Теоријски основ и креирање паралелног корпуса: оцртавање нових перспектива у обради библиотечког материјала

Драгана Столић
stolic@unilib.rs
Универзитетска библиотека „Светозар Марковић” Београд, Србија

РАД ПРИМЉЕН: 24. новембар 2022.
РАД ПРИХВАЋЕН: 31. март 2023.

1. Увод

Публикација Јелене Андоновски *Паралелни српско-немачки корпус књижевних текстова: израда, проналажење информација и семнатички веб*[1] придружује се вредној издавачкој колекцији Универзитетске библиотеке „Светозар Марковић“, покренутој 2016. године под називом „Хомилије“, намењеној публиковању научних монографија, које махом чине прерађене одбрањене докторске дисертације аутора из друштвено-хуманистичких наука, односно дигиталне хуманистике. И овде је реч о прилагођеном тексту дисертације *Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литератног корпуса*, коју је ауторка одбранила на Филолошком факултету у Београду 2020. године.

Публикација представља високопрецизну систематизацију знања и умећа везаних за обраду, опис и истраживање језичких корпуса како би се створиле претпоставке за потоње лексичке, терминолошке и семантичке анализе. Поступно и логично, дело нас води ка разумевању овог специфичног сегмента библиотекарства и информатике, оцртава модел на који начин треба приступити таквом задатку, не заборављајући да укаже на путање даљих истраживања. Уопштено говорећи, овакве публикације помажу запосленима у библиотечко-информационој делатности да јасније сагледају читав домен истраживања који је,

1. Јелена Андоновски: *Паралелни српско-немачки корпус књижевних текстова: израда, проналажење информација и семантички веб*, Београд: Универзитетска библиотека „Светозар Марковић“, 2021.

условљен технолошким напретком и развојем нових алата, омогућио аналитичко задирање у публикације саме, придружујући их доступним надређеним целинама и обезбеђујући раније немогуће приступе књижевним делима и писцима. Као што класично библиотекарство подразумева значајан број разних активности, прикупљања, обраде и презентације публикација како би се кориснику омогућило да дође до жељеног извора информација, тако и нове информационе технологије допуштају анализу и обраду језичке грађе како би се разумела природа језика и његовог коришћења и како би се утабали путеви за лакше превођење и разумевање других језика.

2. Пут ка истраживању корпуса

Иако публикација има више поглавља, условно се могу уочити две крупније целине: једна, коју можемо назвати теоријским односно *појмовним полазиштем* и другу, која је укључила примену свега претходно наведеног на одређени корпус (СрпНемКор), што је подразумевало употребу одређених алата, спровођење претраживања и анализу добијених резултата.

Поменути први део најпре одређује појам *корпуса*, као основе *корпусне лингвистике*, а у оквиру тога – *паралелних корпуса*. У делу се пружа јасна типологија корпуса, који могу бити једнојезични и вишејезични, који се даље могу поделити на *паралелне* и *упоредне*. Паралелни су састављени од текстова на изворном језику и њихових превода на један или више циљаних језика и као такви значајни су за учење страних језика, различита превођења, нарочито машинска, затим терминолошка и лингвистичка истраживања и сл. Истакнуто је да је основни елемент двојезичног корпуса *битекст* или *паралелизовани* текст, где се повезивање текста и његовог превода спроводи на различитим нивоима – од читавог документа, преко поглавља, пасуса до реченице и речи, са циљем *упаривања текстова (паралелизације)* или стварања веза између варијанти јединица превођења и формирања скупа јединица превођења („Translation Unit“).

После осврта на прве видљиве покушаје формирања паралелних корпуса, у раду је читаво поглавље посвећено постигнућима у Србији на овом пољу, пре свега раду Друштва за језичке ресурсе и технологије - JePTex[2]. На тај начин је пружен кратак и сажет, али веома информативан преглед резултата вишегодишњих пројеката.

2. JePTex

Поступак израде паралелних корпуса представљен је у свим корацима, чиме се значајно олакшава разумевање опсега оваквог приступа, где посебну тежину имају анотације, највише оне које укључују коришћење проширивог језика за обележивање (XML – eXtensible Markup Language) и све више присутну, али чини се недовољно прихваћену, иницијативу за кодирање текста – TEI (Text Encoding Initiative). Изабраним сегментима се придодају обележја (етикете) која одређеним речима додељују додатне дескрипторе за врсту речи, канонски облик или морфолошку категорију. Припреми материјала придружује се потом приказ предуслова за проналажење информација где кључну улогу имају правилно додељени метаподаци. Појам „метаподатак“, који практично узима превагу у савременом библиотекарству у односу на класичне библиографске податке, образложен је пажљиво у свим облицима њиховог испољавања (описни, структурални, административни) и приказом основних схема (Dublin Core, METS, MODS).

На самом крају овог, како је речено, фундаменталног сегмента, у којем су формулисане дефиниције и оцртан читав систем средстава неопходних када је у питању стварање једног корпуса, као врхунац и најасложенији појам образлаже се *семантички веб*. Веома стручно, али довољно разумљиво, онај део рада се хвата укоштац са иницијативом „отворених повезаних података“ (Linked Open Data) настојећи да појасни функционисање веба као глобалне базе података чија се структура више не заснива на статичним страницама, већ, кроз механизам за увезивање података на нивоу њиховог значења, тј. на садржајима повезаним према семантичким карактеристикама.

3. Израда српско-немачког паралелног корпуса

После исцрпног приказа и анализе алата и ресурса који се могу користити у стварању паралелног копруса, други део студије (поглавље 6) бави се апликацијом свега наведеног разматрајући могућности, домете или ограничења неких ресурса на датом материјалу. За формирање корпуса коришћено је седам романа различитих српских писаца и исти број дела различитих писаца са немачког говорног подручја, руководећи се критеријумима као што су: награђиваност аутора, доступност дела на оба језика у библиотекама Србије, популарност дела и његова обимност.

Приказани су кораци у стварању корпуса од дигитализације материјала, примене оптичког препознавања текста, структурне

анотације и, најзад, до постављања корпуса у оквир дигиталне библиотеке *Библиша*[3] и обезбеђивања услова за квалитетно претраживање информација, као и за укључивање обрађеног корпуса у мрежу отворених лингвистичких података. Читав приказ јесте аутентична илустрација свих нужних радњи, али и изазова, проблема и ограничења чиме резултати овог рада не губе на значају.

4. Улога библиотека

Ова студија је практично, кроз интеграцију више елемената, и сама креирала један нови „корпус“ појмова и значења који на другачији начин сагледавају текст. У том смислу, чини се да класична библиотечко-информациона делатност нема нужно додирних тачака са таквим приступом. Међутим, посебан квалитет овог рада управо јесте доследно истицање незаобилазне улоге библиотека, које, и поред доминантно „несемантичког“ приступа, морају да чине део једног оваквог процеса, не само због дигитализације где најактивније учествују, или чињенице да поседују грађу која је полазиште у креирању корпуса. Наиме, управо библиотеке, у којима се већ креирају различити метаподаци и исписи, поседују предуслове да учине одговарајући искорак према значењском повезивању добијених података и обезбеђивању потпунијих информација за кориснике.

3. Библиша