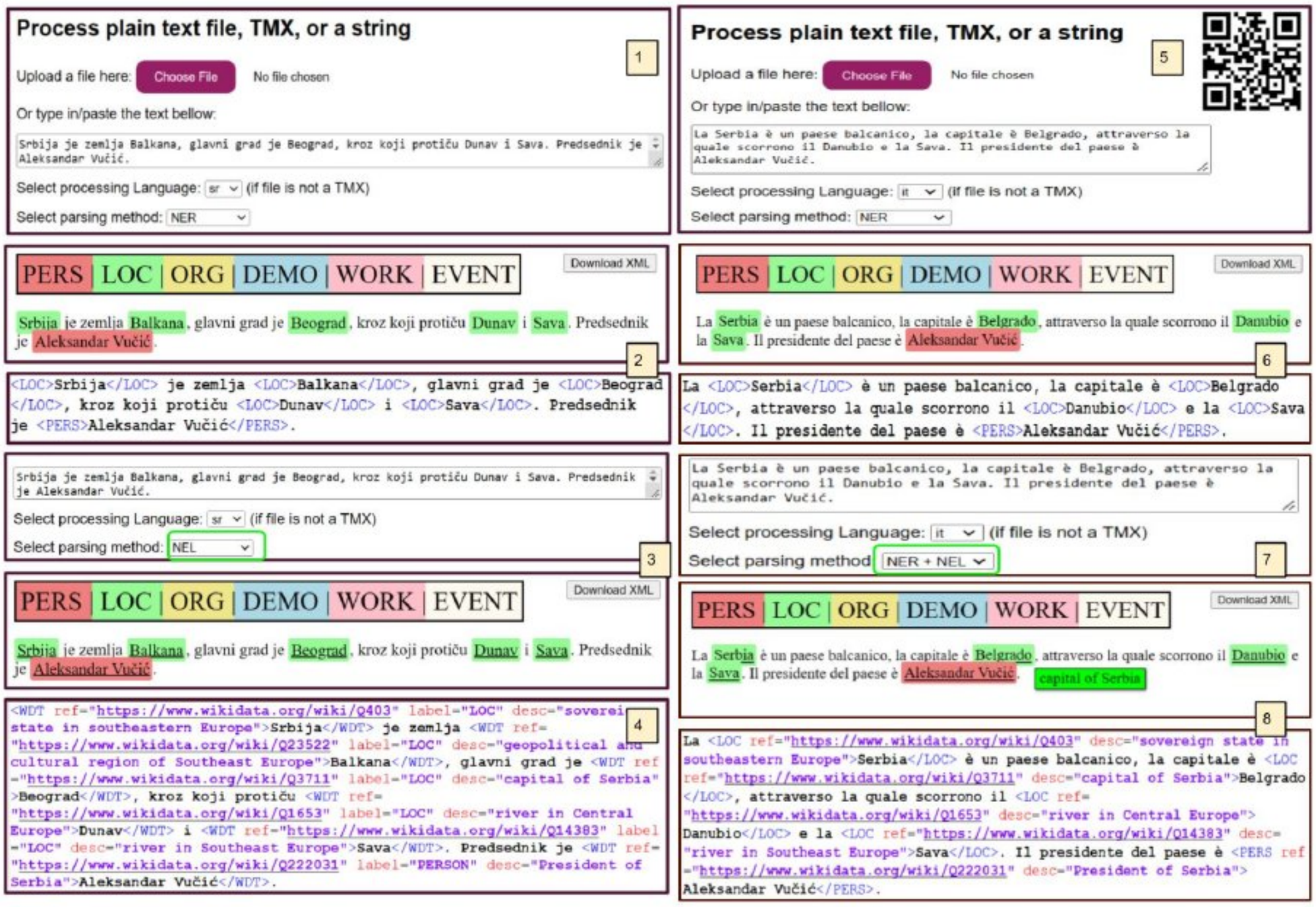

Figure 8. Display of directly entered text processing

The NER+NEL service (number 7) for the Italian language, shows results under number 8, where it can be seen that this service also provides additional information about the entity using the description of the corresponding item in wikidata. It can be seen that the XML documents of these two services (number 4 and number 8) differ by the labels of recognized named entities, ie. NEL service is labeled with the label <WDT>, while the NER+NEL annotation uses the tags described in Table 1. Also, the at-

tributes of these two services differ, the NEL service has one more attribute than the NER+NEL service, which is the *label* attribute, describing tagged named entities in English, regardless of the text language. Descriptions in English are the most common, that is because wikidata is the most developed for English, 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 language. The HTML preview of both services is the same.

Figure 9. Display of bilingual text processing using the NER+NEL service

Figure 9 shows 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 possibility of displaying the description of the item (determiner) America (Q30) is illustrated, where it can be seen that the recognized named entity America is associated by underlined style.

As in the case of previous web services, geoparsing is available for both bilingual and monolingual resources, where only recognized named entities of LOC class by NER+NEL are displayed on the map. Figure 10 shows

geoparsing for monolingual resources for both languages, Serbian (left) and Italian (right).

For different languages (in this case: Italian and Serbian), there may be differences in the recognition of named entities, as well as differences in geoparsing, for several reasons:

1) For the given language, there is no item (headword) for the annotated entity in the knowledge base;
2) The named entity in Serbian is not recognized because the system does not recognize inflected forms for the Serbian language (such as cases different from the nominative singular: Србије, Београду, etc.);
3) The translation equivalents (Serbian and Italian) are not literal, so the named entity does not appear in one of the equivalents (see segment number 6 in Figure 9).

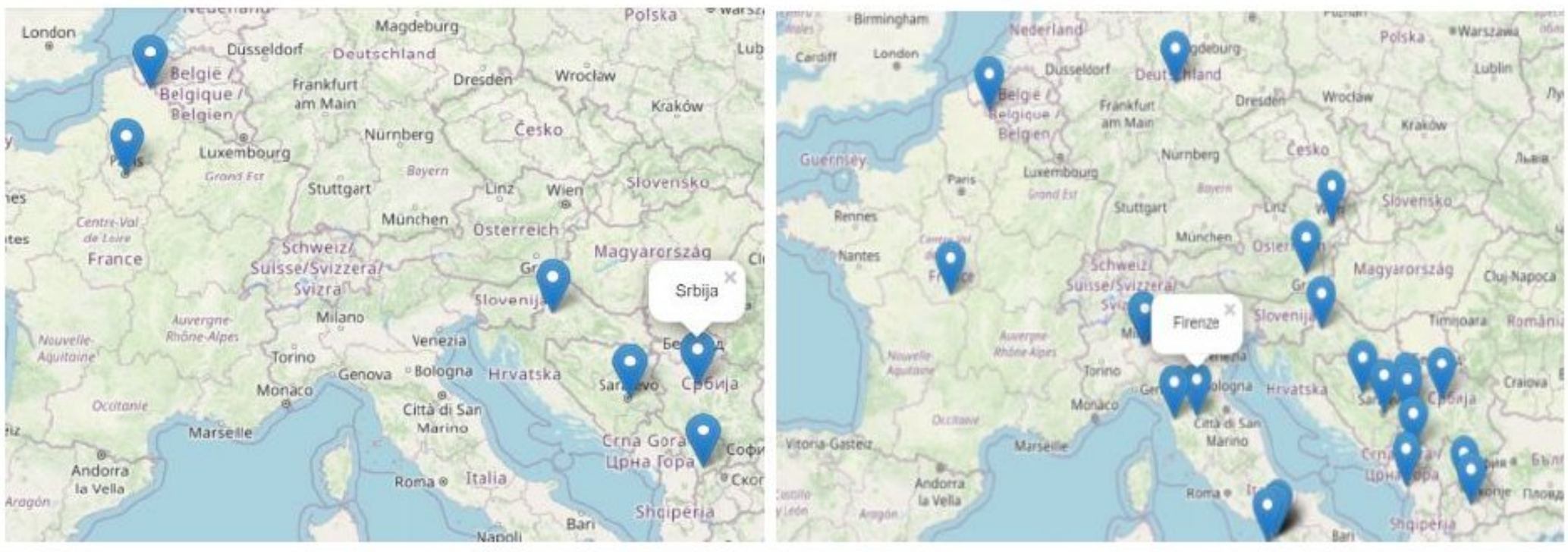

Figure 10. Geoparsing for Serbian language (left) and Italian language (right)

4 Conclusion

In the paper, we presented the results of the project It-Sr-NER: CLARIN compatible NER and geoparsing web services for Italian and Serbian parallel text, supported by the European infrastructure for language resources and technologies CLARIN. The initial motivation for launching a joint project of experts from the University of Turin and JeRTeh, the Society for Language

Resources and Technologies, was to improve the teaching of the Serbian and Italian languages through the creation and publication of a web service for annotating the named entities and displaying them on the map. The lack of specific language technologies for the Serbian language has for years been an obstacle in the application of the results achieved for languages with a larger number of speakers. Individual initiatives that see partial implementation of corpora in teaching are insufficient and do not represent a sufficient stimulus for researchers and teachers in the field of Serbian as a foreign language.

The main primary result of the project is the publication of a set of web services for monolingual and bilingual parallel texts on the CLARIN platform Language Resource Switchboard. At the same time, the project enabled the realization of secondary goals, which are equal in importance to the primary goal, such as the creation and publication of a parallel Italian-Serbian corpus and the construction of a web application and service on the platform of the Society for Language Resources and Technologies JeRTeh. A total of eight services were developed, four each for monolingual and bilingual resources. The text can also be processed by direct input into the provided field at the sentence level or a file entered by the user can be processed. At the same time, linking of named entities with Wikidata and geoparsing are provided. Although Serbian and Italian resources were the focus of the project, the developed services can process texts in 24 languages.

Further research will take place in the direction of organizing additional training in order to popularize web services, as well as their inclusion in teaching. One of the most important goals is to expand the corpus, as well as to improve the model for annotation of named entities and link them to knowledge bases.

Acknowledgment

The authors are thankful for providing the parallelization to: prof. Cvetani Krstev, prof. Duško Vitas, prof. Saša Moderc and Nikola Janković; to reviewers, as well as to the European infrastructure for language resources and technologies CLARIN, on the supported project entitled "It-Sr-NER: Web services for named entities recognition, linking and mapping" within the "Bridging Gaps" call.

References

Delpeuch, Antonin. 2019. *OpenTapioca: Lightweight Entity Linking for Wikidata.* https://doi.org/10.48550/ARXIV.1904.09131.

Granger, Sylviane. 2018. "Has Lexicography Reaped the Full Benefit of the (Learner) Corpus Revolution?" In *Proc. of the XVIII EURALEX International Congress: Lexicography in Global Contexts,* edited by Jaka Čibej et al., 17–24. Ljubljana University Press, Faculty of Arts, July.

Krstev, Cvetana, and Duško Vitas. 2011. "An Aligned English-Serbian Corpus." In *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality),* edited by N. Tomović and J. Vujić, I:495–508. Belgrade: Faculty of Philology, University of Belgrade, December. ISBN: 978-86-6153-005-0.

Moderc, Saša. 2015a. "Elektronski korpus srpskih književnih dela i njihovih prevoda na italijanski jezik" [in Serbian]. 15, *Anali Filološkog fakulteta* 27 (2): 301–316. ISSN: 0522-8468. https://doi.org/10.18485/analiff.2015.27.2.15.

Moderc, Saša. 2015b. "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.

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. "Learning multilingual named entity recognition from Wikipedia." *Artificial Intelligence* 194:151–175. ISSN: 0004-3702. https://doi.org/https://doi.org/10.1016/j.artint.2012.03.006.

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

Perišić, Olja. 2021. "Corpora in the Classroom-the Case of the Serbian Language for Italian Speakers."

Perišić Arsić, Olja. 2018. "L'uso dei corpora nella didattica della traduzione: l'esempio del verbo serbo "prijati" e i suoi traducenti italiani," 49–65.

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian NER&Beyond: The Archaic and the Modern Intertwined." In *Proc. of the Int. Conf. RANLP 2021*, 1252–1260. September. https://aclanthology.org/2021.ranlp-main.141.

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

Stanković, Ranka, Cvetana Krstev, Branislava Šandrih Todorović, and Mihailo Škorić. 2021. "Annotation of the Serbian ELTeC Collection." *Infotheca-Journal for Digital Humanitie* 21 (2): 43–59.

Vitaz, Milica, and Milica Poletanović. 2020. "Data-Driven Learning: The Serbian Case." *EL.LE* (April): 409–422. https://doi.org/10.30687/ELLE/2280-6792/2019/02/009.

Витас, Душко, and Гордана Павловић-Лажетић. 2008. "Ресурси и методе за препознавање именованих ентитета у српском." *Infoteka* 9 (1-2): 33–40.