EUROLAN 2021: Увод у повезане податке у лингвистици – онлајн школа за обуку полазника

УДК 81: 37.018.53

САЖЕТАК: Прва школа за обуку полазника коју је организовала COST акција *NexusLinguarum* одржана је од 8. до 12. фебруара 2021. године са циљем да студенти, истраживачи и стручњаци науче основе лингвистичке науке о подацима. Током обуке полазници су се упознали са широким спектром тема: од семантичког веба, RDF -а и онтологија, до моделирања и претраживања језичких података помоћу најсавременијих онтолошких модела и алата. Школа је одржана у оквиру серије летњих школа EUROLAN-а и организовало ју је виртуелно (онлајн) неколико института; Румунска академија, Истраживачки институт за вештачку интелигенцију у Букурешту и Институт за рачунарске науке, као и Универзитет „Alexandru Ioan Cuza“ у Јашију, Румунија. Школу су похађала 82 полазника.

КЉУЧНЕ РЕЧИ: наука о лингвистичким подацима, повезани подаци у лингвистици, језички подаци, EUROLAN, NexusLinguarum, COST акција, школа за обуку

РАД ПРИМЉЕН: 30. јун 2021.
РАД ПРИХВАЋЕН: 13. јул 2021.

Милан Дојчиновски
milan.dojchinovski@fit.cvut.cz
CTU
Праг, Чешка Република
InfAI при Универзитету у Лајпцигу
Лајпциг, Немачка

Хулија Боске Хил
jbosque@unizar.es
Хорхе Грасија
jogracia@unizar.es
Арагонски институт за инжењерска истраживања (I3A)
Универзитет у Сарагоси
Сарагоса, Шпанија

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

1. Увод

COST акција CA18029 *NexusLinguarum* - European network for Web-centered linguistic data science[1] „Европска мрежа за науку о

1. NexusLinguarum

језичким подацима оријентисаним ка вебу“ започела је активности крајем октобра 2019. Циљ акције *NexusLinguarum* је промовисање проучавања лингвистичких наука о подацима, за које је потребна изградња екосистема вишејезичних и семантички интероперабилних језичких података. Школе су једно од средстава за постизање овог циља, па је стога главни тим акције *NexusLinguarum* организовао школу „Увод у повезане лингвистичке податке,“[2] која је одржана од 8. до 12. фебруара 2021. Школа је имала за циљ промовисање и предавање о основама науке о лингвистичким подацима и сродним технологијама полазницима из академске и шире заједнице. Организована је под окриљем серије летњих школа EUROLAN-а основане 1993. године и обухвата теме које су посебно релевантне за области рачунарске лингвистике и обраде природних језика (ОПЈ). Циљ ове 15. школе EUROLAN-а био је да окупи научнике, наставнике и студенте лингвистике, ОПЈ-а и информационих технологија како би разговарали о принципима и најбољим праксама за представљање, објављивање и повезивање лингвистичких података и питања која чине саставне делове замишљеног вишејезичног и интероперабилног екосистема оријентисаног ка вебу. Овај прилог резимира организацију, садржај и резултате школе обуке и заснован је на извештају Д1.1 акције доступном на вебу.[3]

2. Програм школе

Школа је осмишљена за почетнике, као и за оне који већ имају основно знање из области обухваћених темама. Школа је пружила свеобухватан увод у методологије представљања језичких ресурса коришћењем технологија семантичког веба, заједно са средствима за екстракцију знања из језичких ресурса и његово коришћење помоћу упитних језика и механизама закључивања семантичког веба. У оквиру школе су обрађиване следеће теме:

– Семантички веб и Повезани подаци[4] (Berners-Lee et al. 2006);
– Онтологије: RDF (Resource description framework), RDF Schema (Resource Description Framework Schema, скраћено RDFS, RDF(S), RDF-S, или RDF/S), Web Ontology Language (OWL),[5] etc.);

2. EUROLAN
3. Извештај Д1.1
4. Увод у повезане податке и семантички веб
5. OWL

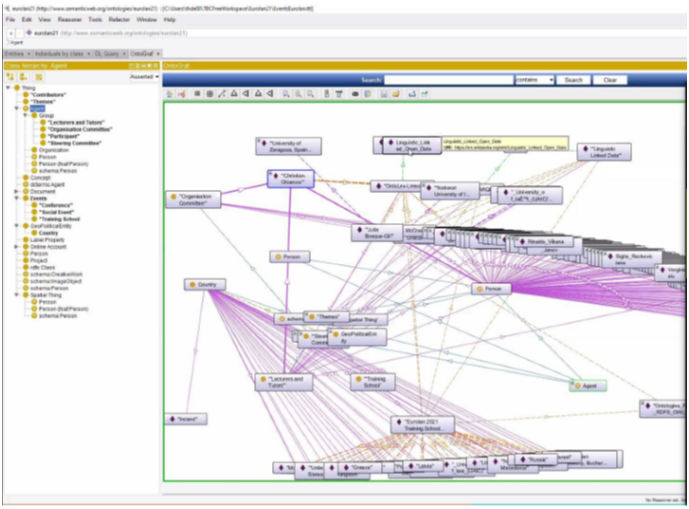
– Упитни језик SPARQL - семантички упитни језик за базе података за претраживање и руковање подацима ускладиштеним у RDF формату; item Метаподаци: DCAT (Data Catalog Vocabulary),[6] VOID (RDF Schema речник за представљање метаподатака о RDF скуповима података, итд.);
– Трансформација и валидација RDF података (Cimiano et al. 2020);
– Лингвистички повезани подаци (Chiarcos et al. 2013);
– Lemon-OntoLex[7] (McCrae et al. 2017; Declerck, Tiberius, and Wandl-Vogt 2017; Stanković et al. 2018)
– Генерисање лингвистички повезаних података (Cimiano et al. 2020);
– Корпуси и повезани подаци (Chiarcos 2012);
– Лингвистичке анотације (Fäth et al. 2020);
– Формат за размену у обради природних језика – NLP Interchange Format (NIF) (Hellmann et al. 2013);
– Алати и апликације лингвистичких повезаних података (Declerck et al. 2020).

Први дан почео је уводном сесијом и кратким уводом у лингвистичке повезане податке (Linguistic Linked Data, LLD), након чега су уследиле сесије посвећене уводу у повезане податке и RDF. Други дан је обухватио теме везане за онтологије, укључујући моделирање знања помоћу онтологија, језике за представљање знања OWL и SKOS,[8] образлагање знања и практичну сесију уз употребу едитора онтологије Protégé.[9] Трећи дан био је посвећен темама у вези са представљањем и испитивањем лексичких података са наменским сесијама о моделу Ontolex-Lemon и упитном језику SPARQL. Четврти дан је укључивао сесије које су дале преглед других лингвистичких речника и речника метаподатака, као и платформе VocBench[10] (Stellato et al. 2020) за моделирање лингвистичких скупова података. У поподневним часовима организовано је виртуелно дружење на коме су учесници могли да виде видео записе лепоте румунске културе, традиције и природе. Пети дан се састојао од три паралелне сесије на различите теме:

6. Data Catalog Vocabulary (DCAT)
7. Lemon - Модел лексикона за онтологије; Лексикон модел за онтологије: Извештај заједнице од 10. маја 2016.
8. SKOS
9. Protégé
10. VocBench: Систем колаборативног управљања за OWL онтологије, SKOS(/XL) тезаурусе, OntoLex-lemon лексиконе и скупове података у генеричком RDF.

(a.) Генерисање/трансформација и повезивање LLD-а,
(b.) Анотације (NIF, веб анотације) (Hellmann et al. 2013) и
(c.) OntoLex проширења: *vartrans* за представљање превода и варијанте термина (засновано на модулу *lemon*-а за превођење (Gracia et al. 2014), *lexicog*[11] - лексикографски модул (Bosque-Gil, Gracia, and Montiel-Ponsoda 2017), *FrAC*[12] - frequency, attestation and corpus Information (Chiarcos et al. 2020) – фреквенције, потрвде и информације о корпусу.

За крај, школа обуке је окончана завршном сесијом где је представљена онтологија учесника, предавача и организатора која је илустровала многе механизме представљања објашњене током недеље.



Слика 1. Онтологија учесника, предавача и организатора школе.

.

Свака од организованих сесија била је пропраћена практичним заједничким и самосталним вежбама. Током практичне сесије, предавачи би предложили вежбу и понудили решавање корак по корак, како би учесници могли да разумеју методологију којом се

11. The OntoLex Lemon Lexicography Module
12. FrAC – Frequency, Attestation and Corpus Information - Ontology-Lexica Community Group

долази до решења. Уз то су представили и основне технологије потребне за решавање задатка. Затим се, током сесије самосталних вежби, од учесника тражило да раде на одређеном задатку попут случајева представљених током практичне сесије, чиме би се упознали са технологијом уведеном у практично окружење. Како су ове сесије биле градиране по сложености, почевши од основних појмова, који су надограђивани тако да су последњег дана детаљно представљене конкретније теме, учесници су имали прилику да стекну чврсте темеље знања пре него што пређу на сложеније сесије. Званични програм школе доступан је на вебу.[13]

Као наставак активности, Друштво за језичке ресурсе и алате JePTex[14] је поставило локалну инсталацију алата VocBench[15], а осим чланова друштва JePTex, користили су је и студенти и наставници програма докторских студија „Интелигентни системи“ Универзитета у Београду за предмет „Представљање знања“ и „Семантички веб“. Модул *Lemon-OntoLex FrAC* примењиван је за представљање уноса из лексикона који се користи за откривање увредљивог обраћања са потврдама из корпуса твитова са анотацијом увредљивих фраза (Jokić et al. 2021).

3. Организација

Због пандемије COVID -19 и ограничења путовања у Европи и шире, школа је одржана онлајн. Следећи традицију EUROLAN-а дугу готово три деценије, која је позната по изврсности академских програма, заједно са дружењима међу професорима и студентима, поред онлајн наставе организован је и низ виртуелних активности са циљем упознавања културе и успостављања ближе интеракције. Присуство је било искључиво онлајн и бесплатно, уз захтев за претходну регистрацију.

Све сесије су биле организоване на платформи за видео конференције. За практичне сесије, било је доступно је неколико виртуелних соба у којима су учесници могли да раде на задацима у мањим групама. Да би охрабрили учеснике да постављају питања и ступају у међусобни контакт, организатори су поставили Slack[16] канал,

13. Програм школе
14. Jerteh
15. Инсталација VocBench-а
16. Slack, Виртуелни центар за сарадњу

као центар за сарадњу, где су предавачи и учесници могли да разјасне све недоумице. Укупан број учесника био је 82, од тога су 52 полазника биле жене, а 30 мушкарци, укључујући и четири учесника из Србије.

За школу су осмишљене различите врсте материјала, укључујући презентације (слајдове)[17] и вежбе[18] пропраћене примерима кода и података.[19] Сви материјали су објављени на интернету и јавно су доступни.

4. Резиме

Школа је пружила драгоцена знања и обучила многе информатичаре и лингвисте о томе како да раде и имају користи од лингвистичких повезаних података. Ово је била прва школа коју је организовала *NexusLinguarum* COST акција из низа обука чије се одржавање планира. Она је имала за циљ да послужи као увод у тему науке о лингвистичким подацима и изгради основу за публику неопходну за похађање наредних школа обуке о напреднијим темама током трајања акције. Сви материјали настали током школе су јавно доступни и заједница их може даље користити. Током завршне сесије, организатори су учесницима дали образац за прикупљање повратних информација о организационим и академским аспектима школе. Резултати показују да су хуманистичке науке, лингвистика и лексикографија имале већу заступљеност међу учесницима него рачунарство, те да је школа била добро фокусирана, тематски уравнотежена и добро организована. Теоријске сесије, подучавање и могућности учења били су веома добро оцењени. С друге стране, због виртуелног начина рада, још увек постоји простор за побољшање у практичним сесијама, организацији друштвених догађаја и могућностима за умрежавање. Стечена знања и вештине ће побољшати развој српских језичких ресурса и помоћи у објављивању више ресурса као лингвистичких повезаних података.

Захвалност

Овај рад је подржала COST акција CA18209 - *NexusLinguarum* “European Network for Web-centred Linguistic Data Science” (NexusLinguarum „Европска мрежа за науку о језичким подацима оријентисаним ка вебу“).

17. Презентације
18. Вежбе
19. Додатни материјали

Литература

Berners-Lee, Tim, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. 2006. "Tabulator: Exploring and analyzing linked data on the semantic web." In *Proceedings of the 3rd international semantic web user interaction workshop,* 2006:159. Athens, Georgia.

Bosque-Gil, Julia, Jorge Gracia, Guadalupe Aguado-de-Cea, and Elena Montiel-Ponsoda. 2015. "Applying the ontolex model to a multilingual terminological resource." In *European Semantic Web Conference,* 283–294. Springer.

Bosque-Gil, Julia, Jorge Gracia, and Elena Montiel-Ponsoda. 2017. "Towards a Module for Lexicography in OntoLex." In *LDK Workshops,* 74–84.

Bosque-Gil, Julia, Jorge Gracia, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez. 2018. "Models to represent linguistic linked data." *Natural Language Engineering* 24 (6): 811–859.

Chiarcos, Christian. 2012. "Interoperability of corpora and annotations." In *Linked Data in Linguistics,* 161–179. Springer.

Chiarcos, Christian, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. 2020. "Modelling frequency and attestations for ontolex-lemon." In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography,* 1–9.

Chiarcos, Christian, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. "Towards open data for linguistics: Linguistic linked data." In *New Trends of Research in Ontologies and Lexical Resources,* 7–25. Springer.

Cimiano, Philipp, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. "Converting language resources into linked data." In *Linguistic Linked Data,* 163–180. Springer.

Declerck, Thierry, John Philip McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel-Ponsoda, Philipp Cimiano, Artem Revenko, Roser Sauri, Deirdre Lee, et al. 2020. "Recent developments for the linguistic linked open data infrastructure." In *Proceedings of the 12th LREC,* 5660–5667.

Declerck, Thierry, Carole Tiberius, and Eveline Wandl-Vogt. 2017. "Encoding lexicographic data in lemon: Lessons learned." In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets. CEURS,* vol. 8.

Fäth, Christian, Christian Chiarcos, Björn Ebbrecht, and Maxim Ionov. 2020. "Fintan-flexible, integrated transformation and annotation engineering." In *Proceedings of the 12th LREC,* 7212–7221.

Gracia, Jorge, Elena Montiel-Ponsoda, Daniel Vila-Suero, and Guadalupe Aguado-De-Cea. 2014. "Enabling Language Resources to Expose Translations as Linked Data on the Web." In *Proceedings of the 9th LREC,* edited by Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland: European Language Resources Association (ELRA), May. ISBN: 978-2-9517408-8-4.

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. "Integrating NLP using linked data." In *International Semantic Web Conference,* 98–113. Springer.

Jokić, Danka, Ranka Stanković, Cvetana Krstev, and Branislava Šandrih. 2021. "A Twitter Corpus and lexicon for abusive speech detection in Serbian." In *Proceedings of the 2021 Language, Data and Knowledge (LDK), 1-3 September in Zaragoza, Spain.*

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The Ontolex-Lemon model: development and applications." In *Proceedings of eLex 2017 conference,* 19–21.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic dictionaries-from file system to lemon based lexical database." In *Proceedings of the 11th LREC - W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018), LREC 2018, Miyazaki, Japan, May 7-12, 2018,* 48–56.

Stellato, Armando, Manuel Fiorelli, Andrea Turbati, Tiziano Lorenzetti, Willem Van Gemert, Denis Dechandon, Christine Laaboudi-Spoiden, Anikó Gerencsér, Anne Waniart, Eugeniu Costetchi, et al. 2020. "VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons." *Semantic Web* 11 (5): 855–881.