SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
KATEDRA ZA ARHIVISTIKU I DOKUMENTALISTIKU

Marta Dević Hameršmit

# Analiza i optičko prepoznavanje rukopisa s herbarijskih etiketa u zbirci Herbarium Croaticum

Diplomski rad

Mentor: dr. sc. Hrvoje Stančić, red. prof.
Neposredni voditelj: Vedran Šegota, dipl. ing. bio. (Botanički zavod, Biološki odsjek, PMF)

Zagreb, srpanj 2018.

# Sadržaj

# 1. Uvod

Volontirajući u herbarijskoj zbirci Herbarium Croaticum Botaničkog zavoda Biološkog odsjeka Prirodoslovno-matematičkog fakulteta u Zagrebu i kroz studentski posao na digitalizaciji herbarijske zbirke u Hrvatskom prirodoslovnom muzeju u Zagrebu susrela sam se s herbarijskim etiketama i problematikom njihovog iščitavanja prilikom prepisivanja njihovog sadržaja u inventarnu knjigu i u bazu podataka Flora Croatica. Studentsku praksu sam provela fokusirajući se na rukopise s herbarijskih etiketa, što je na kraju i dovelo do odabira teme ovog diplomskog rada.

Od velike koristi prilikom iščitavanja etiketa bilo bi i optičko prepoznavanje slova rukopisa, jer pojedini autori etiketa imaju i do nekoliko tisuća etiketa pisanih rukom, a rijetko kad je rukopis lako čitljiv. Herbarijske etikete su prilično stare, mnoge datiraju i iz 19. stoljeća, kada su pisane na različitim jezicima (latinskim, hrvatskim, njemačkim, talijanskim), a i tadašnji hrvatski jezik se razlikuje od današnjeg suvremenog jezika.

U prvom dijelu ovog rada fokus će biti na teoriji optičkog prepoznavanja znakova, dok će u drugom dijelu kroz rezultate i raspravu fokus biti na analizi rukopisa i trenutnim mogućnostima programa otvorenog koda.

## 1.1. Optičko prepoznavanje

Optičko prepoznavanje od velike je važnosti danas kada je prisutna masovna digitalizacija, posebice u knjižnicama. Značajna je količina povijesnih knjiga, časopisa i novina već digitalizirana i dostupna online za pretraživanje i pregledavanje. Situacija je potpuno drugačija kada pogledamo arhive. Iako veliki državni i regionalni arhivi čuvaju veliku količinu arhivskoga gradiva, digitalizacija u njima zaostaje za knjižnicama. Razlog dijelom leži i u nedostatku tehnologije koja se može nositi s rukopisnim dokumentima i pretvoriti ih u tekst koji se može uređivati.

### 1.1.1. Optičko prepoznavanje znakova

Christensson (2018) navodi da je OCR (engl. *optical character recognition*) tj. optičko prepoznavanje znakova tehnologija koja prepoznaje tekst unutar digitalne slike. Obično se koristi za prepoznavanje teksta u skeniranim dokumentima, ali služi i u druge svrhe. OCR softver obrađuje digitalnu sliku lociranjem i prepoznavanjem znakova, kao što su slova, brojevi i simboli. Neki OCR softver će jednostavno izvesti tekst, dok drugi programi mogu pretvoriti znakove u tekst koji se može uređivati izravno na slici. OCR tehnologija može se koristiti za pretvaranje tiskane kopije dokumenta u elektroničku verziju. Skeniranjem dokumenta dobiva se digitalna slika, kao što je npr. TIFF datoteka, učitavanjem slike u OCR program on će prepoznati tekst i pretvoriti dokument u tekstualnu datoteku koja se može uređivati. Neki OCR programi omogućuju skeniranje dokumenta i pretvorbu u dokument za obradu teksta u jednom koraku. Iako je OCR tehnologija izvorno dizajnirana za prepoznavanje tiskanog teksta, može se koristiti i za prepoznavanje i provjeru jednostavnijeg rukopisnog teksta. Na primjer, poštanske usluge kao što je američka poštanska služba USPS, koriste OCR softver za automatsko obrađivanje slova i paketa na temelju adrese. Algoritam provjerava skenirane informacije u bazi podataka postojećih adresa kako bi potvrdio poštansku adresu. Aplikacija za pametni telefon Google Prevoditelj obuhvaća OCR tehnologiju koja radi s fotoaparatom uređaja i pri tome omogućuje uzimanje teksta iz dokumenata, časopisa, znakova iz drugih izvora i prevođenje na drugi jezik u stvarnom vremenu.

### 1.1.2. Inteligentno prepoznavanje znakova

Prema Abbyy Technology portalu Inteligentno prepoznavanje znakova je nadograđena verzija optičkog prepoznavanja znakova. ICR (engl. *intelligent character recognition*) se odnosi na rukom pisane znakove koji su odvojeni, napisani kao pojedinačni znakovi. Polja u koja se pišu znakovi moraju biti strojno čitljivog oblika (Slika 1). ICR ne podrazumijeva prepoznavanje rukopisa.


*Slika 1: Prikaz izgleda polja za strojno čitanje rukom pisanih slova (Izvor: https://abbyy.technology/en:features:ocr:icr)*

Prema ITWissen, inteligentno prepoznavanje znakova ručno pisane tekstove i brojeve pretvara u strojno čitljive znakove i dokumente. Inteligencija prepoznavanja znakova ICR-a ogleda se u korekciji pogrešnih znakova. Kvaliteta rezultata poboljšana je semantičkim korelacijama, jezičnim i statističkim metodama, usporedbom s pohranjenim referentnim uzorcima i ispitivanjem prema rječnicima. U korekciji rezultata, to dovodi do pogrešno tumačenog znaka koji se ispravlja nakon semantičke klasifikacije ili nakon usporedbe s rječnicima. Na primjer, ako ICR softver interpretira "u" kao "a" u riječi "muk", što bi dovelo do riječi "mak", semantički kontekst pretvara slovo "a" u "u", a time ispravlja pogrešno prepoznavanje teksta.

ICR tehnologija prepoznavanja rukopisa polje je koje nije u potpunosti razvijeno, a razina preciznosti koju pružaju mnogi takvi programski paketi nisu baš dobri. Uspoređujući razinu točnosti softvera za inteligentno prepoznavanje znakova u odnosu na softver za optičko prepoznavanje znakova (OCR), zamjetni su nedostatci u kvaliteti u odnosu na OCR softver. Softveri za inteligentno prepoznavanje znakova, koriste neuronske mreže koje zahtijevaju stalno učenje, a treniranjem se povećavaju stope točnosti.

### 1.1.3. Prepoznavanje rukopisa

Optičko prepoznavanje rukopisa u novije vrijeme jedno je od zahtjevnijih i fascinantnijih istraživanja u području obrade slike i prepoznavanju uzoraka.

Amrutiya Hirali, Moghariya Payal i Shah Vatsal (2014) navode da je HCR (engl. *hand written character recognition*) tehnika koja se koristi za prepoznavanje znakova rukopisa. Razlikuju se dvije metode prepoznavanja rukopisa, a to su *online* i *offline* prepoznavanje znakova rukopisa. *Online* prepoznavanje rukopisa uključuje automatsku konverziju teksta koji je napisan na posebnom digitalizatoru ili PDA uređaju, gdje senzor uzima pokrete vrha olovke i pokrete olovke gore-dolje. Takva vrsta olovke poznata kao digitalna tinta može se smatrati digitalnim prikazom rukopisa. Dobiveni signali pretvaraju se u šifre slova koje se mogu koristiti unutar računala i za programe za obradu teksta. *Offline* prepoznavanje rukopisa uključuje konverziju teksta sa slike u šifre slova koje se mogu koristiti na računalu i programima za obradu teksta. Podatci dobiveni na ovakav način smatraju se statičkim prikazom rukopisa.

Ved Prakash Agnihotri (2012) ističe da se sustavima za *offline* prepoznavanje rukopisa koriste aplikacije za razvrstavanje pošte, čitanje dokumenata i prepoznavanje adresa te banke za različite transakcije. Područje prepoznavanja rukopisa i dalje je aktivno za istraživanje novih tehnika koje bi poboljšale točnost prepoznavanja.

Prema Hirali, Payal i Vatsal (2014) prepoznavanje znakova rukopisa uključuje sljedeće korake:

1. Preuzimanje slike (engl. Image acquisition) - potrebna je skenirana slika kao ulazna slika. Format može biti JPEG, BMP, JIF itd.
2. Prethodna obrada (engl. Preprocesing) - ovaj korak uključuje osnovnu obradu slike prije nego što se koristi za prepoznavanje u sustavu. Slika mora biti obrađen na takav način da je sustav može razumjeti. Koraci koji su uključeni u ovu fazu su:
   - Uklanjanje šuma

   Postoji nekoliko razloga zbog kojih se šum pojavi u slikama. Može biti iz mehanizma koji se koristi za dobivanje slike, filmskog zrna ili elektronskog prijenosa slika, a može se ukloniti filtriranjem.
   - Binarizacija

Binarizacija je proces u kojem se prvo sliku RGB boje pretvori u sivu, a sliku sive boje pretvori u binarnu sliku dokumenta. Zatim se svaki piksel na slici pretvara u jedan bit i dodjeljuje mu se vrijednost 0 ili 1 ovisno o srednjoj vrijednosti svih piksela.

- Otkrivanje rubova

Rub je područje gdje se intenzitet objekta drastično mijenja. Često se povezuju rubovi granica objekata na slici kako bi se detekcija ruba koristila za prepoznavanje rubova slika. Rubovi se otkrivaju pomoću algoritama.

- Proširivanje i punjenje

Osnovni učinak korištenja proširivanja na binarnoj slici je postupno povećanje granice područja piksela. Tako površine piksela u prvom planu rastu, a rupe unutar njih postaju manje.

- Obrada slike za ekstrakciju značajki

U ovoj fazi, značajke znakova koje su ključne za njihovo razvrstavanje se ekstrahiraju. Ovo je važna faza, jer djelotvorno ekstrahiranje poboljšava stopu prepoznavanja i smanjuje pogrešnu klasifikaciju.

3. Segmentacija – treći korak je segmentiranje ili rastavljanje riječi i rečenica na znakove ili slova tako da se postavi jasna granica između njih. To je potrebno, jer se algoritmi za prepoznavanje mogu primijeniti samo na pojedini znak ili slovo, a ne i na cijele riječi. Također, svakom znaku se smanjuje rezolucija.
4. Prepoznavanje - u posljednjem koraku, sustav pokušava analizirati i prepoznati znak koji mu je zadan.
5. Naknadna obrada

## 1.2. Herbarij

Herbariji služe za pohranu i čuvanje te za istraživanje i proučavanje raznolikosti biljnih vrsta. Prema Hrvatskom jezičnom portalu herbarij je „zbirka osušenih i prešanih biljki ili njihovih dijelova, a naziva se još i biljnik ili herbar. Herbariji služe prvenstveno

istraživačima koji se bave sistematskom botanikom i taksonomijom te florom nekog područja. Oni omogućavaju uvid u bioraznolikost nekog područja, dok iznimno očuvani primjerci mogu služiti u istraživačke svrhe. Uz znanstvenu ulogu, herbariji imaju važnu ulogu u nastavi iz botanike te u populariziranju znanosti.

### 1.2.1. Herbarijske zbirke

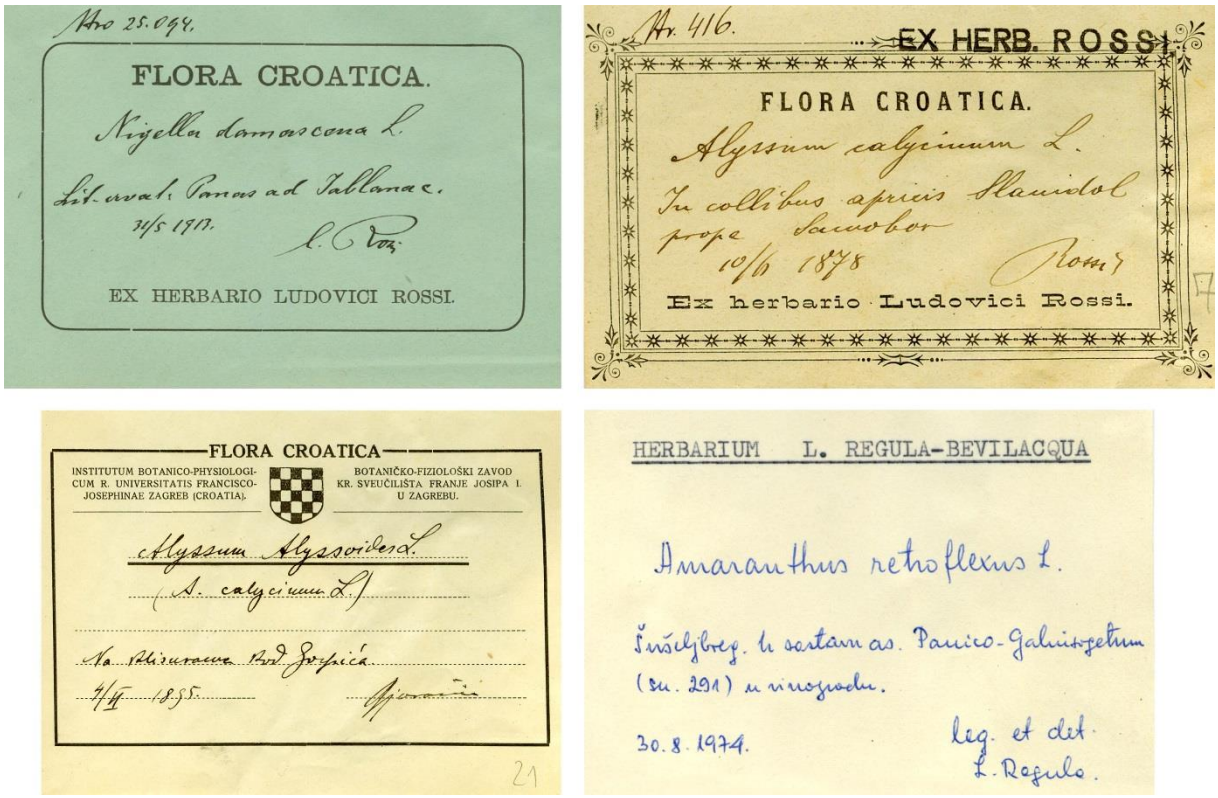
Herbarijske zbirke predstavljaju kulturno blago, jer se radi o povijesnim zbirkama koje su trajni zapis djelatnosti botaničara i istraživača nekog vremena. Razvojem digitalizacije i on-line baza podataka herbarijske zbirke dobivaju na važnosti, jer brzo i jednostavno postaju dostupne velikom broju istraživača i ostalih korisnika. Herbarijske zbirke se najčešće nalaze uz botaničke institucije i prirodoslovne muzeje. U Hrvatskoj ima nekoliko herbarijskih zbirki. Najveća i najstarija je u Botaničkom zavodu Prirodoslovno-matematičkoga fakulteta u Zagrebu. Herbarijske zbirke se označavaju standardiziranim međunarodnim kraticama. Dvije herbarijske zbirke Biološkog odsjeka PMF-a su Herbarium Croaticum (ZA) i Herbarijska zbirka Ive i Marije Horvat (ZAHO).

Digitalizacija i sistematizacija herbarijskih listova ZA i ZAHO zbirke provodi se putem baze Flora Croatica (FCD). Sistematizacija materijala u bazu FCD sprovodi se od 1999. godine, a do 12. lipnja 2018. godine sistematizirano je 23.356 primjeraka iz ZA zbirke i 3.630 primjeraka iz ZAHO zbirke. Godine 2016. počelo se sa skeniranjem herbarijskih listova i do 12. lipnja 2018. skenirano je 3.138 primjeraka iz ZA zbirke i oko 454 primjeraka iz ZAHO zbirke.

### 1.2.2. Herbarijske etikete

Pravilno sastavljena etiketa na sebi ima označen naziv herbarijske zbirke kojoj herbarijski list pripada, ime biljke na latinskom jeziku, nalazište (lokalitet ili mjesto s kojeg je biljka ubrana), zatim stanište ili značajke nalazišta, datum sakupljanja, ime sakupljača i ime osobe koja je biljku determinirala (odredila porodicu, rod, vrstu, a uz to mogu biti navedeni i neki drugi podatci). Na slici 2 su prikazane etikete različitih autora iz Herbarium Croaticum

(ZA) zbirke. Tako npr. na prvoj etiketi piše pripadnost herbariju (ex herbario Ludovici Rossi), u lijevom gornjem kutu piše redni broj (inventarni broj sakupljača) unutar herbarija kojem pripada. Što je taj broj veći, to je biljka kasnije sistematizirana. Osim toga, na etiketi piše i latinski naziv biljke, nalazište, datum sabiranja i potpis. Količina podataka na etiketi ovisi o autoru, tako na nekim etiketama može pisati samo naziv biljke i potpis.




*Slika 2: Primjeri herbarijskih etiketa (Izvor: Baza Flora Croatica)*

# 2. Ciljevi

Ciljevi ovoga rada su:

- analizirati raznolikost rukopisa sakupljača iz zbirke Herbarium Croaticum pohranjenih u bazi podataka Flora Croatica, u daljnjem tekstu FCD,
- izraditi abecede rukopisa najčešćih sakupljača u bazi FCD za potrebe budućih korisnika,
- analizirati mogućnosti optičkog prepoznavanja rukopisa pomoću softvera otvorenog koda dostupnih na internetu.

# 3. Materijali i metode

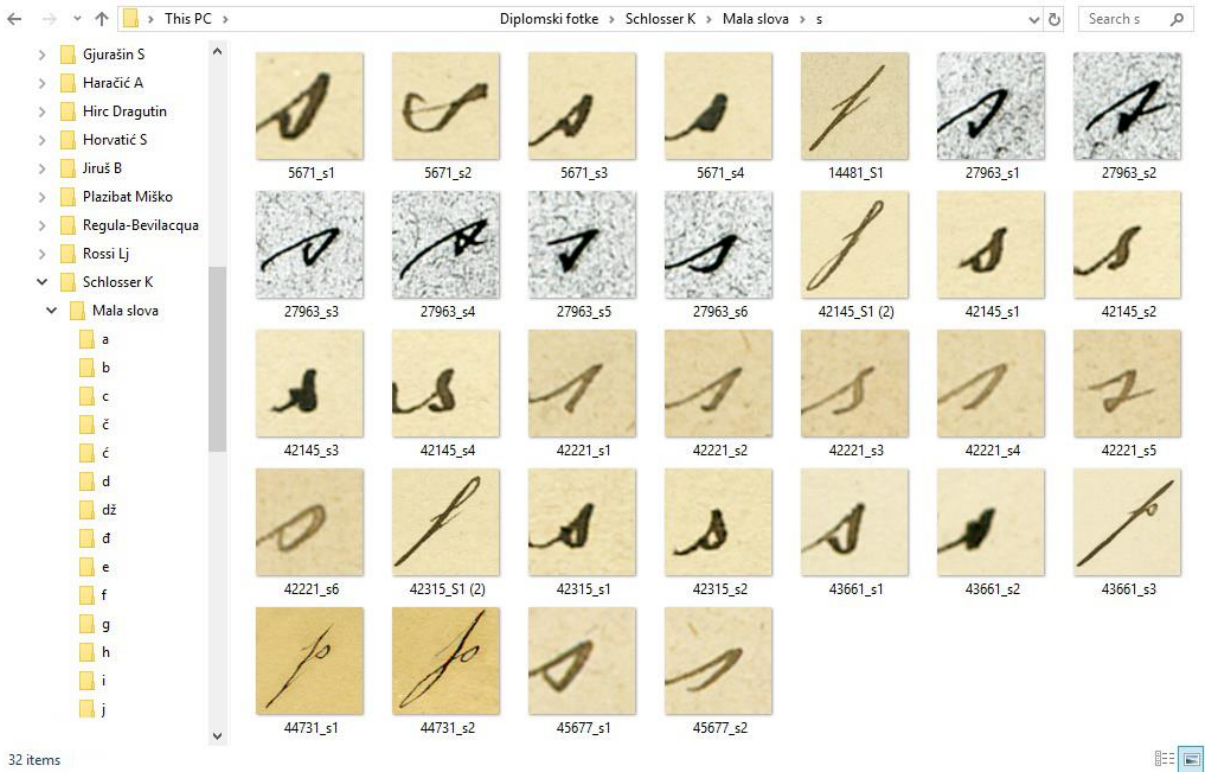
## 3.1. Materijali

U ovom radu korišteni su herbarijski listovi pohranjeni u herbarijskoj zbirki Botaničkog zavoda PMF-a (Herbarium Croaticum). Točnije, korišteni su skenovi dotičnih herbarijskih listova pohranjeni u online bazi podataka Flora Croatica, u TIFF formatu visoke rezolucije (minimalno 300 dpi).


*Slika 3: Herbarijski list iz ZA herbarijske zbirke (Izvor: Baza Flora Croatica)*

## 3.2. Metode

Korištene su digitalne fotografije (skenovi) herbarijskih listova iz baze FCD koje su u nekomprimiranom formatu (TIFF format). Za svakog autora (sakupljača herbarijskih primjeraka) nasumično je izabrano po 10 herbarijskih etiketa s kojih su izrezana sva slova u programu Adobe Photoshop CS4. Sva slova su rezana na veličinu 2x2 cm u rezoluciji 300 dpi, a pohranjena su u TIFF formatu. Slova su od ostalih slova na etiketi izdvojena pojedinačno korištenjem alata za izrezivanje (engl. *crop*) i uklanjanjem susjednih slova alatom za kloniranje (engl. *clone stamp*). Ona slova za koja nije bilo pronađeno minimalno tri primjerka, tražila su se ciljano s etiketa s potrebnim slovima. Etikete se pretraživalo u tražilici FCD-a po određenom slovu, tako što se u tražilicu u polje pretraživanja nalazišta unijelo slovo između dvije zvjezdice npr. slovo *š*. Izrezane fotografije su sistematizirane u direktorije, na način da je svaki autor dobio svoj direktorij, svako slovo unutar direktorija autora svoj direktorij, zasebno velika i mala slova.




*Slika 4: Sistematizacija izrezanih slova pohranjenih na računalu*

# 4. Rezultati i rasprava

U ovom djelu će pojedinačno za svakog autora biti prikazana analiza rukopisa s izrađenom abecedom rukopisa i pratećim slikama. U radu su korištene etikete koje su prethodno već bile pročitane, dešifrirane, ali su usprkos tomu tijekom izrezivanja slova primijećene pogreške te je ponekad bilo potrebno točna slova (riječi) potražiti u tražilici FCD baze ili preko tražilice Google.

Na kraju će biti navedeni programi koji su trenutno dostupni za optičko prepoznavanje.

## 4.1. Analiza rukopisa

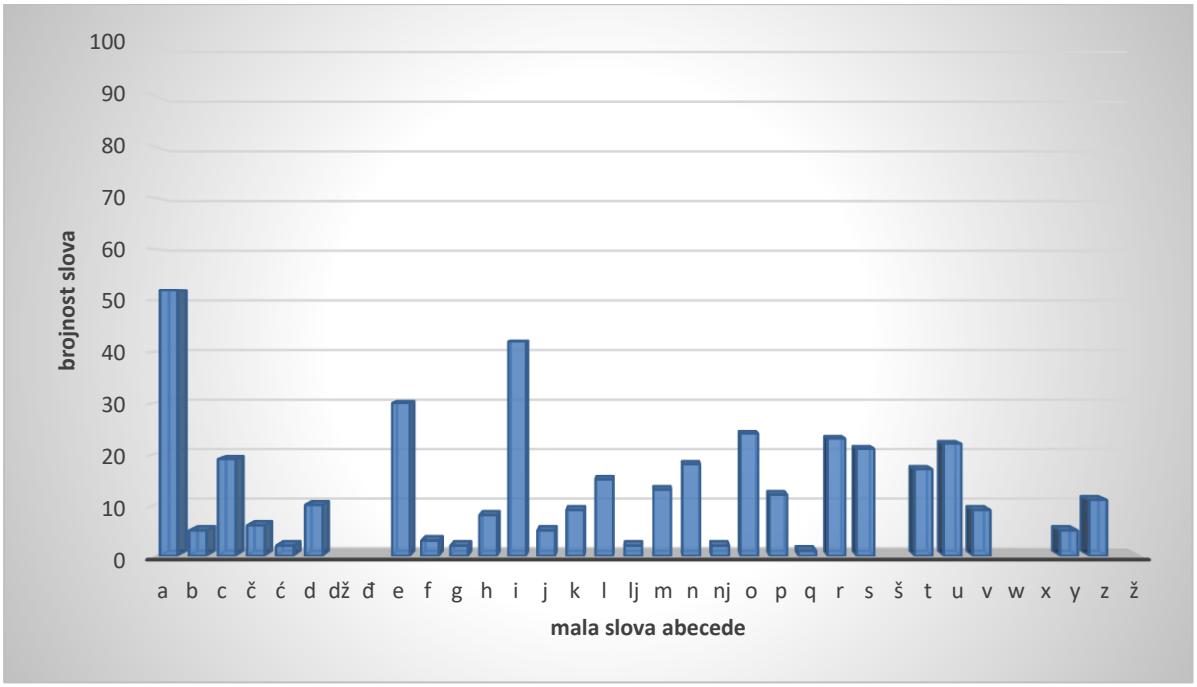
Ukupno je iz baze Flora Croatica preuzeto i korišteno 379 herbarijskih listova odnosno etiketa te je ukupno izrezano 6.020 slova od toga 5.191 malih i 829 velikih slova. Ako se u obzir uzme početnih 100 etiketa (10 etiketa po svakom autoru), tada je iz njih izrezano 5.039 malih i 531 velikih slova. Od ukupnog broja malih slova najviše je bilo slova *a* (648), *i* (498), *e* (422), *r* (348), o (334) te *s* (317). Od ukupnog broja velikih slova najviše je bilo slova *L* (71), *S* (52), *A* (43), *P* (43) te *O* (40). Malo slovo *dž* te velika slova *X* i *Đ* nisu pronađena ni na jednoj etiketi u bazi podataka. Za izrezivanje slova utrošeno je ukupno 77 sati, na traženje dodatnih etiketa za slova kojih u prvotno odabranih 10 etiketa nije bilo minimalno 3 primjerka bilo je potrebno oko 20 sati, a za montažu abeceda oko 8,5 sati. U prosjeku je po jednom slovu utrošeno 1,05 minute vremena bez preuzimanja prvih 10 etiketa za svakog autora.

### 4.1.1. Analiza rukopisa Stjepana Gjurašina

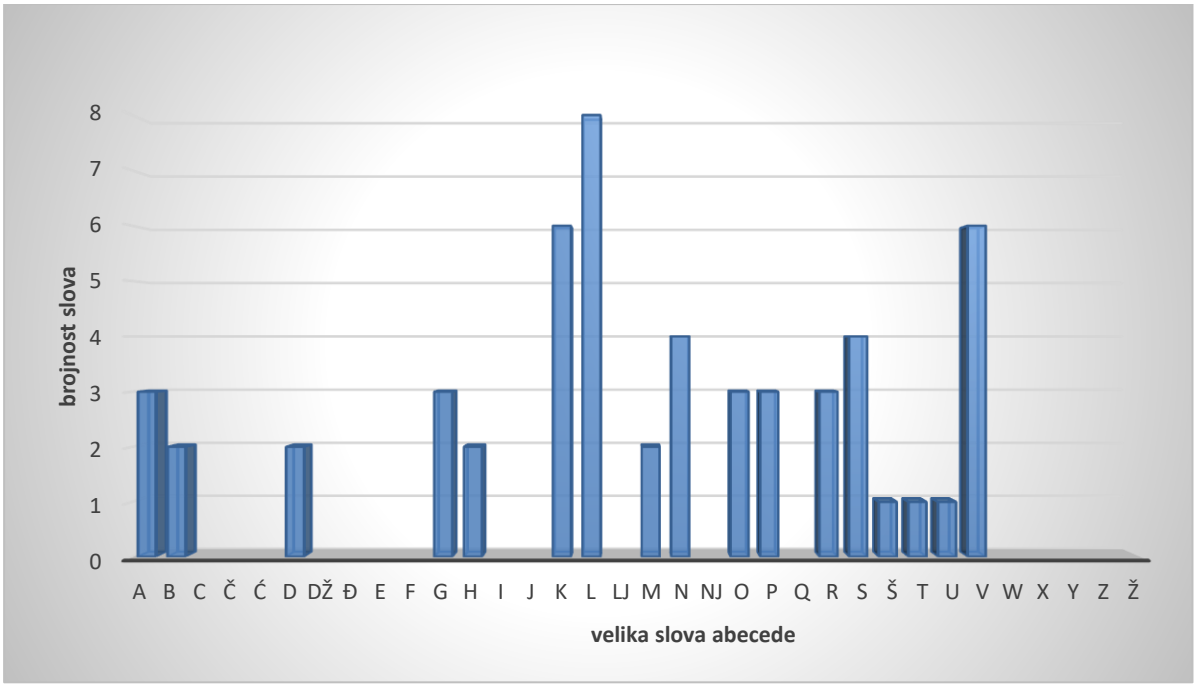
Stjepan Gjurašin hrvatski je botaničar koji je živio od 1867. do 1936. godine. Doktorirao je prirodne znanosti i predavao sistematiku i morfologiju bilja na Filozofskom fakultetu u Zagrebu. Njegovo područje proučavanja su bile floristika i mikologija. Pisao je znanstvene i stručne radove te školske udžbenike.

Ukupno je s 34 etikete izrezano 395 malih te 81 veliko slovo. Najviše je izrezano malih slova a, a brojnost svih malih slova vidljiva je na slici 5. Na slici 6 vidi se da je od velikih slova najzastupljenije slovo *L*, a zatim slova *K* i *V*. Od dostupnih velikih i malih slova složena je abeceda prikazana na slici 7.

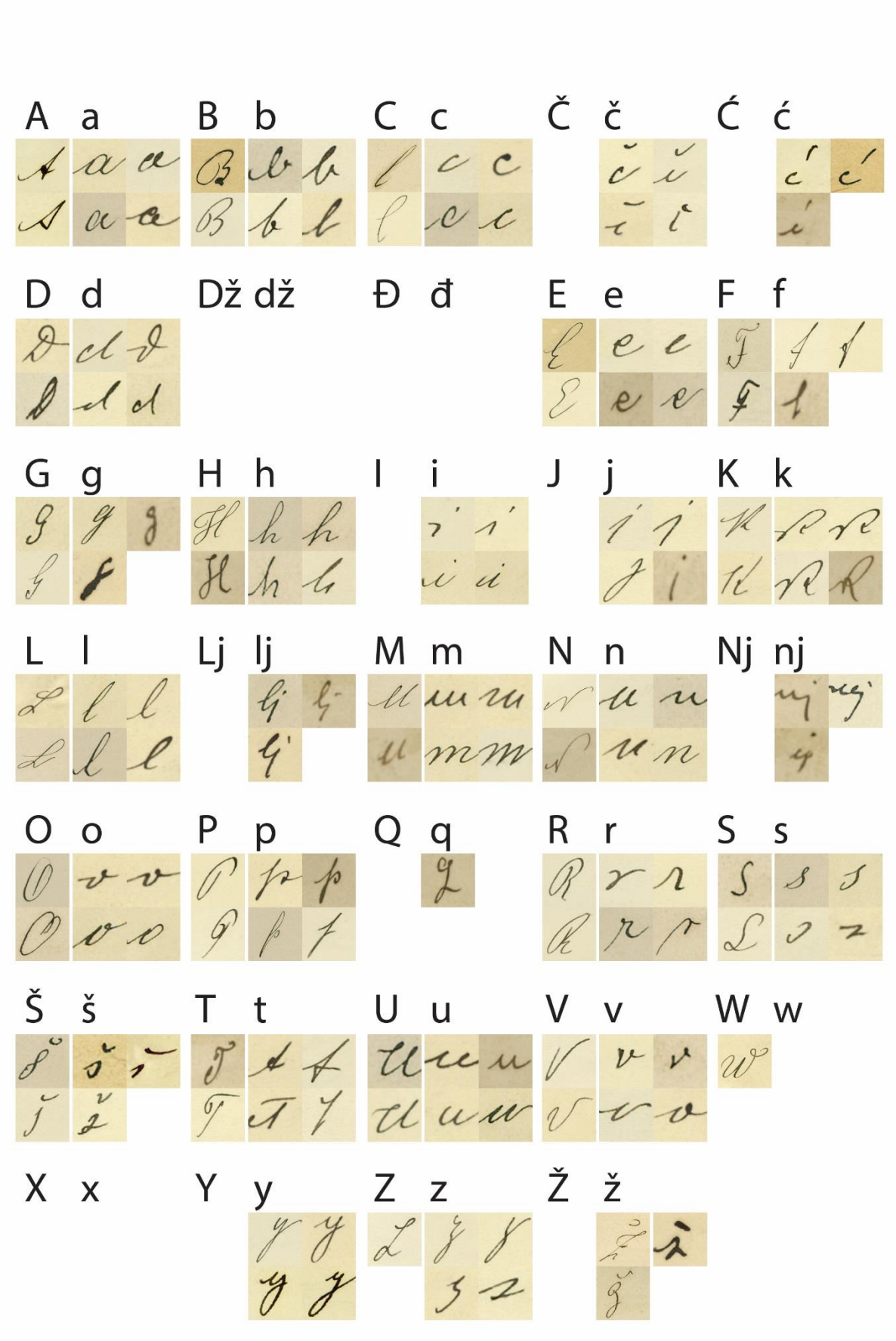
Vizualnim pregledom baze slova rukopisa Stjepana Gjurašina uočeno je da koristi dva načina pisanja, krasopis i obični rukopis. Koristi se pisanim slovima, a tiskana slova ne koristi. Slova nisu konstantne veličine, prevladavaju sitna, ponekad razvučena slova. Malo slovo *k* više sliči velikom slovu *R*. Veliko slovo *M* podsjeća na slovo *u*, slovo *P* više podsjeća na slovo *O*, mala slova t se međusobno dosta razlikuju, slovo *o* ponekad sliči slovu *v*, a *p* slovu *f*, kod slova *r* koristi dva načina njegovog pisanja, što je razumljivo obzirom da sa su etikete iz 19. stoljeća. Rukopis je općenito teže čitljiv.




*Slika 5: Brojnost pojedinih malih slova rukopisa Stjepana Gjurašina generiranih s 10 nasumično izabranih herbarijskih etiketa*




*Slika 6: Brojnost pojedinih velikih slova rukopisa Stjepana Gjurašina generiranih s 10 nasumično izabranih herbarijskih etiketa*



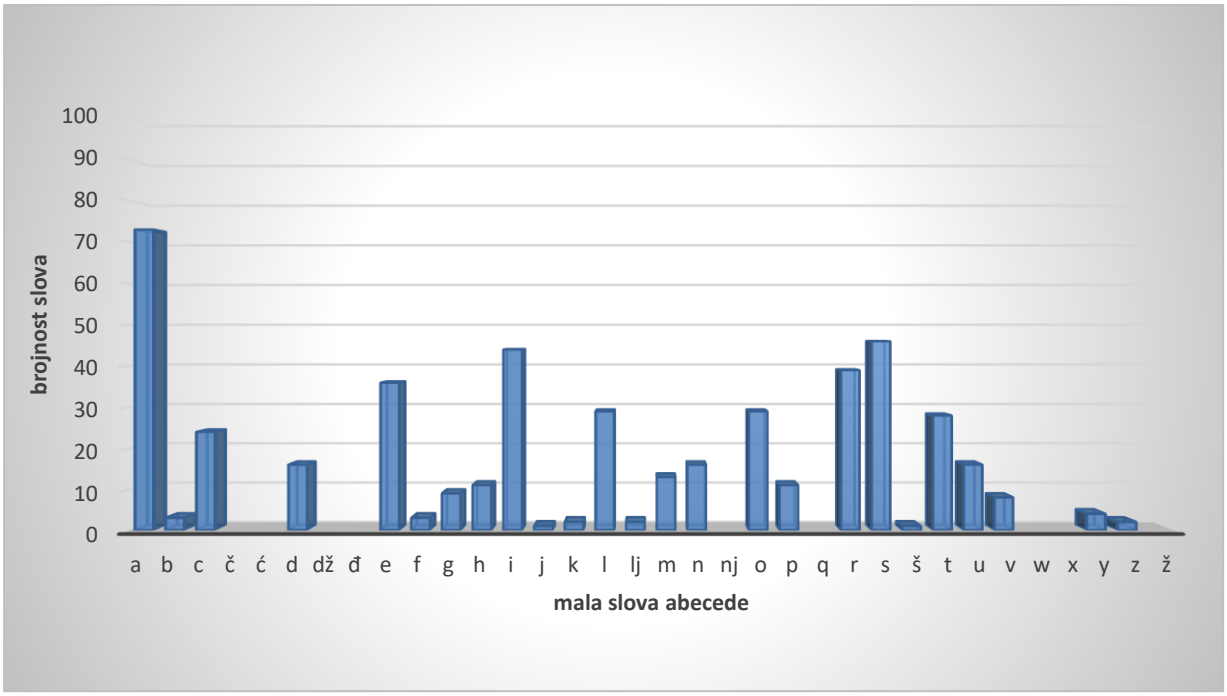

*Slika 7: Abeceda rukopisa Stjepana Gjurašina dobivena iz herbarijskih etiketa*

### 4.1.2. Analiza rukopisa Ambroza Haračića

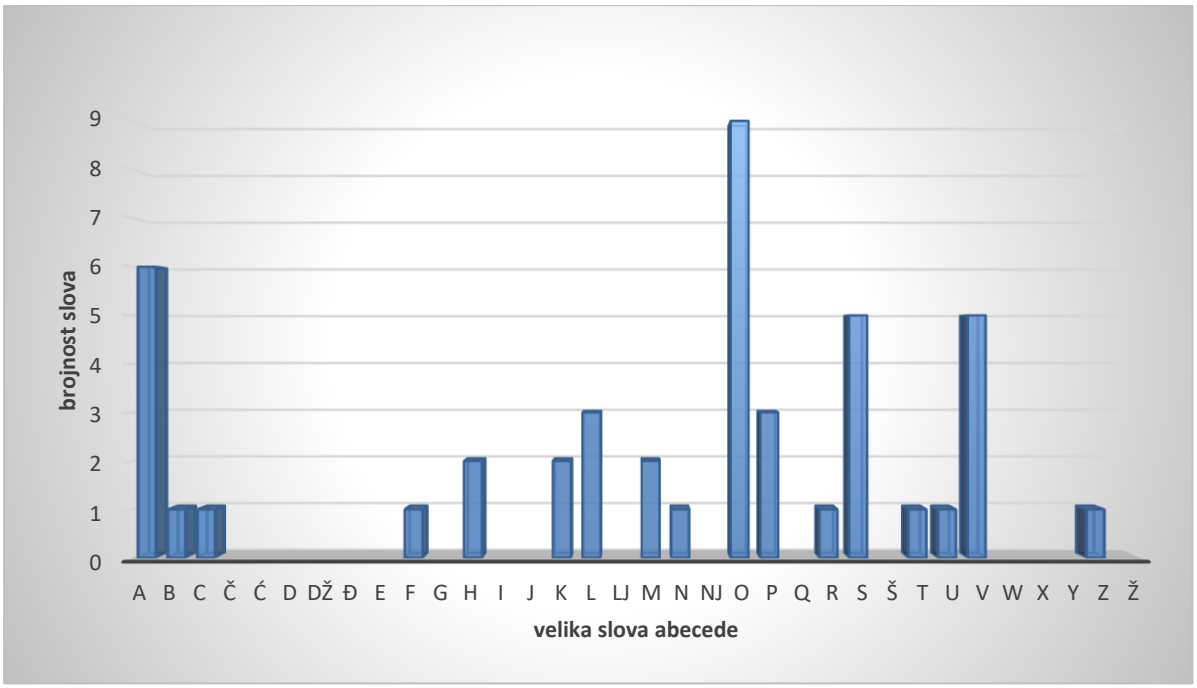
Ambroz Haračić hrvatski je botaničar i meteorolog, rođen 1855., umro 1916. godine. Proučavao je floru Malog Lošinja i okolnih malih otoka, vodio je svoju meteorološku stanicu i vlastitu herbarijsku zbirku, koja je danas u Hrvatskom herbariju Botaničkog zavoda PMF-a u Zagrebu.

Slova su skupljena sa 48 etiketa, ukupno 493 mala te 76 velikih slova, najviše je sakupljeno malih slova *a*, vidljivo na slici 8. Iz slike 9 je vidljivo da je od velikih slova najviše slova *O*. Od skupljenih slova složena je abeceda prikazana na slici 10.

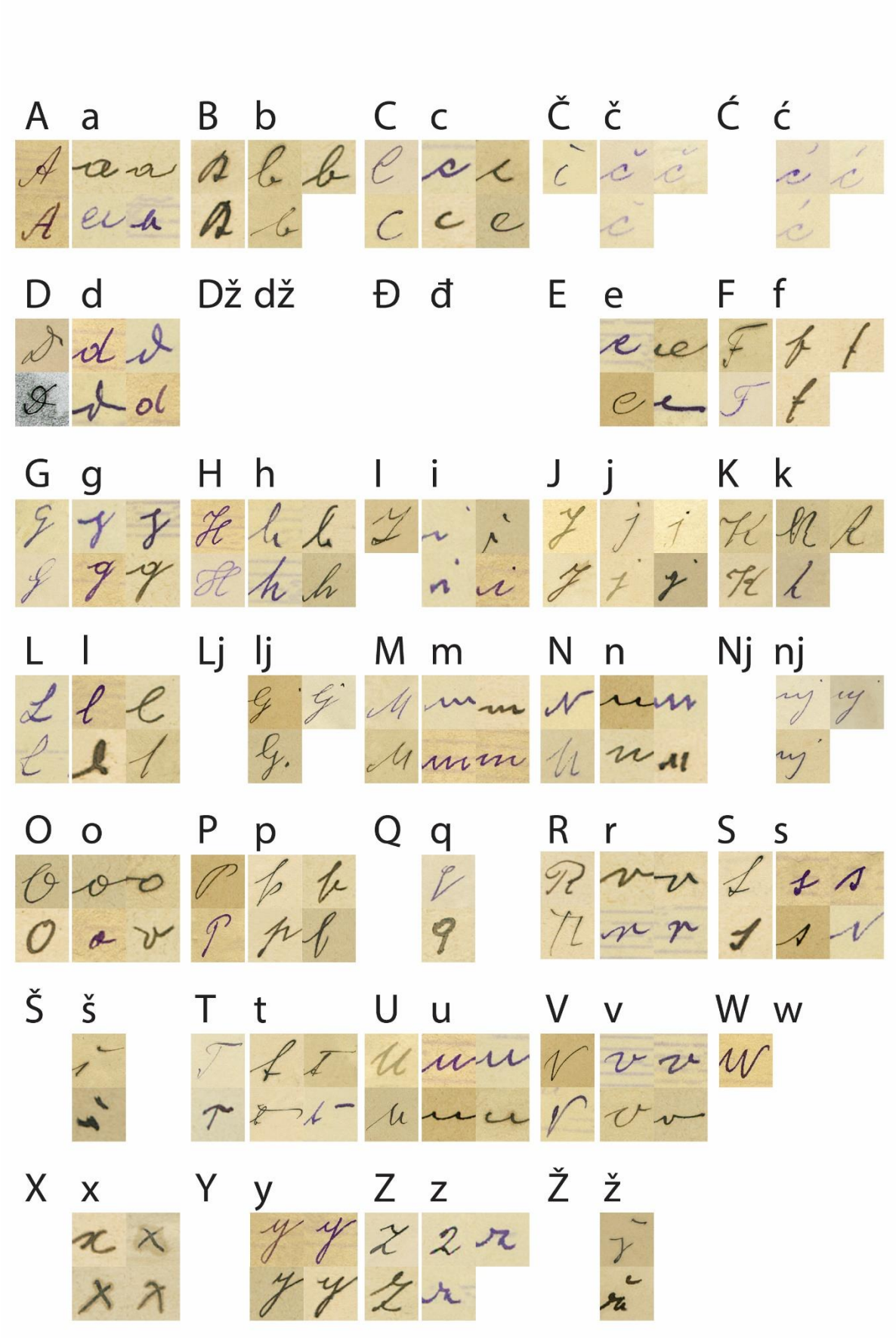
Vizualnim pregledom baze slova rukopisa Ambroza Haračića uočeno je da je rukopis teško čitljiv. Autor piše pisanim slovima, najčešće su mu slova nejednaka i sitna, nagnuta u desno, a između riječi su veći razmaci. Uz hrvatski koristi i talijanski jezik. Točkice i crtice (npr. crtice na slovu *t*) stavlja na kraju napisane riječi, tako da točkica bude iznad nekog drugog slova ili crtica slova *t* pokraj *t* ili je jako duga. Slova *u* i *n* su često jako slična, slovo *p* nekad je sličnije slovu *l* ili *b*. Slovo *r* nekad sliči slovu *v*, slova *o* i *a* često nisu zatvoreni kružići, slovo *a* često pisano iz dva poteza.




*Slika 8: Brojnost pojedinih malih slova rukopisa Ambroza Haračića generiranih s 10 nasumično izabranih herbarijskih etiketa*




*Slika 9: Brojnost pojedinih velikih slova rukopisa Ambroza Haračića generiranih s 10 nasumično izabranih herbarijskih etiketa*



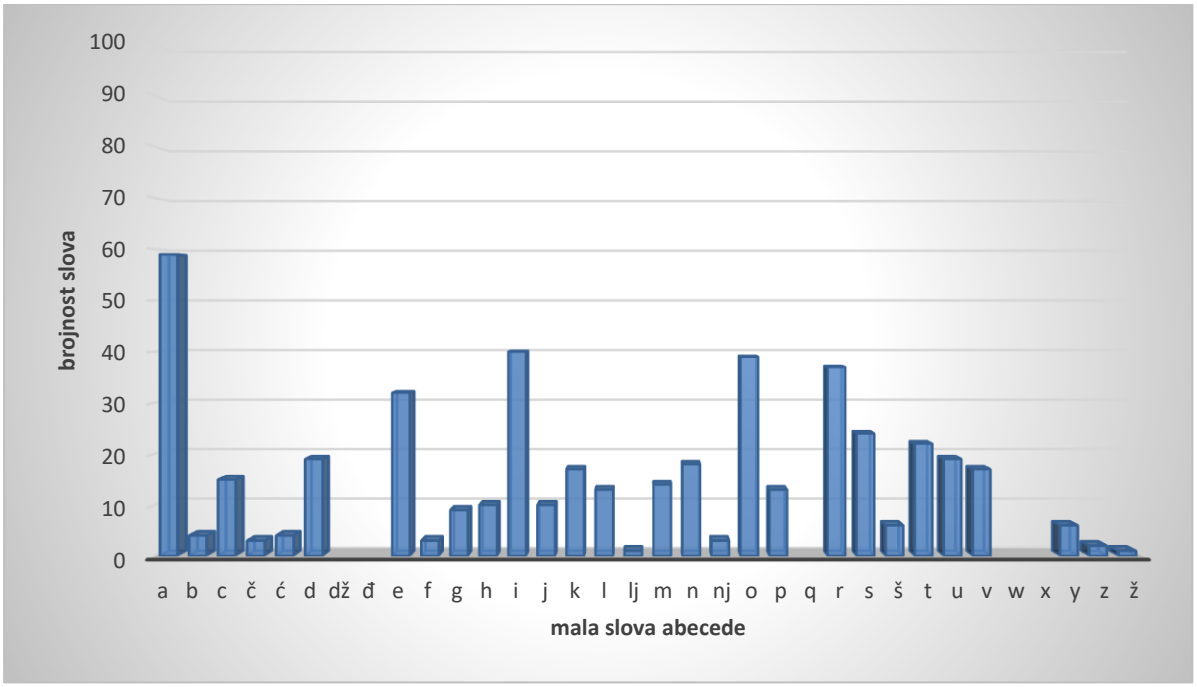

*Slika 10: Abeceda rukopisa Ambroza Haračića dobivena iz herbarijskih etiketa*

### 4.1.3. Analiza rukopisa Dragutina Hirca

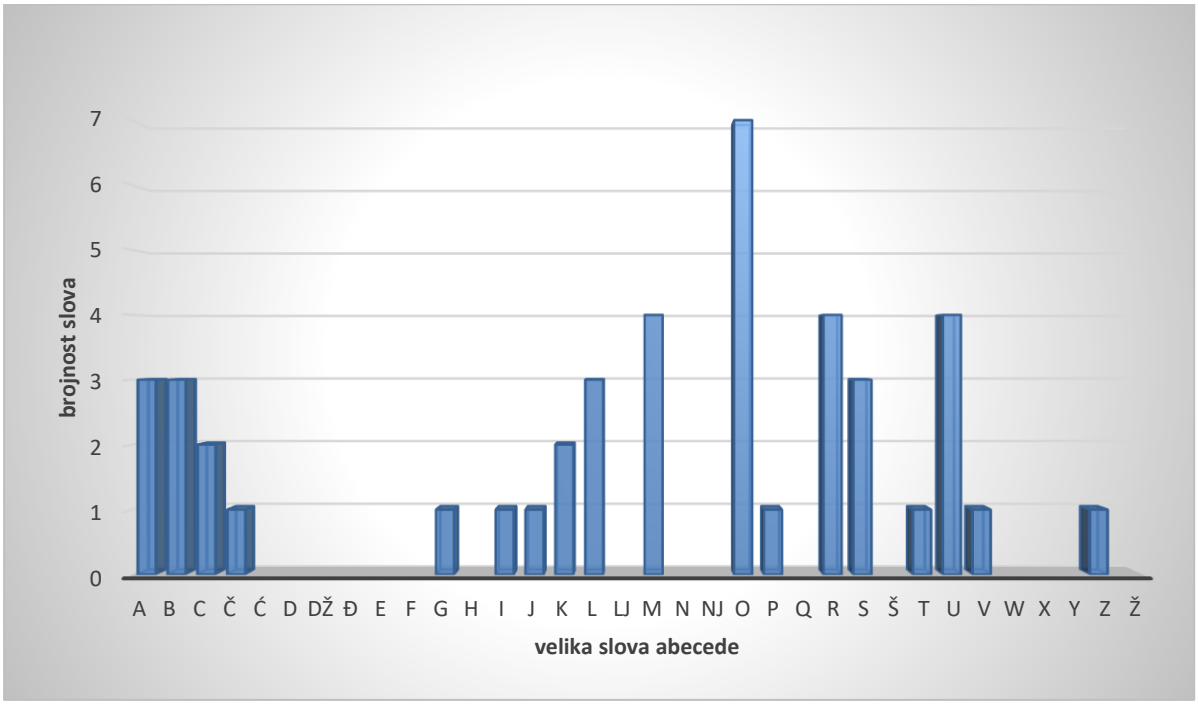
Dragutin Hirc je živio od 1853. do 1921. godine, bio je učitelj, botaničar i zoolog koji se bavio proučavanjem flore, prvenstveno obalnog područja Istre, Kvarnera i otoka, a u manjoj mjeri i ostalih dijelova Hrvatske. Njegova herbarijska zbirka je smještena u herbariju Botaničkog zavoda PMF-a.

Slova su rezana s 40 etiketa, skupljeno je 470 je malih i 77 velikih slova. Od malih slova je najviše slova *a*, a brojnost ostalih slova vidljiva je na slici 11. Od velikih slova najbrojnije je slovo *O*, a brojnost ostalih slova vidljiva je na slici 12.

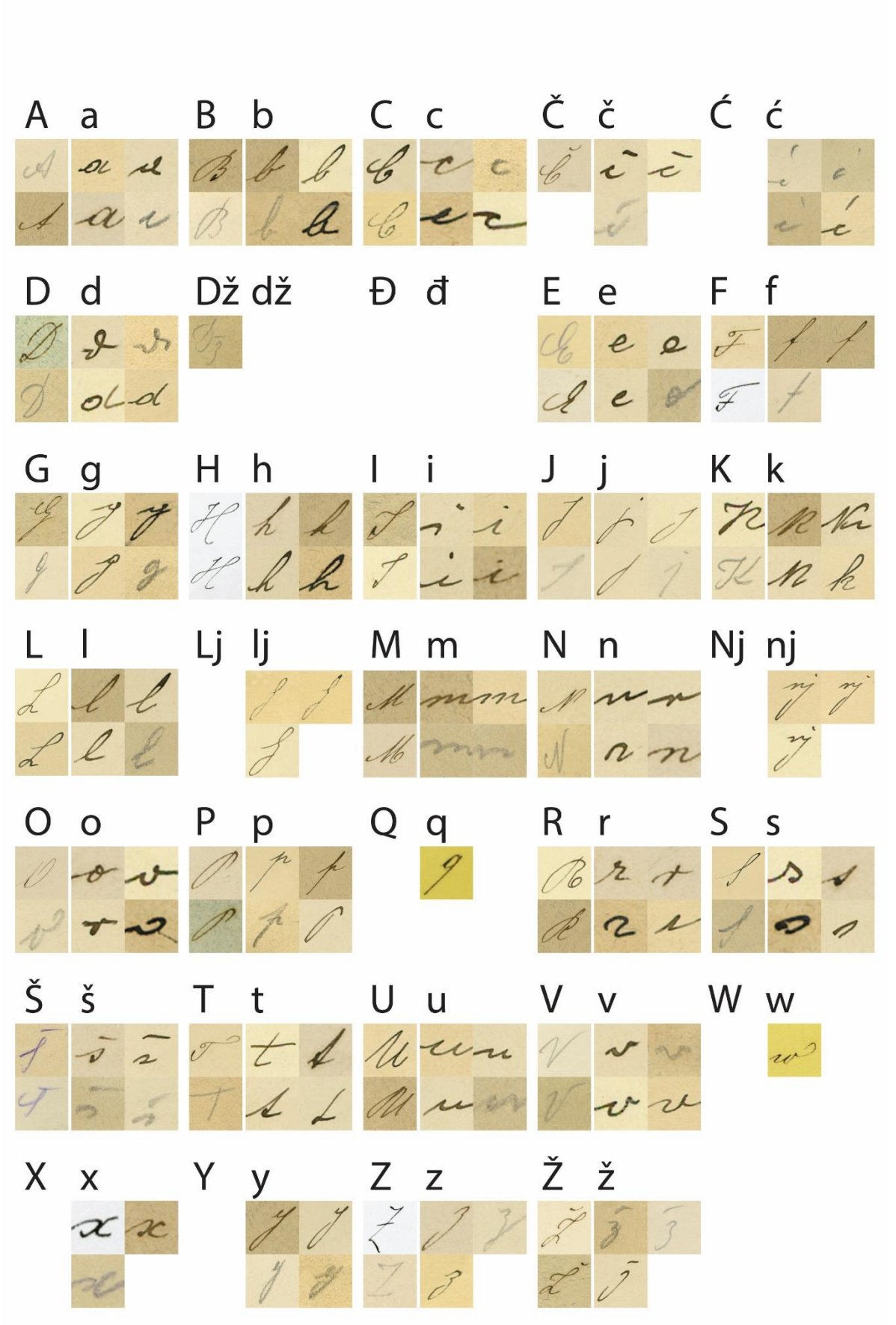
Vizualnim pregledom baze slova rukopisa Dragutina Hirca uočeno je da su neke etikete pisane brzim, teže čitljivim rukopisom, a neke sporijim i jednoličnim, lakše čitljivim slovima. Sukladno tome kada slovo *t* piše iz jednog poteza, ono nema crtice, ali kad je pisano sporije, tada je crtica prisutna. Slova su sitna, uska i znatno ukošena u desno. Riječi imaju dovoljan razmak, dok slova u riječi često imaju malen razmak. Slovo *lj* nema točkicu i više izgleda kao da je jedno slovo. Slovo *r* piše na različite načine, nekad je to samo nekakva krivulja, nekad je kao što se i danas piše, a najčešće je znak karakterističan za 19./20. stoljeće, vidljivo na slici 13.




*Slika 11: Brojnost pojedinih malih slova rukopisa Dragutina Hirca generiranih s 10 nasumično izabranih herbarijskih etiketa*




*Slika 12: Brojnost pojedinih velikih slova rukopisa Dragutina Hirca generiranih s 10 nasumično izabranih herbarijskih etiketa*



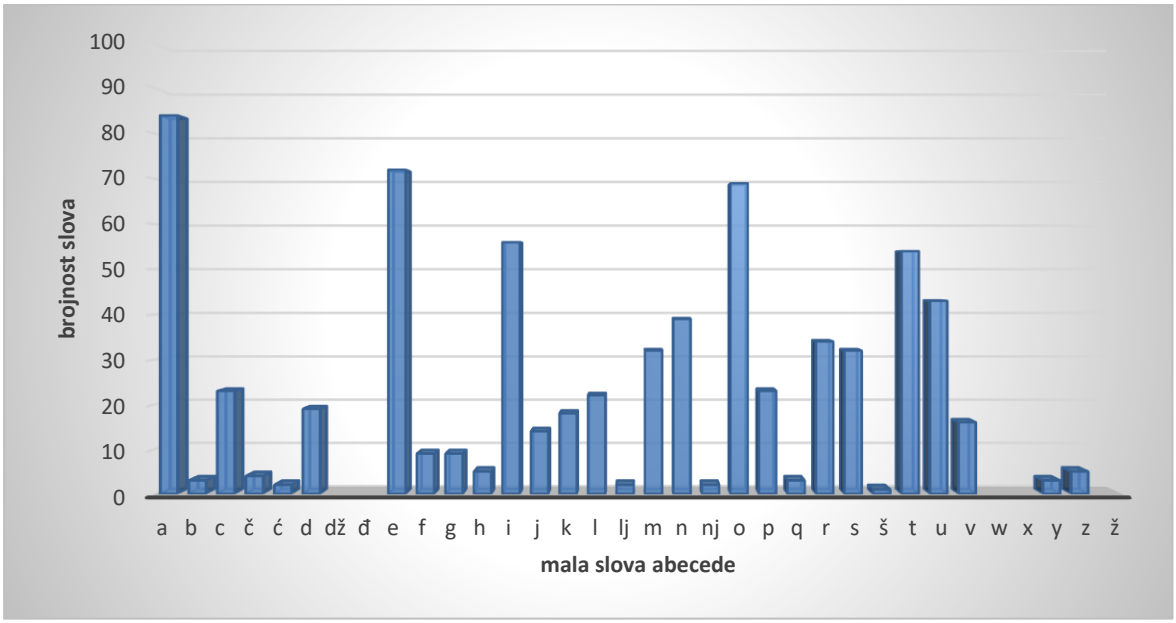

*Slika 13: Abeceda rukopisa Dragutina Hirca dobivena iz herbarijskih etiketa*

### 4.1.4. Analiza rukopisa Stjepana Horvatića

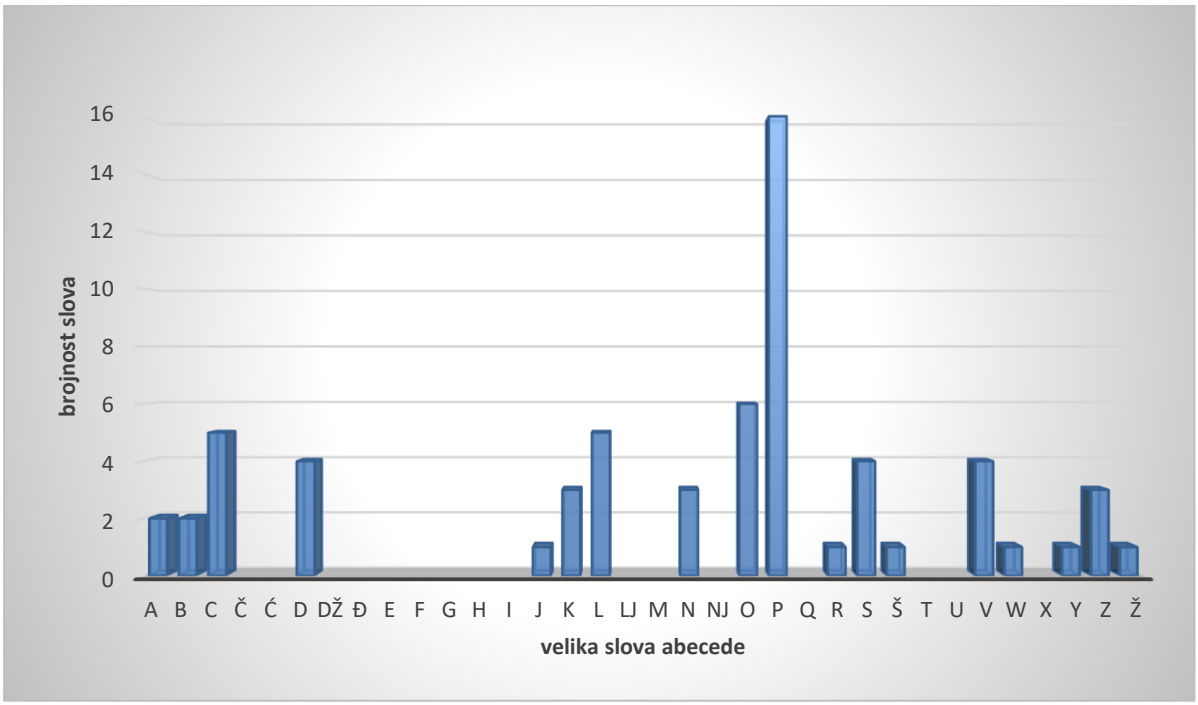
Stjepan Horvatić je živio od 1899. do 1975. godine. Bio je predstojnik Botaničkog zavoda i Botaničkog vrta PMF-a te član HAZU-a. Bavio se istraživanjem livadne i močvarne vegetacije te biljnog pokrova mediteranskog i submeditranskog krškog područja. Napisao je 30-ak znanstvenih radova.

Slova su izrezivana s 15 etiketa te je skupljeno 702 malih slova i 71 veliko slovo, te je napravljena abeceda, vidi sliku 16. Najviše izrezanih malih slova je slova *a*, vidljivo na slici 14, a od velikih slova najviše je izrezanih slova *P*, njih 16, vidljivo na slici 15.

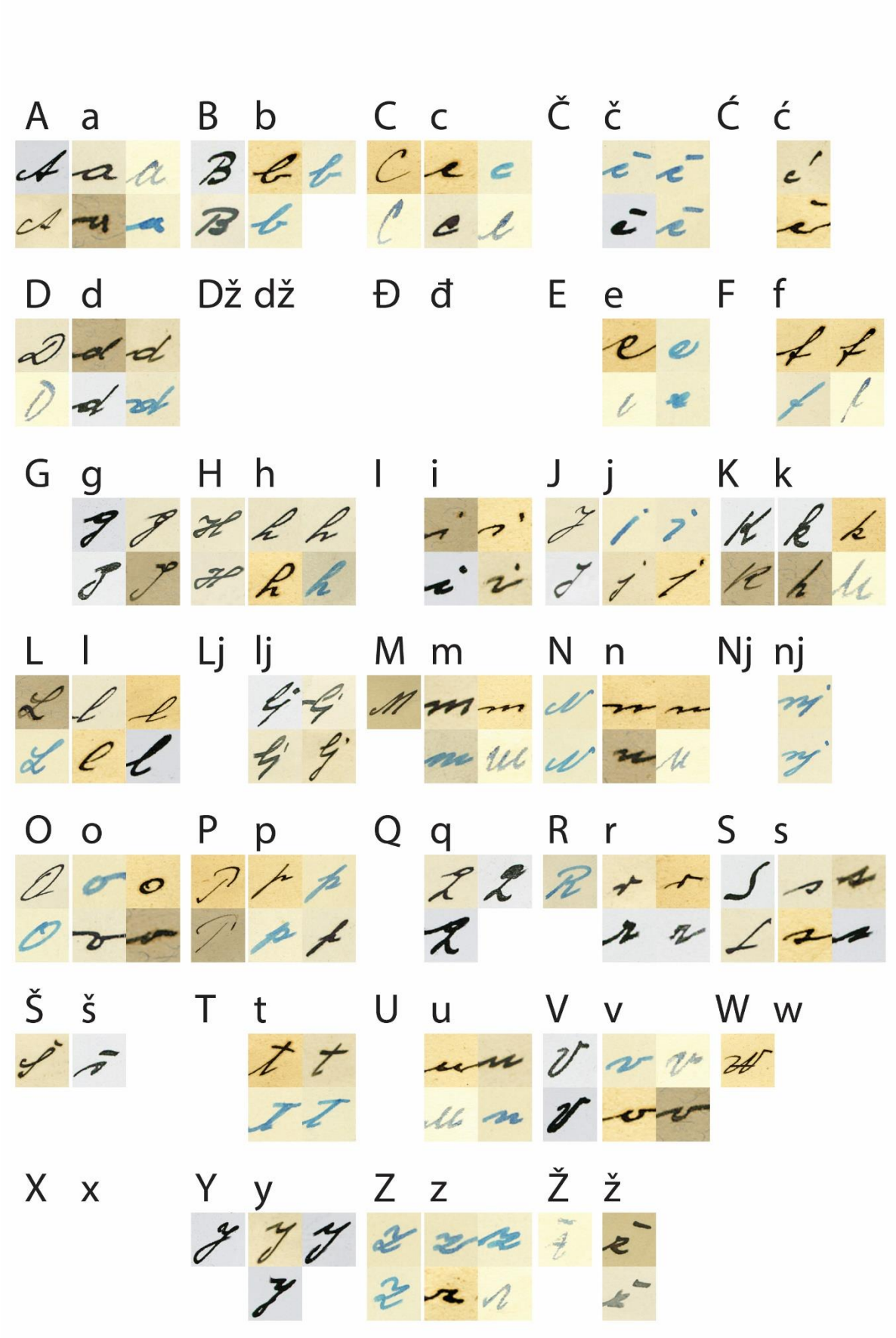
Vizualnim pregledom baze slova rukopisa Stjepana Horvatića uočeno je da piše jednoličnim malim, šiljastim slovima, rukopis mu je znatno ukošen u desno i razvučen. Slova su lako čitljiva, osim kod slova *u*, *m* i *n* koja piše sitno i oštro pa je ponekad teže razaznati koje je slovo. Slova *a* i *o* piše zatvoreno, tako da ih se ne može zamijeniti npr. sa slovima *u* ili *e*. Kod pisanja malog slova *r* koristi dva načina pisanja.




*Slika 14: Brojnost pojedinih malih slova rukopisa Stjepana Horvatića generiranih s 10 nasumično izabranih herbarijskih etiketa*




*Slika 15: Brojnost pojedinih velikih slova rukopisa Stjepana Horvatića generiranih s 10 nasumično izabranih herbarijskih etiketa*



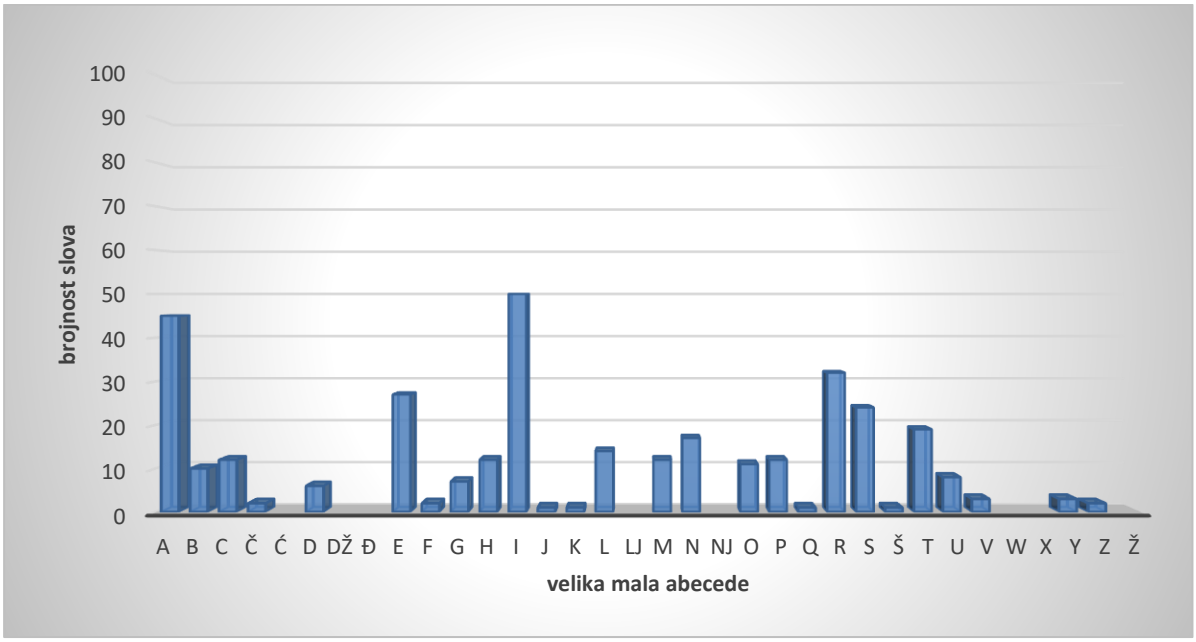

*Slika 16: Abeceda rukopisa Stjepana Horvatića dobivena iz herbarijskih etiketa*

### 4.1.5. Analiza rukopisa Bohuslava Jiruša

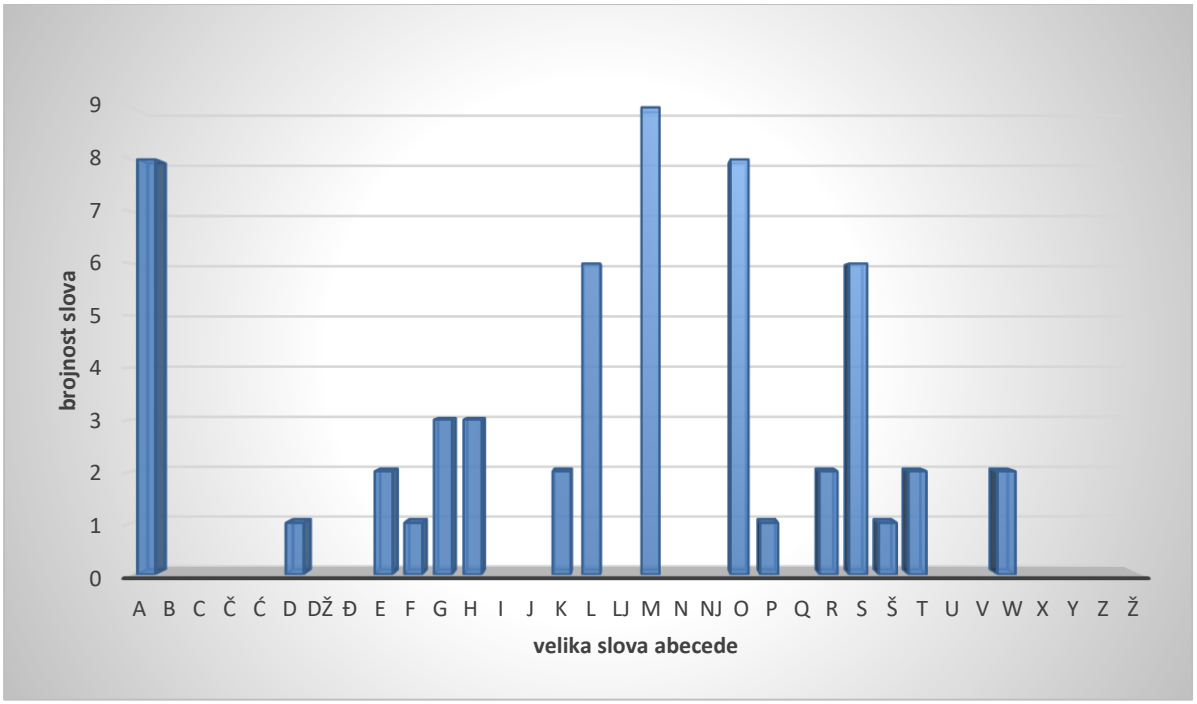
Bohuslav Jiruš rođen je u Pragu, a živio je od 1841. do 1901. godine. Doktorirao je medicinu, a 1875. godine je izabran za prvog redovitog profesora botanike na Sveučilištu u Zagrebu. Utemeljio je botaničko-fiziološki zavod na tadašnjem Mudroslovnom fakultetu, osnovao je herbarijsku zbirku i postavio temelje zavodske biblioteke, a od njega potječe i zamisao o osnivanju Botaničkog vrta.

Za izrezivanje su korištene 24 etikete te je izrezano 354 malih i 81 veliko slovo. Brojnost malih slova vidljiva je na slici 17, a velikih na slici 18. Od izrezanih slova složena je abeceda na slici 18.

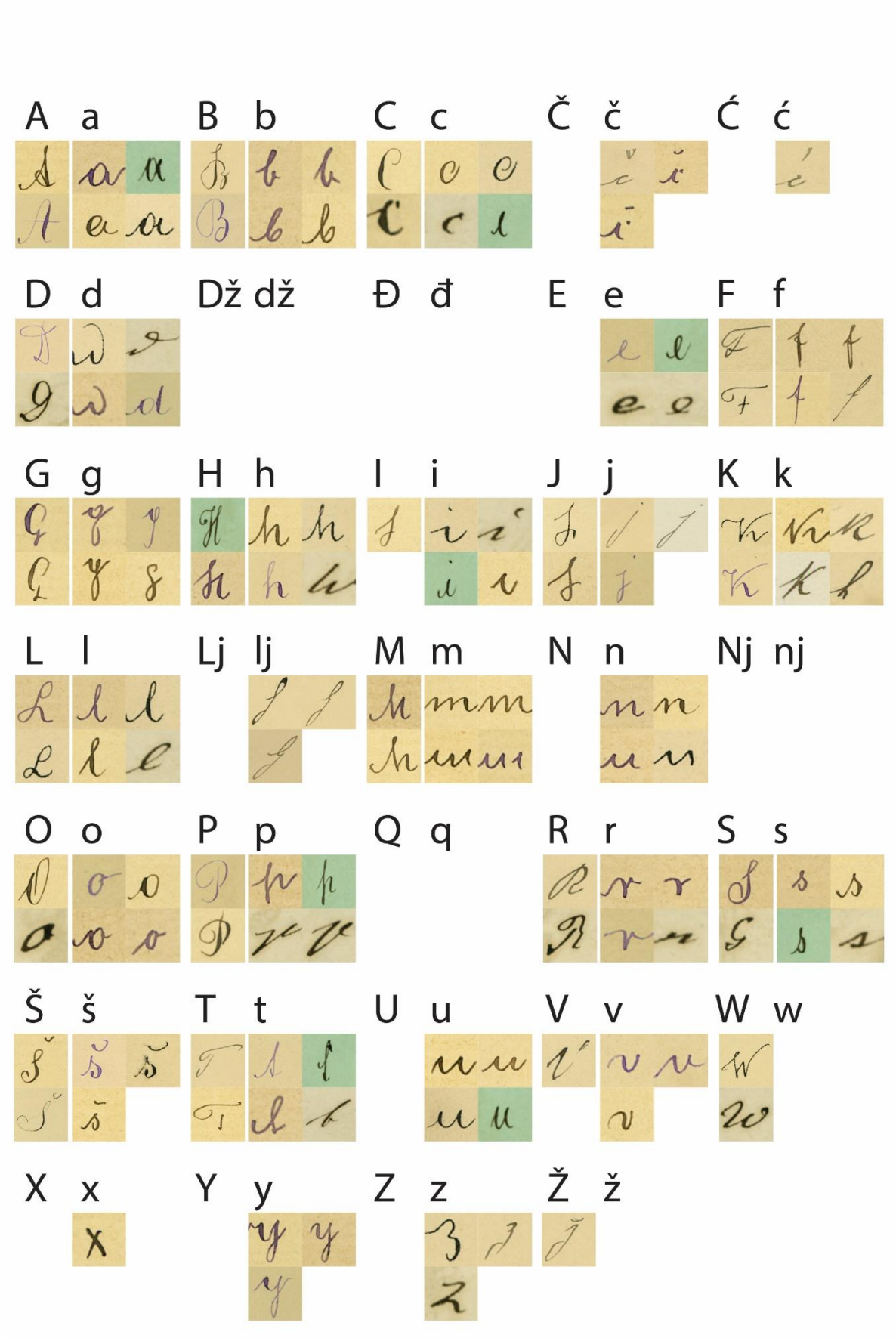
Vizualnim pregledom baze slova rukopisa Bohuslava Jiruša uočeno je da koristi dva različita rukopisa, krasopis i obični rukopis. Kod etiketa koje su pisane krasopisom slova su jednolična, uspravna, velika i okrugla te lako čitljiva. Kod drugog rukopisa, koji je brže pisan, slova su teže čitljiva, ukošena u desno te su česti razmaci između pojedinih slova. Slova *u* i *n* su ponekad ista, što otežava njihovo određivanje. Slovo *lj* piše kao i Dragutin Hirc, *j* nije vidljivo i nema točkice tako da izgleda kao jedno slovo.




*Slika 17: Brojnost pojedinih malih slova rukopisa Bohuslava Jiruša generiranih s 10 nasumično izabranih herbarijskih etiketa*




*Slika 18: Brojnost pojedinih velikih slova rukopisa Bohuslava Jiruša generiranih s 10 nasumično izabranih herbarijskih etiketa*



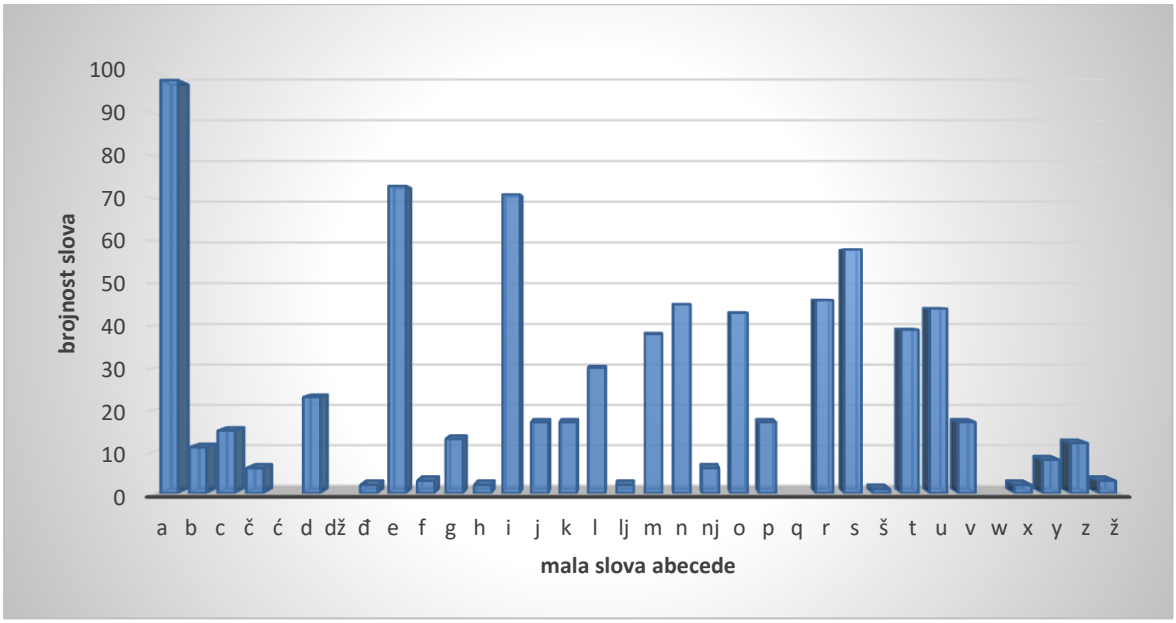

*Slika 19: Abeceda rukopisa Bohuslava Jiruša dobivena iz herbarijskih etiketa*

### 4.1.6. Analiza rukopisa Miška Plazibata

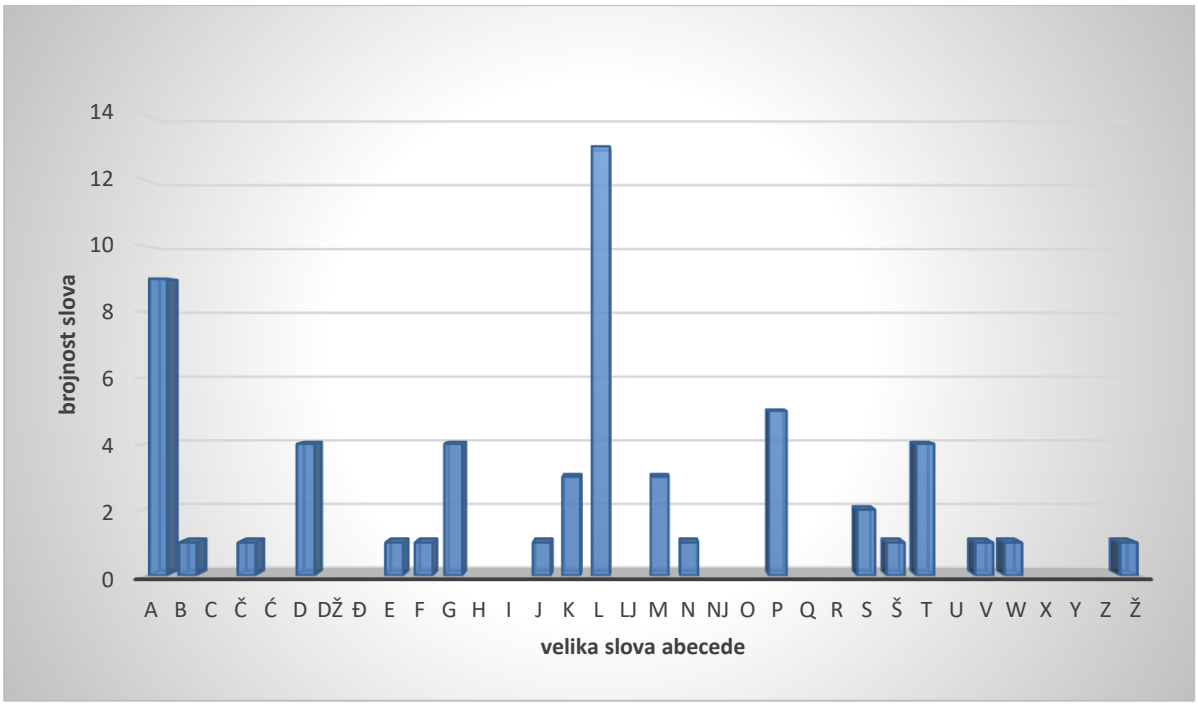
Miško Plazibat hrvatski je biolog, rođen 1949. godine. Bavio se stručnim radom u herbarijskoj zbirci te znanstvenim radom. Objavio je nekoliko radova iz područja virologije, vegetacije i flore. Kao stručni suradnik – voditelj herbarijskih zbirki bio je u stalnom radnom odnosu u Botaničkom zavodu do 2014. godine.

Ukupno je izrezano 773 malih slova, od kojih je najviše slova *a* (njih čak 98), vidljivo na slici 20. Velik slova je skupljeno 81, najviše je slova *L*, vidljivo na slici 21. Od dostupnih slova složena je abeceda na slici 22.

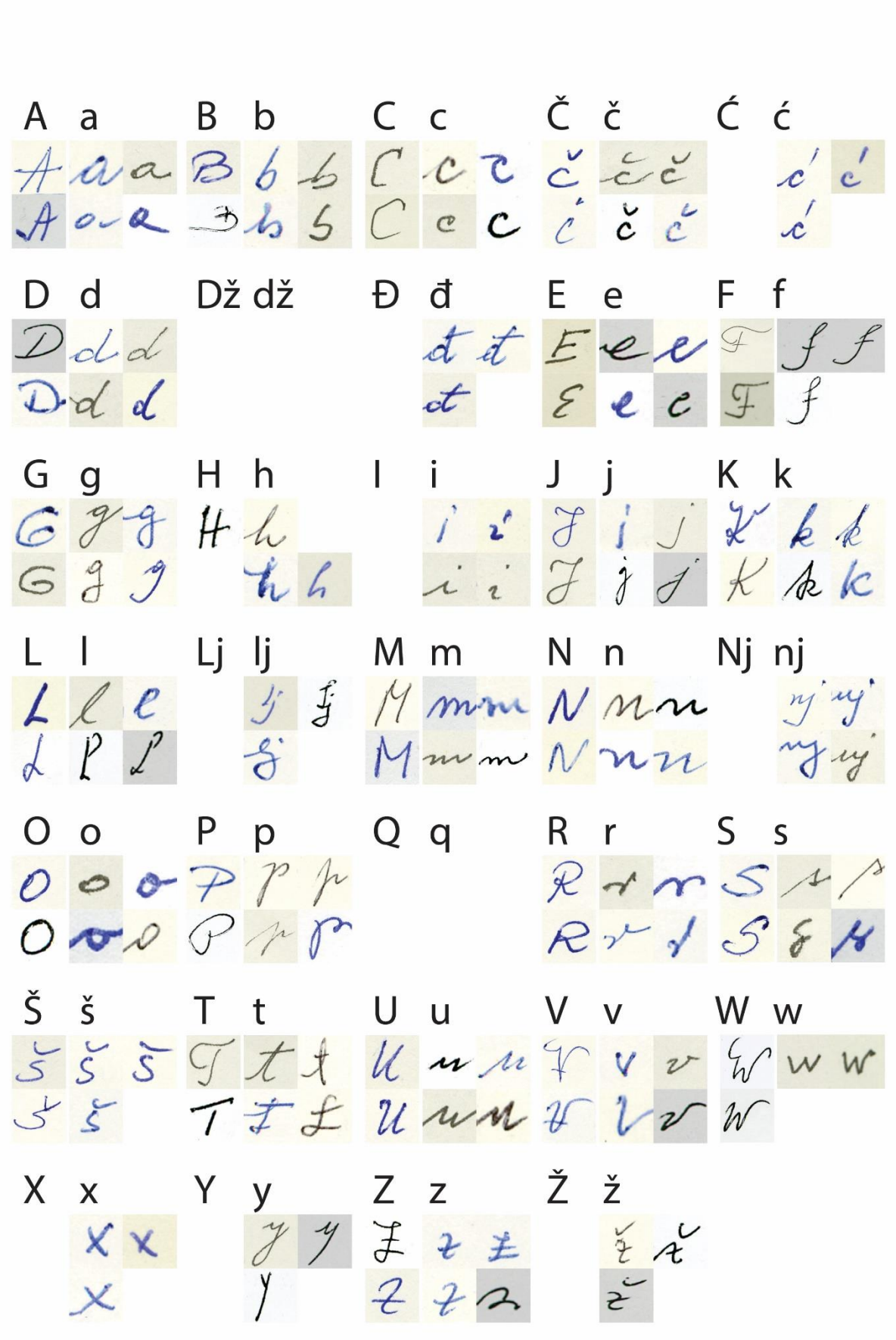
Vizualnim pregledom baze slova rukopisa Miška Plazibata uočeno je da uz pisana slova koristi i velika tiskana slova, rukopis mu je čitljiv, iako slova nisu ujednačene veličine, piše blago nagnuto u desno. Kod pisanja slova *s* koristi u riječima i pisano *s* i tiskano *s*. Preko velikog slova *V* piše crticu na koju se nastavlja sljedeće slovo u riječi.




*Slika 20: Brojnost pojedinih malih slova rukopisa Miška Plazibata generiranih s 10 nasumično izabranih herbarijskih etiketa*




*Slika 21: Brojnost pojedinih velikih slova rukopisa Miška Plazibata generiranih s 10 nasumično izabranih herbarijskih etiketa*



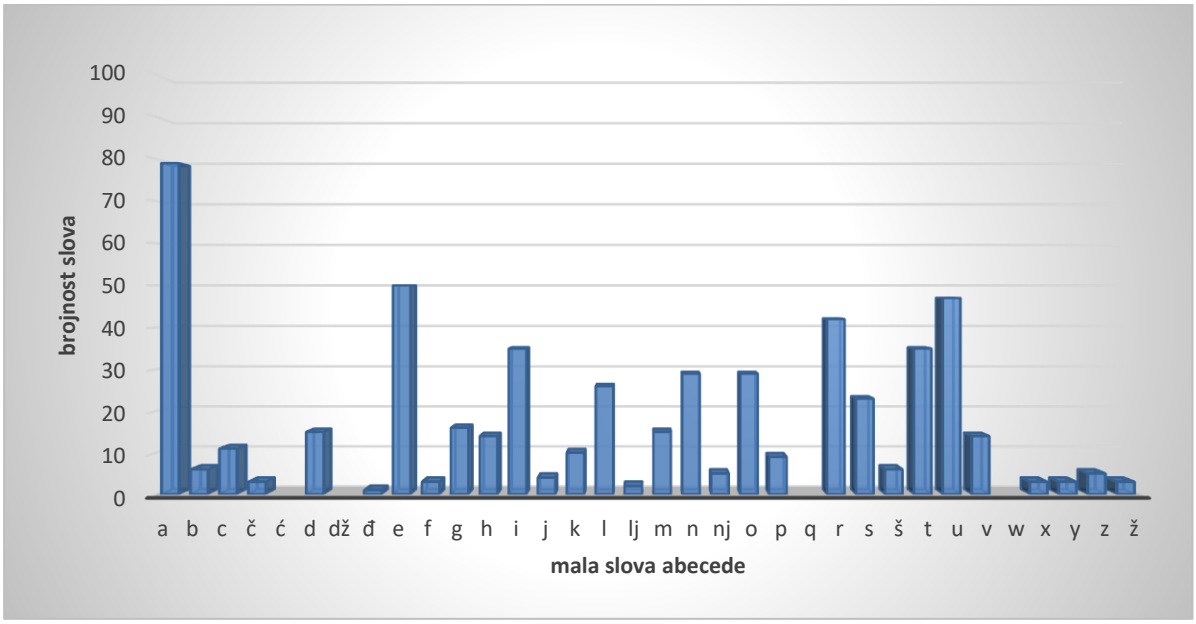

*Slika 22: Abeceda rukopisa Miška Plazibata dobivena iz herbarijskih etiketa*

### 4.1.7. Analiza rukopisa Ljerke Regula-Bevilacqua

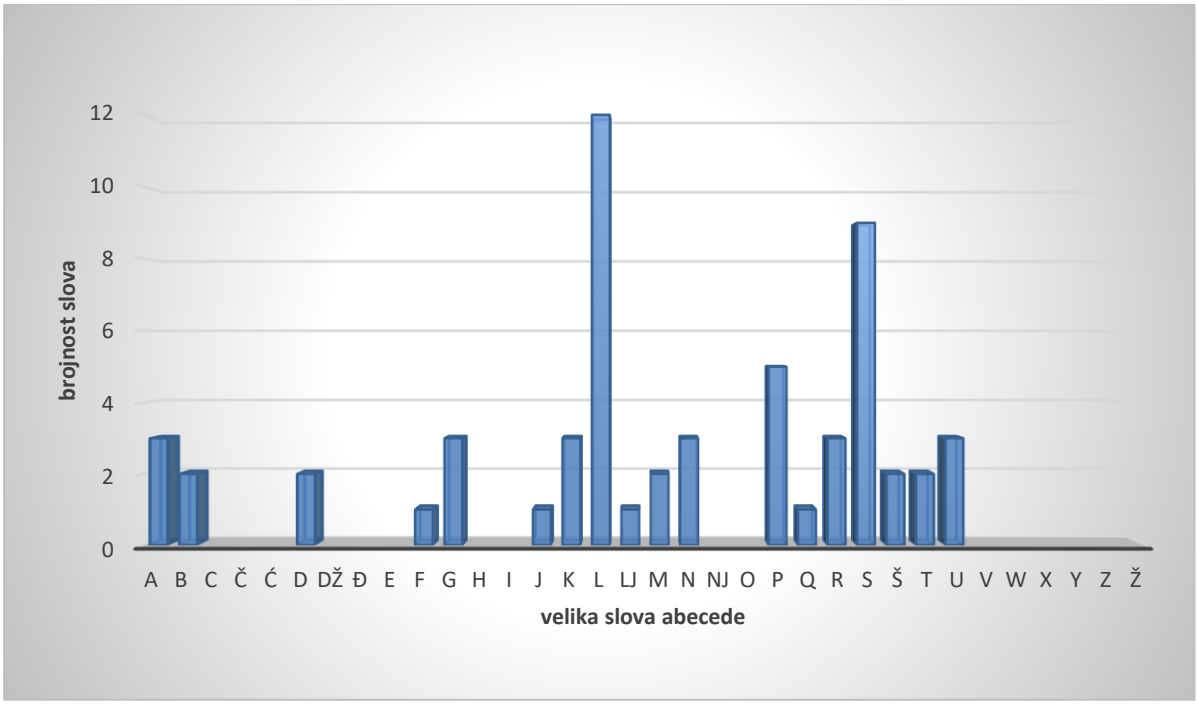
Ljerka Regula-Bevilacqua rođena je u Zagrebu 1934. godine, studirala je i doktorirala biologiju u Zagrebu. Bavila se geobotaničkim istraživanjima. Bila je dugi niz godina ravnateljica Botaničkog vrta u Zagrebu, objavila je 22 znanstvena i oko 20 stručnih radova, većinom u časopisu Acta Botanica Croatica.

Uz 10 osnovnih etiketa za izrezivanje slova korišteno i je i 27 dodatnih etiketa, tako da je s ukupno 37 etiketa izrezano 554 malih slova, od kojih je najviše slova *a*, vidljivo i na slici 23., te 87 velikih slova. Najbrojnije veliko slovo je *L*, vidljivo na slici 24. Na slici 25 prikazana je abeceda od skupljenih slova.

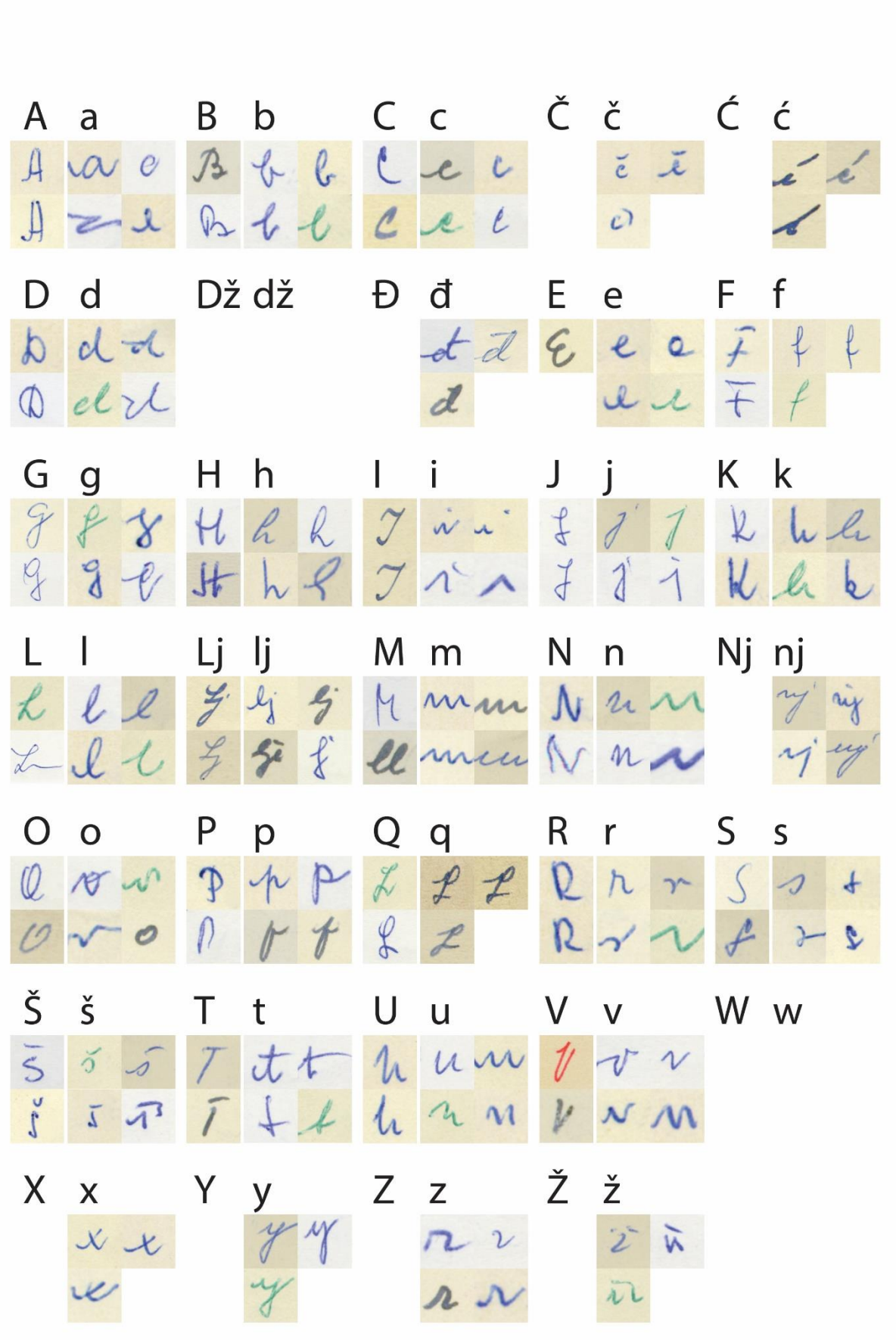
Vizualnim pregledom baze slova rukopisa Ljerke Regula-Bevilacqua uočeno je da piše uspravnim do blago u desno nakošenim rukopisom, slova nisu jednake veličine, mala su i obla. Kod slova *u*, *n* i *r* ponekad je teško odrediti o kojem se slovu radi. Veliko slovo *M* izgleda kao dva mala slova *l*, slovo *a* ponekad sliči slovu *e*, veliko slovo *U* podsjeća na malo slovo *h*, itd.




*Slika 23: Brojnost pojedinih malih slova rukopisa Ljerke Regula-Bevilacqua generiranih s10 nasumično izabranih herbarijskih etiketa*




*Slika 24: Brojnost pojedinih velikih slova rukopisa Ljerke Regula-Bevilacqua generiranih s 10 nasumično izabranih herbarijskih etiketa*



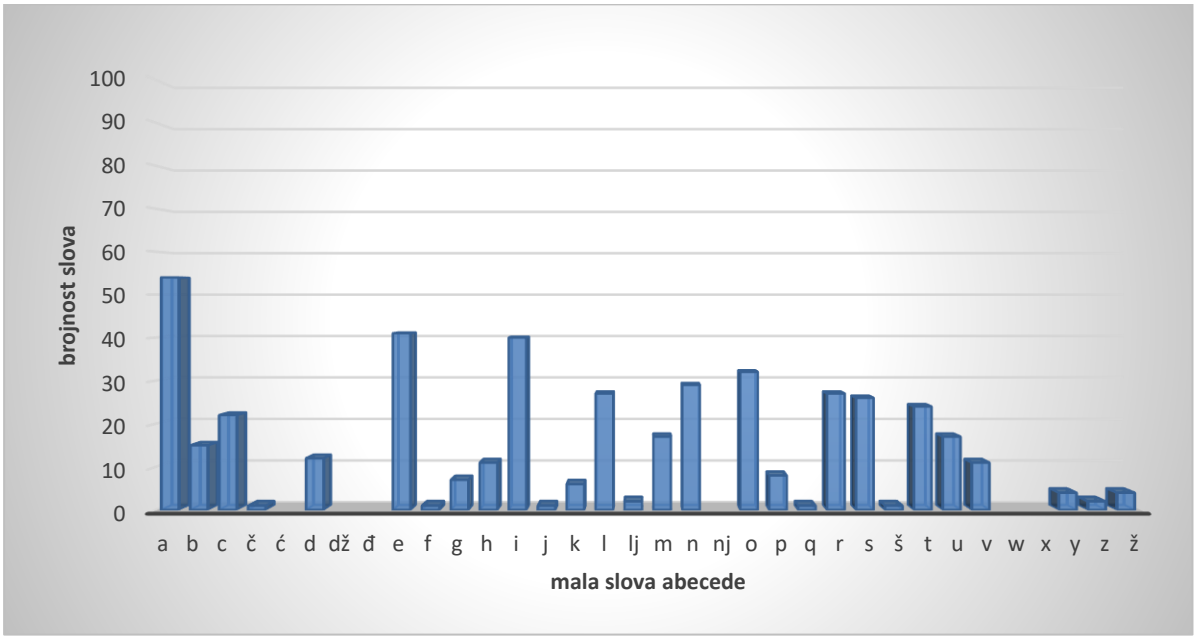

*Slika 25: Abeceda rukopisa Ljerke Regula-Bevilacqua dobivena iz herbarijskih etiketa*

### 4.1.8. Analiza rukopisa Ljudevita Rossia

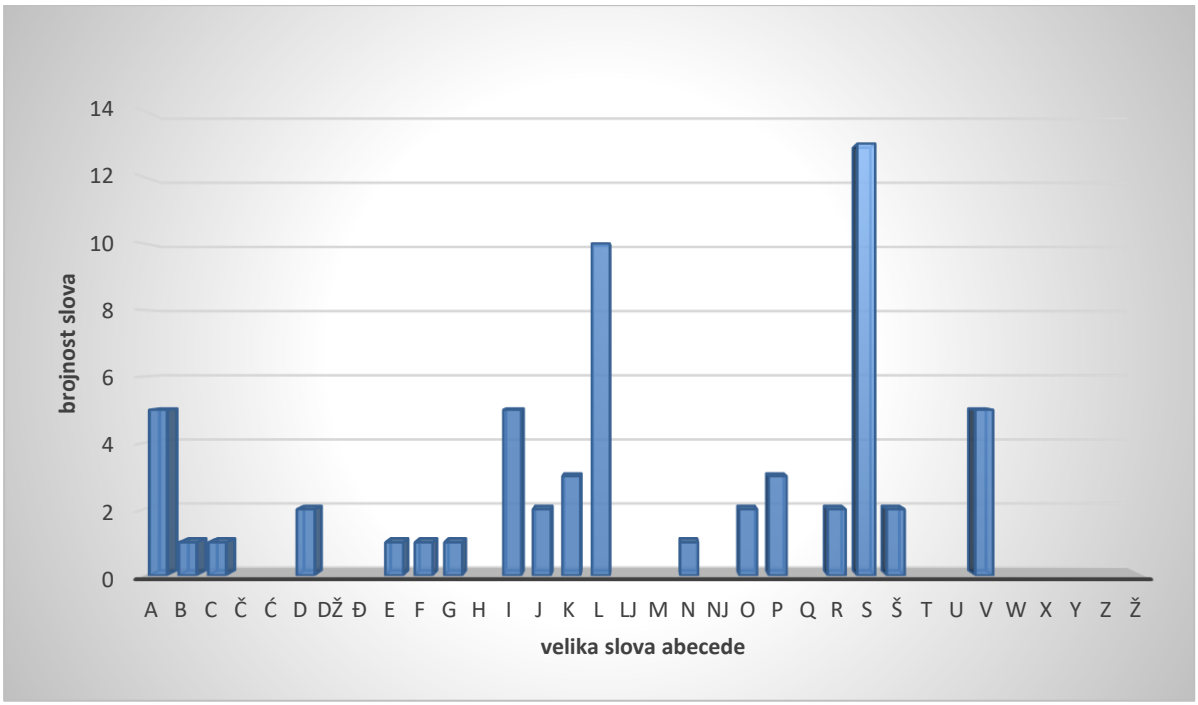
Ljudevit Rossi živio je od 1850. do 1932. godine. Amaterski je počeo sa skupljanjem biljaka i istraživanjem hrvatske flore. Autor je važnog botaničkog djela *Flora Dalmatica*. Posebno je značajno djelo njegova herbarijska zbirka *(Herbarium Croaticum Rossianum)*, koja sadrži oko 30.000 herbarijskih listova, a danas je dio herbarijske zbirke Botaničkog zavoda.

Iz 37 herbarijskih etiketa izrezana su 464 mala slova i 103 velika slova. Od malih slova najviše je slova *a*, vidljivo na slici 26. Od velikih slova prikazanih na slici 27 najviše je izrezano slova *S*, dok mnoga druga slova nedostaju. Na slici 28 prikazana je abeceda od skupljenih slova.

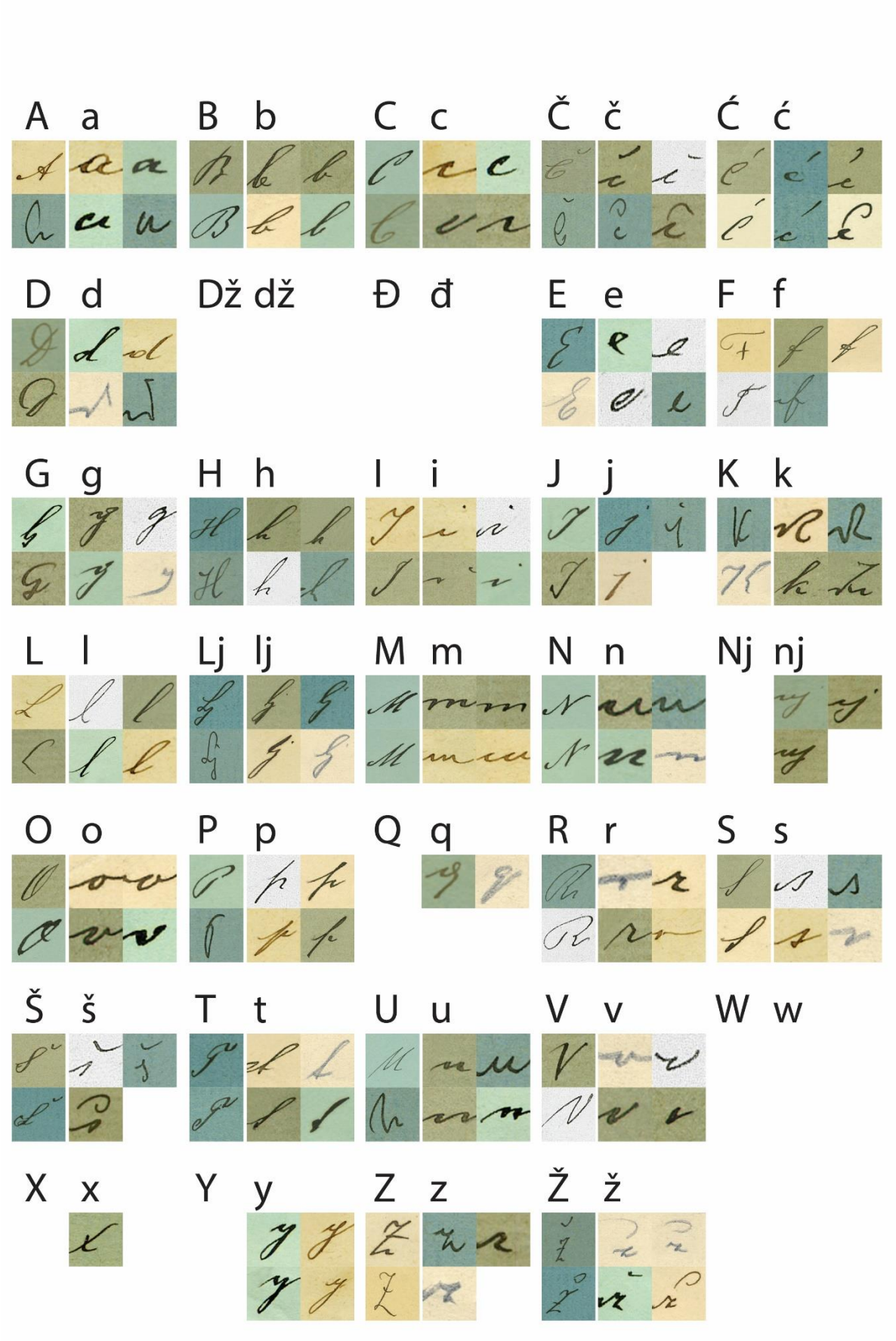
Vizualnim pregledom baze slova rukopisa Ljudevita Rossia uočeno je da piše na latinskom jeziku. Kod pisanja velikih i malih slova razlika u veličini je jako naglašena, mala slova su jako sitna dok su velika višestruko veća uključujući i petlju kod slova *l* i ostalih slova koja su u visini velikih slova. Zbog specifičnog pisanja izrazito sitnih slova rukopis je teže čitljiv, iako je jednolik i slova su konstantne veličine. Primjerice, teško je odrediti koje je koje slovo *u*, *n*, ili *o*. Slovo *a* nije zatvoreno, slovo *t* piše iz jednog poteza te nema crticu, nego je crtica jako nisko i samo je nastavak za sljedeće slovo u riječi.




*Slika 26: Brojnost pojedinih malih slova rukopisa Ljudevita Rossia generiranih s 10 nasumično izabranih herbarijskih etiketa*




*Slika 27: Brojnost pojedinih velikih slova rukopisa Ljudevita Rossia generiranih s 10 nasumično izabranih herbarijskih etiketa*



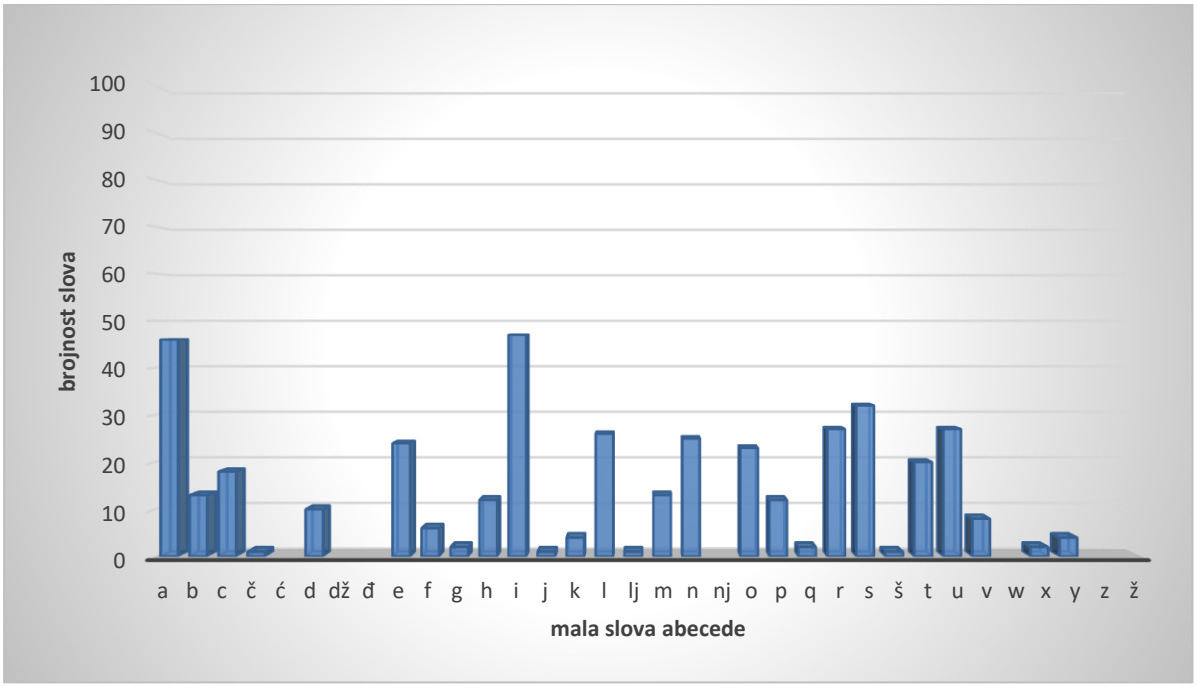

*Slika 28: Abeceda rukopisa Ljudevita Rossia dobivena iz herbarijskih etiketa*

### 4.1.9. Analiza rukopisa Josipa Kalasancija Schlossera-Klekovskog

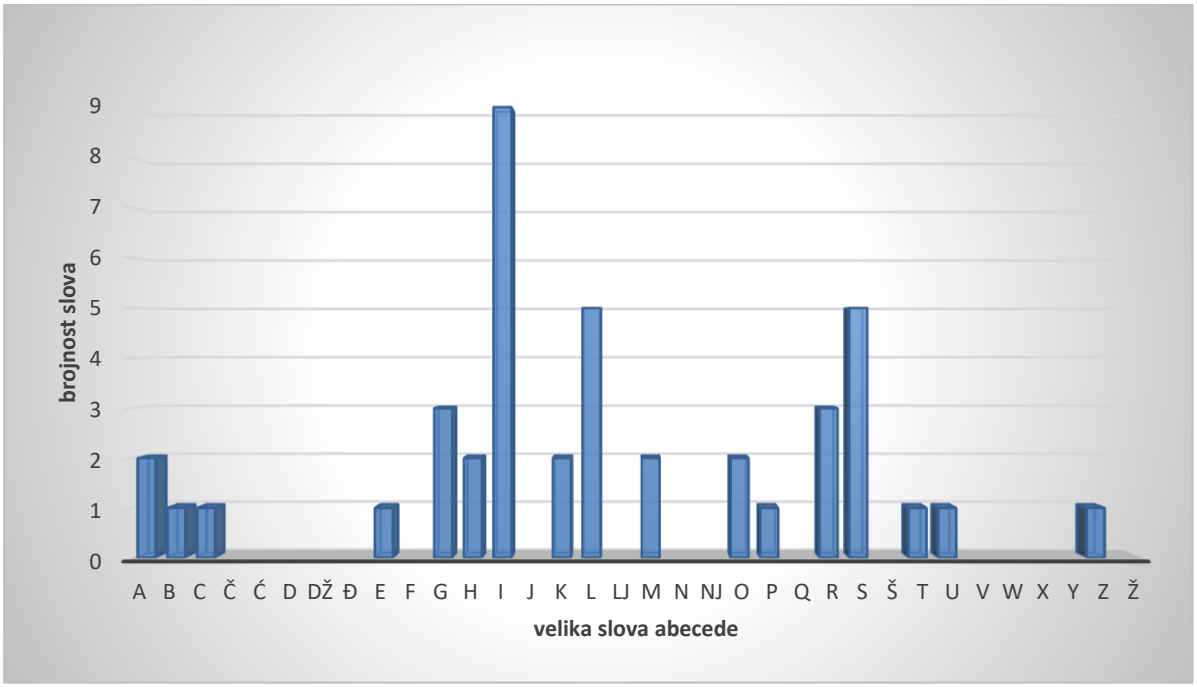
Josip Kalasancije Schlosser-Klekovski, rođen je u Češkoj 1808. godine, završio je medicinu u Beču, a od 1836. živio je u Hrvatskoj. Bavio se ornitologijom i florom Hrvatske, zajedno s Ljudevitom Vukotinovićem je objavio 3 knjige o flori Hrvatske, od kojih je najznačajnija *Flora Croatica*. Dio svoje botaničke zbirke poklonio je Narodnom muzeju Zagrebu, što je postao temelj današnje herbarijske zbirke Herbarium Croaticum.

Slova su izrezivana s ukupno 51 etikete te je izrezano 432 mala slova i 75 velikih slova. Od malih slova najviše je slova *i* (47) te slova *a* (46), što je vidljivo i na slici 29. Od velikih slova također je najbrojnije slovo *I*, vidljivo na slici 30.

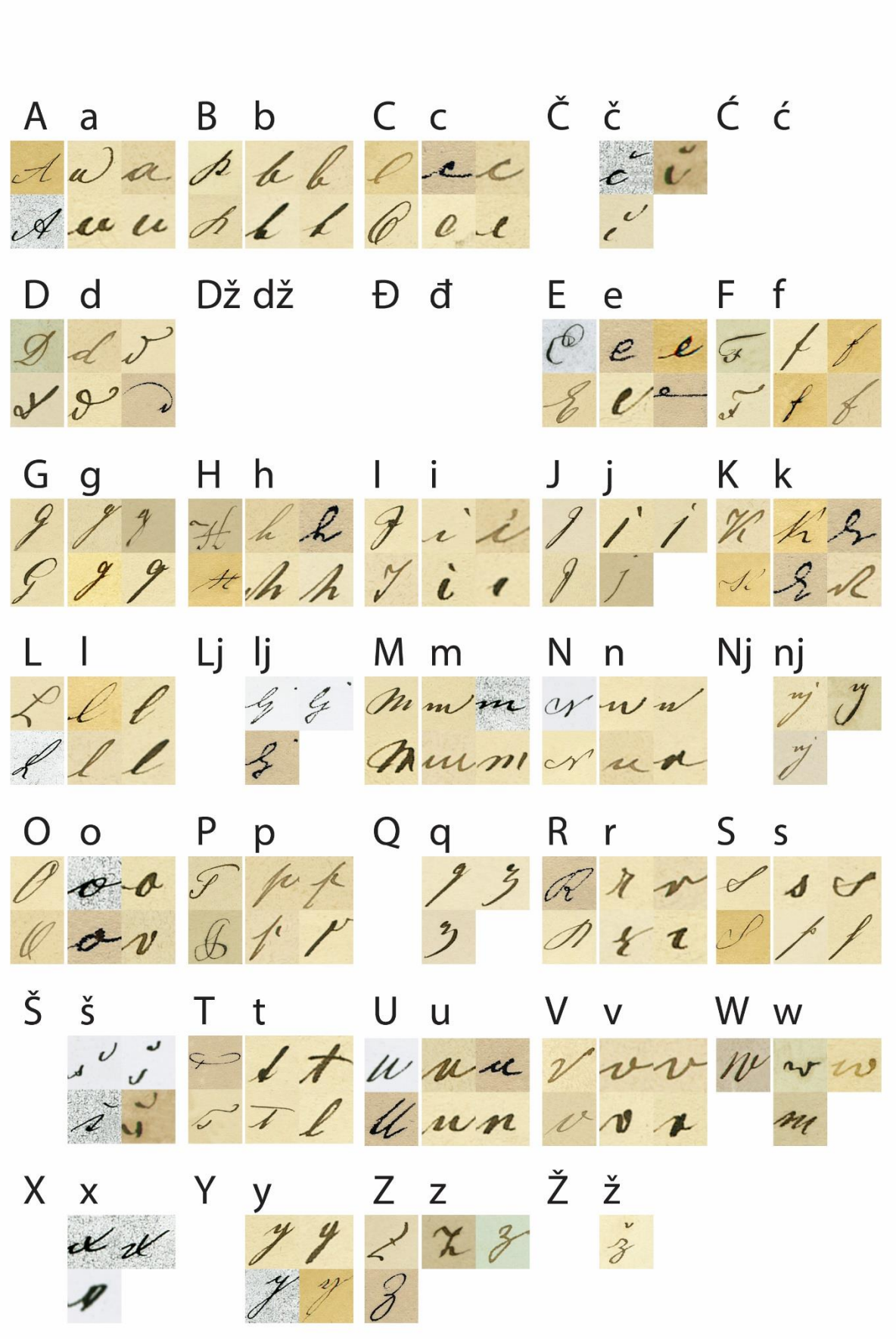
Vizualnim pregledom baze slova rukopisa Josipa Schlossera uočeno je da su mu slova značajno nagnuta u desnu stranu te su velike razlike u veličini malih i velikih slova. Piše na latinskom jeziku, rjeđe na hrvatskom ili njemačkom. Slova u riječi često imaju dosta velik razmak, nisu spojena, slova su sitna i teže ih je pročitati. Među slovima *a, u* i *n* teško je odrediti razliku, jer ih ponekad piše vrlo slično. Veliko slovo *P* ne sliči na tipično slovo *P*, dok malo slovo *k* sliči velikom slovu *R*. Općenito je vrlo velika varijabilnost u pisanju istih slova što je vidljivo i u abecedi na slici 31.




*Slika 29: Brojnost pojedinih malih slova rukopisa Josipa Schlossera generiranih s 10 nasumično izabranih herbarijskih etiketa*




*Slika 30: Brojnost pojedinih velikih slova rukopisa Josipa Schlossera generiranih s nasumično izabranih 10 herbarijskih etiketa*



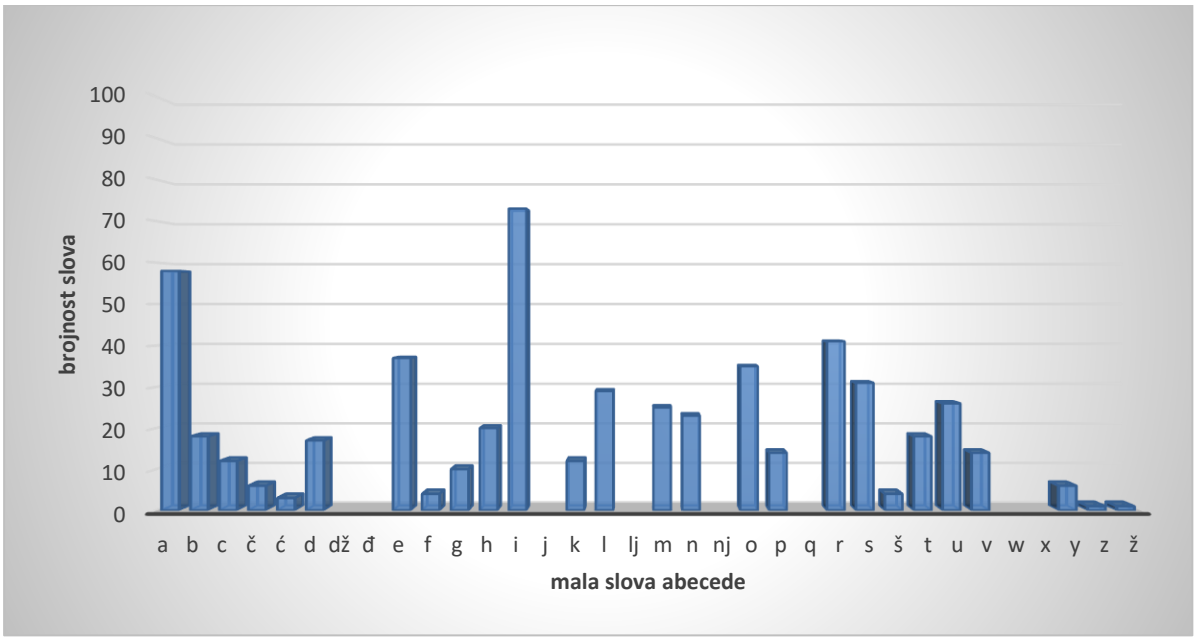

*Slika 31: Abeceda rukopisa Josipa Schlossera dobivena iz herbarijskih etiketa*

### 4.1.10. Analiza rukopisa Ljudevita Vukotinovića

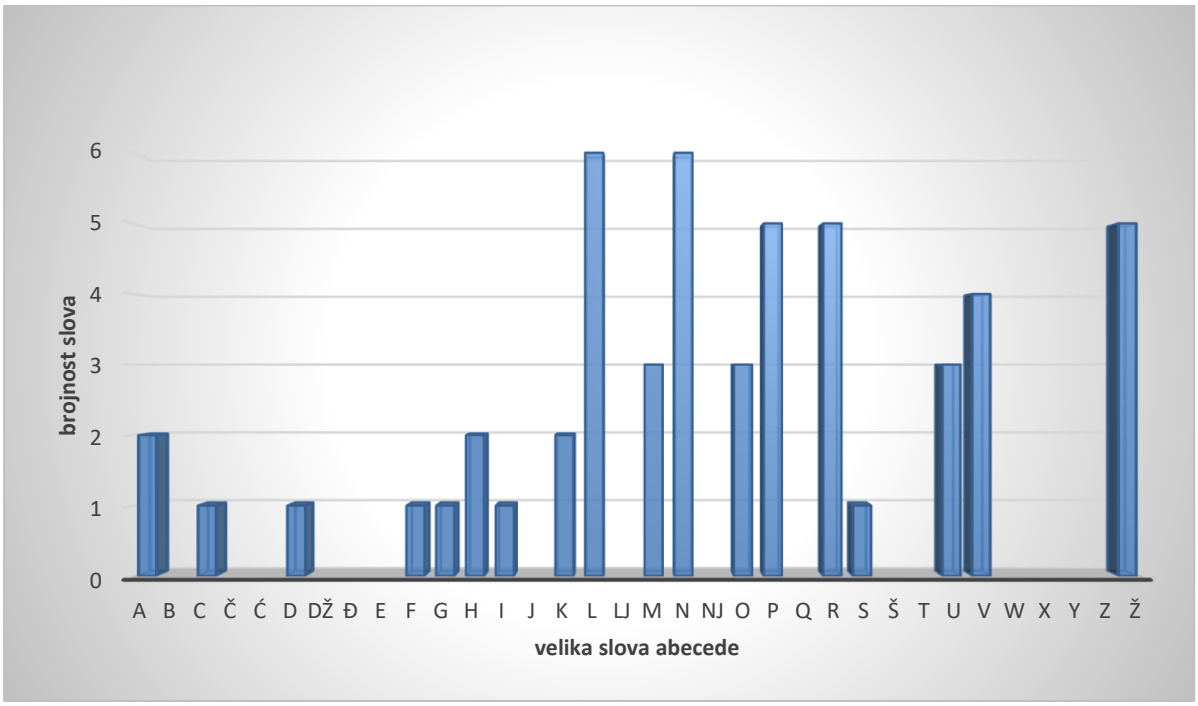
Bio je hrvatski prirodoslovac, živio je u razdoblju od 1813. do 1893. godine. Za vrijeme njegovog rada u Narodnom muzeju utjecao je na povećanje prirodoslovne zbirke i na njezino sređivanje. Baveći se botanikom istraživao je biljni pokrov Hrvatske te objavio 3 djela zajedno s kolegom Josipom Kalasancijem Schlosser-Klekovskim.

Slova su skupljana sa 41 etikete te je skupljeno 554 malih i 88 velikih slova. Od malih slova najviše je skupljeno slova *i*, brojnost ostalih malih slova vidljiva je na slici 32. Kod velikih slova najviše je skupljeno slova *L* i *N* (slika 33). Na slici 34 prikazana je abeceda sastavljena od izrezanih slova.

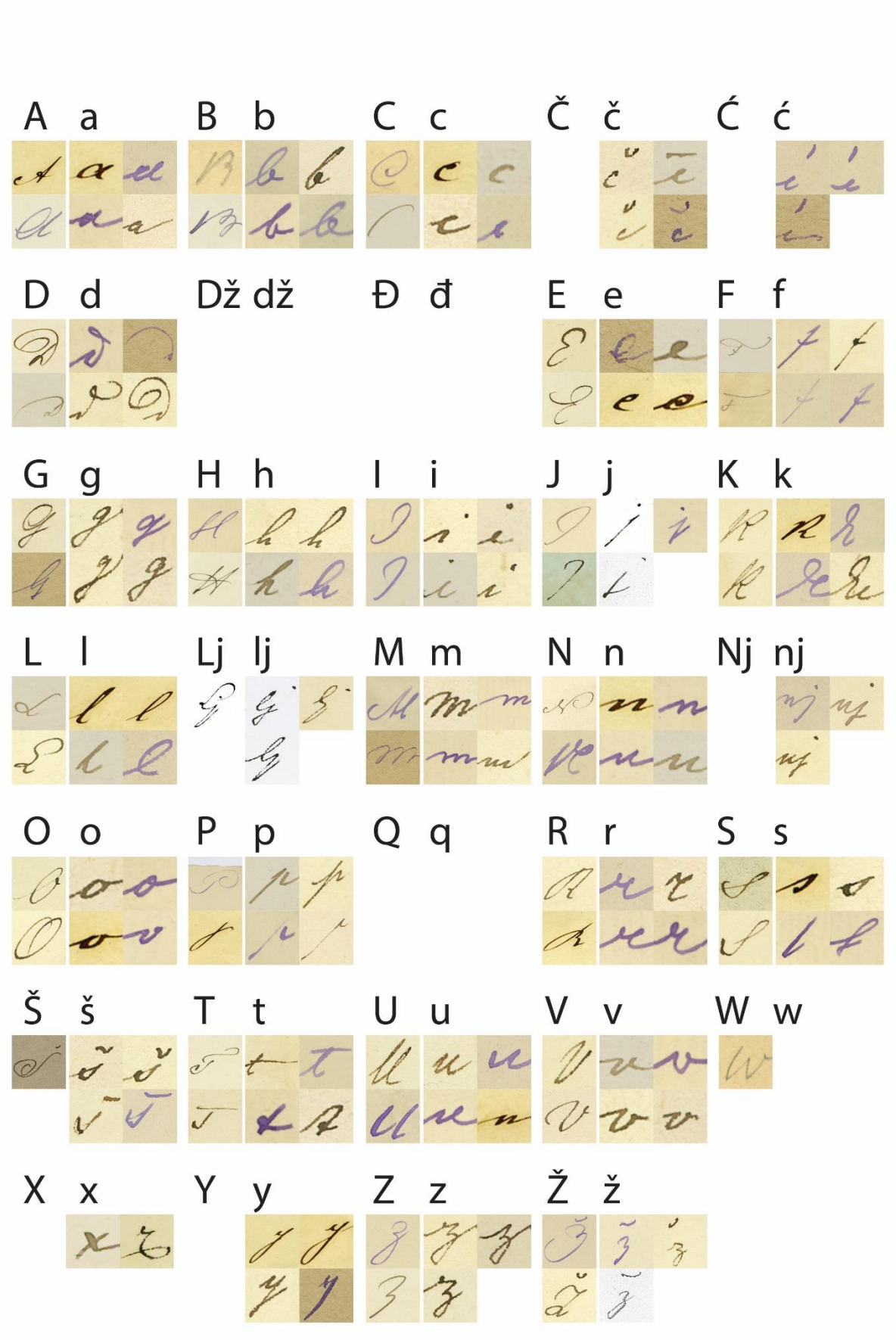
Vizualnim pregledom baze slova rukopisa Ljudevita Vukotinovića uočeno je da slova piše ukošeno u desno, slova su međusobno prilično različita i rijetko da je isto slovo jednako već prethodno napisanom. Etikete su pisane na latinskom i hrvatskom jeziku, rukopis je općenito težak za čitanje. Slovo *d* piše jako specifično s ukrasnim zavijutkom nekad na lijevu, nekad na desnu stranu, dok veliko slovo *D* također ima zavijutak. Općenito je kod velikih slova naglašen produžetak slova ili je napravljen zavijutak, na način da se dio slova proteže preko pola riječi. U pisanju koristi *dj* umjesto slova *đ*. Malo slovo *k* sliči na veliko slovo *R*, a slova *u* i *n* je teško razlikovati.




*Slika 32: Brojnost pojedinih malih slova rukopisa Ljudevita Vukotinovića generiranih s 10 nasumično izabranih herbarijskih etiketa*




*Slika 33: Brojnost pojedinih velikih slova rukopisa Ljudevita Vukotinovića generiranih s 10 nasumično izabranih0 herbarijskih etiketa*




*Slika 34: Abeceda rukopisa Ljudevita Vukotinovića dobivena iz herbarijskih etiketa*

## 4.2. Analiza dostupnih programa za optičko prepoznavanje znakova i rukopisa

Istražujući na Internetu programe za optičko prepoznavanje znakova dobiven je sljedeći popis programa, zasebno za optičko prepoznavanje znakova (OCR), inteligentno prepoznavanje znakova (ICR) te prepoznavanje rukopisa (HTR).

a) OCR programi

Programi za OCR koji se plaćaju:

- BIT-Alpha proizvođača B.I.T. Bureau Ingénieur Tomasi
- FineReader proizvođača ABBYY
- FormPro proizvođača OCR Systeme
- KADMOS best OCR/ICR
- OCRKit za Mac OS i iOS
- OmniPage proizvođača Nuance Communications (ranije: ScanSoft)
- Readiris proizvođača ımage Recognition Integrated Systems Group (I. R. I. S), od 2013 pripada Canon-u
- NSOCR proizvođača Nicomsoft
- ARGUS_Script proizvođača Planet IS GmbH
- Screenworm za Mac OS proizvođača Funchip
- Teleform

OCR programi dostupni kao rješenja u oblaku (SaaS):

- ABBYY Cloud OCR
- Google Cloud Vision (Beta)
- Microsoft Azure Computer Vision API

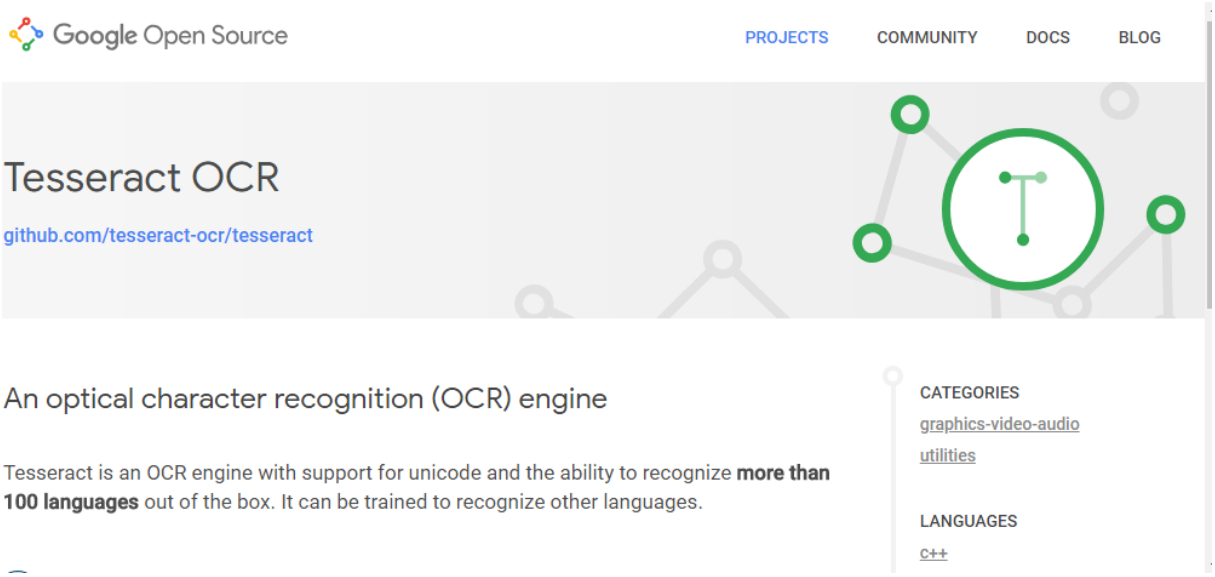
OCR programi dostupni u slobodnom pristupu:

- FreeOCR
- GT Text
- OCRopus
- GOCR

- CuneiForm
- Ocrad
- OCRFeeder
- Tesseract

Primjer OCR programa:

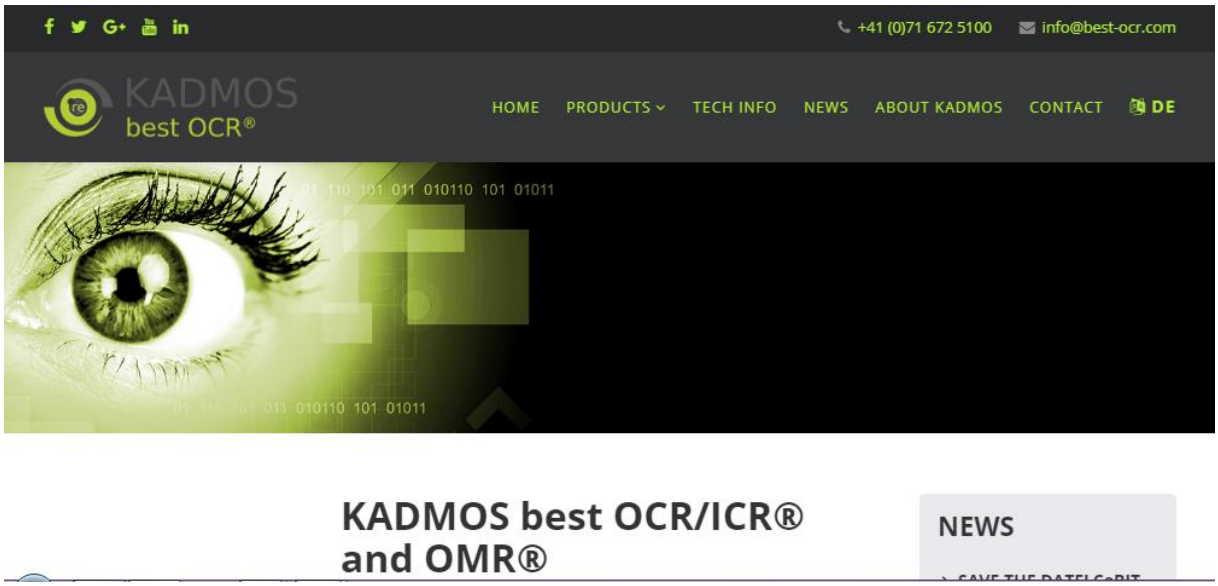
Tesseract je softver za prepoznavanje optičkih znakova kreiran za različite operativne sustave. Softver je otvorenog koda, besplatno dostupan, objavljuje se pod Apache licencom, Verzija 2.0, a od 2006. godine razvoj sponzorira Google.




*Slika 35: Izgled dijela početne stranice Tesseract OCR (Izvor: https://opensource.google.com/projects/tesseract)*

b) ICR programi

- **Kadmos best OCR/ICR** prepoznaje ručno i strojno pisane tekstove (ICR i OCR), podržava sve uobičajene operacijske sustave poput Windows Vista/7/8, Linux Suse32 / Ubuntu 64, Android, Apple OSX, Apple iOS, Windows Mobile i drugih.




*Slika 36: Izgled dijela početne stranice Kadmos best OCR (Izvor: https://www.best-ocr.com/products/kadmos-best-ocr-icr-and-omr.html)*

- **A2iA's intelligent character recognition (ICR) technology**

  Prepoznaje izolirane znakove razlikovanjem pojedinačnih oblika i veličina unutar svakog klasifikatora znakova - prepoznavanjem značajki kao što su krivulje, petlje i linije - i njihovo organiziranje na logičan i djelotvoran način prema korisničkim specifikacijama upravljanja informacijama.
- **Leadtools**

c) HTR programi

- **Transkribus i READ projekt**

  **Recognition and Enrichment of Archival Documents projekt**

  Transkribus (Read.transkribus) je platforma za automatsko prepoznavanje, transkripciju i pretraživanje rukopisnih povijesnih dokumenata. Dio je programa READ (Recognition and Enrichment of Archival Documents) koji je sufinanciran iz EU fondova. Glavni zadatak READ projekta je učiniti arhivsko gradivo pristupačnijim kroz razvoj i diseminaciju sustava za HTR.

Transkribus (Read.transkribus) nudi niz alata za automatsku obradu dokumenata, kao što su:

- Prepoznavanje ručno pisanog teksta (HTR)
- Analiza slike

- Razumijevanje dokumenta
- Otkrivanje ključnih riječi
- Prepoznavanje optičkih znakova (OCR) pomoću programa ABBYY Finereader11

Cilj Transkribusa, prema Wikipedii, je pružiti podršku svima koji su uključeni u transkripciju povijesnih, tiskanih ili rukopisnih dokumenata, prvenstveno znanstvenicima, institucijama kao što su arhivi i knjižnice, ali i svim zainteresiranim koji se bave istraživanjem svojih predaka, poviješću i sl. Transkribus je namijenjen i programerima koji su u potrazi za novim izazovima. Platforma je besplatna i dostupna svim korisnicima, a s njom upravlja Sveučilište u Innsbrucku tj. skupina DEA (njem. Digitalisierung und Elektronische Archivierung; hrv. Digitalizacija i elektroničko arhiviranje). Europska unija podržava rad platforme kroz financiranje iz programa za istraživanja, FP7 projekt tranScriptorium i Horizon 2020. Platforma je razvijena zahvaljujući tehničkom veleučilištu u Valenciji.

Osnovni koncepti su (Wikipedia):

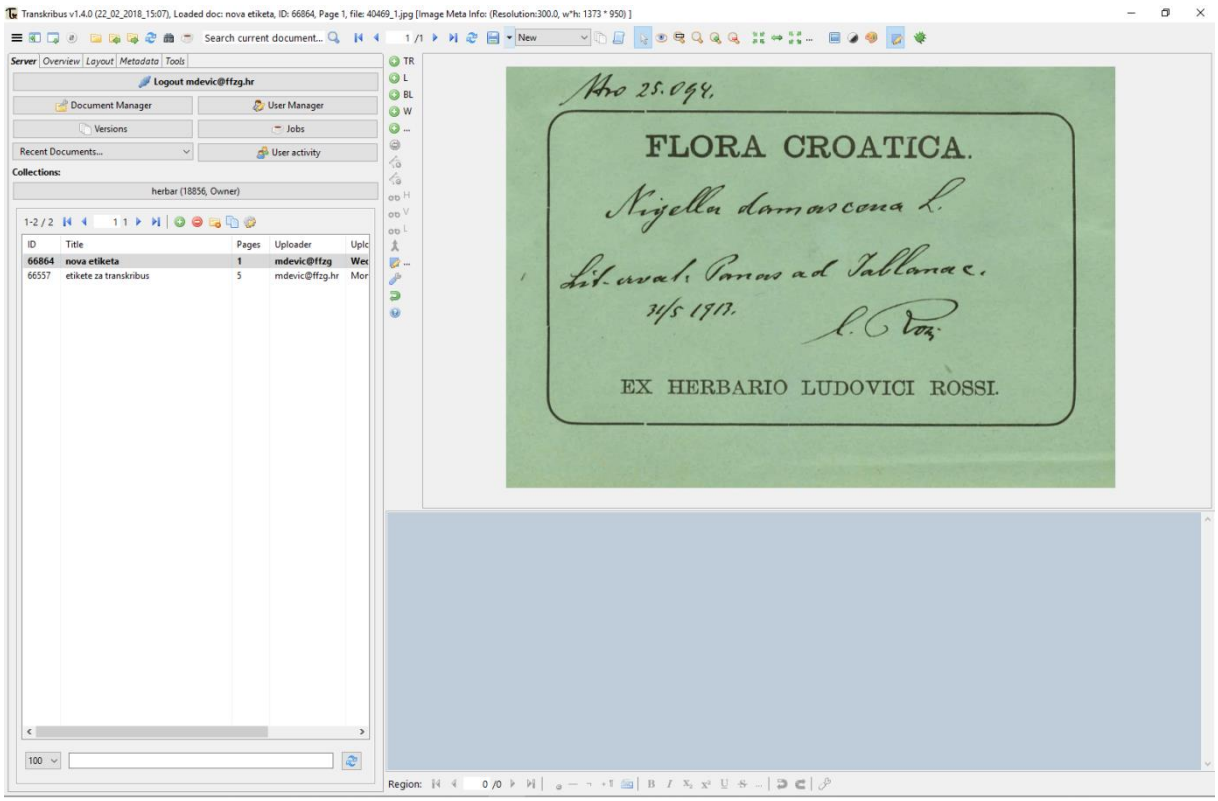
- Transkribus je besplatan program, nakon registracije ga se može preuzeti, ne spada u „opensource“ programe.
- Transkribus nudi "zaštićeno" područje, dokumenti koji se podižu na sustav samo su nama na raspolaganju, nisu vidljivi drugima osim ako ih podijelimo s nama poznatim korisnicima.
- Nema transkripcije bez prethodne segmentacije. Da bi rukopis bio prepoznat, dokumente se prethodno mora podijeliti u tekstualne regije i dodati osnovne linije ili linije područja. To se može odraditi djelomično automatski ili djelomično ručno. Ručna transkripcija može se pokrenuti samo ako postoji prethodno napravljena automatska segmentacija.
- Transkribus se mora trenirati. U pravilu je potrebno trenirati i ručno napisati 50 stranica dokumenta, kako bi se preostale stranice dokumenta ili zbirke automatski prepoznale.
- Transkribus integrira usluge. Ne radi na lokalnom računalu, jer je prepoznavanje rukopisnih dokumenata vrlo zahtjevan proces za koji su potrebna jaka računala.
- Transkribus treba „hranu“. Što se više dokumenata obrađuje na platformi, to je više podataka dostupno za treniranje programa.

- Transkribus je više od softvera za automatsko prepoznavanje. On je platforma dizajnirana tako da su dostupni i drugi oblici korištenja kao npr. mogućnost stvaranja digitalnih izdanja povijesnih dokumenata.
- Transkribus također čita tiskane dokumente, one koji su pisani gotičkim pismom.
- Transkribus je dio sustava virtualnog oblaka.

Korisničko sučelje Transkribusa vidljivo na slici 37 sastoji se od 5 različitih elemenata (Wikipedia):

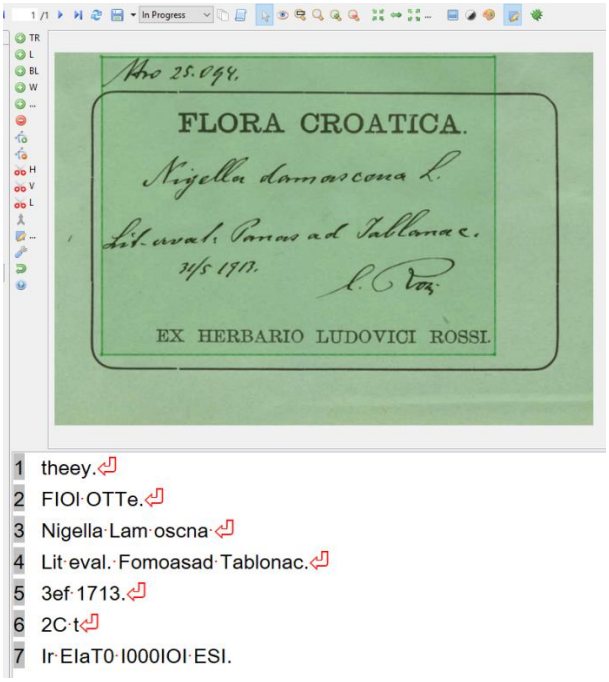
- Trake izbornika - nalazi se na vrhu.
- Kartice s lijeve strane - postojeće podkartice uglavnom pružaju informacije i postavke te služe za kretanje između stranica.
- Kartice s desne strane - pružaju razne alate.
- Područje slike (engl. Canvas) - uključujući traku izbornika, prikazuje sliku trenutne stranice i segmentirane dijelove teksta, retke, riječi.
- Područje teksta (engl. Editor) uključujući pripadajuću traku izbornika - omogućuje transkribiranje, ispravljanje, uređivanje itd. Područje teksta je izravno povezano s područjem slike.

Prije transkripcije, kao što je već spomenuto, potrebno je sliku segmentirati. Prvo se segmentiranje provodi automatski, klikom na gumb, te program napravi točkice koje su povezane linijom i tako za svaki redak teksta, vidljivo na slici 38. Nakon toga se segmentacija može raditi i ručno. Ukoliko je list teksta bio ukrašen, program će i to prepoznati kao nekakav tekst te je potrebno izbrisati, odnosno dodati linije gdje je to potrebno, te se tek nakon toga provodi transkripcija. Na žalost, kod dostupne verzije Transkribusa, skinute s interneta, opcija *train* nije bila dostupna te nije bilo moguće provesti treniranje. Bez treniranja su rezultati jako loši, što je vidljivo i na slici 39 u dostupnoj verziji koja je skinuta s interneta.




*Slika 37: Izgled sučelja programa Transkribus*


*Slika 38: Prikaz segmentacije u programu Transkribus*




*Slika 39: Prikaz transkripcije teksta u programu Transkribus*

# 5. Zaključak

U novije vrijeme kada sve više informacija primamo korištenjem moderne tehnologije od velike važnosti su baze podataka, tj. podaci pohranjeni u njima i dostupni za pretraživanje, istraživanje i preuzimanje. Prije nego što podaci dođu u baze podataka pred njima je dug put, a pri tome veliku važnost ima digitalizacija. Danas kada digitalizacija već široko rasprostranjena pojava, a sve više institucija radi na tome da gradivo bude dostupno što većem broju korisnika, od velike je važnosti i optičko prepoznavanje znakova. Ono olakšava korištenje i iskorištavanje digitaliziranog gradiva. Kada su u pitanju knjige koje su tiskane, situacija je puno bolja nego kada je u pitanju rukom pisani materijal. Istražujući u okviru rada mogućnosti koje se trenutno nude, uočen je velik nesrazmjer između optičkog prepoznavanja znakova, inteligentnog prepoznavanja znakova i prepoznavanja rukopisa. Trenutno su na tržištu brojni OCR programi od kojih su mnogi dostupni i besplatno. Kod prepoznavanja rukopisa situacija je bitno lošija te tehnologija još uvijek nije na toj razini da bi se rukopise moglo jednostavno pretvarati u strojno čitljiv i upotrebljiv oblik.

U okviru praktičnog dijela rada provedena je analiza rukopisa 10 najčešćih autora iz herbarija Herbarium Croaticum u bazi Flora Croatica. Ukupno je za potrebnu analizu te izradu abecede za svakog pojedinog autora preuzeto 379 herbarijskih etiketa. Nakon početnih 10 nasumce izabranih etiketa za svakog autora s kojih su izrezana sva slova bilo je potrebno uzeti dodatne etikete kako bi se prikupila slova koja nedostaju. Kod većine autora je s 10 inicijalnih etiketa sakupljeno mnogo samoglasnika malih slova. Kod velikih slova najčešće su se pojavljivala slova *L* i *S*. Najrjeđa slova su bila *dž, đ, Dž, Đ, Č, Ć, Y, X*. Autori čije etikete potječu iz 19. stoljeća koriste se krasopisom i običnim rukopisom te pri pisanju etiketa koriste i druge jezike (latinski, talijanski, njemački). Samo kod jednog autora su prisutna velika tiskana slova na etiketama. Kako većina piše pisanim i teže čitljivim slovima, otežano je dešifriranje slova te će abecede rukopisa dobivene kroz ovaj rad zasigurno pomoći i olakšati rad budućim korisnicima herbarijske zbirke Herbarium Croaticum.

# 6. Popis slika

# 7. Literatura


A2ia, dostupno na: https://www.a2ia.com/, 17.6.2018.

Abbyy technology portal, dostupno na: https://abbyy.technology/en:features:ocr:icr, 17.6.2018.

Agnihotri, V. P. (2012). Offline Handwritten Devanagari Script Recognition. *I.J. Information Technology and Computer Science*, 8:37-42. Preuzeto sa: http://www.mecs-press.org/ijitcs/ijitcs-v4-n8/IJITCS-V4-N8-4.pdf , (13.6.2018).

Christensson, P. (2018). *OCR Definition*, dostupno na: https://techterms.com/definition/ocr, 15.6.2018.

Enciklopedija leksikografskog zavoda Miroslav Krleža, dostupna na http://enciklopedija.hr, 11.6.2018.

Flora Croatica Database, dostupna na https://hirc.botanic.hr/fcd/, 12.6.2018.

Herbarium Croaticum, dostupan na: http://herbariumcroaticum.biol.pmf.hr, 11.6.2018.

Hirali S. A. , Payal P. M. i Vatsal H. S. (2014). Handwritten character recognition. *International Journal of Advance Engineering and Research Development*, 1(4): 1-6. Preuzeto sa: http://www.ijaerd.co.in/papers/finished_papers/ijaerd%2014-037.pdf, 13.6.2018

Hrvatski jezični portal, dostupan na: http://hjp.znanje.hr, 11.6.2018.

Hrvatsko botaničko društvo, dostupno na: http://www.hbod.hr/hr, 11.6.2018.

ITWiessen, dostupno na: https://www.itwissen.info/Intelligente-Zeichenerkennung-intelligent-character-recognition-ICR.html, 17.6.2018.

Leadtools, dostupno na: https://www.leadtools.com/, 17.6.2018.

Read.transkribus, dostupno na: https://read.transkribus.eu, 16.6.2018.

Tesseract, dostupno na: https://opensource.google.com/projects/tesseract, 18.6.2018.

Wikipedia. dostupno na: https://transkribus.eu/wikiDe/index.php/Hauptseite, 19.6.2018.

# Sažetak

**Analiza i optičko prepoznavanje rukopisa s herbarijskih etiketa u zbirci Herbarium Croaticum**

Rad obuhvaća teorijski i praktični dio pripreme i procesa analiziranja rukopisa i stvaranja baze slova. Cilj rada je analizirati raznolikost rukopisa sakupljača iz zbirke Herbarium Croaticum pohranjenih u bazi podataka Flora Croatica te stvoriti bazu slova rukopisa 10 najčešćih sakupljača biljaka iz zbirke *Herbarium Croaticum* Botaničkog zavoda Biološkog odsjeka PMF-a u Zagrebu. Svrha ovog rada je olakšati i ubrzati korisnicima pretraživanje, čitanje i dešifriranje rukopisa kroz stvaranje abecede za svakog od 10 izabranih autora. Kroz teorijski dio, uz pojmove herbarija, herbarijske zbirke i etikete rad se bavi i optičkim prepoznavanjem znakova i rukopisa (OCR-om, ICR-om i HCR-om) te mogućom primjenom u digitalizaciji herbarijskih zbirki. U okviru rezultata praktičnog dijela prikazani su koraci koji su potrebni za ostvarenje takvog cilja kao i programi koji se mogu koristiti.

**Ključne riječi:** rukopis, herbarijska etiketa, baza podataka, optičko prepoznavanje znakova i rukopisa

# Summary

**Analysis and Optical Recognition of Handwriting from Herbarium Labels in the Herbarium Croaticum Collection**

This thesis consists of theoretical and practical part of preparation process of analysis of handwritings and creation of database of characters. The goal is to analyze the diversity of handwriting collections from the *Herbarium Croaticum* collection stored in the Flora Croatica database and to create a database of the handwriting characters from the 10 most common plant collectors from the *Herbarium Croaticum* collection of the Botanical Division of the Biology Department, Faculty of Science, University of Zagreb. The aim of this thesis is to facilitate and speed up the search, reading and decrypting handwriting process for users by creating alphabets for each of the 10 selected authors. In the theoretical part, the concepts of herbarium, herbarium collection and label are explained and examined. It also deals with the optical recognition of characters and manuscripts (OCR, ICR, HCR) and the possible application in the digitization of herbarium collections. Within the practical part of the results,

the steps that are needed to achieve this goal as well as the programs that can be used are analyzed and presented.

**Keywords:** handwriting, herbrium label, database, optical recognition of characters and manuscripts