Теоријско-методолошки оквир за изградњу електронског корпуса српских средњовековних повеља и писама

Владимир Поломац

vladimir.polomac@gmail.com

v.polomac@filum.kg.ac.rs

ФИЛУМ

Претходна истраживања

– Пре-електронски корпус историјски корпус српског језика Ђорђа Костића (1957-1962)

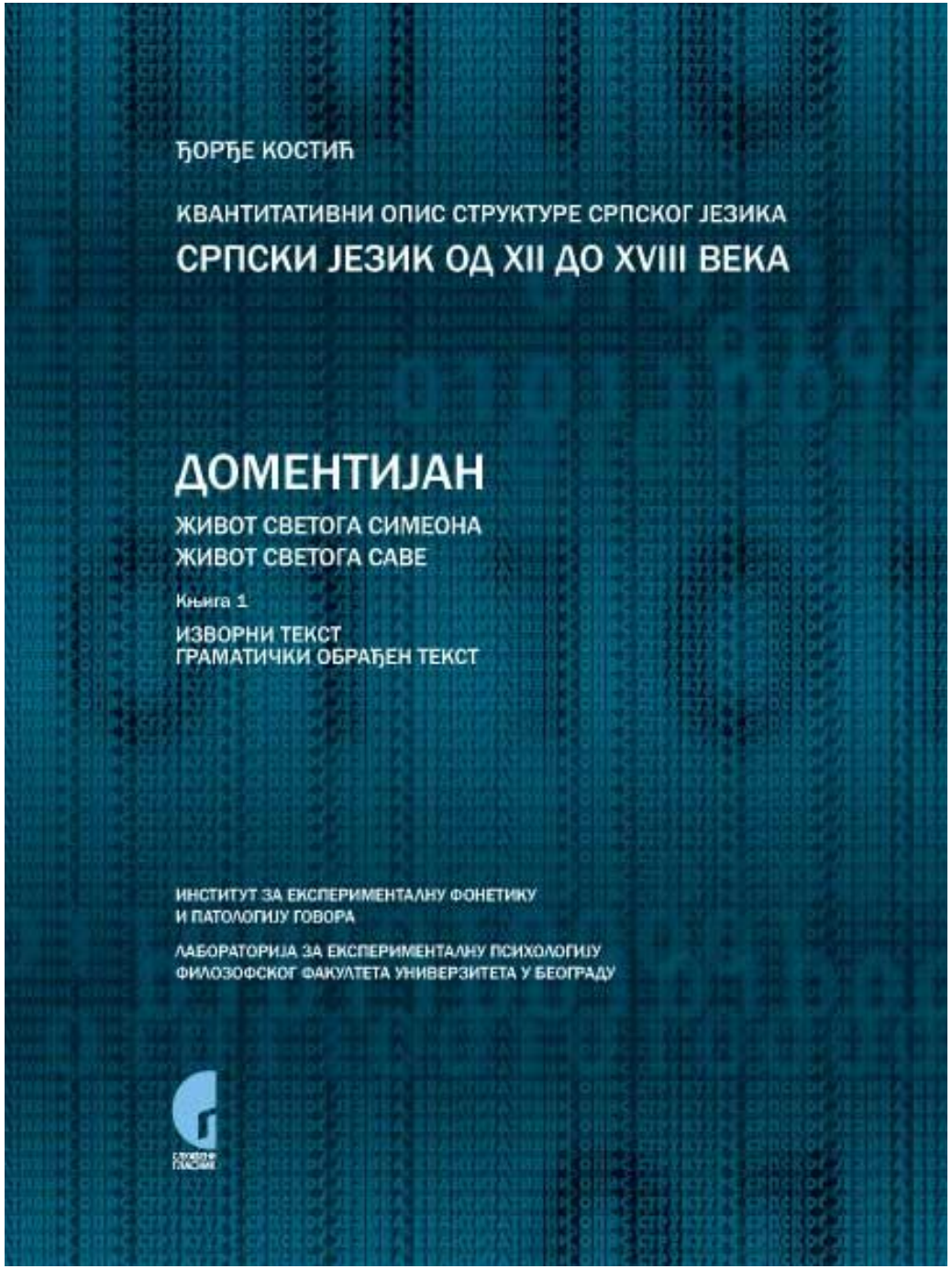
– Пројекат је обновљен деведесетих година XX века иницијативом његовог сина Александра Костића

– Осавремењен је систем граматичке анотације

– Објављено је 8 феквенцијских речника

– Корпус није јавно доступан

– Селекција текстова из корпуса не репрезентује српски језик XII до XIX века


ЂОРЂЕ КОСТИЋ
КВАНТИТАТИВНИ ОПИС СТРУКТУРЕ СРПСКОГ ЈЕЗИКА
СРПСКИ ЈЕЗИК ОД XII ДО XVIII ВЕКА
ДОМЕНТИЈАН
ЖИВОТ СВЕТОГА СИМЕОНА
ЖИВОТ СВЕТОГА САВЕ
Књига 1
ИЗВОРНИ ТЕКСТ
ГРАМАТИЧКИ ОБРАЂЕН ТЕКСТ
ИНСТИТУТ ЗА ЕКСПЕРИМЕНТАЛНУ ФОНЕТИКУ
И ПАТОЛОГИЈУ ГОВОРА
ЛАБОРАТОРИЈА ЗА ЕКСПЕРИМЕНТАЛНУ ПСИХОЛОГИЈУ
ФИЛОЗОФСКОГ ФАКУЛТЕТА УНИВЕРЗИТЕТА У БЕОГРАДУ


ЂОРЂЕ КОСТИЋ
КВАНТИТАТИВНИ ОПИС СТРУКТУРЕ СРПСКОГ ЈЕЗИКА
СРПСКИ ЈЕЗИК ОД XII ДО XVIII ВЕКА
ТЕОДОСИЈЕ ХИЛАНДАРАЦ
ЖИВОТ СВЕТОГА САВЕ
ИЗВОРНИ ТЕКСТ
ГРАМАТИЧКИ ОБРАЂЕН ТЕКСТ
ФРЕКВЕНЦИЈСКИ РЕЧНИК

Историјски електронски корпуси словенских језика

– Историјски поткорпус руског националног корпуса

https://ruscorpora.ru/search?search=CgQyAggPMAE%3D

– Око 14 милиона речи

– Староруски текстови

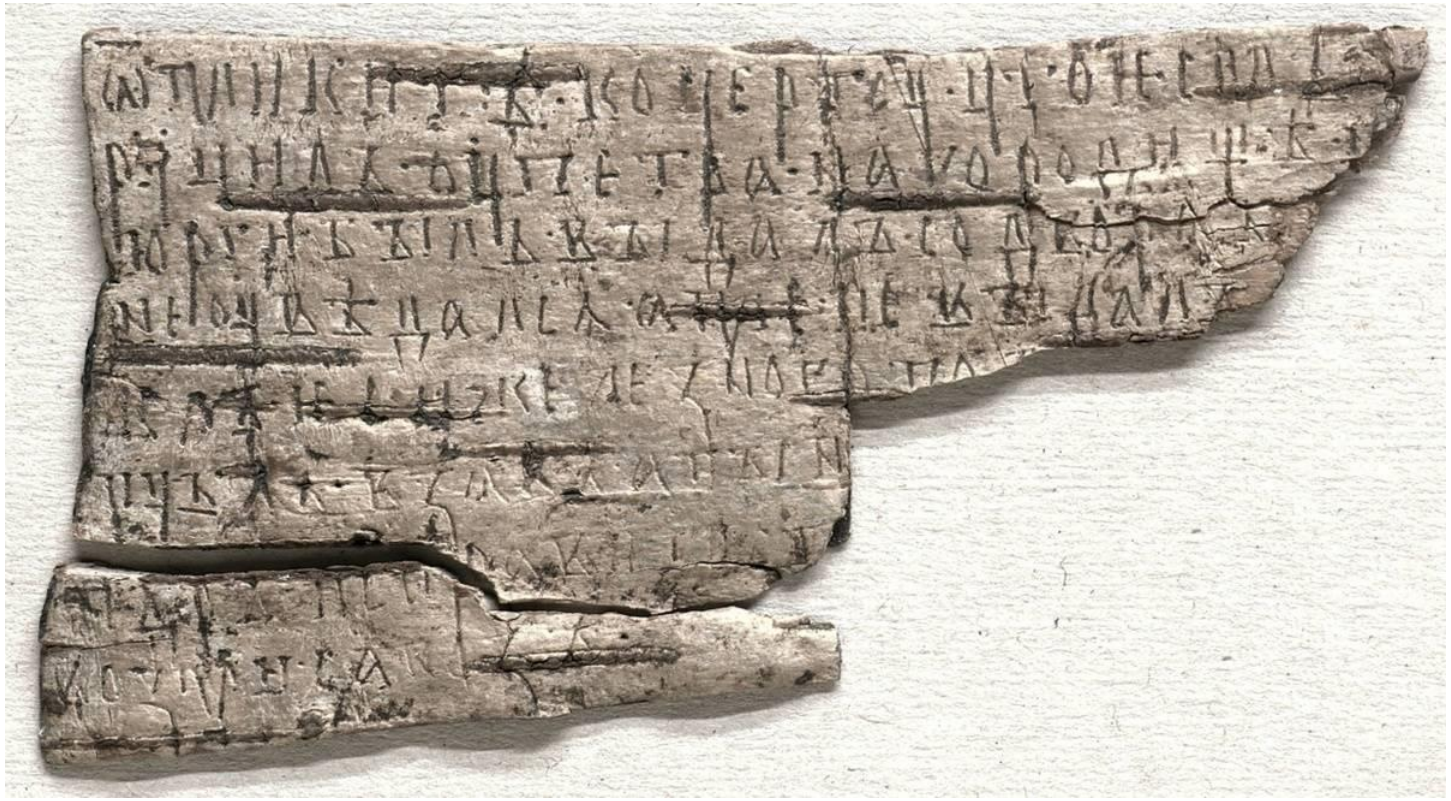
– Средњеруски текстови

– Повеље на брезовој кори

– Епиграфика

– Руски црквенословенски

– RRuDI



Историјски корпус доњолужичког језика

– DOTKO https://www.dolnoserbski.de/korpus/

Историјски корпус горњолужичког језика

– HOTKO https://wiki.korpus.cz/doku.php/en:cnk:hotko

Дијахрони поткорпус чешког националног корпуса
https://wiki.korpus.cz/doku.php/en:cnk:diakorp

Електронски корпус пољских текстова XVII и XVIII века

https://www.korba.edu.pl/query_corpus/

Дигитални корпус старопољских текстова (до 15. века)

https://ijp.pan.pl/en/publikacje-i-materialy/zasoby/

Историјски поткорпус словачког језика

https://korpus.juls.savba.sk/hks_en.html

Старословенски језик

– **PROIEL** – паралелни корпус превода Новог Завета у старим индоевропским језицима: http://foni.uio.no:3000/users/sign_in

– **TOROT** – https://nestor.uit.no – наставак PROIEL корпуса (старословенски и староруски текстови)

– **Корпус хрватске редакције старословенског језика (у припреми)**

– **CroDi Регенсбуршки дијахрони корпус хрватског језика**

https://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/netzwerke/regensburger-korpora/index.html

Историјски корпус словеначког језика

https://nl.ijs.si/imp/index-en.html#corpus

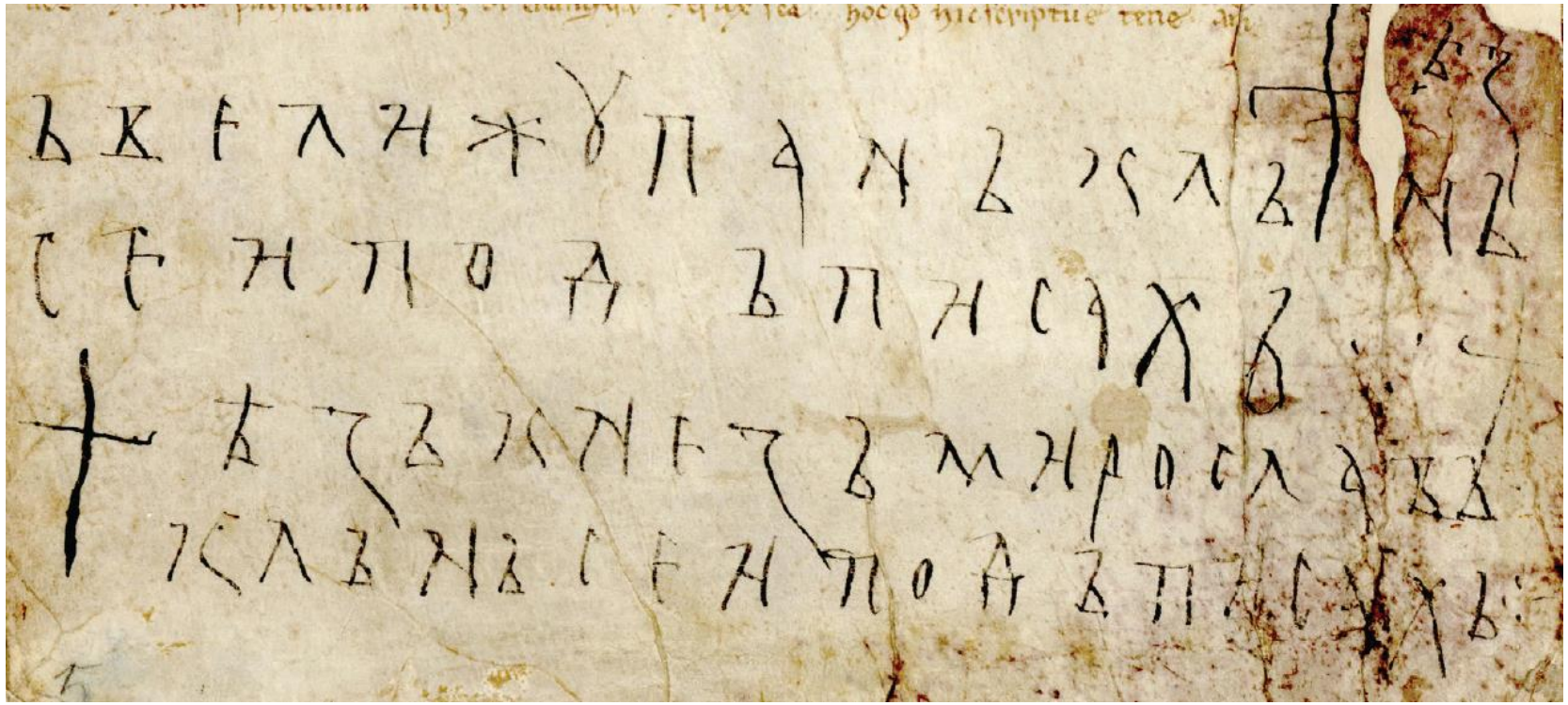
Дефинисање историјског корпуса српског језика

– Пројекат *Речник српског језика XII-XVIII века* Матице српске (руководилац академик Јасмина Грковић-Мејџор)

– Доња хронолошка граница: крај XII века

– Горња хронолошка граница: крај XVIII века

(грађа за Речник САНУ)



Дефинисање историјског корпуса српског језика

Критеријум језика и хомогена диглосија

– Старословенски језик

– Српскословенски језик

– Српски језик XII-XVIII века

– Штокавско наречје

Критеријум писма и вероисповести писара

– ћирилица, латиница; без обзира на вероисповест писара

Критеријум имена језика

Језичка ситуација након распада СФРЈ

Средњовековне повеље и писма као најзначајнији поткорпус

Хронолошки критеријум

– Српске средњовековне повеље и писма (XII-XVI век)

– Душанов законик (старији преписи из XIV и XV века)

– Рударски законик деспота Стефана Лазаревића

(препис из XVI века)

– Роман о Троји (XV век)

Обим и географски критеријум

– Око 2 000 хиљаде текстова из средњовековне Србије (Зета, Хум, Рашка), Босне и Дубровника и страних канцеларија (Турска, Угарска, Албанија, Молдавија, Бугарска)

Средњовековне повеље и писма као најзначајнији поткорпус

Средњовековне повеље

– Акта којима владар или властелин утврђује права или поседе онога у чију је корист повеља издата

– Документи са трајним, али не и општим важењем

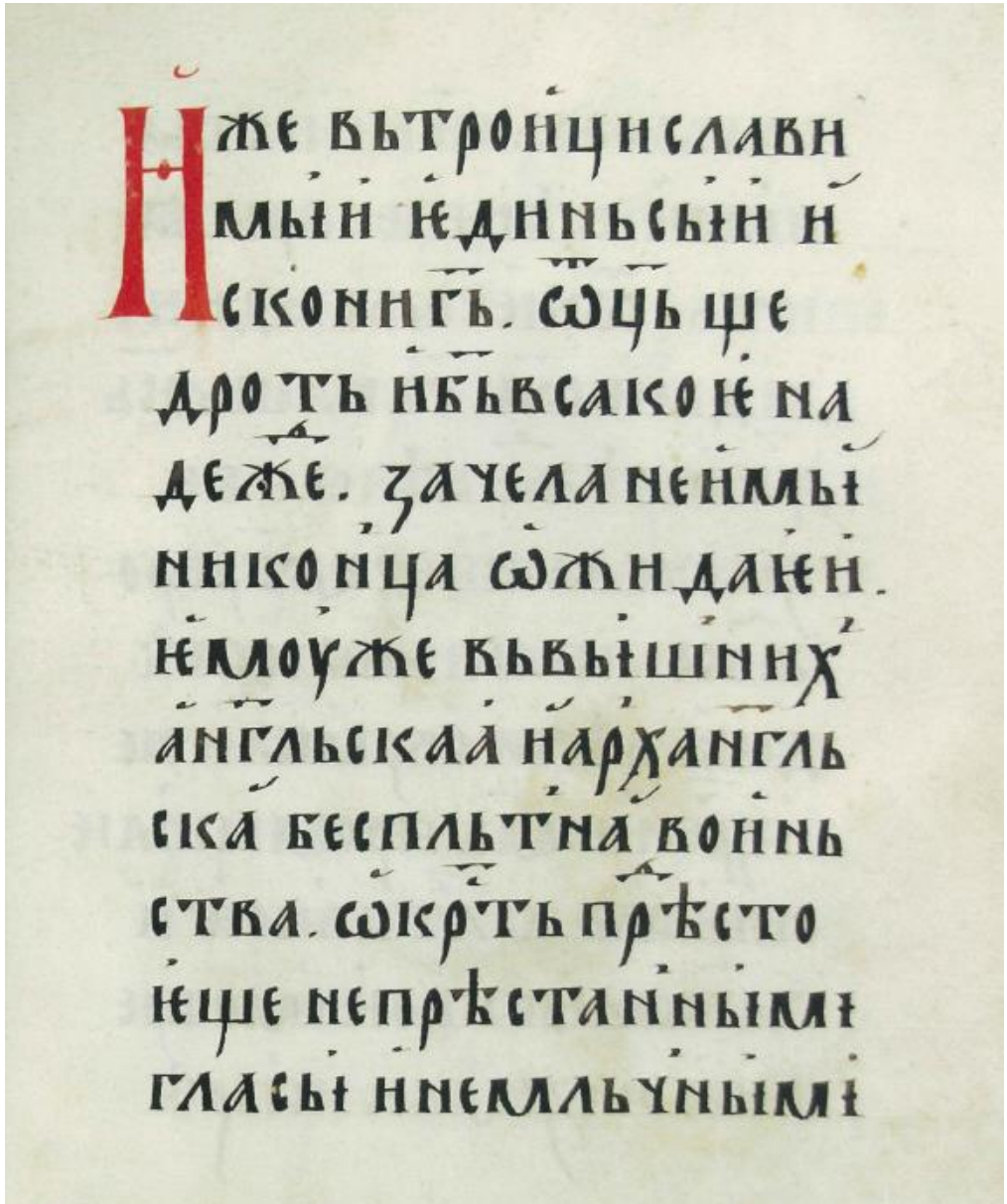
– Историјски архив у Дубровнику, Архив Хиландара, Архиви других светогорских манастира

– Ћирилица, српски (штокавски) језички израз, српскословенски

– Језички израз зависи од садржаја и намене

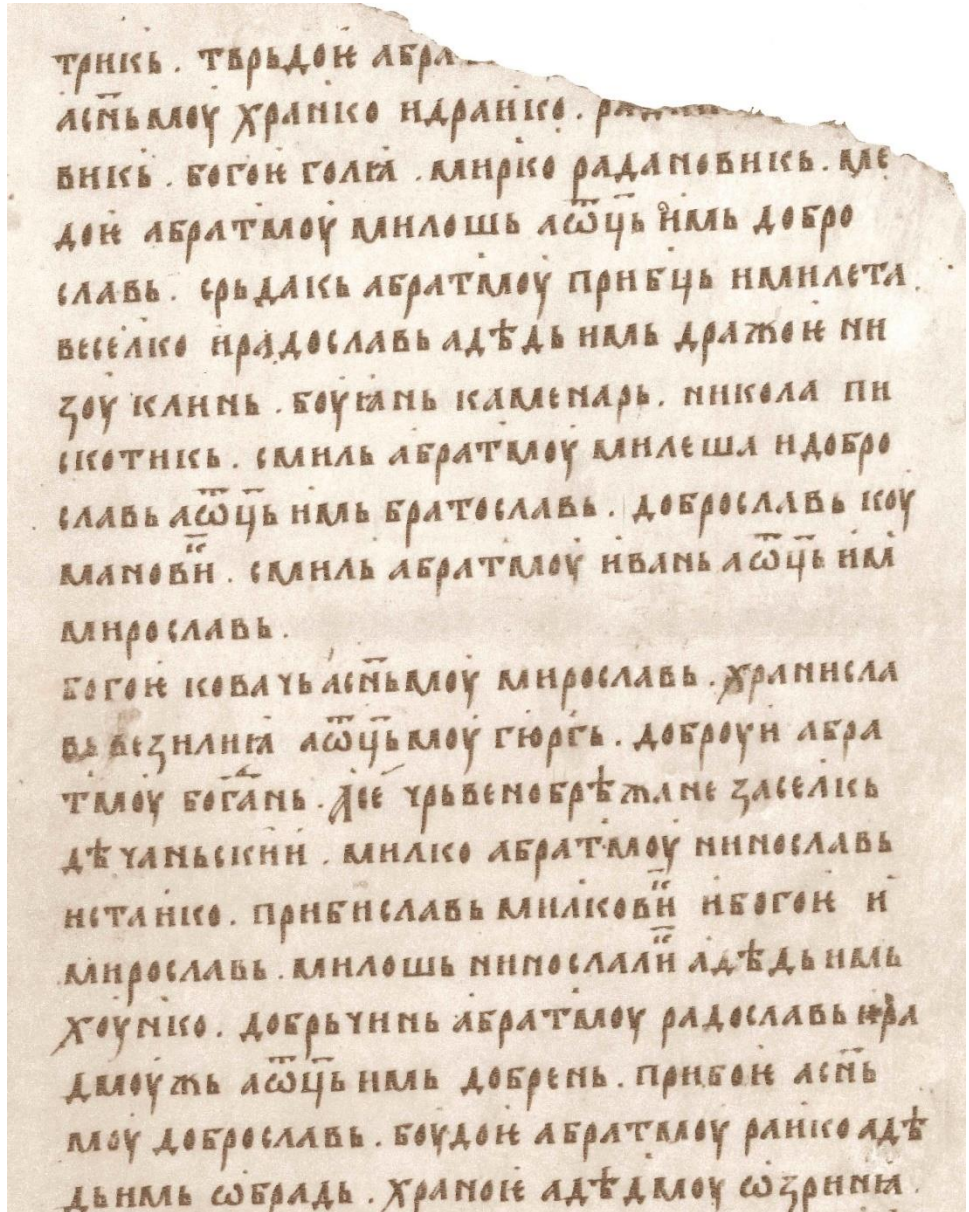
– Различитог обима и изгледа (од неколико редова до обимних повеља у облику рукописне књиге

Бањска хрисовуља



Иже вьтроицнославн
ымыи иединьсыи и
сконнгъ. ѡцъ ще
дро тъ ибьвсакоѥ на
деже. зачела неимы
никонца ѡжидаѥи.
ѥмоуже вьвышних
англьскаа нархангль
ска бесплътна воинь
ства. ѡкрътъ прѣсто
ѥще непрѣстаннымь
гласы инемльчнымь

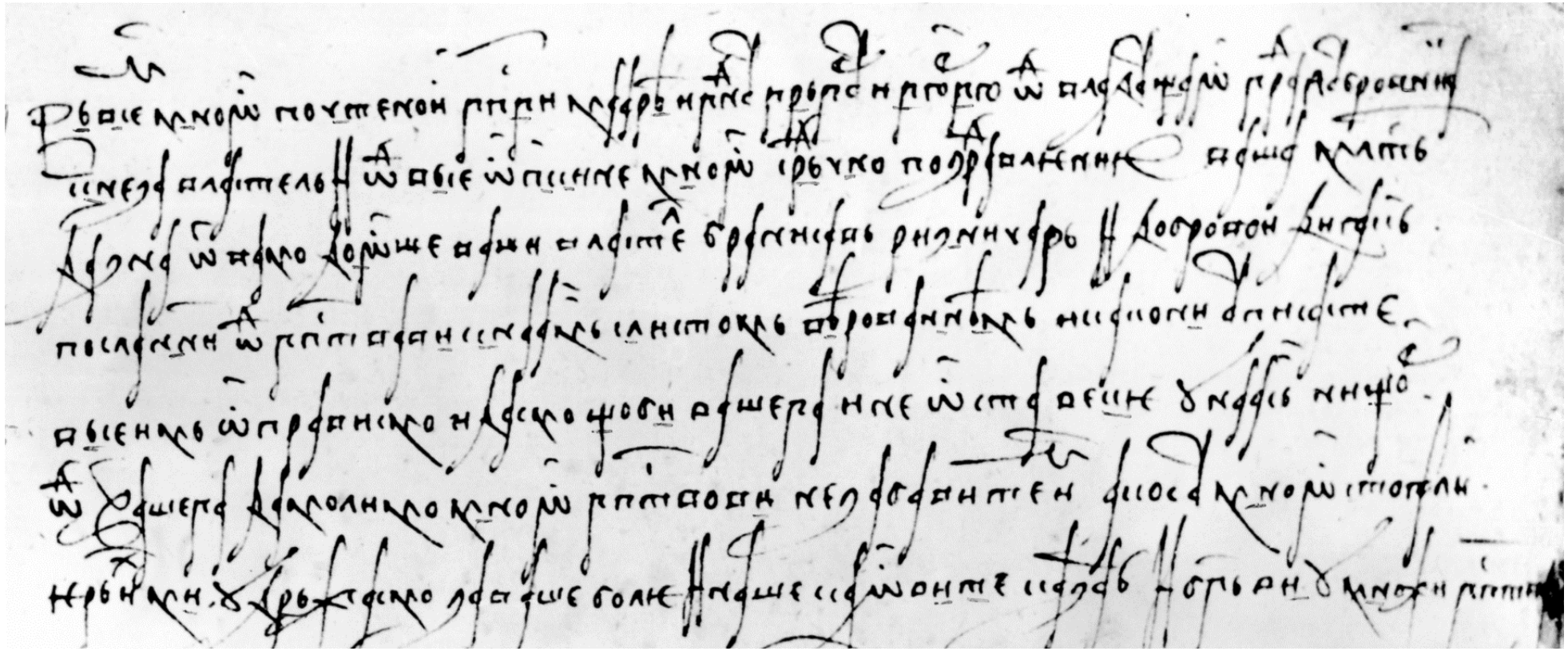
Дечанска хрисовуља



трнкь. тврьдок лбра
лсньмоу хранико ндранико. р
вникь. богок голѣи. мнрко радамовникь. кле
док лбратмоу милошь лѡцъ имь добро
слвь. срьдакь лбратмоу прибцъ имилета
весѣлко прадослвь адѣдь имь дражок ни
зоу клинь. боуиань камснарь. никола пи
скотникь. смиль лбратмоу милешл ндобро
слвь лѡцъ имь братослвь. доброслвь коу
мановн. смиль лбратмоу ивлнь лѡцъ им
мирослвь.
богок ковачь лсньмоу мирослвь. жранисла
въ безднна лѡцъмоу гюрѣь. доброун лбра
тмоу богань. дѣе чрьвеноврѣмане зделѣкь
дѣчаньскин. миліко лбратмоу нинославь
нстаніко. прибиславь миліковн ибогок и
мирослвь. милошь нинослалн ладѣдь имь
хоуніко. добрьчинь лбратмоу радослвь ирл
дмоужь лѡцъ имь добрень. прибок лсньмоу
доброслвь. боудок лбратмоу раніко ладѣ
дьимь ѡбрадь. храноіе ладѣдмоу ѡзрнна

Средњовековне повеље и писма као најзначајнији поткорпус

Средњовековна писма (краћи текст, 1 лист, брзописна ћирилица, пословноправна преписка са Дубровником)



Општи циљеви формирања корпуса

– Теоријско-методолошки оквир за израду референтног дијахроног корпуса српског језика

– Израда корпусно заснованог историјског речника српског језика

– Израда корпусно заснованог историјског речника старосрпских личних имена

– Израда корпусно засноване историјске граматике српског језика

– Корпусно засноване дијахроне студије српског језика

– Корпусно заснована историјско-компаративна проучавања словенских језика

Општа начела формирања корпуса и селекције текстова

Начело репрезентативности и балансираности корпуса

– Целина сачуване писане заоставштине

Начело поузданости

– Документа позната на основу оригинала рукописа, микрофилма или фотографије; документа позната само на основу издања?

Селекција текстова према дипломатичком критеријуму

– Оригинал, копија, препис, превод?

Селекција текстова према критеријуму језика?

Израда регистра текстова и дефинисање метаподатак

– Регистар текстова за *РСЈ XII-XVIII века* Матице српске

Пример:

Д XII 1 – *Потпис великог жупана Стефана Немање и кнеза Мирослава на њиховој латинској повељи упућеној Дубровнику* (27. IX 1186) • (Ђорђић 1990: 249 ⇒ ПП 1; оригинал Δ П. Ивић и Јерковић 1981: № 1)

– Регистар текстова за корпус:

Пример:

№1 – *Потпис великог жупана Стефана Немање и кнеза Мирослава на њиховој латинској повељи упућеној Дубровнику* (27. IX 1186)

Скраћеница: №2
Наслов: Повеља босанског бана Кулина Дубровнику
Ауктор: Бан Кулин
Дестинатар: Дубровник
Дипломатички статус: Оригинал
Општи тип: Повеља
Наменски тип: Уговор
Датум настанка: 29. VIII 1189.
Век настанка: XII
Место настанка: /
Регион: Босна
Канцеларија: Канцеларија првих банова
Писар: Радоје дијак
Издање: Мошин–Ћирковић–Синдик 2011: 49–52
Снимак: Ђорђић 1991: 65

Израда регистра текстова и дефинисање метаподатака

Дипломатички статус: оригинал, копија, препис, превод

Општи типови: повеља, писмо

Наменски типови докумената: даровнице, уговори, признанице, нотификације, пуномоћја, пресуде, пропуснице, налози, тестаменти

Регион: Зета, Рашка, Хум, Босна, Дубровник, Далмација; Албанија, Турска, Молдавија, Угарска, Бугарска

Канцеларије: Немањићи, Котроманићи, Косаче ...

Принципи преношења текстова у електронски облик

– Преносе се цели текстови, а не одломци

– Тежи се верности оригиналу, уз минималну ортографску нормализацију

– Фонт BukyVede усаглашен са Unicode стандардом

– Скраћенице и надредна слова се доносе према оригиналу

– Задржава се оригинална интерпункција (нормализује се употреба тачке на средини реда)

– Не преносе се савремена правила о писању великог слова

– Из текста су изостављени акценатски знаци

Пример едициоког поступка

† ѹ име wца и сн̃а : и ст̃ога дх̃а ѣ бань : бо|[2]сьньски кѹлинь : присезаю т҃ е҃бѣ кнеже |[3] крьвашѹ : и вьсѣмь граѣамь дѹбровьч|[4]амь : правы : приѣтель : быти вамь |[5] ѡ̂ь селѣ : и до вѣка : и правь гои дрьжати |[6] сь вами : и правѹ : вѣрѹ : до кола сьмь живь : вь|[7]си дѹбровьчане : кире ходе : по моемѹ владани|[8]ю : трьгѹюке : годѣ си кто : хоке : крѣвати : го|[9]дѣ си кто мине : правовь вѣровь и правымь : сr̃ь|[10]дьцемь : дрьжати є : безь вьсакое зледи : раз|[11]вѣ що ми кто : да с҃воевь воловь пок̂онь : и да имь |[12] не бѹде : ѡ̂ь моихь : чьстьниковь : силе : и до колѣ : |[13] ѹ мне бѹдѹ : дати имь : сьвѣть : и помокь какоре : и се|[14]бѣ коликоре моге : безь вьсега : зьлога примы|[15]сла : тако ми бж̃е помагаи : и сие ст̃о еваньгелие : |[16] ѣ радое : диѣкь бань : писахь сию : книгѹ : повеловь |[17] бановь : ѡ̂ь рожьс҃тва : хв̃а : тисѹка : и сьто : и wсм|[18]ьдесеть : и деветь : лѣть : мѣ̃сеца : авьгѹста : |[19] ѹ дьвадесети : и деветы : дн̃ь : ѹсѣчение : гла|[20]ве : иwвана : крьстителa ⁘

Начела токенизације

– Речи су одвојене белином

– Симболи су одвојени белином

– Знаци интерпункције су одвојени белином (тачка, две тачке ...)

– Текстови од средине XIV века познају и запету (одваја се белином од речи)

– Проблем вишечланих токена (нпр. *Ново Брдо, ко год* и сл.)?

Начела лематизације

Реконструише се лема према претпостављеном стању српског језика X/XI века

– Прасл. назални вокали су замењени српским гласовима /е/ и /о/

Пример: стсл. десѧть, стсрп. десет; стсл. рѫка, стсрп. рука

– Прасл. полугласник у слабом положају се изгубио

Пример: стсл. кьто, чьто, стсрп. кто, што

– Задржава се стсрп. јаки полугласник: дьн, Петьр, јесьм ...

– Задржава се стсрп. вокал „јат“: вѣтьр, вѣра, мѣсец

– Задржава се стсрп. вокал „јери“: быти, босаньскыи, ...

– Задржава се стсрп. вокално /л/: Влк, јаблка, стлп, ...

Проблем различитих хронолошких слојева приликом лематизације

Прва половина XIII века

быти > бити, босьнскыи > босански

XIII век

мѣсец > месец, мисец

XIV век

мѣсец > мјесец

Крај XIV века/почетак XV века

дьн > дан

Влк > Вук

Проблеми лематизације

Српскословенске и српске варијанте:

Српскословенски предлог и префикс **вь**, **вьз,** од XIV века **ва**, **ваз**

Српски предлог и префикс **у**

Српскословенски облици са **шт** и **жд** наспрам српских облика **ћ** и **ђ**: **ношт**, **межда**, **свъшта** према **ноћ**, **међа**, **свећа**

Српскословенски **вьс**, од XIV века **вас**, српски **сав**

Српскословенски **аз**, српски **јаз** или **ја**

Концепт „хиперлеме“!

Проблеми анотације

Опсег анотације

– Врсте речи (POS tagging)

– Морфосинтаксички опис (MSD tagging)

– Синтаксичке релације?

Избор сета ознака за анотацију

– Universal Dependencies или Multext East?

Прилагођавање сета ознака старосрпском/српскословенском језику

– Српски и/или старословенски сет ознака?

UD за старосрпски/српскословенски

Прилагођавање сета ознака за старословенски језик

Ознаке за врсте речи (PoS tags)

NOUN Именице
PROPN Властите именице
PRON Заменице
DET Детерминатори
ADJ Придеви
NUM Бројеви
VERB Глаголи
AUX Помоћни глаголи
ADV Прилози
ADP Предлози
INTJ Узвици
PART Речце
CCONJ Координирани везник
SCONJ Субоординирани везник
PUNCT Интерпункција
SYMB Симбол
X Остало

UD за старосрпски/српскословенски

Прилагођавање сета ознака за старословенски језик

https://universaldependencies.org/cu/index.html

– DET није део српске граматичке традиције?

– Користи се ознака PART (нема је у старословенском сету)

– Обележава се интерпункција PUNCT

– Ознака SYMB користи се за обележавање крста:

† ѣзь вели жѹпань клнь се и подьписахь

† ѣзь кнезь мирославь клнь се и подьписахь

† ѹ име wца и сн̃а : и ст̃ога дх̃а

– Ознака X користи се за ретке нејасне примере (најчешће писарске грешке)

UD за старосрпски/српскословенски

Прилагођавање сета ознака за старословенски језик

https://universaldependencies.org/cu/index.html

– DET није део српске граматичке традиције?

– Користи се ознака PART (нема је у старословенском сету)

– Обележава се интерпункција PUNCT

– Ознака SYMB користи се за обележавање крста:

† ѣзь вели жѹпань клнь се и подьписахь

† ѣзь кнезь мирославь клнь се и подьписахь

† ѹ име wца и сн̃а : и ст̃ога дх̃а

– Ознака X користи се за ретке нејасне примере (најчешће писарске грешке)

UD за старосрпски/српскословенски

Различити приступи у означавања морфолошких категорија у старословенском и српском сету ознака

- Аниматност?
- Неодређени и одређени придевски вид?
- Аорист и имперфекат?
- Сложени глаголски облици?
- Партиципи?

Објављивање корпуса?

SketchEngine?

NoSketchEngine?

Посебна веб апликација?

Тренутно стање и планови

– Припрема модела *Logothet* за аутоматско рашчитавање повеља и писама помоћу софтверске платформе *Transkribus*

– Дефинисање UD тагсета и припрема сета за тренирање модела за аутоматску лематизацију и анотацију помоћу алата *Stanzatagger* или помоћу алата *UDPipe* (материјал чине ручно лематизирани и анотирани текстови XII и XIII века)

– Успостављање процеса рада у припреми корпусних фајлова

– Пробна верзија корпуса (са текстовима XII и XIII века) би требало да буде објављена до краја 2024. године, а презентована на МКС-у у Паризу 2025. године

Хвала на пажњи!

Питања, коментари, сугестије?

Могућности за сарадњу?

vladimir.polomac@gmail.com