Josip Mihaljević
Institut za hrvatski jezik i jezikoslovlje
Okrugli stol – Terminološki izazovi u 21. stoljeću
Zagreb, 29. listopad 2020.

Korpus *Jene*

*Ova je prezentacija izrađena u okviru projekata *Hrvatsko jezikoslovno nazivlje – Jena* (Struna-2017-09-05) i *Hrvatski mrežni rječnik – Mrežnik* (IP-2016-06-2141), koje u cijelosti financira Hrvatska zaklada za znanost.

Korpus

- Korpus je zbirka tekstova prirodnoga jezika sastavljena po određenome kriteriju, skup jezičnih odsječaka (tekstova) koji su odabrani i skupljeni prema jasnim kriterijima radi dobivanja određenoga jezičnog uzorka.
- za hrvatski jezik ne postoji specijalizirani korpus jezikoslovnoga nazivlja
- većina se korpusnih istraživanja hrvatskoga jezika provodi na dvama općim hrvatskim korpusima:
  - *Hrvatskoj jezičnoj riznici*
  - *Hrvatskome jezičnom korpusu (hrWaC-u)*
- ti korpusi, međutim, obuhvaćaju malo tekstova koji pripadaju znanstvenomu stilu, pa stoga nisu prikladni za terminološka istraživanja

Korpus *Jene*

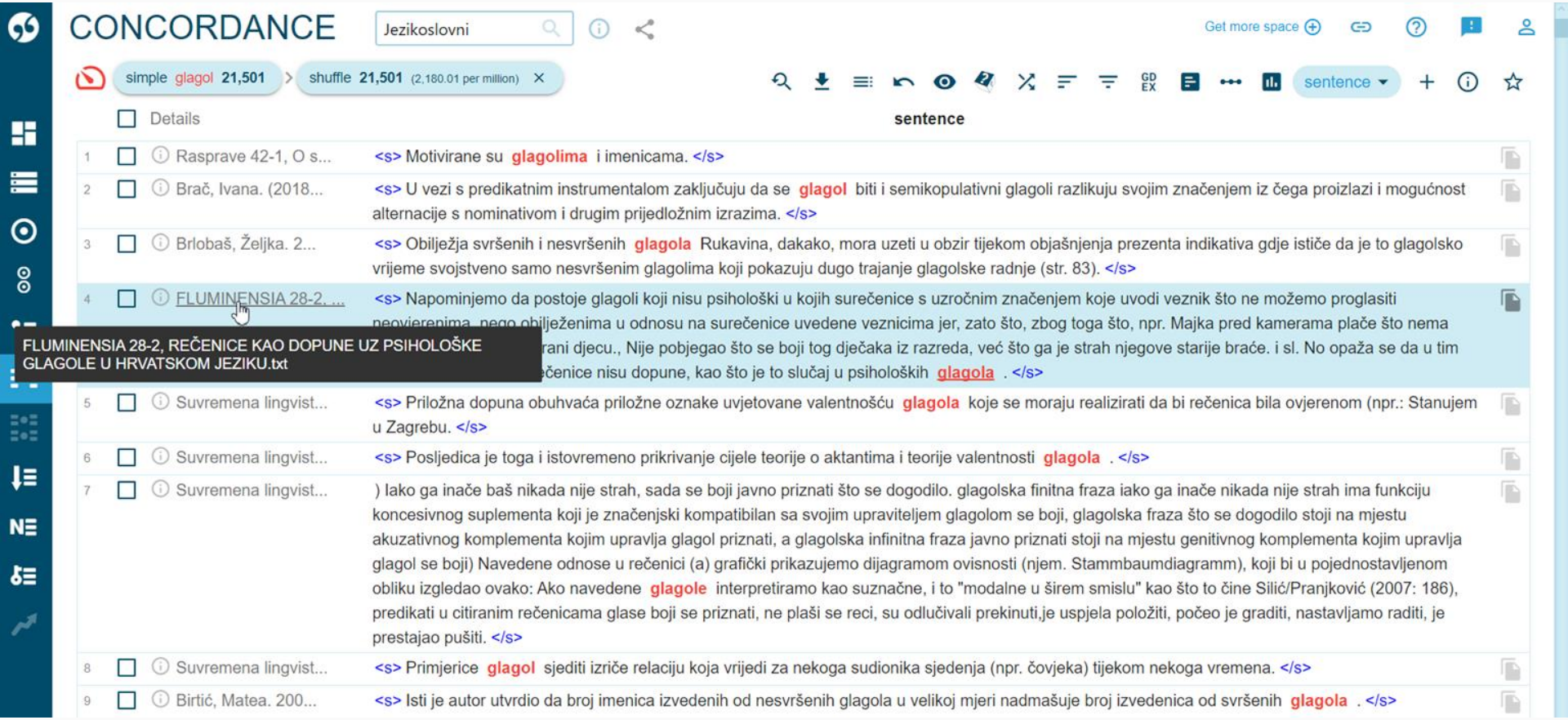
- U okviru projekta *Hrvatsko jezikoslovno nazivlje – Jena* izgrađuje se i specijalizirani jezikoslovni korpus.
- taj je korpus
  - tekstni
  - jednojezični (hrvatski jezik)
  - posebni (obuhvaća tekstove iz jezikoslovnih članaka i monografija)
  - sinkronijski
  - sastoji se od tekstova čiji su autori izvorni govornici hrvatskoga jezika
- korpus je sastavljen s pomoću programa Sketch Engine te pristup korpusu mogu zatražiti osobe koje imaju AAI@EduHr korisnički račun

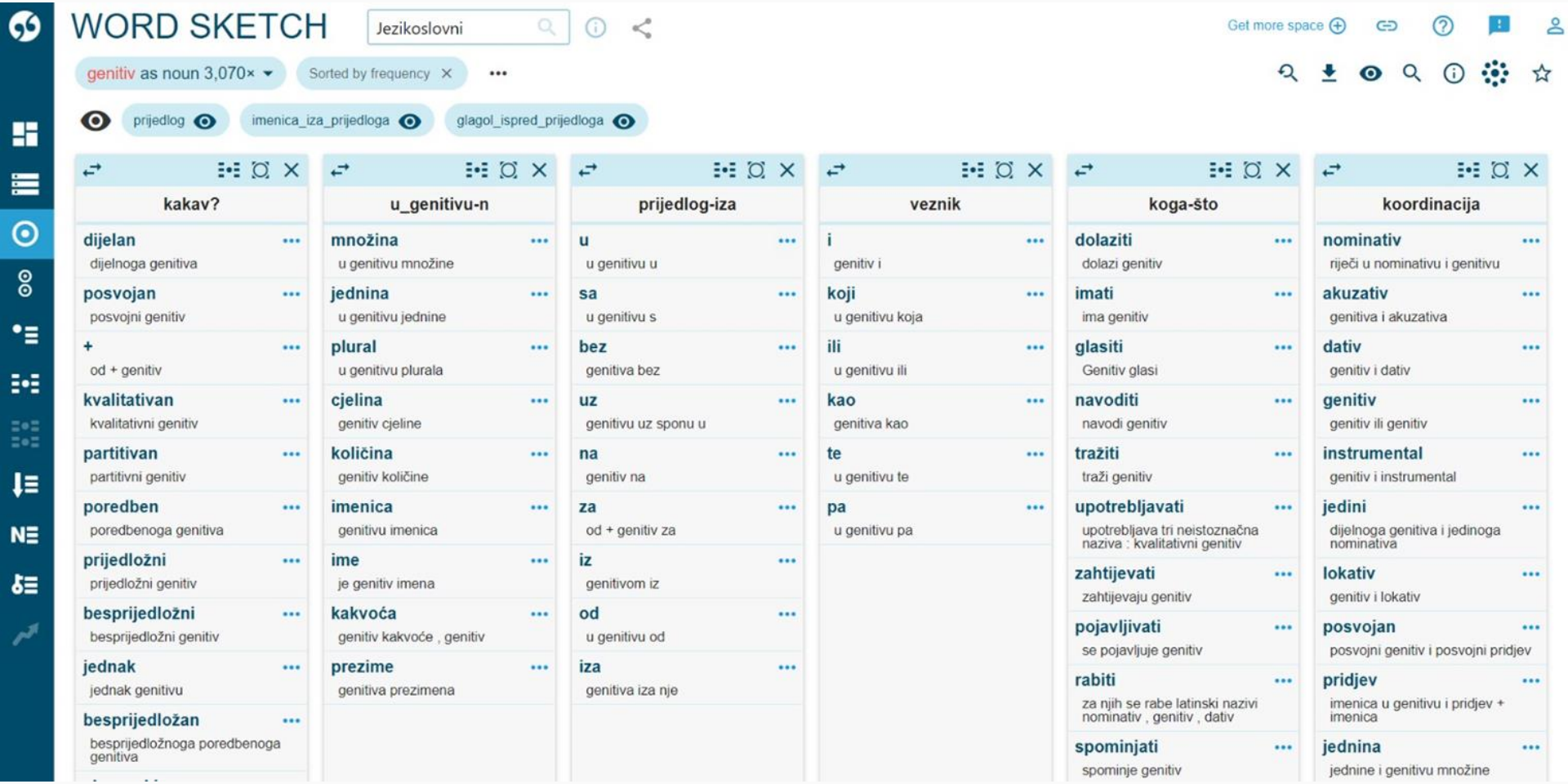
Korpus sadržava

- 1811 izvora
- 310 231 rečenica
- 10 103 069 pojavnica
  - 837 jedinstvenih pojavnica
- 7 858 801 riječi
  - 545 220 jedinstvenih riječi

Izvori korpusa *Jene*

- kako bi jezikoslovni korpus bio reprezentativan, uključeno je više različitih izvora različitih autora
- ti izvori uglavnom sačinjavaju dostupne izvore na portalu *Hrčak*:
    - Fluminensia
    - Hrvatski jezik
    - Rasprave
    - Suvremena lingvistika
    - Folia onomastica Croatica
    - Filologija
    - Jezikoslovlje
    - …
- ostale izvore zasad sačinjavaju monografije u izdanju Instituta za hrvatski jezik i jezikoslovlje:
    - Hrvatska školska gramatika
    - Instrumental u hrvatskom jeziku
    - Unutarnja struktura od glagolskih imenica u hrvatskome jeziku
    - ...


CONCORDANCE
Jezikoslovni
Get more space
simple glagol 21,501
shuffle 21,501 (2,180.01 per million)
sentence
Details
sentence
1 Rasprave 42-1, O s... <s> Motivirane su glagolima i imenicama. </s>
2 Brač, Ivana. (2018... <s> U vezi s predikatnim instrumentalom zaključuju da se glagol biti i semikopulativni glagoli razlikuju svojim značenjem iz čega proizlazi i mogućnost alternacije s nominativom i drugim prijedložnim izrazima. </s>
3 Brlobaš, Željka. 2... <s> Obilježja svršenih i nesvršenih glagola Rukavina, dakako, mora uzeti u obzir tijekom objašnjenja prezenta indikativa gdje ističe da je to glagolsko vrijeme svojstveno samo nesvršenim glagolima koji pokazuju dugo trajanje glagolske radnje (str. 83). </s>
4 FLUMINENSIA 28-2. ... <s> Napominjemo da postoje glagoli koji nisu psihološki u kojih surečenice s uzročnim značenjem koje uvodi veznik što ne možemo proglasiti neovjerenima, nego obilježenima u odnosu na surečenice uvedene veznicima jer, zato što, zbog toga što, npr. Majka pred kamerama plače što nema
FLUMINENSIA 28-2, REČENICE KAO DOPUNE UZ PSIHOLOŠKE GLAGOLE U HRVATSKOM JEZIKU.txt
rani djecu., Nije pobjegao što se boji tog dječaka iz razreda, već što ga je strah njegove starije braće. i sl. No opaža se da u tim
ečenice nisu dopune, kao što je to slučaj u psiholoških glagola . </s>
5 Suvremena lingvist... <s> Priložna dopuna obuhvaća priložne oznake uvjetovane valentnošću glagola koje se moraju realizirati da bi rečenica bila ovjerenom (npr.: Stanujem u Zagrebu. </s>
6 Suvremena lingvist... <s> Posljedica je toga i istovremeno prikrivanje cijele teorije o aktantima i teorije valentnosti glagola . </s>
7 Suvremena lingvist...) Iako ga inače baš nikada nije strah, sada se boji javno priznati što se dogodilo. glagolska finitna fraza iako ga inače nikada nije strah ima funkciju koncesivnog suplementa koji je značenjski kompatibilan sa svojim upraviteljem glagolom se boji, glagolska fraza što se dogodilo stoji na mjestu akuzativnog komplementa kojim upravlja glagol priznati, a glagolska infinitna fraza javno priznati stoji na mjestu genitivnog komplementa kojim upravlja glagol se boji) Navedene odnose u rečenici (a) grafički prikazujemo dijagramom ovisnosti (njem. Stammbaumdiagramm), koji bi u pojednostavljenom obliku izgledao ovako: Ako navedene glagole interpretiramo kao suznačne, i to "modalne u širem smislu" kao što to čine Silić/Pranjković (2007: 186), predikati u citiranim rečenicama glase boji se priznati, ne plaši se reci, su odlučivali prekinuti,je uspjela položiti, počeo je graditi, nastavljamo raditi, je prestajao pušiti. </s>
8 Suvremena lingvist... <s> Primjerice glagol sjediti izriče relaciju koja vrijedi za nekoga sudionika sjedenja (npr. čovjeka) tijekom nekoga vremena. </s>
9 Birtić, Matea. 200... <s> Isti je autor utvrdio da broj imenica izvedenih od nesvršenih glagola u velikoj mjeri nadmašuje broj izvedenica od svršenih glagola . </s>


WORD SKETCH
Jezikoslovni
Get more space
genitiv as noun 3,070×
Sorted by frequency
prijedlog
imenica_iza_prijedloga
glagol_ispred_prijedloga
kakav?
dijelan
dijelnoga genitiva
posvojan
posvojni genitiv
+
od + genitiv
kvalitativan
kvalitativni genitiv
partitivan
partitivni genitiv
poredben
poredbenoga genitiva
prijedložni
prijedložni genitiv
besprijedložni
besprijedložni genitiv
jednak
jednak genitivu
besprijedložan
besprijedložnoga poredbenoga genitiva
u_genitivu-n
množina
u genitivu množine
jednina
u genitivu jednine
plural
u genitivu plurala
cjelina
genitiv cjeline
količina
genitiv količine
imenica
genitivu imenica
ime
je genitiv imena
kakvoća
genitiv kakvoće , genitiv
prezime
genitiva prezimena
prijedlog-iza
u
u genitivu u
sa
u genitivu s
bez
genitiva bez
uz
genitivu uz sponu u
na
genitiv na
za
od + genitiv za
iz
genitivom iz
od
u genitivu od
iza
genitiva iza nje
veznik
i
genitiv i
koji
u genitivu koja
ili
u genitivu ili
kao
genitiva kao
te
u genitivu te
pa
u genitivu pa
koga-što
dolaziti
dolazi genitiv
imati
ima genitiv
glasiti
Genitiv glasi
navoditi
navodi genitiv
tražiti
traži genitiv
upotrebljavati
upotrebljava tri neistoznačna naziva : kvalitativni genitiv
zahtijevati
zahtijevaju genitiv
pojavljivati
se pojavljuje genitiv
rabiti
za njih se rabe latinski nazivi nominativ , genitiv , dativ
spominjati
spominje genitiv
koordinacija
nominativ
riječi u nominativu i genitivu
akuzativ
genitiva i akuzativa
dativ
genitiv i dativ
genitiv
genitiv ili genitiv
instrumental
genitiv i instrumental
jedini
dijelnoga genitiva i jedinoga nominativa
lokativ
genitiv i lokativ
posvojan
posvojni genitiv i posvojni pridjev
pridjev
imenica u genitivu i pridjev + imenica
jednina
jednine i genitivu množine

Problemi pri izradi korpusa

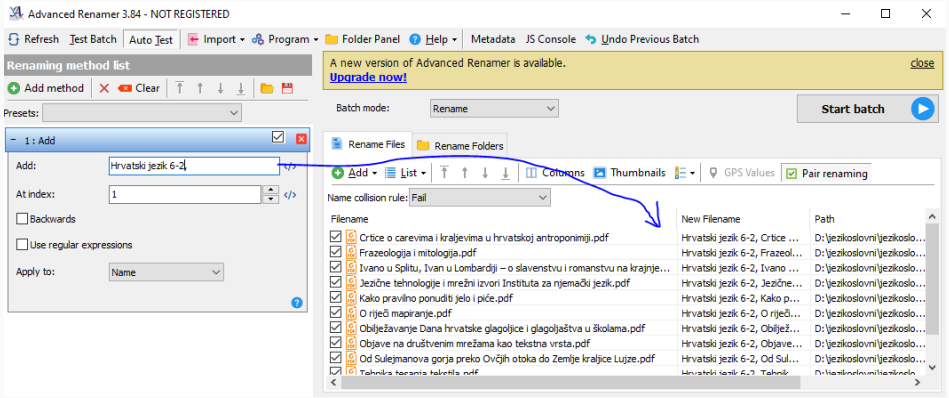
- mnogi tekstovi u člancima i knjigama nisu računalno čitljivi
  - treba ih OCR-irati
  - potrebno je još pregledati kvalitetu OCR-a i ispraviti pogreške OCR-a
- potrebno je očistiti tekstove prije nego se učitaju u program za izradu korpusa:
  - slike, grafikone, tablice te druge grafičke objekte maknuti jer se ne prikazuju pravilno
  - izbaciti podatke iz zaglavlja i podnožja dokumenta
  - izbaciti literaturu
  - ispraviti nepoznate znakove u tekstu (npr. **?** u *ù*)
  - maknuti nepotrebne razmake i spojnice
  - paziti na jednojezičnost dokumenata (zbog izrada skica riječi te smanjenja pogrešaka)

Faze izrade korpusa *JENE*

1. pronalaženje odgovarajućih izvora
2. preimenovanje izvora
3. automatsko pročišćavanje teksta
4. pregled teksta te ispravljanje pogrešaka
5. učitavanje datoteka u Sketch Engine

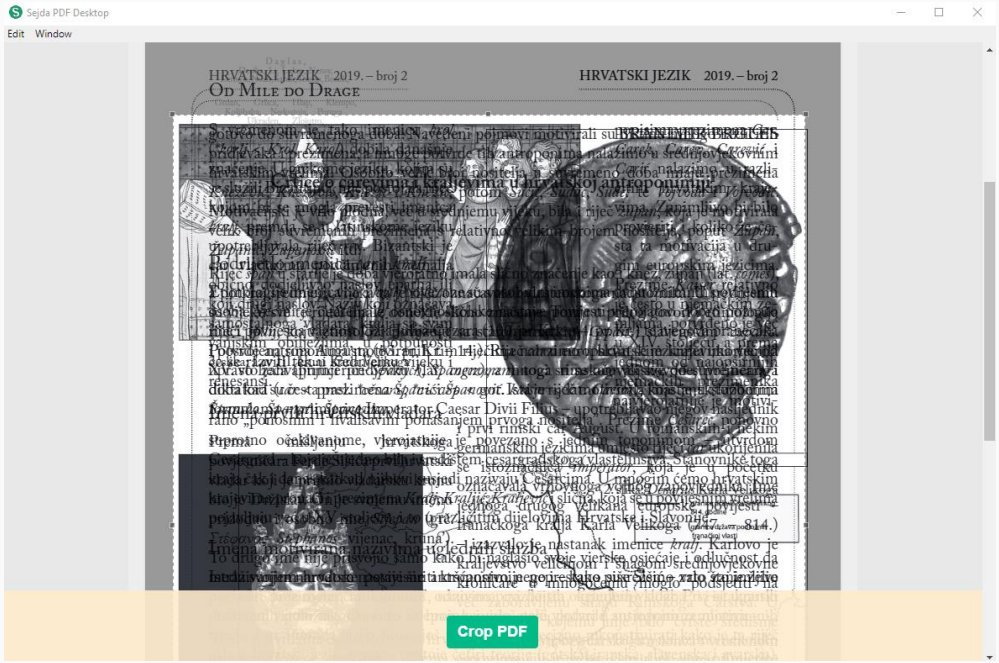
Pronalaženje i preimenovanje odgovarajućih izvora

- članci s *Hrčka* preuzeti su jedan po jedan, pri čemu je svaki članak preimenovan tako da je naziv datoteke istovjetan nazivu članka
- datoteke u svojemu nazivu osim naslova članka sadržavaju i ime časopisa te godište i broj časopisa
- da bi se više članaka moglo usporedno brže preimenovati, upotrijebljen je besplatni program Advanced Renamer



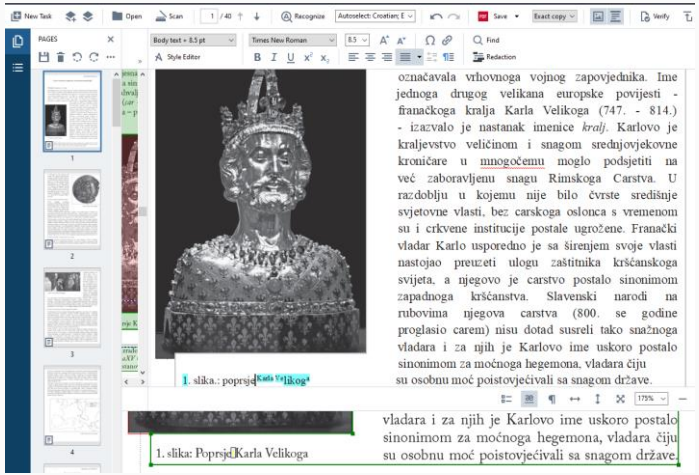
Automatsko pročišćavanje teksta

- zaglavlje i podnožje dokumenta izrezano je s pomoću programa Sejda PDF



Automatsko pročišćavanje teksta

- u slučaju potrebe za OCR-om koristi se program Abbyy FineReader 14 (bitna je da program ima podršku za hrvatski jezik)
- u istome programu provedeno je prebacivanje u Word dokumente (može izvojiti bilješke te ostale elemente unutar strukture dokumenta)

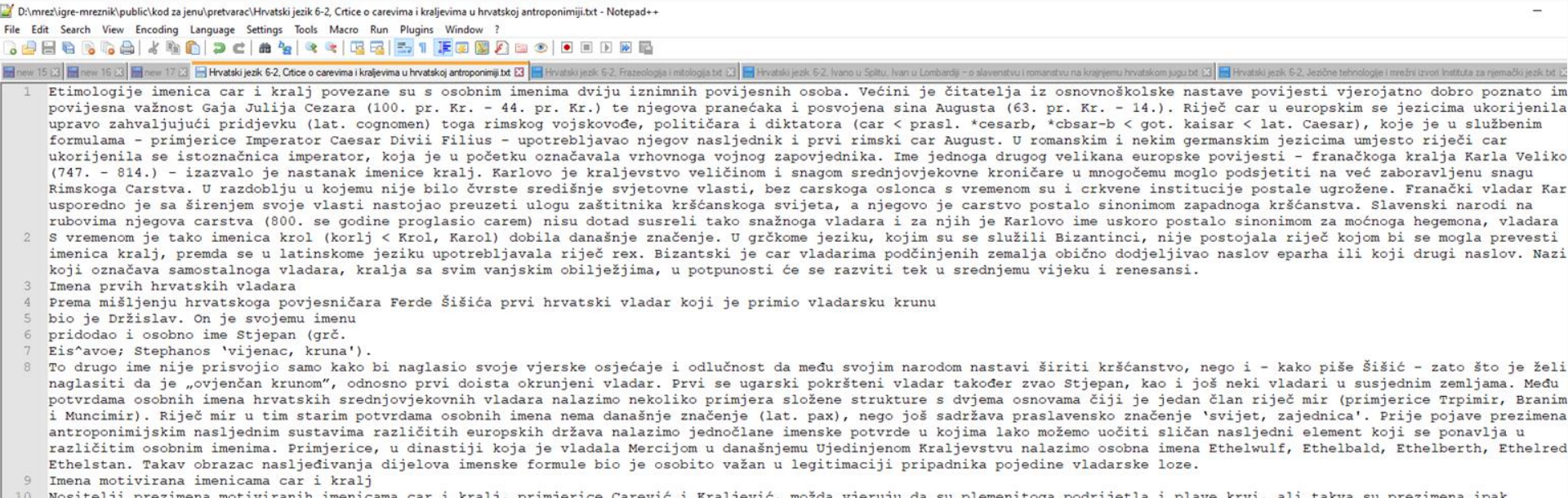


Automatsko pročišćavanje teksta

- u dokumentima u Wordu s pomoću makronaredbe moglo se unutar više dokumenata maknuti bilješke, slike, tablice, naslove poglavlja itd. (cilj je sravnati tekst kako bi bio pregledan u korpusu)
- nakon toga s pomoću Python skripte napravilo se prebacivanje datoteka iz Worda u .txt
    - također se automatiziralo brisanje literature, nepotrebnih razmaka i spojnica te ispravljanje neprepoznatljivih znakova

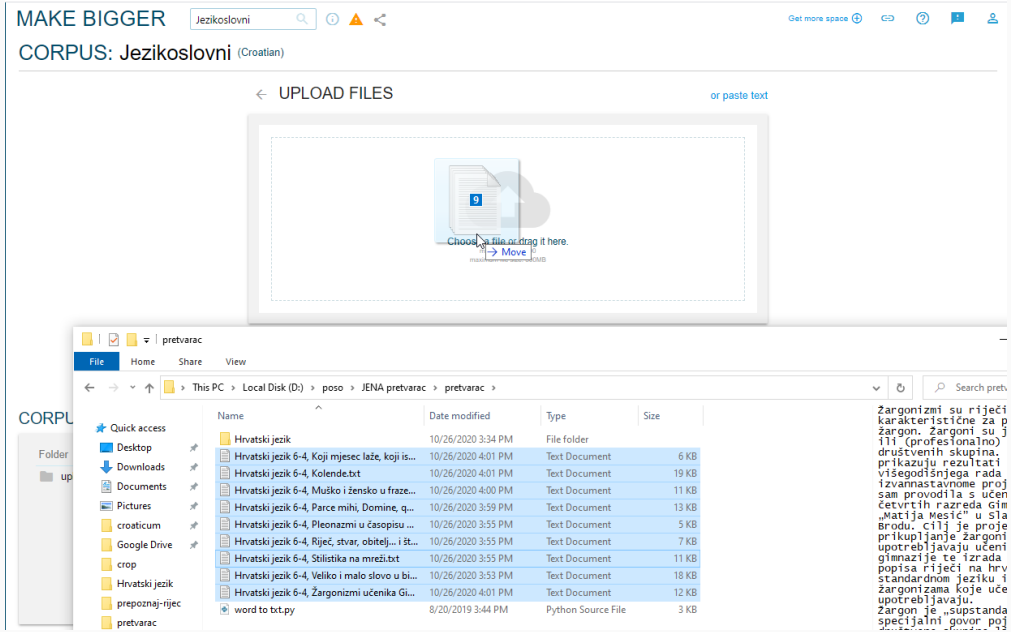
Pregled teksta te ispravljanje pogrešaka

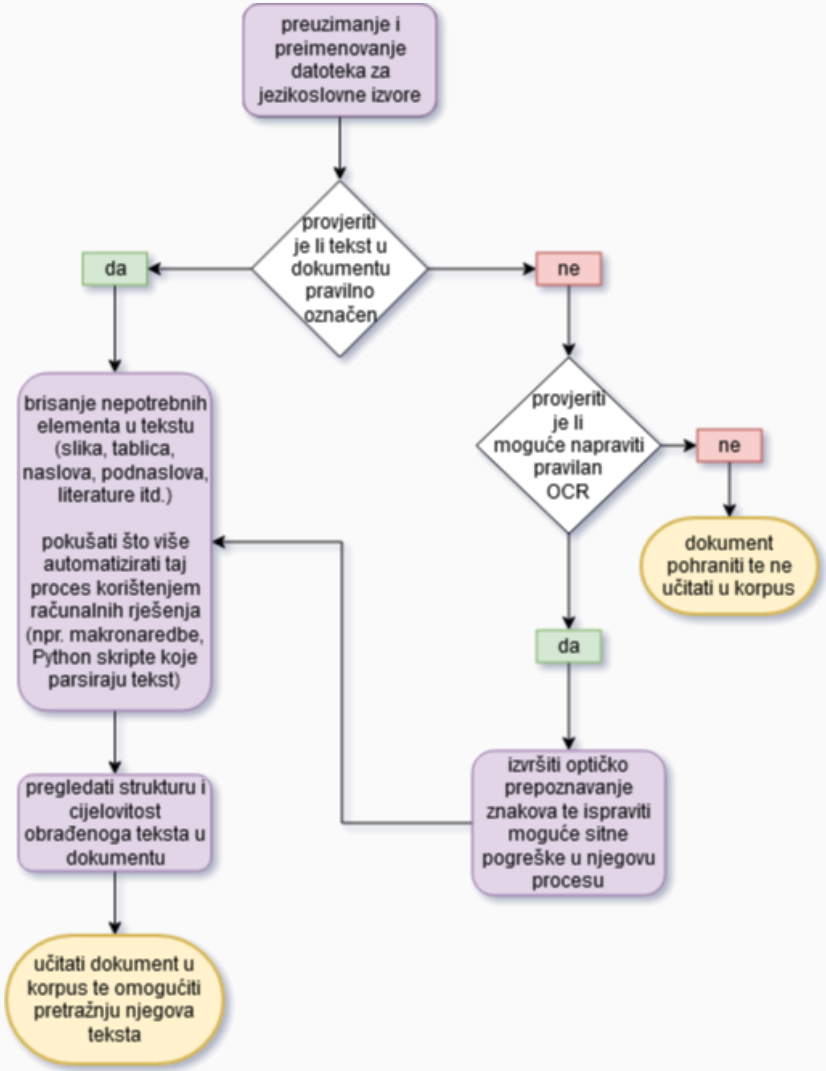
- .txt datoteke mogu se pregledati s pomoću programa Notepad++
- najčešće na početku i kraju dokumenta mogu se naći još neki dijelovi koje treba ispraviti



Učitavanje datoteka u Sketch Engine

- učitati sve .txt datoteke u program i izvršiti kompajliranje korpusa
- koristiti se Serbo-Croatian (MTE) 1.4 za izradu skica riječi




preuzimanje i preimenovanje datoteka za jezikoslovne izvore
provjeriti je li tekst u dokumentu pravilno označen
da
ne
brisanje nepotrebnih elementa u tekstu (slika, tablica, naslova, podnaslova, literature itd.)
pokušati što više automatizirati taj proces korištenjem računalnih rješenja (npr. makronaredbe, Python skripte koje parsiraju tekst)
provjeriti je li moguće napraviti pravilan OCR
ne
dokument pohraniti te ne učitati u korpus
da
izvršiti optičko prepoznavanje znakova te ispraviti moguće sitne pogreške u njegovu procesu
pregledati strukturu i cijelovitost obrađenoga teksta u dokumentu
učitati dokument u korpus te omogućiti pretražnju njegova teksta

Literatura

- Hudeček, Lana i dr. 2018. Pojmovnik. *Hrvatski mrežni rječnik – Mrežnik*. Pristupljeno 3. siječnja 2020. (http://ihjj.hr/mreznik/page/pojmovnik/6/).
- Marković, Mario; Mihaljević, Josip; Mihaljević, Milica. 2020. Kako pronaći jezikoslovni naziv. *Hrvatski jezik* 7/1. 18–22.

Josip Mihaljević
Institut za hrvatski jezik i jezikoslovlje
Okrugli stol – Terminološki izazovi u 21. stoljeću
Zagreb, 29. listopad 2020.

Korpus *Jene*

*Ova je prezentacija izrađena u okviru projekata *Hrvatsko jezikoslovno nazivlje – Jena* (Struna-2017-09-05) i *Hrvatski mrežni rječnik – Mrežnik* (IP-2016-06-2141), koje u cijelosti financira Hrvatska zaklada za znanost.