Владимир Поломац[*]
Универзитет у Крагујевцу[**]
Филолошко-уметнички факултет
Одсек за филологију, Катедра за српски језик

ТЕОРИЈСКО-МЕТОДОЛОШКИ ОКВИР ЗА ИЗГРАДЊУ ЕЛЕКТРОНСКОГ КОРПУСА СРПСКИХ СРЕДЊОВЕКОВНИХ ПОВЕЉА И ПИСАМА

1. На потребу формирања електронског корпуса старосрпског језика на принципима поузданости, репрезентативности и претраживости указано је тек у скорије време у Павловић 2009. Основни теоријско-методолошки принципи за израдњу електронског корпуса српских средњовековних повеља и писама, као најважнијег поткорпуса будућег референтног историјског електронског корпуса српског језика, детаљније су размотрени у Polomac 2021. Најважнији резултат овога рада представља (а) дефинисање историјског корпуса српског језика, као и места средњовековних повеља и писама у њему, посебно с обзиром на принципе репрезентативности, поузданости и балансираности историјских корпуса, (б) дефинисање принципа за селекцију текстова повеља и писама за корпус, као и (в) дефинисање основних принципа за припрему и приређивање текстова средњовековних повеља и писама за корпус. У ослонцу на овај рад, у наставку истраживања занимало нас је како се у теоријско-методолошки оквир за изградњу корпуса могу укључити алати засновани на принципима вештачке интелигенције и машинског учења. Основни циљеви рада могу се свести на следеће: а) креирање алата за аутоматско рашчитавање текста српских средњовековних повеља и писама помоћу HTR (енгл. *Handwritten Text Recognition*) технологије; б) успостављање теоријског оквира за креирање почетног скупа података за обуку модела за аутоматску обраду (токенизацију, лематизацију и морфосинтаксичку анотацију) текста српских средњовековних повеља и писама помоћу алата UDPipe (Straka, Hajič et al. 2016; Straka, Straková 2022).

2. За обуку прве верзије генеричког HTR модела помоћу софтверске платформе *Transkribus* коришћени су скупови података из Бањске хрисовуље, трећег примерка Дечанске хрисовуље, писама Портине српске канцеларије и повељама и писама XII–XV века из Дубровачког архива. Карактеристике прве верзије генеричког (општег) HTR модела *Logothet 1.0* за српске средњовековне повеље и писма приказане су у следећој табели.

Табела 1. Карактеристике генеричког HTR модела *Logothet 1.0*

Број страна на скупу за обуку	Број речи на скупу за обуку	Број епоха за обуку	Проценат погрешно препознатих карактера (CER)
565	79 187	112	5,6%

Наведена табела показује како се помоћу овога HTR модела аутоматски може добити рашчитан текст српских средњовековних повеља и писама писаних различитим типовима ћирилице (устав, полуустав, брзопис) са око 94–95% тачности. Мали

[*] Електронска адреса: v.polomac@filum.kg.ac.rs

[**] Рад је настао у оквиру међународног билатералног пројекта *Креирање AI модела за аутоматску обраду српких средњовековних рукописа* који финансирају Министарство науке, технолошког развоја и иновација Републике Србије и Немачка служба за академску размену (DAAD).

проценат грешака (око 5–6%) најчешће се односи на размак међу речима, затим на надредна слова, титле и специфичне лигатуре, а сасвим ретко и појединачна слова. Уз помоћ овога HTR модела процес преношења текста у машински читљив облик, а самим тим и даља обрада текста и припрема корпуса, вишеструко се може убрзати. Посебно треба указати и на то да је ово само прва верзија HTR модела, те да ће напредак у рашчитавању текста донети нове скупове података за обуку на основу којих ће се перформансе HTR модела усавршавати.

3. Скорашњи развој алата UDPipe (Straka, Hajič et al. 2016; Straka, Straková 2022) омогућава аутоматску обраду (токенизацију, лематизацију и морфосинтаксичку анотацију) текста не само за велики број модерних језика света већ и за многе класичне језике (нпр. старогрчки, латински, готски, акадски, санскрит или коптски), као и за историјске варијетете индоевропских језика (нпр. старофранцуски или староруски). UDPipe омогућава аутоматску обраду српског језика (Samardžić, Starović et al. 2017) и великог броја других словенских језика, као и два историјска варијетета: старословенског језика, као и староруског језика, на основу више модела заснованих на материјалу различитих текстова (Zeman 2015: 152–153; Zeman, Kosek et al. 2023: 215). Иницијални кораци у аутоматској обради помоћу овога алата недавно су проведени и за старочешки језик (Zeman, Kosek et al. 2023). У ослонцу на ова истраживања, приступили смо креирању почетног скупа података за обуку UDPipe модела за аутоматску обраду српских средњовековних повеља и писама. За реализацију овога циља било је најпре неопходно дефинисати теоријски оквир за токенизацију и лематизацију текста, као и скуп ознака за морфосинтаксичку анотацију у складу са теоријским оквиром универзалних зависности (енгл. *Universal Dependencies*).

3.1. Основна начела за токенизацију текста српских средњовековних повеља и писама подударна су са начелима коришћеним приликом припреме скупа података за старословенски језик у оквиру пројекта PROIEL (Eckhoff, Bech et al. 2018): а) токени су одвојени размаком, б) нису присутни токени састављени од више речи (енгл. *multi-word tokens*). За разлику од старословенског, у нашем скупу података присутни су и интерпункцијски знаци.

3.2. Приликом лематизације текста српских средњовековних повеља и писама придржавали смо се следећих општих начела: а) начело ортографске нормализације, б) начело реконструкције језичког стања са краја XII и почетка XIII века, и в) начело чувања варијаната лема које рефлектују језичке одлике старосрпског и српскословенског језика. Начело ортографске нормализације подразумевало је да су леме наведене савременом српском ћирилицом којој је додато само неколико специфичних слова (ь, ѣ и ы) којима су означавани гласови присутни у старосрпском и српскословенском вокалском систему са краја XII и почетка XIII века. У складу са овим, друго начело подразумевало је реконструкцију леме према претпостављеном фонолошком систему заједничком за старосрпски и српскословенски језик са краја XII и почетка XIII века, без обзира на каснији језички развој (нпр. рука, дьн, јесьм, мѣсец, вѣра, быти, Влк, вьс, и сл.). Будући да се у текстовима повеља и писама употребљавају и старосрпски и српскословенски језик, треће начело подразумевало је чување варијаната лема које рефлектују њихове специфичне фонетске одлике (нпр. аз према јаз или ја; межда према међа или меја, и сл.).

3.2. Будући да опсег анотације електронског корпуса српских средњовековних повеља и писама у почетној фази подразумева само врсте речи и морфолошке одлике, неопходно је дефинисати скуп ознака за анотацију према теорији универзалних зависности (енгл. *Universal Dependencies*).

У оквиру ове теорије употребљава се седамнаест универзалних ознака за врсте речи: ADJ (придев), ADP (предлог), ADV (прилог), AUX (помоћни глагол), CCONJ (координирани везник), DET (детерминатор), INTJ (узвик), NOUN (именица), NUM (број), PART (речца), PRON (заменица), PROPN (властита именица), PUNCT (интерпункцијски знак), SCONJ (субординирани везник), SYM (симбол), VERB (глагол) и X (остало). За разлику од скупа ознака развијеног за потребе морфосинтаксичке анотације старословенског језика у коме изостају PART, PUNCT и SYM, у нашем корпусу присутне су све ознаке. Посебан коментар најпре захтева категорија детерминатора DET, неуобичајена у традиционалним граматичким описима словенских језика, уведена за означавање придевских заменица како би се осигурала доследност анотације са старословенским и модерним словенским језицима (Zeman 2015: 153–154). Проблем означавања партиципа, карактеристичан и за друге словенске језике (Zeman 2016: 145), решен је тако што се тежило њиховом доследном означавању као VERB. Одступање од овога начела било је оправдано само у случајевима када није било могуће реконструисати глаголску лему (нпр. *светопочивши*, *блаженопочивши* и сл.). У оваквим примерима коришћена је ознака ADJ, али је уз морфолошке одлике карактеристичне за придеве додата и ознака која упућује на партиципско порекло VerbForm=Part, чиме се омогућава претраживање свих партиципских форми, независно од тога да ли су њихове леме означене као VERB или ADJ. Ознака SYM коришћена је најчешће за обележавање крста који у повељама и писмима долази у функцији симболичке инвокације, док је ознака X најчешће употребљавана за обележавање оштећених и нејасних делова текста.

Адаптација старословенског скупа ознака за морфолошке одлике проведена је с обзиром на језичке посебности корпуса, али и с обзиром на специфичне потребе претраживања и екстракције грађе из корпуса. У категорији именица помоћу Polarity=Neg обележене су именице са лексичком негацијом. Мали број непромењивих именица (нпр. *кир*, *кира*) обележен је помоћу InflClass=Ind. Будући да је један од општих циљева креирања електронског корпуса и аутоматска екстракција грађе за историјски речник старосрпских имена, у категорији властитих именица додата је ознака NameType са вредностима NameType=Geo (географско име), NameType=Giv (лично име), NameType=Sur (презиме), NameType=Pat (патроним), NameType=Zoo (зооним) и NameType=Nat (етноним). У категорији личних заменица првог и другог лица, повратне заменице за сва лица и именичких заменица *кто* и *што* изостављена је ознака Gender. Будући да су придевске заменице означене као DET, њихова врста је означена помоћу PronType=Dem (показне), PronType=Rel (релативне), PronType=Int (упитне), PronType=Ind (неодређене), PronType=Neg (одричне) и PronType=Tot (опште). Посесивне заменице нису обележене као посебна врста придевских заменица већ као Poss=Yes|PronType=Prs. Морфосинтаксичким одликама придева обележеним и у старословенском скупу података додали смо Poss=Yes за посесивне придеве и Polarity=Neg за придеве са лексичком негацијом. Бројеви редовно долазе са ознаком NumForm која може имати вредност NumForm=Word за бројеве написане речима и NumForm=Cyrill за слова у бројној вредности. Врсте бројева обележене су ознаком NumType са вредностима NumType=Card (основни бројеви), NumType=Ord (редни бројеви) и NumType=Sets (збирни бројеви). Најважније одступање од старословенског скупа података у категорији глагола тицало се обележавања аориста и имперфекта. Скуп ознака модификован је по угледу на старочешки (Zeman 2023: 218) тако да се презент, аорист и имперфекат разликују по ознаци Tense која може имати вредности Tense=Pres (презент), Tense=Past (аорист) и Tense=Imp (имперфекат). Сва три облика носе ознаку Mood=Ind која их разликује од императива који је обележен са

Mood=Imp. Активни и пасивни партиципи презента и претерита, поред ознаке VerbForm=Part и ознака за падеж (Case), род (Gender) и број (Number) обележени су и помоћу Tense=Pres или Tense=Past и Voice=Act или Voice=Pass. Као и код придева, и краћи облици партиципа обележени су као Variant=Short. За партицип перфекта који се налази у склопу перфекта, плусквамперфекта, футура II и потенцијала коришћена је ознака VerbForm=PartRes, као и у старословенском скупу података. Проблем обележавања сложених глаголских облика само је делимично решен код потенцијала и футура. У првом случају помоћни глагол који гради потенцијал посебно је обележен ознаком Mood=Cnd, а у другом случају инфинитив у склопу футура, поред ознаке VerbForm=Inf, носи и ознаку Tense=Fut. Претраживање перфекта, плусквамперфекта и футура II за сада је могуће само преко VerbForm=PartRes, након чега би требало да следи ручно филтрирање примера.

3.3. Креирање скупа података за обуку модела започели смо хронолошки, од осам најстаријих сачуваних повеља с краја XII и почетка XIII века. Аутоматску обраду наших повеља започели смо помоћу UDPipe2 модела за старословенски језик. У резултату овога процеса за сваку повељу смо добили посебан фајл у *CoNLL-U* формату који је ручно коригован у *CoNLL-U* едитору (Heinecke 2019). Након ручне корекције аутоматски обрађених повеља добијен је почетни скуп података који се састојао од 142 реченице и 3531 токена. С обзиром на величину нашег скупа података и одсуство анотације синтаксичких релација у њему за обуку модела био нам је на располагању само алат UDPipe1. Посебно обучени UDPipe1 модели са нашим скупом података дају боље резултате у анотацији врста речи у односу на претходни обучени UDPipe2 модел за старословенски, али не и боље резултате у односу на претходно обучене UDPipe2 моделе за староруски језик. То значи да се наставак обраде повеља и писама XIII века и ширење скупа података за обуку модела у првој фази може засновати на UDPipe2 моделима за староруски језик, након чега би дошла ручна корекција обрађеног текста помоћу *CoNNL-U* едитора. Након достизања критичног броја обрађених токена (најмање 10 000) и ширења опсега анотације и на синтаксичке релације, добили бисмо скуп података за обуку UDPipe2 модела за старосрпски језик, чиме би се процес обраде текста за електронски корпус српских средњовековних повеља и писама значајно убрзао.

4. У раду је показано како се процес изградње електронског корпуса српских средњовековних повеља и писама може убрзати уз помоћ информационих технологија заснованих на принципима вештачке интелигенције и машинског учења. За потребе преношења текста у машински читљив облик обучили смо генерички (општи) HTR модел *Logothet 1.0* у софтверској платформи *Transkribus*. Модел је обучен на материјалу српских средњовековних повеља и писама XII–XVI века писаних различитим типовима ћирилице (устав, полуустав и брзопис) са одличним квантитативним и квалитативним показатељима. Проценат погрешно рашчитаних карактера износи око 5,6%, а најчешће грешке односе се на размак међу речима, надредна слова, титле и специфичне лигатуре. За потребе обраде текста српских средњовековних повеља и писама било је неопходно најпре успоставити теоријски оквир за токенизацију, лематизацију и морфосинтаксичку анотацију, након чега је креиран почетни скуп података (3 531 токен) за обуку модела помоћу алата UDPipe1. Будући да перформансе овога модела нису биле боље од претходно обучених модела за староруски језик помоћу алата UDPipe2, неопходно је даље ширење скупа података за обуку модела, као и опсега анотације додавањем ознака за синтаксичке релације, како би у крајњем исходу могао бити обучен UDPipe2 модел за старосрпски језик. Овакав модел би значајно убрзао обраду текста српских средњовековних повеља и писама и

изградњу и објављивање пилот верзије електронског корпуса који би обухватио повеље и писма XII–XIII века.

Кључне речи: историја српског језика, корпусна лингвистика, историјски електронски корпус, српске средњовековне повеље и писма, аутоматска обрада текста, вештачка интелигенција, машинско учење.

Цитирана литература

Павловић, С. „Електронско претраживање старосрпског језичког корпуса у светлу стандардизације старословенске ћирилице.” У: *Стандардизација старословенског ћириличког писма и његова регистрација у Уникоду*. Г. Јовановић, Ј. Грковић-Мејџор, З. Костић, В. Савић (ур.). Београд: Институт за српски језик САНУ (2009): 135–146.

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O. E., Johndal, M. „The PROIEL treebank family: a standard for early attestations of Indo-European languages.” *Language Resources and Evaluation*, 52 (2018): 29–65.

Heinecke, J. „*ConlluEditor*: a fully graphical editor for universal dependencies treebank files.” In: *Proceedings of the Third Workshop on Universal Dependencies*. A. Rademaker, F. Tyers (eds.). Paris: Association for Computational Linguistics (2019): 87–93.

Polomac, V. „Towards Fundamental Principles for Creating Electronic Corpus of Serbian Medieval Charters and Letters.” *Scripta&e-Scripta*, 21 (2021): 41–53.

Samardžić, T., Starović, M., Agić, Ž., Ljubešić, N. „Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages.” In: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, R.Yangarber (eds.). Valencia: Association for Computational Linguistics (2017): 39–44. <https://aclanthology.org/W17-1407/> 31.1.2025.

Straka, M., Hajič, J., Straková, J. „UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing.” In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds.). Portorož: European Language Resources Association (ELRA) (2016): 4290–4297 <https://aclanthology.org/L16-1680/> 31.1.2025.

Straka, M., Straková J., *UDPipe 2*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). Prague: ÚFAL, 2022. <http://hdl.handle.net/11234/1-4816> 31.01.2025.

Zeman, D. „Slavic Languages in Universal Dependencies.” In: *Natural Language Processing, Corpus linguistics, E-Learning*. K. Gajdošová, A. Žáková (eds.). Lüdenscheid: Ram-Verlag (2015): 151–163.

Zeman, D. „Universal Annotation of Slavic Verb Forms.” *The Prague Bulletin of Mathematical Linguistics*, 105 (2016): 143–193.

Zeman, D., Kosek P., Březina M., Pergler J., „Morphosyntactic Annotation in Universal Dependencies for Old Czech.” *Jazykovedný časopis*, 74/1 (2023): 214–222.