It-Sr-NER: Веб сервиси за препознавање и повезивање именованих ентитета у тексту и њихово приказивање на веб карти

УДК 81'322.3

САЖЕТАК: У раду ће бити представљени резултати пројекта "It-Sr-NER: Web services for named entities recognition, linking and mapping", у ком су учествовали тимови Универзитета у Торину и Друштва за језичке ресурсе и технологије JePTex, а чији циљ је био развој веб сервиса It-Sr-NER за обележавање именованих ентитета у тексту и њихово приказивање на карти. Именовани ентитети у овим сервисима су имена особа, места, организације, демоними (етници), догађаји и ауторска дела.

КЉУЧНЕ РЕЧИ: паралелни корпуси, именовани ентитети, препознавање именованих ентитета, повезивање именованих ентитета, геопарсирање, српски, италијански, веб сервиси.

РАД ПРИМЉЕН: 27. новембар 2022.
РАД ПРИХВАЋЕН: 31. јануар 2023.

Оља Перишић
olja.perisic@unito.it
Универзитет у Торину
Департман за стране језике, књижевност и модерне културе
Торино, Италија

Ранка Станковић
ranka.stankovic@rgf.bg.ac.rs
Универзитет у Београду
Рударско-геолошки факултет
Београд, Србија

Милица Иконић Нешић
milica.ikonic.nesic@fil.bg.ac.rs
Универзитет у Београду
Филолошки факултет
Београд, Србија

Михаило Шкорић
mihailo.skoric@rgf.bg.ac.rs
Универзитет у Београду
Рударско-геолошки факултет
Београд, Србија

1. Увод

Недостатак алата и ресурса који омогућавају анотацију, истраживање и анализу двојезично поравнатих (паралелизованих) италијанско-српских текстова је била основна мотивација и инспирација за покретање оваквог пројекта. У дидактици страних језика у Србији

бележи се готово потпуно одсуство корпусних алата у настави (Vitaz and Poletanović 2019), док се на индивидуалном плану и путем личних иницијатива у настави српског језика као страног у Италији и италијанског језика у Србији показало да су корпуси у настави многоструко значајни и да их студенти радо прихватају и примењују у заједничком раду, али и самосталним истраживањима (Модерц 2015; Perišić 2021). Богата и истанчана морфологија српског језика предвиђа деклинацију топонима и осталих именованих ентитета које страни студенти нису увек у стању да препознају и сведу на њихов основни облик. Неки од разлога су јединствени наставци за мушки и средњи род у већини падежа, присуство појединих топонима у форми множине, тзв. *pluralia tantum* (Пљевља, Дивчибаре итд.), али и фонетска транскрипција страних имена као и неке ортографске недоследности (Витас and Павловић-Лажетић 2008).

У оквиру позива "Bridging Gaps" европске инфраструктуре за језичке ресурсе и технологије CLARIN[1] (Common Language Resources & Technology Infrastructure), тим стручњака Универзитета у Торину и Друштва за језичке ресурсе и технологије JePTex удружио је снаге и развио веб сервисе за обележавање именованих ентитета у тексту, обезбедио њихово повезивање са википодацима,[2] као и геопарсирање, тј. геолоцирање препознатих локација и њихово приказивање на мапи. Именовани ентитети у овим сервисима су имена особа, места, организације, демоними (етници), догађаји и ауторска дела.

Основни циљ пројекта је реализација и публиковање веб апликација и сервиса за једнојезичне и двојезичне паралелне текстове у оквиру инфраструктуре CLARIN, као и на платформи Друштва за језичке ресурсе и технологије JePTex. Пројекат предвиђа и креирање и публиковање верзије италијанско-српског корпуса од 10.000 сегмената, то јест, екстрахованих и упарених реченица преузетих из класичних дела италијанске и српске књижевности. Резултати пројекта не ограничавају се на српско-италијанску језичку комбинацију, већ се развијени сервиси могу применити на обраду текстова на чак двадесет и четири светска језика.

Иницијатор пројекта и вођа тима је Ољa Перишић, професорка на Универзитету у Торину (*Università degli Studi di Torino, Dipartimento di Lingue e Letterature Straniere e Culture Moderne*) где предаје српски језик. Испред JePTex-а, развојем сервиса је руководила проф. Ранка

1. CLARIN
2. Wikidata

Станковић у сарадњи са проф. Душком Витасом. Више података о пројекту је доступно и на веб страни Друштва за језичке ресурсе и технологије JePTex.[3]

2. Паралелни корпус

У глотодидактици, тј. настави страних језика, паралелни корпуси показали су се као незаобилазни инструмент у циљу усвајања морфосинтаксе и лексике. Паралелно проматрање два или више језика олакшава контрастивну анализу, односно уочавање сличности и разлика језичких структура захваљујући великом броју примера реченица у контексту, што је још у раној фази корпусне лингвистике уочио Џон Синклер: „Језик изгледа прилично другачије уколико се посматра велики број примера у исто време" (Sinclair 1991, 100).[4] У дидактици превођења паралелни корпуси омогућавају диференцијацију блиских преводних еквивалената и дефинисање полисемичне лексике која је често запостављена или површно обрађена у билингвалним речницима (Moderc 2015; Perišić Arsić 2018). Истовремено, у пракси се бележи да је број репрезентативних преводилачких (паралелних корпуса) чак и за веће светске језике недовољан (Granger 2018).

Из свих ових разлога у првој фази пројекта било је неопходно креирати италијанско-српски корпус од исечака романа са 10.000 поравнатих сегмената (реченица) из десет романа. Романи су представљени узорцима у којима су сегменти измешани како би се избегли проблеми са ауторским правима. Романи италијанских писаца представљени у корпусу су: Умберто Еко, *Име руже*; Карло Колоди, *Пинокијеве авантуре*; Елена Феранте, *Прича о онима који одлазе и онима који остају*; Луиђи Пиранделo, *Један, ниједан и сто хиљада*. Српски писци су заступљени са пет романа: Иво Андрић, *Аникина времена* и *На Дрини ћуприја*; Борисав Станковић, *Нечиста крв*; Бранислав Нушић, *Општинско дете: роман једног одојчета*; Данило Киш, *Башта, пепео*. Имајући у виду да је основни задатак пројекта обележавање именованих ентитета, у корпус су укључени и преводи на италијански и српски романа *Пут око света за осамдесет дана* Жила Верна.

3. It-Sr-NER CLARIN compatible NER and geoparsing web services for parallel texts

4. У оригиналу: "The language looks rather different when you look at a lot of it at once"

Романи су поравнати и припремљени у формату TMX (Translation Memory eXchange) коришћењем апликације ACIDE, интегрисаног окружења за развој паралелних корпуса (Obradović, Stanković, and Utvić 2008; Krstev and Vitas 2011).

```
<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>
```

Слика 1. Пример TMX излазног документа.

Слика 1 приказује прве две преводне јединице, обележене етикетама `<tu>` (translation unit), у оквиру којих се налазе преводни еквиваленти обележени етикетом `<tuv>` (translation unit variant). Упарени, односно поравнати сегменти на италијанском и српском су нумерисани (`n1`, `n2`,...) и сваки има атрибут који указује на језик: `xml:lang="it"` или `xml:lang="sr"`. Процес паралелизације апликацијом ACIDE, осим генерисања TMX излазног документа генерише и HTML приказ који се може видети на слици 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.

Слика 2. Пример паралелизованих сегмената преводних еквивалената на италијанском и српском језику у HTML формату.

Корпус је објављен у ILC4CLARIN B центру, са јединственим идентификатором[5], тако да је видљив преко VLO.[6] Компримовани облик корпуса садржи више делова: поравнате двојезичне, али и појединачне, једнојезичне верзије. Аутоматски обележени именовани ентитети (на начин објашњен у Одељку 3.) су такође саставни део публикованог корпуса. Корпусу се такође може приступити са радне *github* локације.[7]

Осим за овог корпуса који се може преузети, шири корпус који садржи комплетне романе и из којег је извучена публикована верзија од 10.000 сегмената доступан је за претрагу на дигиталној библиотеци Библиша.[8] Карактеристика библиотеке *Библиша* је унапређена претрага са могућностима морфолошког и семантичког проширења упита за српски језик.

metadata	**n3330** E si riunivano per lo più dietro la casa, presso le finestre della stanza grande, perchè lì 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.**"

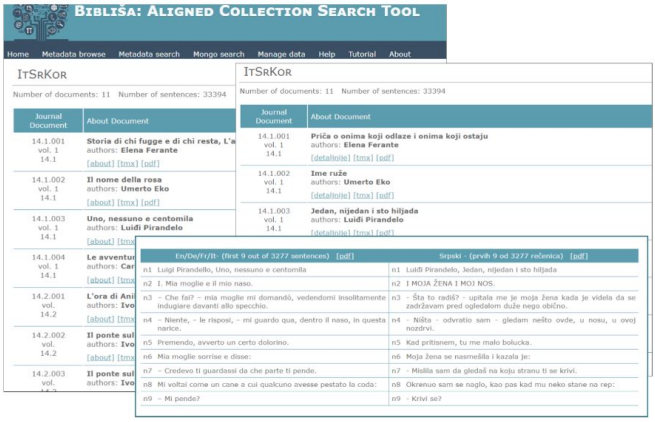
Слика 3. Резултат претраге проширивањем упита у Библиши.

На слици 3 је приказан панел на ком се може видети пример резултата семантичког проширења упита „брат" који је аутоматски проширен синонимом „бата", а потом су додати и морфолошки облици свих упитних речи. На слици 3 се може видети неколико поравнатих сегмената који су добијени као резултат постављеног упита. Како је упит

5. It-Sr-NER у ILC4CLARIN B центру
6. Virtual Language Observatory
7. It-Sr-NER на github-у
8. Библиша

постављен на српском језику, у извученим сегментима су кључне речи истакнуте.

Опција за преглед колекција паралелних докумената (слика 4) доступна је са ограниченим бројем сегмената за све кориснике и са већим бројем сегмената само за регистроване кориснике.



Слика 4. Прелиставање колекције паралелних докумената ItSrKor у Библиши

Паралелни корпуси су драгоцени за студије превођења, а контрастивна лингвистика и једноставна употреба конкорданци знатно олакшавају проучавање међујезичких појава. Студенти италијанског језика Универзитета у Београду, где проф. Саша Модерц предаје, као и студенти српског језика у Торину, где проф. Оља Перишић предаје, користиће развијене ресурсе у будућој настави, с обзиром на то да су отворени и самим тим доступни и другим студентима и истраживачима.

3. Имплементација сервиса

It-Sr-NER сервиси,[9] похрањени на репозиторијуму CLARIN са јединственим идентификатором,[10] не само што обрађују једнојезичне

9. It-Sr-NER сервиси
10. It-Sr-NER сервиси у репозиторијуму CLARIN

текстове (на 24 језика), већ успешно обележавају и двојезичне текстове у облику преводилачких меморија у формату TMX.

Осим саме изградње веб сервиса, крајњи циљ је био интеграција у европску инфраструктуру за језичке ресурсе и технологије CLARIN, конкретно на платформу Language Resource Switchboard.[11] Примарни циљ је било обележавање именованих ентитета за италијански и српски језик, али је проширено и на друге језике за које су постојали компатибилни spaCy[12] модели. За обележавање именованих ентитета коришћени су модели који су тренирани уз помоћ библиотеке spaCy, преузети за сваки језик са одговарајућег репозиторијума.[13] За италијански језик преузет је модел `it_core_news_sm-3.4.0`,[14] обучен на аутоматски креираном корпусу за препознавање именованих ентитета WikiNER,[15] заснованом на тексту и структури Википедије (Nothman et al. 2013), док је за српски језик имплементиран модел који је трениран на корпусу старих српских романа SrpCNNER (Šandrih Todorović et al. 2021), преузет са платформе European Language Grid (ELG).[16]

Поред препознавања именованих ентитета, циљ апликације је био и њихово повезивање са ставкама у отвореној бази знања Википодаци. Универзитетска библиотека у Манхајму (Universitätsbibliothek Mannheim, скр. UB Mannheim) је развила spaCy-OpenTapioca[17] коришћењем OpenTapioca[18] чији је један од задатака препознавање именованих ентитета, као и њихово повезивање са концептима у википодацима (Delpeuch 2019). Изворни код сервиса и апликације је доступан на репозиторијуму github. Коришћење пакета spacyOpenTapioca омогућава сервисима It-Sr-NER не само да препознају и обележавају именоване ентитете већ и да их повезују са ставкама у википодацима. Коначан резултат је веб сервис који омогућава геопарсирање, тј. приказивање препознатих именованих ентитета који постоје у википодацима на карти.

Развијено је 8 веб сервиса, по 4 за једнојезичне и двојезичне ресурсе, који омогућавају:

11. Language Resource Switchboard
12. spaCy
13. spaCy модели
14. NER модел за италијански језик
15. WikiNER
16. ELG
17. spaCyOpenTapioca
18. OpenTapioca

1. NER – препознавање именованих ентитета према класама из табеле 1 обученим језичким моделима библиотеке spaCy;
2. NER+NEL – повезивање препознатих и обележених ентитета са википодацима, што омогућува коришћење функције сервиса spacyOpenTapioca (само)за препознати именовани ентитет, односно за текст унутар етикете);
3. NEL – препознавање и повезивање именованих ентитета са википодацима ослањајући се на систем spacyOpenTapioca, при чему се препознати именовани ентитет обележава етикетом `<WDT>`, а класа именованог ентитета атрибутом *label*;
4. Геопарсирање – за препознате именоване ентитете класе LOC који постоје у википодацима, проналажење географске ширине и дужине (геолоцирање) коришћењем библиотеке *geopy*,[19] за чим следи њихово приказивање на карти коришћењем библиотеке *folium*.[20]

Етикета	Опис класе ентитета
PERS	Имена, презимена, надимци и њихове комбинације (стварних људи и измишљени ликови, укључујући богове и свеце).
LOC	Континенти, земље, региони, насељена места, ороними, водене површине, имена небеских тела, градске локације.
ORG	Називи компанија, политичких партија, образовних институција, спортских тимова, болница, музеја, библиотеке, угоститељских објеката, сакралних објеката.
DEMO	Становници држава, градова, региона или етничке групе; придеви изведени из назива локација.
EVENT	Називи догађаја који се редовно понављају или су се десили једном али имају своје име: празници, природне катастрофе, револуције, битке, ратови.
WORK	Наслови књига, драма, песама, слика, скулптура, новина.

Табела 1. Класе ентитета.

19. Geopy
20. Folium

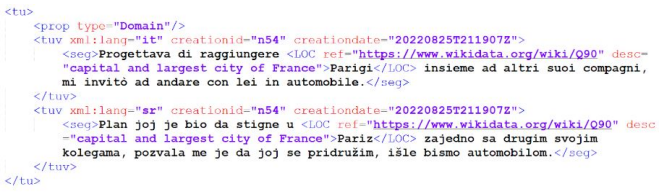
Модели за поједине језике користе често различите етикете за обележавање класа именованих ентитета. Тако, на пример, за класу просторних локација се обично користи LOC, али се може наћи и GPE за геополитичке ентитете (енглески модел), потом LC (кореански модел), као и placeName и geogName (пољски модел). Због хармонизације етикета класа именованих ентитета између модела за различите језике, све коришћене ознаке класа које означавају локације и геополитичке ентитете (GPE, LC, placeName, geogName) преименоване су у јединствену ознаку LOC. Мапирање је урађено систематки за све класе и може се видети у конфигурационој датотеци.[21]

Такође, етикета PERS ентитета класе којој припадају особе, је постављена као основна ознака у коју су се мапирале одговарајуће ознаке из других модела: PER (италијански), PRS (шведски), PERSON (македонски), persNAME (пољски). Етикета NORP (енг. nationalities or religious or political groups) националности, политичке и верске групе из јапанског и финског модела, потом NAT_REL_POL из румунског мапирани су у етикету класе DEMO којима су обележени демоними, односи етници (Stanković et al. 2021). Пошто неки језички модели имају много богатији скуп класа именованих ентитета, на пример, енглески има 18 класа а румунски 16, у конфигурационој датотеци је дефинисана колона са списком етикета које се игноришу.

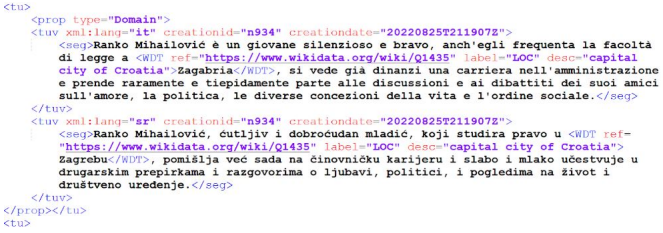
Поред поменутих класа са којима се повезују именовани ентитети, веб сервиси NER+NEL и NEL пружају информације о типу ентитета (атрибут *label*), опису (атрибут *desc*) и вези ка бази знања Википодаци (атрибут *ref*). Већ је поменуто да улаз може бити једнојезични или двојезични текст. У случају обраде двојезичних ресурса улазна датотека мора бити валидан TMX документ. На слици 5 приказан је излаз сервиса NER+NEL за двојезичне ресурсе, док је на слици 6 приказан излаз NEL сервиса, такође за двојезичне ресурсе, који препознате именоване ентитете повезује са ставкама у википодацима коришћењем библиотеке spacyOpenTapioca за оба језика.

Сви резултати пројекта: програмски код, веб сервиси, веб апликација, као и паралелни корпуси, објављени су са отвореним лиценцама, и као такви се могу слободно користити у истраживачке и комерцијалне сврхе.

21. конфигурациона датотека



Слика 5. Приказ излаза NER+NEL за двојезичне ресурсе у TMX формату.



Слика 6. Приказ излаза NEL за двојезичне ресурсе у TMX формату.

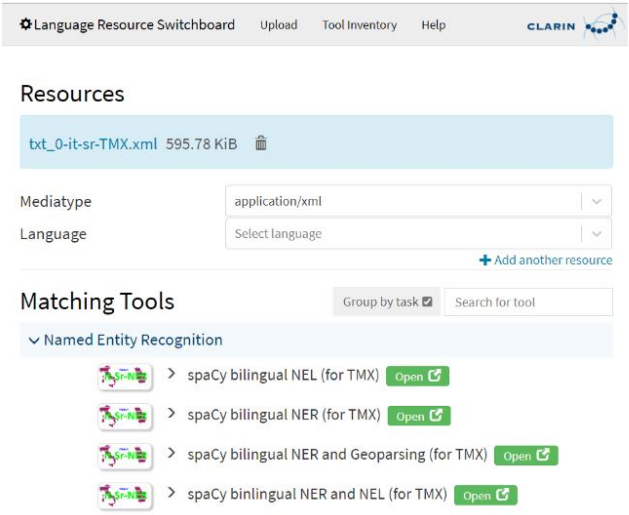
4. Начини коришћења

Описани веб сервиси су доступни преко веб апликације на адреси Друштва за језичке ресурсе и технологије,[22] као и употребом модула *requests* у оквиру Python програма[23] на следећи начин:

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

22. Сервиси за именоване ентитете JePTeха
23. https://colab.research.google.com

Слика 7 приказује интегрисане веб сервисе на платформи Language Resource Switchboard. Улазни подаци се могу проследити у виду XML датотеке у случају двојезичних ресурса, док се за једнојезичне може проследити текстуална датотека, а текст се може и директно унети у предвиђено поље формулара веб апликације.



Слика 7. Приказ интегрисаног веб сервиса на инфраструктури CLARIN-а.

Развијени сервиси омогућавају два различита формата приказа резултата обраде, HTML и XML документ као што је илустровано на следећем примеру. На слици 8 представљен је приказ обраде текста унетог директно у поље за унос. На левој страни слике унет је текст на српском језику, а на десној страни на италијанском.

За оба језика приказан је избор сервиса NER (број 1 за српски и број 5 за италијански), а као резултат рада, представљени су HTML и XML документи (број 2 за српски и број 6 за италијански). Користећи сервис NEL (број 3) за српски језик, представљен је излаз у облику HTML и XML докумената (број 4), где се може уочити да овај сервис

Слика 8. Приказ обраде директно унетог текста.

у HTML приказу пружа и опис ентитета преузет из википодатака позиционирањем на обележени ентитет на генерисаној HTML страни.

Сервис NER+NEL (број 7) за италијански језик, приказује резултате под бројем 8, при чему се види да и овај сервис пружа додатне информације о ентитету користећи опис одговарајуће ставке у википодацима. Може се уочити да се XML документи ова два сервиса (број 4 и број 8) разликују по етикетама препознатих именованих ентитета, тј. сервис NEL обележава етикетом `<WDT>`, док NER+NEL обележава етикетама описаним у табели 1. Такође, атрибути етикета ова два сервиса се разликују, па сервис NEL има један атрибут више од сервиса NER+NEL, а то је атрибут *label*, у ком се даје опис именованог ентитета, који је увек на енглеском, без обзира на језик текста. Описи на енглеском језику су најбројнији, односно википодаци су најбогатији за енглески језик, тако да је он у овој верзији имплементиран. У наредним верзијама сервиса приступ ће бити вероватно промењен тако да језик описа одговара језику самог текста. HTML приказ оба сервиса је исти.

На слици 9 приказан је пример резултата обраде помоћу сервиса NER+NEL на CLARIN платформи *Language Resource Switchboard* двојезичног текста, отпремљеног у виду TMX документа. Могућност приказа описа је илустрована ставком (одредницом) *Америка* (*Q30*),

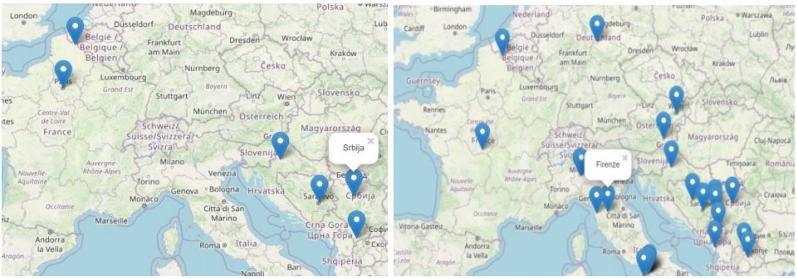
Слика 9. Приказ обраде двојезичног текста коришћењем сервиса NER+NEL.

при чему се може уочити да је препознати именовани ентитет *Америка* повезан по томе што је подвучен.

Као и у случају веб сервиса, геопарсирање је доступно и за двојезичне и за једнојезичне ресурсе, при чему се на карти приказују само именовани ентитети класе LOC које препознаје NER+NEL. На слици 10 је приказано геопарсирање за једнојезичне ресурсе за оба језика, српски језик (лево) и италијански језик (десно).

За различите језике (у овом случају: италијански и српски) могу наступити разлике у препознавању именованих ентитета, као и разлике у геопарсирању, из више разлога:

1. За дати језик не постоји у бази знања ставка (одредница) за обележени ентитет;
2. Именовани ентитет у српском није препознат зато што систем за повезивање са википодацима заснован на библиотеци *spaCyOpenTapioca* не препознаје флективне облике за српски језик (попут падежа различитих од номинатива једнине: Србије, Београду, итд.);
3. Преводни еквиваленти (српски и италијански) нису дословни, тако да се именовани ентитет не појављује у једном од еквивалената (видети сегменте број 6 на слици 9: име реке *Дрина* се појављује у италијанском тексту, али не и у српском).



Слика 10. Геопарсирање за српски језик (лево) и италијански језик (десно).

5. Закључак

У раду смо приказали резултате пројекта *It-Sr-NER: CLARIN compatible NER and geoparsing web services for Italian and Serbian parallel text*, подржаног од стране европске инфраструктуре за језичке ресурсе и технологије CLARIN. Полазна мотивација за покретање заједничког пројекта стручњака са Универзитета у Торину и JePTex-a, Друштва за језичке ресурсе и технологије, било је унапређење дидактике српског и италијанског језика путем креирања и публиковања веб сервиса за обележавање именованих ентитета и њихово приказивање на мапи. Недостатак појединих језичких технологија за српски језик годинама представља кочницу и препреку у примени резултата достигнутих за језике са већим бројем говорника јер су појединачне иницијативе којима се делимично имплементирају корпуси у настави недовољне и не представљају довољан стимуланс за истраживаче и наставнике из области српског језика као страног.

Примарни резултат пројекта је публиковање скупа веб сервиса за једнојезичне и двојезичне паралелне текстове на CLARIN платформи *Language Resource Switchboard*. У исто време пројекат је омогућио реализацију секундарних циљева, који се по својој важности изједначавају са примарним циљем, као што су креирање и публиковање паралелног италијанско-српског корпуса и изградња веб апликације и сервиса на платформи Друштва за језичке ресурсе и технологије JePTex. Развијено је укупно осам сервиса, по четири за једнојезичне и двојезичне ресурсе. Текст се уз то може обрађивати директним уносом у предвиђено поље на нивоу реченица или се може обрадити датотека унета од

стране корисника. У исто време обезбеђено је повезивање именованих ентитета са википодацима и геопарсирање. Иако су у фокусу пројекта били српски и италијански ресурси, развијени сервиси могу да обраде текстове на 24 језика.

Даљи правци истраживачког рада одвијаће се у смеру организовања додатних обука у циљу популаризације веб сервиса, као и њихово укључивање у наставу. Као један од најважнијих циљева је проширивање корпуса, као и унапређење модела за обележавање именованих ентитета и повезивање са базама знања.

Захвалност

Аутори се захваљују проф. др Цветани Крстев, проф. др Душку Витасу, проф. др Саши Модерцу и Николи Јанковићу, студенту докторских студија, на помоћи у припреми паралелног корпуса, као и европској инфраструктури за језичке ресурсе и технологије CLARIN, на подржаном пројекту у оквиру позива "Bridging Gaps" – "It–Sr–NER: Web services for named entities recognition, linking and mapping".

Литература

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

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

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

Moderc, Saša. 2015. "Su un modo di tradurre l'avverbio serbo "inače" in italiano: il caso dell'equivalente "altrimenti"." *Italica Belgradensia* 2015 (1): 62–79. ISSN: 0353-4766. https://doi.org/10.18485/italbg.2015.1.4.

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.

Obradović, Ivan, Ranka Stanković, and Miloš Utvić. 2008. "Integrisano okruženje za pripremu paralelizovanog korpusa." In *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen,* edited by Branko Tošović, 563–578. Münster, Germany, LITVerlag.

Perišić, Olja. 2021. "Corpora in the Classroom – the Case of the Serbian Language for Italian Speakers." In *New Trends in Slavic Studies,* edited by S. Jose et al. S. Cuadros, 126–137. URSS. ISBN: 978-5-396-01059-8.

Perišić Arsić, Olja. 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. https://doi.org/10.18485/italbg.2018.1.3.

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian NER & Beyond: The Archaic and the Modern Intertwined." In *Deep Learning Natural Language Processing Methods and Applications – Proc. of the Int. Conf. RANLP 2021,* edited by G. Angelova et al., 1252–1260. https://doi.org/10.26615/978-954-452-072-4_141.

Sinclair, John. 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 Humanities* 21 (2): 43–59. https://doi.org/doi.org/10.18485/infotheca.2021.21.2.3.

Vitaz, Milica, and Milica Poletanović. 2019. "Data-Driven Learning: The Serbian Case." *EL.LE* 8 (2): 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." *Инфотека* 9 (1–2): 33–40.

Модерц, Саша. 2015. "Електронски корпус српских књижевних дела и њихових превода на италијански језик." *Анали Филолошког факултета* 27 (2): 301–316. ISSN: 0522-8468. https://doi.org/10.18485/analiff.2015.27.2.15.