Шести пленарни и завршни састанак мреже *NexusLinguarum*

РАД ПРИМЉЕН: 28. јун 2024.
РАД ПРИХВАЋЕН: 29. јун 2024.

Ранка Станковић
ranka@rgf.rs

ORCID: 0000-0001-5123-6273

Универзитет у Београду
Рударско-геолошки факултет
Београд, Србија

COST Акција CA182091[1] - Европска мрежа за науку о лингвистичким подацима засновану на вебу (*European network for Web-centered linguistic data science (NexusLinguarum)*[2]) започета је у октобру 2019. године и завршила се у априлу 2024. Управни одбор акције NexusLinguarum бројао је 69 чланова из 38 земаља. У радним групама (WG) учествовало је 213 чланова из 39 земаља. Главни циљ Акције био је да подстакне синергију широм Европе између лингвиста, информатичара, терминолога и других актера из индустрије и друштва, како би се анализирало и унапредило поље науке о лингвистичким подацима. Технологије повезаних података (LD), у комбинацији са техникама обраде природног језика (NLP) и вишејезичним језичким ресурсима (двојезични речници, вишејезични корпуси, терминологије итд.), препознате су као потенцијал за стварање екосистема који би омогућио транспарентан проток информација између различитих лингвистичких извора података на више језика, решавајући при томе проблем семантичке интероперабилности.

Шести пленарни састанак COST Акције CA18209 – Европске мреже за науку о лингвистичким подацима засновану на вебу (*NexusLinguarum*) – одржан је у Атини 20. и 21. марта 2024. године. Био је то хибридни састанак, праћен низом додатних догађаја поводом завршетка Акције,

1. CA182091
2. NexusLinguarum

као што су снимање лекција које ће бити коришћене за MOOC (*Massive Open Online Course*) курсеве и излагање постера.

Први дан био је посвећен достигнућима, резултатима и исходима. Поднети су извештаји за сваку радну групу (WG). Резултате WG1: *Језички ресурси засновани на повезним подацима* представио је др Милан Дојчиновски. Патриција Мартин Чозас (Patricia Mart´ın Chozas) известила је о резултатима WG2: *NLP сервиси засновани на повезним подацима.* Дагмар Громан (Dagmar Gromann) представила је активности WG3: *Подршка науци о лингвистичким подацима*, док је преглед WG4: Студије случаја и примене изложила Сара Карваљо (Sara Carvalho).

Током презентације главних резултата акције NexusLinguarum одржана су инспиративна излагања и представљени постери. Марија Пија ди Буоно (Maria Pia di Buono) говорила је о коришћењу великих језичких модела (Large Language Models, LLM) у оквиру лингвистичких повезаних података (*Linguistic Linked Data*, LLD). Мађеј Огродничук (Maciej Ogrodniczuk) представио је предлог пројекта под називом „Универзални дискурс“. Тема излагања Макса Јонова (Max Ionov) била је лингвистички повезани отворени подаци (LLOD) за интероперабилне морфолошке описе, док је Инеке Шурман (Ineke Schuurman) известила о знаковним језицима и LLOD-у.

Резултат (deliverable) 4.3 са Завршним извештајем о активностима, студијама случаја и применама (Carvalho and Kernerman 2024) пружио је свеобухватан опис девет студија случаја реализованих у последњем извештајном периоду и њихове имплементације. Студије случаја и апликације коришћене су за тестирање и валидацију релевантних методологија, технологија и стандарда Акције. Студију случаја о јавном здрављу представила је Ана Острошки Анић. Златни стандард за категоризацију увредљивог језика према LLOD шеми представила је Ана Банчковска (Anna Ba¸czkowska). Флорентина Армаселу (Florentina Armaselu) описала је студију случаја LLODIA (LLOD за дијахроне анализе), док је Гиедре Валунаите Олескевицине (Giedre Valunaite Oleskevicien˙e) представила студију случаја у области друштвених наука. Примену дубоког учења за анализу лингвистичких података представио је Atanas Hristov (Atanas Hristov). Hugo Гонсало Оливера (Hugo Gon¸calo Oliveira) презентовао је *MultiLexBATS: Description and Analogy Completion*: опис и допуну аналогија (Gromann et al. 2024). Паола Маронђију (Paola Marongiu) говорила је о студији о детекцији лексичко-семантичких промена у латинском језику, са посебним фокусом на медицински изразе у латинском.

Након ових излагања уследило је излагање постера, у оквиру ког је Радован Гарабик представио искуства у коришћењу LLOD-а за иницијално формирање двојезичних речника. Верџиника Барбу Митителу (Verginica Barbu Mititelu) приказала је румунске ресурсе за повезане податке развијене током NexusLinguarum акције. Ранка Станковић говорила је о приступу и искуствима у повезивању NIF (формат за размену у обради природног језика) корпуса са речником сагласним са *Ontolex-Lemon* моделом на платформи *Wikibase*. Катерина Здравкова представила је резултате решавања флективне вишезначности македонских придева, док је Димитар Трајанов изложио приступ аутоматизацији процеса израде речника.

Представљени су и резултати краткорочних научних мисија (STSM) и финансирања за виртуелну мобилност. Студију случаја из области сајбер безбедности представила је Сигита Раковичиене (Sigita Rackevi čiene), приказавши развој језичких ресурса из те области. Другу тему у вези са сајбер безбедношћу представио је Кристијан Кијаркос (Christian Chiarcos), говорећи о конверзији и повезивању лингвистичких скупова података. Манџола Заселари (Manjola Zacellari) дала је преглед израде новог *UD-Treebank* корпуса за албански језик, док је Анас Фахад Кан (Anas Fahad Khan) говорио о STSM пројекту о ознакама домена у повезаним лексикографским ресурсима. Александра Томазуска (Aleksandra Tomaszewska) резимирала је Иницијативу за вишејезичку анотацију дискурса: циљеве, изазове и могућности (*Multilingual Discourse Annotation Initiative*, MDAI).

Другог дана, Рут Коста (Rute Costa) представила је Заједнички академски курикулум за науку о лингвистичким подацима, развијен као заједнички напор у циљу припреме предлога Европској унији за Erasmus Mundus заједнички мастер програм (*Academic Common Curriculum on Linguistic Data Science*). Прва фаза укључивала је развој пројекта Повезивање лингвистике са науком о подацима (*Linking Linguistics to Data Science*, LL2DS), усмереног на успостављање иновативног Erasmus Mundus заједничког мастер програма, који би био међународно препознат по интеграцији лингвистике и науке о подацима. Овај пројекат водио је ка другој фази, у оквиру које је предат потпуно развијен предлог EMJM програма (Costa and Garcia 2024).

Представљени су и план и програм са заједничком агендом за будућа истраживања у области науке о лингвистичким подацима. У првом делу идентификовани су и анализирани бројни изазови у области науке о лингвистичким подацима (LDS), нарочито у оквиру LLOD-а. Главни

проблеми обухватају баријере за приступ технологији, одрживост, покривање постојећих модела репрезентације, метаподатке, међујезично повезивање, недовољно заступљене језике и вишејезичност. Посебно је истакнута веза између LLOD-а и нових великих језичких модела (LLMs), а представљен је и конкретан план за наставак активности започетих у оквиру *NexusLinguarum* COST Акције, кроз реализацију предложеног плана и програма.

Припрема Смерница и најбоље праксе за LLOD (Martín Chozas et al. 2024) резултат је заједничког подухвата у оквиру COST Акције NexusLinguarum и W3C заједнице за најбоље праксе у вези са вишејезичким повезаним отвореним подацима (Best Practices for Multilingual Linked Open Data, BPMLOD), усмереног на израду смерница и препорука за повезивање података и сервиса између језика. Први скуп смерница фокусиран је на ЛЛОД, а следећи на интеграцију LLOD-а са сервисима обраде природног језика (NLP). Свеобухватне препоруке за LLOD и LLOD-свесне NLP сервисе, заједно са даљим развојем BPMLOD заједнице и текућим активностима на ажурирању смерница, обезбеђују одрживост резултата и након завршетка COST Акције *NexusLinguarum*.

Захвалност

Овај рад је реализован уз подршку Акције COST CA18209 – Европска мрежа за науку о лингвистичким подацима засновану на вебу (*NexusLinguarum*).

Литература

Carvalho, Sara, and Ilan Kernerman. 2024. *Final Activity Report (months 25-54). Working Group 4: Use Cases and Applications.* https://nexuslinguarum.eu/wp-content/uploads/2024/04/Deliverable-D4.3-WG4-Final-Activity-Report_compressed.pdf.

Costa, Rute, and Jorge Garcia. 2024. *Academic Common Curriculum on Linguistic Data Science - LDS.* https://nexuslinguarum.eu/wp-content/uploads/2024/04/Deliverable-D3.3-common-curriculum-1.pdf.

Gromann, Dagmar, Hugo Gonçalo Oliveira, Lucia Pitarch, and et al. 2024. "MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* 11783–11793. ELRA Language Resource Association. https://aclanthology.org/2024.lrec-main.1029.pdf.

Martín Chozas, Patricia, Milan Dojchinovski, Katerina Gkirtzou, Anas Fahad Khan, and Andon Tchechmedjiev. 2024. *Guidelines and Best Practices on Linguistic Linked Open Data.* https://nexuslinguarum.eu/wp-content/uploads/2024/03/Deliverable-D1.4-Guidelines-and-Best-Practices-on-LLOD.pdf.