

Projekt MONOGRAF: autocestom do opsežnoga povijesnog elektroničkog korpusa hrvatskoga jezika

Današnje jezikoslovlje ne može bez korpusâ, koje je jednostavno i ponešto neprecizno, ali za naše svrhe dovoljno dobro opisati kao elektroničke skupove tekstova. Hrvatska povijesna tropismenost (latinica, glagoljica i ćirilica) i različite grafije znatno otežavaju stvaranje povijesnih elektroničkih korpusa hrvatskoga jezika (pustit ćemo da dvije priložene slike govore o složenosti nekadašnje latinične grafije), pri čemu je grafija izbor slova ili skupova slova nekoga jezika uređen prema potrebama toga jezika. Stoga se u Institutu za hrvatski jezik (IHJ) uz potporu Europske unije provodi projekt *Razvoj i primjena modela za normalizaciju grafije starih latiničnih tiskanih tekstova* (MONOGRAF; dalje ćemo u tekstu projekt zvati Monograf).¹ Projektom se izgrađuje model kojim će se stariji latinični tiskani tekstovi povezati u pouzdan i javno dostupan elektronički korpus, što je jedan od temeljnih ciljeva projekta. Latinica je odabrana pred ostalim pismima jer je u povijesti bila složena i izazovna, a danas su tekstovi starih latiničnih knjiga lako dostupni u velikim količinama.

Projekt se provodi od 1. siječnja 2024. do 31. prosinca 2027. godine, uz potporu instrumenta NextGenerationEU u okviru Nacionalnoga plana oporavka i otpornosti. Voditelj je projekta Vuk-Tadija Barbarić, a suradnica Marijana Horvat. Monograf se nadovezuje na prethodni institutski projekt koji je trajao od 2021. do 2023. godine i postavio temelje za današnji.

Čemu normalizacija grafije?

Grafija u počecima tiskarstva nije bila stabilna i općeprihvaćena kao danas. Čitanje starih tiskanih knjiga današnjim bi prosječnim čitateljima bio izazovan zadatak, i to bez obzira na to je li riječ o čitanju latinice, glagoljice ili ćirilice. Normalizirati u kontekstu projekta znači ujednačiti, standardizirati i prilagoditi pisani oblik teksta određenoj normi,

¹ Naslov i dio sadržaja članka oslanja se na predavanje što sam ga održao zadarskim studentima kroatistike 19. ožujka 2025. Posebno zahvaljujem Gordani Čupković što mi je to omogućila u okviru projekta *Ćirilični i glagoljični apokrifni molitvenici iz 17. i 18. st.*, koji financira Zaklada Adris. Suradnja Monografa s tim projektom bila je moguća zato što je jedna od knjiga predviđenih za obradu bila latinična knjiga Josipa Banovca *Blagosov od polja* (1767.), s kojom se sadržajno uspoređuju ćirilični i glagoljični molitvenici.

Ovaj rad nastao je u okviru projekta *Razvoj i primjena modela za normalizaciju grafije starih latiničnih tiskanih tekstova* (MONOGRAF), koji u Institutu za hrvatski jezik financira Europska unija – NextGenerationEU. Izneseni stavovi i mišljenja pripadaju samo autoru i ne odražavaju nužno službena stajališta Europske unije ili Europske komisije. Ni Europska unija ni Europska komisija ne mogu se smatrati odgovornima za njih.

a u našem je slučaju riječ o prilagodbi današnjemu pravopisu, prije svega današnjemu standardnom latiničnom slovopisu. Takav pristup omogućuje čitateljima lakši pristup tekstu, a jezikoslovcima jednostavnije provođenje računalnih analiza. Dakle, u projektu *Monograf model* označava sustav ili skup pravila razvijenih za normalizaciju grafije starih latiničnih tiskanih tekstova, što podrazumijeva prilagodbu pisanoga oblika tih tekstova suvremenim pravopisnim normama, čime se olakšava njihova čitljivost, analiza i digitalna obrada.

Stari problemi

Dosad nismo rekli ništa pretjerano novo. Cijela hrvatska dosadašnja djelatnost filološkoga (tekstološkoga) objavljivanja starih rukopisnih i tiskanih djela počiva upravo na takvoj normalizaciji, koja se uglavnom dijeli na dvije različite prakse, transkribiranja i transliteriranja, ali ovdje nema potrebe ulaziti u detaljno objašnjenje postupaka. I jedan i drugi postupak ponovnim objavljivanjem čuvaju stari tekst u novome obliku. Danas se u Hrvatskoj takvi tekstovi objavljuju gotovo isključivo latinicom, dok je u povijesti bilo drukčijih praksa (npr. ono što je u staro doba bilo zapisano glagoljicom ili ćirilicom, u nova se izdanja prenosilo istim tim pismima). Time je otklonjena jedna bitna prepreka uspostavi povijesnoga korpusa hrvatskoga jezika. Ako je sve objavljeno latinicom, zašto jednostavno ne skupimo sve te tekstove na jedno mrežno mjesto, omogućimo njihovo sustavno pretraživanje i proglasimo to korpusom? Rješenje se čini jednostavnim, ali u praksi je s provedbom takve zamisli povezano mnogo problema. Prije svega, ti tekstovi pripadaju različitim autorima i izdavačima. Pod autorima ovdje zapravo mislimo na priređivače starih djela, filologe, koji polažu prava na svoja djela, a ne na izvorne autore tekstova, koji imaju danas pravo da ih se ne prešućuje, no ne puno više od toga jer su živjeli prije više stotina godina. Drugi od niza problema koji bismo istaknuli jesu zahtjevi struke, tj. korpusnoga jezikoslovlja. Da bi nešto u lingvistici bilo prihvatljivo kao korpus, tekstovi prikupljeni u njemu moraju načelno biti međusobno usporedivi. Tu bismo npr. odmah naišli na problem usporedivosti dvaju tekstova od kojih je jedan priređen transkribiranjem, a drugi transliteriranjem. Štoviše, ako se dva djela koja su priređena transkribiranjem razlikuju u nekim načelima po kojima su priređena, opet se postavlja nezgodno pitanje njihove usporedivosti. Takvi tekstovi uvijek mogu činiti elektroničku zbirku tekstova, ali u smislu korpusnoga jezikoslovlja oni stvaraju određene tehničke probleme i mogu se smatrati samo korpusom u širem smislu. Ako želimo uspostaviti korpus u užemu smislu, treba znati da jezikoslovci, ako postoji potreba, dijele korpusne na manje cjeline s obzirom na kombinaciju jezikoslovnih kriterija i istraživačkih ciljeva (registri, funkcionalni stilovi; korpus dječjega govora, korpus tekstova određene struke i sl.).

Što je onda novo u Monografu?

Ako smo vas naveli da mislite da je cilj Monografa s pomoću normalizacije jednoliko prirediti mnoštvo tekstova prema suvremenoj pravopisnoj normi za potrebe stvaranja

povijesnoga korpusa hrvatskoga jezika, trebate znati da to nije posve točno (ubrzanje i unapređenje tekstološke produkcije IHJ-a samo je dodatni cilj projekta i neovisan je o stvaranju korpusa). Monograf se u tome slučaju ne bi ni po čemu bitnome razlikovao od prethodnih pokušaja uspostave takvoga korpusa. Građa za predviđeni korpus bit će izvornim latiničnim grafijama vjerno prepisani tekstovi, koji će uključivati čak i izvorne pogreške. Što se tiče samoga priređivanja tekstova, normalizacija će biti u njemu provedena samo na najtrivijalnijoj razini – svi će tekstovi biti prepisani istim fontom. O vrsti pogrešaka koje će se tolerirati u samome korpusu odlučit će se nakon istraživanja u završnoj fazi projekta. No, već se i prije kraja projekta i uspostave korpusa digitalizirani tekstovi sa svim svojim pogreškama dijele putem mrežne stranice projekta <https://monograf.jezik.hr/> i putem akademskih mreža. Jedan specifičan način dijeljenja teksta u izvornoj grafiji može se naći na stranici <https://marulic.hr/>, na kojoj je *Judita* Marka Marulića objavljena na tri načina – izvornim slovopisom, transkripcijom i kao suvremena prilagodba.

Bitna novost sastoji se u tome što će se na Monografu razviti posebna mogućnost za pretraživanje tekstova u korpusu. Korisnik će upisivati svoj „normalizirani” upit u polje za pretragu, a kao rezultat dobit će sve ono što je razvijeni model za pretragu prepoznao kao prikladne parove u izvornim grafijama svih pojedinačnih tekstova koji se nalaze u korpusu. Pojednostavnjeno, ako u polje za pretraživanje upišete riječ *ljubav*, kao rezultat pretraživanja dobit ćete sve moguće načine na koje je ta riječ zabilježena u korpusu, npr. *gliubav*, *gliubau*, *glubav*, *lyubav*..., pri čemu ćete zapravo u svojim pretragama dobivati cijele isječke iz tekstova u kojima se nalazi tražena riječ. Međutim, ne možete očekivati da ćete za upit *riječ* kao rezultat dobiti *rič* ili *reč*. Očekivanja morate uskladiti s osnovnim znanjima o raznolikosti naših povijesnih tekstova, osobito dijalektnoj. Dakle, korpus neće biti normaliziran prema današnjoj pravopisnoj normi, nego je upravo korisnik taj koji svojim upitom unosi normalizirani element, pri čemu se onda taj element raslojava na sve u korpusu zastupljene izvorne latinične grafije. Može se reći da će se korpus normalizirati na zahtjev korisnika, ali samo u postupku dolaska do rezultata, skriveno od očiju korisnika. No, pustimo da sljedeća slika izgovori svojih 1000 riječi.

Koristi su od toga višestruke. Jezikoslovci će imati korpus usporedivih povijesnih tekstova uz zgodan alat koji će im omogućiti da ne moraju biti stručnjaci za svaki pojedinačni tekst koji analiziraju. Filolozi će općenito lako moći dolaziti do informacija koje traže u korpusu (ili će se, pak, lako moći uvjeriti da ih ondje nema), a i svi zainteresirani za povijest hrvatskoga jezika moći će lakše pristupati naizgled teškim tekstovima. Dok se za potrebe funkcioniranja modula za pretraživanja korpusa može uvjetno reći da normalizacija ide obrnutim putom, za potrebe ubrzavanja tekstološke djelatnosti normalizacija doista ide očekivanim putom, tj. u smjeru nastanka uz pomoć računala normaliziranih povijesnih tekstova prema utvrđenim pravilima i ograničenjima za raspodjelu slova. Zapravo, to se u IHJ-u već provodi.



Grafematička korpusna pretraga

Unesi riječ ili dio riječi standardnim pravopisom (npr. življenje, kršćanin) kako bi pretražio/la sve moguće grafematske varijante u korpusu.

Upit:

gubljenja



Rezultati (KWIC):

Ukupno: 13

- Nedigliu Septuagelima. Od izghubgliegnea Dufcaa. nà liftu
- nebi poznao fvoga izghubgliegnea, çiniga izghubit ftrah,
- bialce chier od izghubgliegnea s' jednim olcczem.
- put od tvoga izghubgliegnea, s' koim igelc,
- krixem od Adamova izghubgliegnea odkupiti, i toje
- začina jednoga gniova izghubgliegnea, veli S. Agufctin.
- S. Andriju. OD IZGHUBGLIEGNIA VRIMENA. Onni udigl
- ù mifto od izghubgliegnea, i zàtochiefc bitti
- Ovochie fctetu od izghubgliegnea Razum ofugenoga promifcgliati
- svoju srichiu, izvan izghubgliegnea gniove Dufce, jefu
- Nedigliu Septvagexime. OD IZGHUBGLIEGNIA DUSCAA. Mnogifu zvani,
- Boxju? Zlamegnije gniova izghubgliegnea, veli Ifukarf. Propterea
- slidech dillo tvoga izghubgliegnea? Znaj dobro, dà

1. slika: Prvi rezultati rada na modelu za pretraživanje projektnoga korpusa; za pokusni tekst uzeli smo jednu knjigu Josipa Banovca

Kako učiniti da algoritam za normalizaciju grafije proradi?

Sâm algoritam koji će računalno omogućiti pretraživanje kakvo smo gore opisali nije toliko teško zamisliti. Problem je što ostvarivanju njegova uspješna funkcioniranja mora prethoditi mukotrpan posao digitaliziranja starih latiničnih knjiga. Kao da već to po sebi nije dovoljno, potrebno je i analizirati grafijski sustav svake pojedinačne, i to cijele knjige (kako bi se postigla što veća preciznost modela), pa izvedena pravila i ograničenja u raspodjeli slova (grafematička analiza) prevesti u računalno uporabljiv oblik. Za razvoj

modela predviđena je digitalizacija desetak knjiga. Knjige se digitaliziraju tehnologijom optičkoga prepoznavanja znakova (engl. OCR, *optical character recognition*) i analiziraju prema već ustaljenom postupku (dosad opisanom u nekoliko publikacija u vezi s projektom):

- provjera teksta digitaliziranoga s pomoću optičkoga prepoznavanja znakova prema izvorniku (tj. provjera prema cjelovitome izvorniku)
- grafematička analiza s pomoću posebnoga jezikoslovnog računalnog programa, pri kojoj se otklanjaju naknadno uočene pogreške (selektivna provjera prema izvorniku)
- provjera popisa riječi koje se samo jedanput pojavljuju u tekstu, a polazi od pretpostavke da će pogrešno strojno čitanje određene riječi najčešće proizvesti jedinstvenu („nepostojeću”) riječ (selektivna provjera prema izvorniku)
- provjera prema kritičkome izdanju teksta (ako takvo postoji).

Zašto se vozimo autocestom, a ne letimo još puno brže zrakom?

Za razumijevanje Monografa nužno je ući u motivaciju za njegov nastanak. Oba suradnika na Monografu imaju ukupno znatno iskustvo u „klasičnome” priređivanju starih hrvatskih rukopisnih i tiskanih tekstova, i to ne samo latiničnih. Odlika takvoga pristupa priređivanju starih tekstova za objavljivanje jest velik utrošak vremena za relativno malo koristi, a kao što smo prije dali naslutiti, izgledi da će tako priređeno djelo postati dijelom kojega povijesnog elektroničkog korpusa hrvatskoga jezika nisu dobri. Prema tomu, jezikoslovna zajednica (hrvatska i međunarodna) nema u ovome trenutku od toga koristi koliko bi prema našem mišljenju trebala i morala imati. Klasični je pristup dakle ekvivalentan ručnoj zaprezi. Zamišljali smo da će nam, što se brzine tiče, Monograf donijeti let zrakoplovom do cilja, ali u međuvremenu je na velika vrata došla umjetna inteligencija, čije granice još nisu dovoljno ispitane na projektu, pa se sada i svako stajanje na odmoristima na autocesti doživljava s previše nestrpljenja. Ipak, moramo biti zadovoljni jer je činjenica da je vožnja autocestom dovoljno velik napredak u odnosu na vučenje ručne zaprege. Nekada smo mogli samo sanjati da ćemo na zaslonu računala moći gledati i analizirati slike kao što je sljedeća.

iaa kako poghani nyednih terpliaunih praud nyemaiuchi daleko od Prauicze blude,i lištör takoue
 od Prauicze blude,i lištör takoue praude chinee, koieße nih poghlaiui vyde. Zatoßu vßaki narod
 i ponomonogha, i Wgherßki Krali i vßa ob china teh orßaghau koteri pod Wgherßku Corunu Bliffe
 rßku Corunu Bliffe neke napraue ktem praudam prinachinili i naprauili, i toßeie uidelo liucztuu i obchin
 achinili i naprauili, i toßeie uidelo liucztuu i obchine haßnouito i praudenno biti. Zaato ghdabih iaz
 iofche i poghani vßake fele knighe imaiu domaa chim Buoym iezikom popyßane. Po opomeneny nekih dobri
 od Bofiegha poroda 1514 leto zeuße ghoßpode i obchine volium poterdil) na Blouienßki iezik kolikoßamg
 Blouenßki te knighe chtali. Koiegha iaaz hoteß vchiniti ißampati Vaffegha Ghoßpocztua STAMPAREM koieg
 i ghoßpoczke nature noßiti paßku na odychenie riechi Bofie, i na obrambu od niepriatelieu Buoie
 a obrambu od niepriatelieu Buoie domouine: kotero chiny vaffe ghoßpocztuo ghledaiuchi i toßu dobra perua
 domouine: kotero chiny vaffe ghoßpocztuo ghledaiuchi i toßu dobra perua kerßtianßka ghoßpoda chinyla. A
 ghledaiuchi i toßu dobra perua kerßtianßka ghoßpoda chinyla. A akobi i na nikogha vaffe ghoßpocztuo
 vmre. I zaato kakoße branienie i Bofie ryechi j orßagha, ghoßpode doßtoi, Takoei potreбно nim
 praude znati da prauo znaiu pod Bobum budu chim praude chiniti: kryue kaftyghati, a nekryue bran
 da prauo znaiu pod Bobum buduchim praude chiniti: kryue kaftyghati, a nekryue braniti. Mozibit
 kaafu modus i fundamentum kakoße imaiu praude chiniti. Zaato koi razume fundamentum laghlie i drugih
 truud zmoium Blusbum zadobro pryne. Bogh Wßemoghu chi na vnogha lieta v dobrom zdraviu i
 lieta v dobrom zdraviu i v dobroj Brechi zaußem Buoym draghim ghoßpochkim rodod V. Ghoßpoc
 ghoßpochkim rodod V. Ghoßpocztuo zdersy. V Nedeliß chiu v petak po Lourencheue ghdabiße pyßalo od
 urencheue ghdabiße pyßalo od naroda nefleghe zueli chitela iefulla M.D.LXXVIII leto. V: G:
 daie na ghriehe vun fentenciu. A po milofchi mertuchliui i theßi kaftyghu ghriehe: da raunim
 po prauici pokafygha. A neka da po milofchi proßty, koteriße vßaghdar ochima Buoima Bofiegha
 neka da po milofchi proßty, koteriße vßaghdar ochima Buoima Bofiegha opytania boiechi i Btrahuiuchi
 riße vßaghdar ochima Buoima Bofiegha opytania boiechi i Btrahuiuchi i vßakomße poßlu boy, da

2. slika. Rezultat jedne pretrage djela *Decretum* (1574.) Ivana Pergoßića u izvornoj grafiji.

Umjesto zaključka: Čemu dijeljenje resursa?

Stvaranje nacionalnih reprezentativnih korpusa svakako je velik događaj za jezikoslovnu zajednicu određene države, i ne samo jezikoslovnu. Ali moramo se pomiriti s činjenicom da korpusi dolaze i odlaze. Njihova prisutnost na internetu ovisi o činjenicama koji nisu trajni – prije svega o financiranju, a dostupnost im varira. Korpus koji će nastati na Monografu sigurno će biti godinama dostupan na internetu, ali tko može obećati cijelo stoljeće? I za sto godina, nadamo se, Hrvate će zanimati kako se i što pisalo u stoljećima neposredno nakon izuma tiska. Sudeći prema navikama današnjega svijeta, izgledno je da će im informacije trebati jako brzo, puno brže nego nama kako bi zaokupile njihovu pozornost. Zato bez gotovo ikakvih ograničenja sve svoje resurse (tekstove) dijelimo i dijelit ćemo na još mnogo mrežnih mjesta (digitalnih arhiva) kako bismo osigurali da se građa koju smo prikupili uvijek može ponovno objediniti i upotrebljavati na korist naše jezične i kulturne zajednice.