Обука за читање на даљину – *Exploring ELTeC: Use-Cases for Information Extraction and Analysis*

Александра Марковић
aleksandra.markovic@isj.sanu.ac.rs
Институт за српски језик САНУ
Београд, Србија

РАД ПРИМЉЕН: 28. мај 2022.
РАД ПРИХВАЋЕН: 8. јун 2022.

У Београду је од 22. до 24. марта одржана радионица *Exploring ELTeC: Use-cases for Information Extraction and Analysis* (*Истраживање ELTeC-а: примери екстракције и анализе информација*). Радионица је била посвећена обуци за истраживање ELTeC-а (ELTeC је скраћеница за *European Literary Text Collection* – Колекција текстова европске књижевности). Ово је била завршна обука у оквиру COST акције (16204) *Distant Reading for European Literary History* (Читање на даљину за европску књижевну историју).

Радионица је била организована на Рударско-геолошком факултету Универзитета у Београду, а организатори су били проф. др Ранка Станковић с Рударско-геолошког факултета (Универзитет у Београду), проф. др Цветана Крстев с Филолошког факултета (Универзитет у Београду),[1] као и мр Јоана Бишук (Joanna Byszuk) са Института за пољски језик Пољске академије наука. Одржана је у хибридном облику, па су учесници могли да бирају да ли ће учествовати уживо или онлајн. Уживо су учествовали полазници из Србије, Словеније, Румуније, Белгије и Литваније, док су онлајн учествовали још и полазници из Аустрије, Британије, Португала и Мађарске. За учешће нису били постављени формални захтеви у смислу образовања, али је, с обзиром на интензитет и обим програма, било препоручено поседовање основних информатичких вештина. Радионица је била

1. Цветана Крстев и Ранка Станковић заслужне су за израду српске колекције текстова. Заинтересовани за српску колекцију романа дигитализованих у оквиру акције *Distant Reading* могу прочитати више о томе у часопису Инфотека, бр. 21/2, посвећеном у целости српској колекцији дигитализованих романа, док самој колекцији могу приступити на github-у акције.

намењена истраживачима из земаља укључених у пројекат *Distant Reading*, заинтересованим за дигитално истраживање књижевности, корпусну и рачунарску лингвистику, теорију књижевности, као и њихову методолошку примену ван граница националних традиција.

Представљени су практични приступи екстраховању информација и анализи текстуалних података, нарочито из корпуса ELTeC развијеног у оквиру COST акције. Међу обрађеним темама били су различити аспекти рада са именованим и географским ентитетима: њихово препознавање (NER, *Named entities recognition*) и екстраховање (NEE, *Named entities extraction*), радом са Вики подацима (Wiki-ELTeC data), повезивање (историјских) података у Nodegoat-у (апликацији посебно израђеној за истраживања у хуманистици),[2] семантичка анализа помоћу вектора речи и језичких модела, као и поређење корпуса помоћу стилометрије.

Међу предавачима је било познатих имена, учесника у акцији; поменућемо нека од њих: Christof Schöch (професор дигиталне хуманистике на Универзитету у Триру, руководилац COST акције *Distant Reading for European Literary History*; Maciej Eder (директор Института за пољски језик Пољске академије наука); Diana Santos (Универзитет у Ослу), Fotis Yannidis (професор дигиталне хуманистике на Универзитету у Вирцбургу), проф. др Цветана Крстев (проф. информатике на Катедри за библиотекарство и информатику Филолошког факултета УБ, у пензији) и проф. др Ранка Станковић (професор информатике на Катедри за примењену математику и информатику Рударско-геолошког факултета УБ).

Радионица је била организована у девет модула и од учесника се очекивало да похађају сваки од њих.[3]

1. Кристоф Шех (Christof Schöch) одржао је (онлајн) уводно предавање на тему: *What is ELTeC all about?* (Шта је то ELTeC?). Говорио је о циљевима акције акције *Distant Reading for European Literary History*, с посебним освртом на структуру основног резултата овог пројекта – вишејезичне колекције европских књижевних текстова (ELTeC).
2. Маћеј Едер, Јоана Бишук, Артјомс Шела (Maciej Eder, Joanna Byszuk, Artjoms Šeļa) одржали су предавање (онлајн): *Exploring and comparing ELTeC corpora with stylometry* (Истраживање и поређење корпуса ELTeC помоћу стилометрије).

2. Nodegoat
3. Заинтересовани могу погледати програм радионице.

3. Дијана Сантос (Diana Santos) говорила је на тему (онлајн): *NER exploitation and analysis* (Препознавање именованих ентитета (NER) – експлоатација и анализа).
4. Бенедикт Перак (Benedikt Perak) одржао је предавање (уживо): *ELTeC Data Analysis, Representation of the Geo-Entities and Interlinking with Knowledge bases* (Анализа података из ELTeC-а, представљање географских ентитета и повезивање с базама знања).
5. Фотис Јанидис и Леонард Конле (Fotis Jannidis, Leonnard Konle) говорили су на тему (онлајн): *Semantic analysis using word embeddings and language models* (Семантичка анализа помоћу вектора речи и језичких модела).
6. Ранка Станковић и Милица Иконић Нешић одржале су (уживо) сесију посвећену Вики подацима: Увод у Вики податке; схема Вики-ELTeC (сви метаподаци из заглавља заједно са основним ликовима који се помињу у романима, њиховим односима, локацијама; процедура и кораци за попуњавање Вики-ELTeC података; предефинисана испитивања помоћу SPARQL упита. Вежбе: попуњавање Вики података.
7. Денис Морел, Ерик Лапорт и Цветанав Крстев (Denis Maurel, Eric Laporte) припремили су (онлајн) излагање на тему: *Unitex for processing of literary text: the case of NER automata. Enriching ELTEC texts by Named Entity Recognition using CasSys to parse texts with Unitex graph cascade of finite state transducers in different languages* (Unitex за обраду књижевних текстова: случај аутомата за препознавања именованих ентитета. Обогаћивање текстова из ELTEC-а именованим ентитетима препознатим Unitex-овим алатом CasSys за парсирање текстова на различитим језицима помоћу каскада коначних трансдуктора).
8. Џеси Лабов, Пим ван Бре, Герт Кеселс (Jessie Labov, Pim van Bree, Geert Kessels) имали су (онлајн) излагање на тему: *ELTec in Nodegoat – Introduction to the Nodegoat interface and how it works with this kind of data,* (ELTec на платформи Nodegoat – Увод у интерфејс Nodegoat-а и начин обраде текстуалних података).
9. Пим ван Бре, Герт Кеселс имали су (онлајн) излагање на тему: *Using Nodegoat for working with the ELTeC data* (Употреба Nodegoat-а за рад са подацима из ELTeC-а), у коме су се фокусирали посебно на NER и представљање начина за обогаћивање тих података повезивањем са отвореним изворима података).

Радионица је била интензивна, веома информативна и динамична. Материјали потребни за рад и вежбе били су постављени на платформи GitHub.[4] Организација и вођење кроз одабране теме били су одлични, теме занимљиве, а вежбе добро осмишљене (чак и за оне који нису довољно информатички обучени, као ауторка овог приказа, која је и сама учествовала на радионици).

4. Материјали са радионице