UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Barbara Suhadolc

# Statistična analiza slovenskih besedil

DIPLOMSKO DELO

VISOKOŠOLSKI ŠTUDIJSKI PROGRAM PRVE STOPNJE
RAČUNALNIŠTVO IN INFORMATIKA

Ljubljana, 2013

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Barbara Suhadolc

# Statistična analiza slovenskih besedil

DIPLOMSKO DELO

VISOKOŠOLSKI ŠTUDIJSKI PROGRAM PRVE STOPNJE
RAČUNALNIŠTVO IN INFORMATIKA

Mentor: doc. dr. Dejan Lavbič

Ljubljana, 2013

*Besedilo je oblikovano z urejevalnikom besedil Microsoft Word 2010.*

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*

*Tržaška 25*
*1001 Ljubljana, Slovenija*
*telefon: 01 476 84 11*
*01 476 83 87*
*faks: 01 426 46 47*
*01 476 87 11*
*www.fri.uni-lj.si*
*e-mail: dekanat@fri.uni-lj.si*

Št. naloge: 00479/2013
Datum: 11.04.2013

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **BARBARA SUHADOLC**

Naslov: **STATISTIČNA ANALIZA SLOVENSKIH BESEDIL**
**STATISTICAL ANALYSIS OF SLOVENIAN TEXTS**

Vrsta naloge: Diplomsko delo visokošolskega strokovnega študija prve stopnje

Tematika naloge:

Angleški jezik je bil že podrobno statistično obdelan z vidika raziskav frekvenc posameznih besed, črk, n-gramov ipd. Veliko manj analiz je bilo opravljenih na slovenskem jeziku, ki ga govori zelo majhen del svetovne populacije. V okviru diplomske naloge zato izvedite analizo več korpusov in sicer od leposlovja, poezije, Wikipedije, spletnih blogov do člankov. V okviru analize izvedite analizo pogosto uporabljanih besed, dolžine besed, črk (samoglasniki, soglasniki, položaj črk), n-grame (2, 3 in 4) ter izvedite analizo tudi z lematizacijo besed. Rezultat diplomske naloge naj bodo tudi identificirana področja uporabe vaših rezultatov statistične analize slovenskega jezika ter primerjava z angleškim jezikom.

Mentor:

doc. dr. Dejan Lavbič

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

Dekan:

prof. dr. Nikolaj Zimic

# IZJAVA O AVTORSTVU DIPLOMSKEGA DELA

Spodaj podpisana Barbara Suhadolc, z vpisno številko 63080190, sem avtorica diplomskega dela z naslovom:

*Statistična analiza slovenskih besedil*

S svojim podpisom zagotavljam, da:

- sem diplomsko nalogo izdelala samostojno pod mentorstvom doc. dr. Dejana Lavbiča,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,

V Ljubljani, dne 12. 09. 2013 Podpis avtorja:

# KAZALO

## KAZALO GRAFOV

## KAZALO TABEL

# POVZETEK

Slovenščina se je že mnogokrat srečala z analizo, vendar večina njih sloni na slovnični obravnavi – uporaba določenih spolov in sklanjatev. Same statistične analize pojavitve črk, kombinacij le teh in podobnih frekvenčnih izpisov je malo in se večinoma osredotočajo na osnove, kot so črke ali same besede.

Za boljši pregled nad tovrstnimi podatki je potrebna obširnejša analiza, ki se ne ustavi le pri osnovah kot frekvenca samih črk, temveč se spusti v podrobnosti, kot so frekvenca položajev tovrstnih črk, kombinacij med črkami, ipd..

Besedila so pomemben del vsakodnevnega življenja, saj so najlažji zapis dogodkov, govora in izražanja misli. Z njimi se človek sreča povsod: od reklamnih zapisov, do pogodb. Seveda so tudi velik del kulture in prostega časa, ki se ga preživi za knjigo, internetom, pri pošiljanju sporočil itd.. Zaradi te vloge, je sama fizika besedila pogosto spregledana, njena pozornost prenesena na pomen izraza. Vseeno pa je vse sestavljeno iz skupine črk, kombinacij le teh, kombinacij teh kombinacij itd.

Zaradi tovrstnih dejstev naloga vsebuje analizo slovenskega jezika, ki se je osredotočala bolj na fizično ureditev besed kot pomensko. To je bilo doseženo s pomočjo grafov in tabel, ki so olajšali prikaz rezultatov pridobljenih iz krajših, enostavnejših programov. Ciljna zanimanja so bila že vnaprej določena – k temu so pripomogle številne analize drugih, širše govorečih jezikov – vendar se je analiza na zanimivejših mestih poglobila in osredotočila na podrobnosti. Za vsako ugotavljanje je bil posebej sprogramiran program za lažji nadzor nad pravilnostjo rezultatov. Vsakemu rezultatu je sledila odločitev o boljšem prikazu le tega – primerjanje med seboj, naštevanje, prikaz. Besedilo je bilo lematizirano; opravljena je bila analiza tudi nad takšnim korpusom. Na koncu naloge so podane ideje za uporabo pridobljenih podatkov, korpusi in napisani programi pa so bili objavljeni tudi na internetu za lažji dostop do njih.

# ABSTRACT

Slovenian language was analysed many times, but most of this research focuses on the grammatical side – use of genders and declension. The statistical part of comparisons of letters, letter combinations and other results of frequency is lack.

For better overview of this information a wider analysis is needed. It should not stop just with one letter frequency, but focus on the details as the position of these letters, combinations, etc..

Text is an important part of everyday life for describing events, recording speech and thought. A person can see it everywhere: from commercial ads to contracts. It is a big part of culture and free time a man uses for reading a book, exploring internet, sending messages and so on. Because of this the physics of the text is usually not spoken about, its attention focused more on the meaning of the expression. Still, everything is pieced together by letter groups, their combinations and combinations of those combinations.

For these causes, the research centred around Slovenian language and not so much as meaningful, but physical part of words that make text. To do that, number of graphs and tables that made the results easier to display were used. These were work of many short programs. The main line of work was already pre-chosen by a large number of already done analyses of different languages. Still, when the analysing hit some interesting point, it focused on it for a deeper research. Every graph or table display of the results was an outcome of a written program that made the control over the result correctness easier. For each outcome a decision had to be made to either show it in a comparison to others, make a simple display of it or make it a part of some visually pleasant graph. The text was lemmatised and analysed in that shape. At the end of the paper the ideas of result use are lined. Corpuses and written programs were uploaded on the internet for easier access.

# 1. UVOD

Slovenščina je jezik, ki ga v svetu govori izredno majhen procent ljudi (približno 3% svetovne populacije). Z leti in tehnologijo se je tudi njena uporaba nekoliko izkrivila z uporabo angleških izrazov in podobnih 'bližnjic' do razumevanja. Slovenščina pa je še vedno naš materni jezik, ki pa si zasluži enakih raziskav kot vsak drugi.

S tem smo prišli do izbire teme za naše diplomsko delo, saj smo želeli dati tudi temu, malo govorečemu jeziku enako pozornost kot ga dobivajo drugi, obširnejši jeziki. Kot materni jezik naše države z bogato razvojno zgodovino se zadnje čase njegova uporaba manjša in meša z drugimi jeziki, različnim žargonom in na sploh izgublja svoj pomen. Zato se nismo odločili za pomensko analizo besedil, temveč za fizično. Filozofska fakulteta je s pomočjo svojih študentov naredila že veliko raziskav na temo slovenske slovnice, medtem pa se na internetu pojavljajo frekvence uporabe črk v našemu jeziku (recimo Wikipediji). Za hitri pregled osnov je to sicer dovolj, ni pa dovolj obširno za potrebo po detajlnih informacijah.

Če sprogramiramo program, ki ugotavlja neko besedo, je mogoče da bo v določenem številu ugibanj s pomočjo tabele frekvenc črk našel pravilno rešitev. Če ima program na voljo sezname najpogostejših položajev teh črk v besedi, bo besedo uganil veliko prej. Ravno tako je enaka tabela primerna za preprost iskalnik, ki nima pregleda nad rezultati. Tako lahko pri vpisu iskalnega regularnega izraza "mo*a" iz statistike ugotovi ali je večja možnost, da iskalec išče besedo "mora" ali "moka" ali kakšno tretjo kombinacijo črk. Tako lahko mogoče rezultate prikaže po vrsti glede na frekvenco predvidene sestavljene besede. Ob raznih ugankah ali igricah, kjer moramo anagram sestaviti v prvotno besedo, si ravno tako lahko pomagamo s takšnimi ugotovitvami.

Nekoč so tovrstne analize počeli ročno, zato so bili korpusi veliko manjši in zato rezultati manj zanesljivi, a z razvojem različnih računalniških aplikacij lahko preskočimo takšne težave. Zato smo sprejeli to prednost, sposobnosti programiranja, in izkoristili elektronsko dobo, da smo analizirali slovenska besedila tudi številčno (kako pogosto so uporabljene posamezne besede, kolikšna je njihova povprečna dolžina, ipd.). Možnosti takšnega analiziranja je ogromno.

Namesto, da bi uporabili en sam korpus, smo jih uporabili več in jih med seboj primerjali. Za pridobitev rezultatov smo se polastili programiranja v programerskem jeziku *Python* in namreč verzijo *2.7* ter njegove knjižnice *os*, *sys* in *urllib2*. Za urejanje dolgih tabel smo uporabili Microsoftov *Excel 2010*, kjer smo izkoristili možnosti formul za posamezno

celico in tako iz nekaj osnovnih podatkovnih baz dobili rezultate, ki jih bomo uporabili v tej diplomski nalogi.

# 2. ŽE OPRAVLJENE RAZISKAVE

Kot že omenjeno imamo na področju slovenščine že vrsto raziskovalnih nalog in krajših statistik. Večina teh se ne ukvarja z fiziko slovenske besede. Zelo pomanjkljivo se kakšne statistične informacije nahajajo tudi na slovenski Wikipediji, pod temo *Najpogostejše slovenske besede* [1] in pa *Frekvence črk* [1], drugače pa je veliko takšnih raziskav mogoče videti v – in o – drugih jezikih.

Nekaj raziskav se ukvarja z uporabo slovenščine v primernih stilih pisanja (Primer na referenci: [2]). Ta stil je poleg častnikov in revij potreben tudi pri glasbi, saj se v besedilih pesmi nahajajo narečne besede, ki ohranjajo ritem in dajo občutek domačnosti [3]. Seveda je stil pomemben za vsakega avtorja, ki bi rad obdržal svojo skupino bralcev.

Tema govorečega slovenskega jezika je ena izmed najbolj popularnih, in tudi najlažje najdenih na internetu. Problem vedno manj govorjene pravilne slovenščine predvsem skrbi Ministrstvo za kulturo, ki ima objavljenih kar nekaj raziskovalnih nalog, kot so neslovenska imena trgovin in pa pomembnost slovenščine v šolskem učnem načrtu [4]. Opravljenih je bilo tudi že nekaj sestav samostalniških sistemov in njihovih naglasov glede na regijo [2].

Programi za lematizacijo slovenščine so večinoma pomankljivi in dajejo nepravilne rezultate. Jeziku je zaradi njegove komplikacije (različni spoli, končnice, spreganje, časi, predpone) težko določiti nek osnoven ključ za pretvarjanje. Sami smo za lematizacijo uporabili projekt *LemmaGen* (*http://lemmatise.ijs.si/*), ki nam je od vseh preizkušenih predstavil še najboljše rezultate in možnost uporabe velike datoteke za pretvorbo.

# 3. OPIS IN ZGRADBA KORPUSOV

Ker smo v naši statistični analizi uporabili več različnih korpusov, smo morali poiskati veliko slovenskega besedila na spletu. Izogibali smo se forumom, saj se tam besedila ponavljajo, imajo zelo slabo besedišče in nasploh postajajo vedno bolj angleška. Za leposlovje in poezijo so bile zelo koristne spletne strani, ki so sicer v ozadju že vsebovale slovenske korpuse, a jih ni bilo mogoče doseči. Skozi takšne strani smo nato izbrskali besedila, ki so jih uporabljali za zgradbo svojih korpusov in nadaljevali od tam [3]. Pri poeziji so nam ravno tako prav prišle strani, ki vsebujejo besedila slovenskih hitov (*slolyrics.com* [4] in *pesmi.si* [5]). *Digitalna knjižnica Slovenije* [6] nam je omogočila dostop do kvalitetnih daljših besedil visoke kulturne vrednosti. Pri člankih smo se polotili malo bolj znanih spletnih strani novic (*24ur.com* [7], *žurnal24.si* [8] in *delo.si* [9]), ter še nekaj drugih koncev interneta, kjer smo iskali malo bolj znanstvene članke. Spletni blogi in pa Wikipedija sta pa že sama po sebi bila dober vir sama zase.

Ker smo slovensko leposlovje in prozo iskali po zelo različnih straneh, nismo imeli možnosti sprogramirati programa, ki bi lahko avtomatsko kopiral besedilo v .txt datoteko. Pri blogih, člankih in Wikipediji pa smo spogramirali program, ki je to počel, in ki smo ga lahko hitro preurejali glede na potrebe spletnih strani.

Nekaj problemov smo imeli z našimi šumniki in naglašenimi besedami zaradi različnih kodiranj, zato smo morali v program vključili preprost prevajalnik iz ASCII v Unicode kodiranje.

## 3.1. Leposlovje

Najbolj obširen del slovenskih besedil je prav gotovo kulturno leposlovje – knjige naših znanih in manj znanih pisateljev in pisateljic. Posamezno enoto proze predstavlja natančno en stavek. V takšno skupino spadajo lirika (izpoved v prozi), epika (pripoved v prozi) in dramatika (igra). Vse našteto je bilo tudi vključeno v naš korpus, večinoma pa smo za primere vzeli že zgodovinsko znane pisatelje. Besedila, ki smo jih uporabili v naši raziskavi, so bila različnih dolžin in žanrov. V veliko pomoč pri iskanju besedil so nam bili drugi korpusi (*SRC SAZU*) [3] in pa *Digitalna knjižnica Slovenije* [6], kjer smo lahko našli celotna dela pisateljev. Naš korpus leposlovja je vseboval 1.349.484 besed.

## 3.2. Poezija

Poezija se od leposlovja razlikuje po osnovni enoti. Nasprotno od proze, tu posamezno enoto predstavlja en verz. Znano je, da poezija teži k rimi in krajši dolžini. Poznamo lirično (izpoved v poeziji), epično (pripoved v poeziji) in ljudsko (se je prinašala skozi zgodovino v ustnem izročilu) poezijo. Poleg znanih poetov smo v naš korpus vključili tudi besedila slovenskih pesmi oziroma hitov. V korpusu se je nahajalo 293.860 besed.

## 3.3. Članki

Članki se od poezije in proze razlikujejo po besedišču in tudi sami vsebini. Gre za dnevne novice in zanimivosti – informacije o trenutnih dogajanjih. Novice se sprašujejo: *kaj*, *kdo*, *kje* in *kako* [10]. V našem korpusu jih predstavljajo članki iz strani *24ur* [7] in pa *Delo* [9]. Članki raznih analiz in združitve že znanih informacij in znanstvenih odkritij pa so bili izvzeti iz *Članki.net* [11], *Revije Ventil* [12], spletne strani *Univerzitetnega rehabilitacijskega inštituta Republike Slovenije - Soča* [13] in pa spletne strani *National Geographic Slovenija* [14]. Za analiziranje člankov smo le te prekopirali v .txt datoteke. Za tiste članke, ki se niso nahajali v PDF-jih smo sprogramirali program, ki je to počel in je mimogrede izvzel slike in dodatne nastavitve (različna velikost črk, odebeljeni fonti), tako da smo dobili le gole besede, ki smo jih nato naprej uporabili za našo statistično analizo. V korpusu člankov se je nahajalo 231.915 besed.

## 3.4. Spletni blogi

Angleški slovar (*Dictionary.com*) opiše blog kot "Spletna stran, ki vsebuje pisateljeve – ali skupine pisateljev – izkušnje, ugotovitve, mnenja itd., in pogosto vsebuje slike ali linke do drugih strani." [15] Uradno ime bloga je "webblog" oziroma naš izraz "spletni blog". Pisatelj bloga se imenuje *bloger*. Naš korpus vsebuje bloge iz spletnega združenja blogov, *Slovenski blogi (sloblogi.com)* [16]. Izbrali smo najbolj popularne bloge, vendar smo preskočili tiste z malo pisne vsebine. Kot večino blogov po svetu, tudi slovenski blogi vsebujejo veliko politike, mnenje o trenutni situaciji v državi in blogerjev pogled na ostale dogodke po svetu. Za slovenske bloge pričakujemo, da bodo vsebovali bolj preprost jezik, saj je namenjen celotni populaciji, ne glede na spol, starost ali izobrazbo. Število besed v tem korpusu je 441.154.

## 3.5. Wikipedija

Kot zanimivost smo se lotili tudi prispevkov na slovenski Wikipediji [1]. Gre za neprofitno spletno enciklopedijo, ki dnevno dobiva vse več prispevkov. Ker gre za slovensko verzijo Wikipedije, so prispevki nekoliko krajši in bolj pomanjkljivi kot recimo tisti v angleški verziji, vendar so ravno tako zelo informativni. Zaradi zapletene html sestave strani, smo se poslužili pythonove knjižnice *beautifulsoup,* in namreč *bf4* [17], s pomočjo katere smo odstranili označbe in dobili le pomembni tekst. Ker gre za enciklopedijo pričakujemo veliko različnega besedišča, znanstvenih besed in tujih črk. Korpus Wikipedije vsebuje 380.292 besed.

# 4. ANALIZA BESED

## 4.1. Najbolj uporabljene besede

V programerskem jeziku *Python* smo spisali program, ki je iz vsakega dela odstranil ločila in velike črke spremenil v male, nato pa za vsako delo posebej naredil svojo tabelo besed in njihovo ponovitev. To je storil s štetjem posamezne besede v slovarju, ki je bila vstavljena v le tega pri svoji prvi pojavitvi. Na koncu je spisani program vse tabele združil in nam podal končno število vseh uporabljenih besed ter število ponovitev posamezne besede. Tako smo dobili le gole besede, ki pa so še vedno bile nelematizirane. Ker se najbolj uporabljene besede med korpusi razlikujejo, smo raje naredili tabele za vsak posamezen korpus (Tabela 1 in Tabela 2).

| | Leposlovje (1.349.484 besed) | | Poezija (293.860 besed) | | Članki (231.915 besed) | |
|---|---|---|---|---|---|---|
| | **Beseda** | **Št. ponovitev** | **Beseda** | **Št. ponovitev** | **Beseda** | **Št. ponovitev** |
| **1.** | je | 77355 | je | 8026 | in | 6476 |
| **2.** | in | 44110 | in | 8021 | je | 5518 |
| **3.** | se | 35442 | v | 6846 | v | 5176 |
| **4.** | v | 23095 | se | 6470 | na | 3500 |
| **5.** | da | 21191 | ne | 3815 | za | 3124 |
| **6.** | na | 17158 | na | 3376 | so | 2396 |
| **7.** | ne | 14139 | da | 2893 | se | 2172 |
| **8.** | so | 13681 | pa | 2330 | da | 2135 |
| **9.** | pa | 13168 | si | 2245 | ki | 2113 |
| **10.** | bi | 10564 | mi | 2111 | z | 2012 |
| **11.** | ni | 9661 | sem | 2094 | pri | 1866 |
| **12.** | sem | 8687 | za | 2044 | s | 1817 |
| **13.** | z | 8480 | so | 1978 | po | 1518 |
| **14.** | za | 8282 | ti | 1970 | pa | 1464 |
| **15.** | ga | 7587 | bi | 1944 | tudi | 1322 |
| **16.** | po | 7583 | ko | 1934 | kot | 951 |
| **17.** | ki | 7383 | z | 1736 | smo | 911 |
| **18.** | še | 7336 | ki | 1710 | ne | 904 |
| **19.** | s | 6553 | bo | 1634 | of | 887 |
| **20.** | tako | 6550 | še | 1633 | the | 881 |

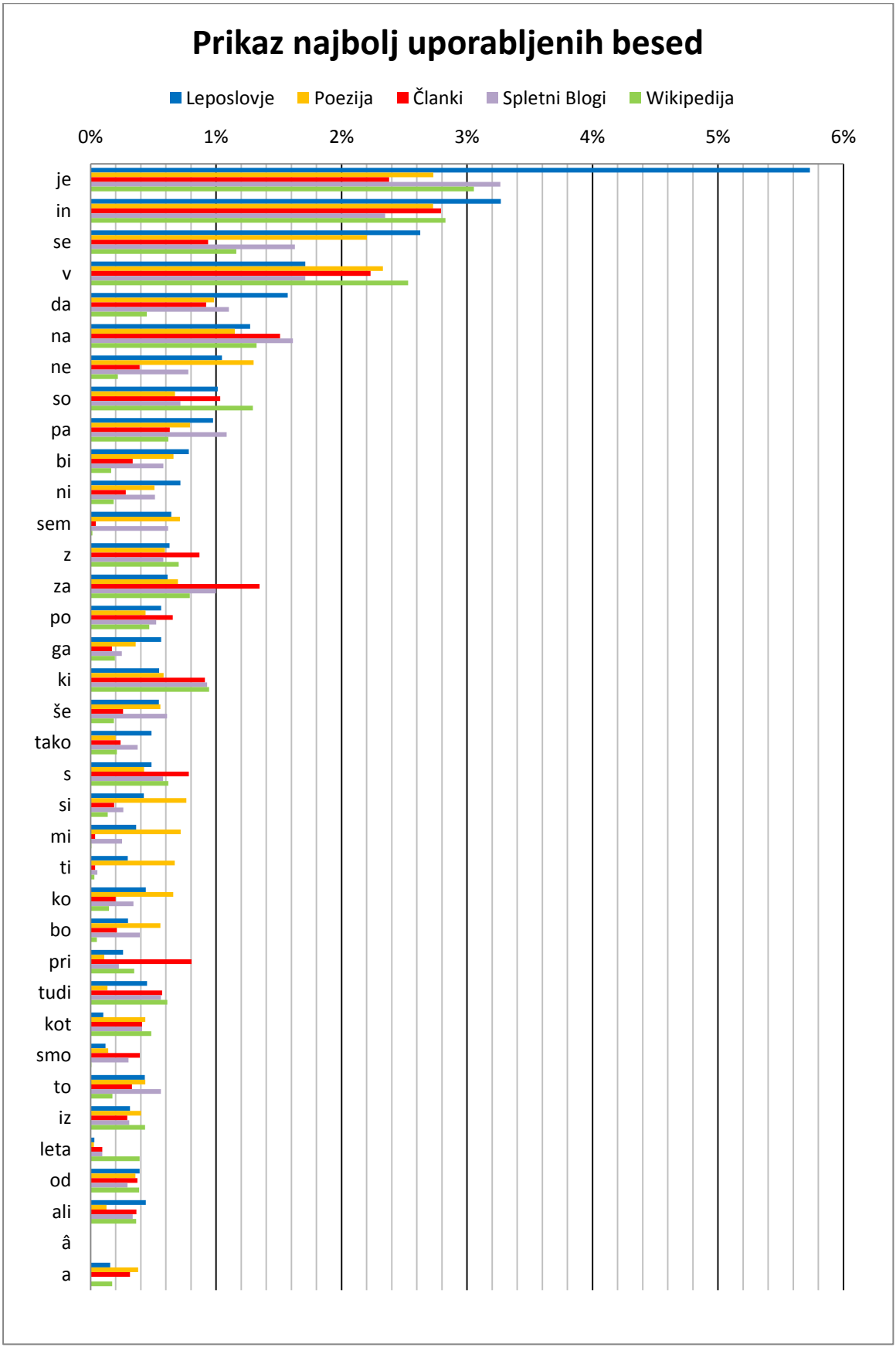
**Tabela 1: Tabele najbolj uporabljenih besed (v leposlovju, poeziji in člankih)**

| | Spletni blogi (441.154 besed) | | Wikipedija (380.292 besed) | |
|---|---|---|---|---|
| | **Beseda** | **Št. ponovitev** | **Beseda** | **Št. ponovitev** |
| **1.** | je | 14402 | je | 11615 |
| **2.** | in | 12496 | in | 10755 |
| **3.** | v | 10357 | v | 9625 |
| **4.** | se | 7545 | na | 5029 |
| **5.** | na | 7182 | so | 4915 |
| **6.** | da | 7113 | se | 4413 |
| **7.** | pa | 4859 | ki | 3588 |
| **8.** | za | 4782 | za | 3007 |
| **9.** | ki | 4405 | z | 2671 |
| **10.** | ne | 4097 | s | 2356 |
| **11.** | so | 3437 | pa | 2355 |
| **12.** | sem | 3160 | tudi | 2323 |
| **13.** | še | 2721 | kot | 1838 |
| **14.** | z | 2687 | po | 1776 |
| **15.** | bi | 2562 | da | 1705 |
| **16.** | s | 2560 | iz | 1653 |
| **17.** | to | 2545 | leta | 1485 |
| **18.** | tudi | 2469 | od | 1468 |
| **19.** | po | 2468 | ali | 1377 |
| **20.** | ni | 2306 | pri | 1320 |

**Tabela 2: Tabeli najbolj uporabljenih besed (v spletnih blogih in korpusu Wikipedije)**

Vidimo, da beseda *je* daleč najbolj uporabljena v vseh korpusih. Gre za stavčno obliko besede *biti*, ki jo SSKJ (2008) preprosto opiše kot: "izraža materialno ali duhovno navzočnost v stvarnosti" [18]. Prav tako so zelo pogosti vezniki (predvsem *in*), predlogi in svojilni zaimki. Na splošno pa so vse besede kratke in sestavni deli večine stavkov, ki brez njih ne bi imeli pomena.

Iz vseh zgornjih besed smo naredili graf (Graf 1) vseh petih korpusov, v kateri se nahaja procent vsake besede glede na število njenih ponovitev v posameznem korpusu.




Graf 1: Prikaz procenta vsake besede, glede na posamezen korpus

## 4.2. Vezniki

Zaradi popularnosti besede *in*, smo se nekoliko bolj poglobili v veznike, ki združujejo stavke v povedi. Poznamo priredne in podredne. Priredni povezujejo dva enakovredna stavka, medtem ko so podredni odvisni od glavnega in sami ne morejo stati v besedilu. Naredili smo tabelo veznikov (Tabela 3) in sešteli njihovo uporabo glede na vrsto (priredni in podredni). Predvidevamo, da bo podrednih več, saj omogočajo bolj tekoče besedilo – predvsem v leposlovju in blogih.

| VRSTA VEZNIKA | VEZNIK | Leposlovje | | Poezija | | Članki | | Blogi | | Wikipedija | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Vsota | | Vsota | | Vsota | | Vsota | | Vsota |
| PRIREDNI | in | 3,27 | 7,1 | 2,73 | 5,4 | 2,79 | 5,5 | 2,83 | 6,8 | 2,83 | 5,3 |
| | pa | 0,98 | | 0,79 | | 0,63 | | 1,10 | | 0,62 | |
| | ter | 0,18 | | 0,02 | | 0,19 | | 0,12 | | 0,18 | |
| | tako | 0,49 | | 0,20 | | 0,24 | | 0,39 | | 0,21 | |
| | pa | 0,98 | | 0,79 | | 0,63 | | 1,10 | | 0,62 | |
| | oziroma | 0,00 | | 0,00 | | 0,06 | | 0,05 | | 0,05 | |
| | a | 0,16 | | 0,38 | | 0,31 | | 0,22 | | 0,17 | |
| | toda | 0,08 | | 0,01 | | 0,01 | | 0,01 | | 0,01 | |
| | vendar | 0,10 | | 0,05 | | 0,06 | | 0,07 | | 0,10 | |
| | ampak | 0,05 | | 0,02 | | 0,02 | | 0,16 | | 0,02 | |
| | marveč | 0,00 | | 0,00 | | 0,00 | | 0,00 | | 0,00 | |
| | vendar | 0,10 | | 0,05 | | 0,06 | | 0,07 | | 0,10 | |
| | namreč | 0,02 | | 0,00 | | 0,04 | | 0,05 | | 0,01 | |
| | temveč | 0,01 | | 0,00 | | 0,01 | | 0,02 | | 0,01 | |
| | kajti | 0,03 | | 0,01 | | 0,00 | | 0,01 | | 0,00 | |
| | torej | 0,04 | | 0,01 | | 0,03 | | 0,06 | | 0,03 | |
| | zato | 0,10 | | 0,07 | | 0,11 | | 0,16 | | 0,10 | |
| | zatorej | 0,01 | | 0,01 | | 0,00 | | 0,00 | | 0,00 | |
| | tako | 0,49 | | 0,20 | | 0,24 | | 0,39 | | 0,21 | |
| | tedaj | 0,08 | | 0,02 | | 0,00 | | 0,01 | | 0,02 | |
| PODREDNI | da | 1,57 | 3,4 | 0,98 | 3,2 | 0,92 | 2,6 | 1,61 | 4,1 | 0,45 | 1,9 |
| | ki | 0,55 | | 0,58 | | 0,91 | | 1,00 | | 0,94 | |
| | kdaj | 0,03 | | 0,07 | | 0,02 | | 0,03 | | 0,00 | |
| | kar | 0,23 | | 0,24 | | 0,16 | | 0,34 | | 0,12 | |
| | čeprav | 0,02 | | 0,03 | | 0,02 | | 0,05 | | 0,04 | |
| | kjer | 0,08 | | 0,11 | | 0,08 | | 0,11 | | 0,09 | |
| | če | 0,28 | | 0,32 | | 0,16 | | 0,34 | | 0,08 | |
| | ker | 0,17 | | 0,15 | | 0,08 | | 0,26 | | 0,07 | |
| | čeravno | 0,00 | | 0,00 | | 0,00 | | 0,00 | | 0,00 | |
| | ko | 0,44 | | 0,66 | | 0,20 | | 0,36 | | 0,15 | |
| SKUPAJ | | 10,5 | | 8,6 | | 8,1 | | 10,9 | | 7,2 | |

**Tabela 3: Tabela uporabe veznikov (vse vrednosti so v %)**

Ker je prirednih veznikov mnogo več, je tudi njihova vsota prišla večja. Vendar pa je bilo naše predvidenje uporabe podrednih stavkov v leposlovju in blogih pravilno. Wikipedija vsebuje zelo malo podrednih stavkov v primerjavi s prirednimi, saj gre za več naštevanja. Medtem ko ostali očitno ohranjajo krajše povedi, pa se avtorji leposlovja in blogov poslužujejo daljših povedi, kar je razvidno iz višjega procenta veznikov.

## 4.3. Najbolj uporabljene besede z informacijsko vrednostjo

Skoraj vse do sedaj izpostavljene besede so bila tako imenovana *polnila* (v angleščini *stopwords*). Zaradi tega razloga, smo v programerskem jeziku *Python* spisali program, ki vsako posamezno tabelo besed in njihovih ponovitev ločil na dve: tisto, ki je vsebovala le polnila in tisto, ki je vsebovala le besede z nekakšno informacijsko vrednostjo.

Med prepisovanjem v posamezno tabelo smo kot zanimivost sešteli pojavitve teh besed. Nato smo jih primerjali s številom vseh besed in tako dobili, kolikšen procent posameznega korpusa tabeli predstavljata (Tabela 4). Najmanj polnil pričakujemo pri korpusu Wikipedije, zaradi njene kratke in jedrnate vsebnosti. Nasprotno, jih največ pričakujemo pri leposlovju.

| **BESEDE** | **VSEBNOSTNI PROCENT (v %)** | | | | |
|---|---|---|---|---|---|
| | **Leposlovje** | **Poezija** | **Članki** | **Blogi** | **Wikipedija** |
| Polnila | 51 | 46 | 35 | 46 | 32 |
| Besede z informacijsko vrednostjo | 49 | 54 | 65 | 54 | 68 |
| **SKUPAJ** | **100** | **100** | **100** | **100** | **100** |

**Tabela 4: Procent vseh besed z informacijsko vrednostjo in polnil**

Polnila predstavljajo kar polovico vseh besed pri leposlovju, sledijo ji poezija in spletni blogi. Kot predvideno se najmanj polnil uporablja v Wikipediji. Cilj člankov je prav tako jedrnatost, saj bralci želijo iz novic te hitro razbrati.

Ponovno smo izpostavili dvajset najbolj uporabljenih informacijsko vrednih besed. Dobili smo spodnje tabele (Tabela 5 in Tabela 6). Pri častnikih, korpusu Wikipedije in blogih smo naleteli na angleške besede. Ker so le te pri častnikih v prevedenem delu povzetka, smo jih preskočili. Ravno tako smo jih preskočili v korpusu Wikipedije, saj so oglaševani del strani – ogrodje – in zato niso del samega članka. V blogih smo jih pustili, saj so del 'internetnega jezika'.

| | Leposlovje (1.349.484 besed) | | Poezija (293.860 besed) | | Članki (231.915 besed) | |
|---|---|---|---|---|---|---|
| | **Beseda** | **Št. ponovitev** | **Beseda** | **Št. ponovitev** | **Beseda** | **Št. ponovitev** |
| **1.** | sam | 1713 | dan | 633 | slika | 337 |
| **2.** | oči | 1482 | svet | 538 | let | 241 |
| **3.** | dobro | 1358 | sam | 452 | leta | 216 |
| **4.** | dan | 1286 | srce | 447 | dela | 205 |
| **5.** | gospod | 1237 | noč | 436 | bolnikov | 203 |
| **6.** | bog | 1219 | oči | 405 | glede | 197 |
| **7.** | rekel | 1195 | čas | 403 | hoje | 196 |
| **8.** | oče | 1088 | bog | 325 | št | 194 |
| **9.** | imel | 1081 | dni | 275 | otrok | 194 |
| **10.** | roko | 1060 | gre | 274 | rehabilitacija | 184 |
| **11.** | človek | 1005 | glej | 266 | test | 159 |
| **12.** | ima | 958 | življenje | 260 | letn | 158 |
| **13.** | glavo | 908 | pesem | 249 | stran | 156 |
| **14.** | mati | 902 | pesmi | 243 | najbolj | 152 |
| **15.** | prišel | 885 | nazaj | 239 | dne | 146 |
| **16.** | šel | 869 | vem | 225 | ima | 140 |
| **17.** | dolgo | 868 | dekle | 225 | vrednosti | 139 |
| **18.** | obraz | 847 | en | 215 | amputaciji | 138 |
| **19.** | videl | 812 | lepo | 214 | protezo | 138 |
| **20.** | janez | 804 | dva | 213 | objavil | 136 |

**Tabela 5: Tabele najbolj uporabljenih besed z informacijsko vrednostjo (v leposlovju, poeziji in člankih)**

| | Spletni blogi (441.154 besed) | | Wikipedija (380.292 besed) | |
|---|---|---|---|---|
| | **Beseda** | **Št. ponovitev** | **Beseda** | **Št. ponovitev** |
| **1.** | janša | 567 | leta | 1485 |
| **2.** | danes | 527 | glej | 496 |
| **3.** | sds | 496 | del | 455 |
| **4.** | dan | 466 | povezave | 442 |
| **5.** | gre | 437 | zunanje | 423 |
| **6.** | leta | 412 | ima | 422 |
| **7.** | tek | 397 | org | 410 |
| **8.** | dobro | 385 | ljubljana | 374 |
| **9.** | ima | 362 | sl | 360 |
| **10.** | čas | 351 | št | 359 |
| **11.** | the | 345 | vzpostavljeno | 348 |
| **12.** | sam | 343 | km | 340 |
| **13.** | skupaj | 327 | države | 335 |
| **14.** | najbolj | 319 | cerkev | 328 |
| **15.** | pot | 308 | članek | 320 |
| **16.** | slovenija | 297 | pridobljeno | 319 |
| **17.** | strani | 297 | let | 1435 |
| **18.** | poti | 288 | mesto | 1355 |
| **19.** | časa | 280 | ime | 1345 |
| **20.** | del | 271 | snovi | 268 |

**Tabela 6: Tabeli najbolj uporabljenih besed z informacijsko vrednostjo (v spletnih blogih in korpusu Wikipedije)**

Zanimivo je, da bi bilo iz vsake tabele tudi brez njenega naslova mogoče prebrati iz katerega korpusa je bila izvzeta. Nasprotno od prejšnjih tabel (Tabela 1 in Tabela 2) se tu besede popolnoma razlikujejo med seboj.

Iz leposlovja je mogoče videti uporabo besed, ki so uporabljene za opis oseb (glavnih in stranskih likov zgodb) – realnost njih ni pomembna: *sam, gospod, bog, oče, človek, mati* in celo *janez*. Zadnja celo podpira splošno znanje, da je Janez najbolj popularno ime v Sloveniji. Da gre za pripoved je ravno tako razvidno iz opisa 'akcij' teh akterjev: *rekel, imel, ima, prišel, šel* in *videl*.

Poezija, podobno kot leposlovje, sloni na starejših besedah, vendar pa je izražanje nekoliko bolj romantično. Oseb je tokrat manj: *sam, bog* in *dekle* – navadno muza pesnika. Kot romantične izraze pa lahko naštejemo: *svet, srce, čas, pesem, pesmi, oči, noč* in *lepo.*

Članki slonijo na praktičnih besedah, večinoma takšnih, ki se pojavljajo v naslovu članka, oziroma časopisa, tokrat spletne strani, iz katerih smo članek vzeli. Takšna beseda je *rehabilitacija*. Poleg tega so tu še elementi, ki so navadno ne-tekstovni del ter na katere se posamezen članek očitno sklicuje: *slika, št* in *stran*. Nato pa so seveda tu še teme samih člankov: *let, leta, dela, bolnikov, hoje, otrok, test, vrednosti in amputaciji*.

Kar se tiče blogov, je njihova definicija, da vsebuje veliko političnih tem, očitno prava. Besede, ki nam to potrdijo so: *sds, slovenija, janša*. Tu imamo še angleško besedo *the*, ki jo iz tega korpusa nismo izvzeli, saj je tokrat del samega glavnega besedila. Verjetno gre sklicevanje na razne multimedijske uspešnice. Najdemo tudi besedo, ki dokazujejo da so blogi del izražanja o osebnih doživetij in mnenj: *sam*.

Korpus Wikipedije dokazuje naravo le te. Kot vsaka enciklopedija vsebuje veliko krajšav: *sl, št* in *km*; elementov, ki prikazujejo sklicevanje na določeno stvar: *glej, povezave, članek* in *pridobljeno*; stvari, o katerih posamezna stran sploh govori: *leta, ljubljana, države, cerkev, let, mesto* in *ime*.

# 5. DOLŽINE BESED

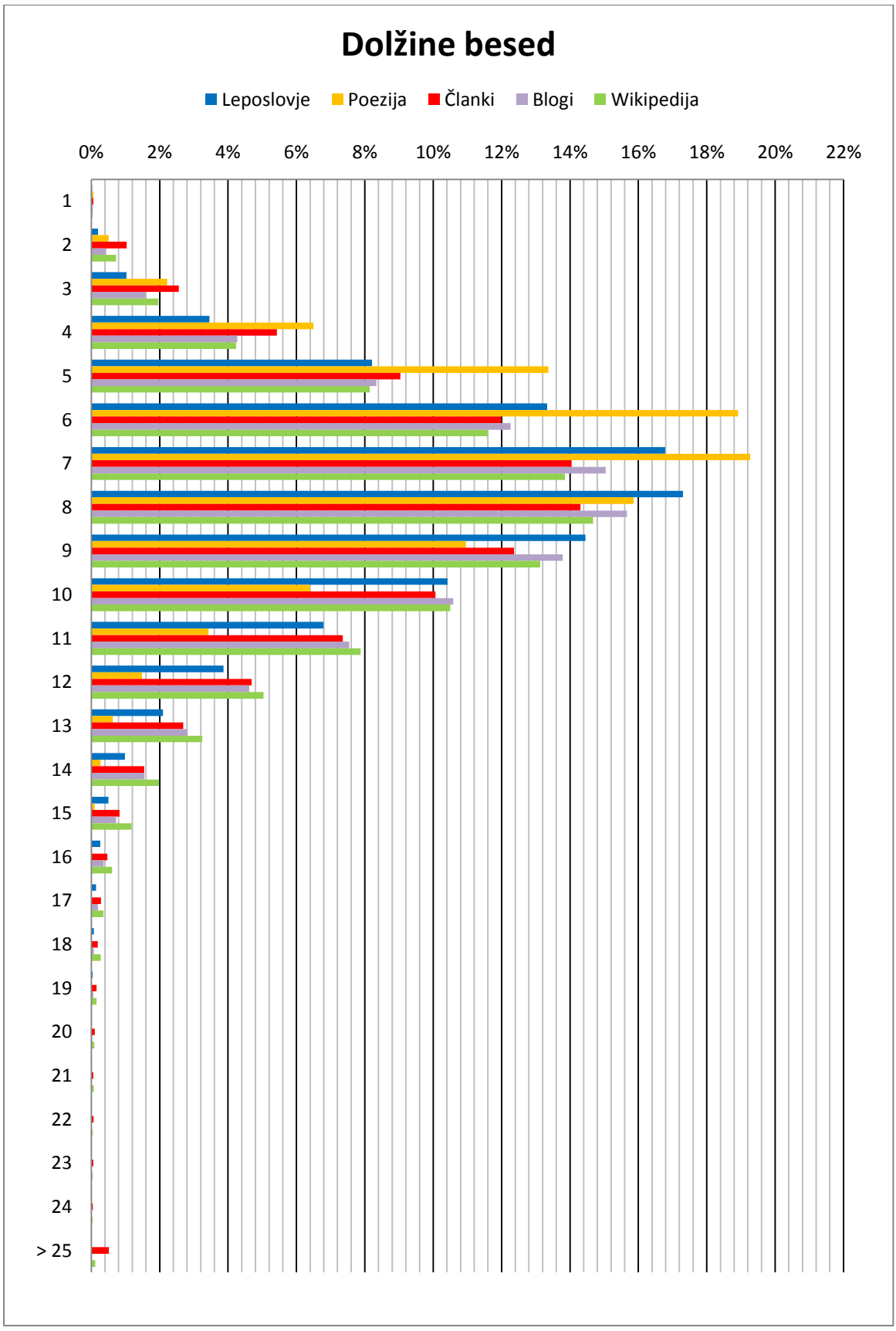
Poleg samih besed smo želeli vedeti tudi njihovo dolžino. Tu smo uporabili naše že narejene korpuse. To smo storili zato, ker nas število posameznih besed ne zanima več, poleg tega pa bi številke prišle večje, in zato izračuni procentov manj zanesljivi. Iz teh besed smo naredili slovar njihovih dolžin v programskem jeziku *Python*. Ta slovar smo nato prepisali v novo tekst datoteko, kjer smo lahko pridobljene informacije uporabili za nadaljnje raziskovanje. Spodnja tabela (Tabela 7) nam prikazuje število različnih besed v posameznem korpusu.

| Leposlovje | Poezija | Članki | Blogi | Wikipedija |
|---|---|---|---|---|
| 102.623 | 44.092 | 40.698 | 60.106 | 63.781 |

**Tabela 7: Število različnih besed v posameznem korpusu**

Seveda je bilo pričakovano, da bo leposlovje imelo najbolj razvito besedišče. V poeziji se besede mnogokrat ponavljajo, saj se pesniki polastujejo že preizkušenih besed zaradi omejitve zlogov in rim. Wikipedija ima več različnih besed zaradi strokovnih izrazov in imen mest, držav, bolezni ipd. Blogi so očitno besedno veliko bolj razviti od člankov, čeprav sta njuna namena podobna (podati bralcu informacije o dogajanju).

S tabelo dolžin in številom besed smo lahko za vsak korpus izračunali delež besed z določeno dolžino. Pridobili smo naslednji graf (Graf 2).




Graf 2: Graf števila besed glede na njihovo dolžino

Število besed glede na dolžino v leposlovju hitro narašča vse tja do dolžine osem (8) črk. Potem začne upadati nekoliko počasneje. Kot zanimivost smo pogledali katere besede so enake ali daljše od dvaindvajsetih (22) črk. Dobili smo naslednje:

- "petnajststošestinpetdesetem" –gre za število. Uporabljena je bila v Thanbiti Kumi, ki jo je napisal Ivan Pregelj.
- " petindevetdesetprocenten" – znova oblika števila. Beseda je bila uporabljena v prevodu Kako vzgojiti očeta, ki jo je originalno napisal Jordan Horowitz.
- Ostale dolge besede se pojavljajo v Ena dolga predguvor, ki jo je napisal Primož Trubar. To je besedilo, kjer se slovenščina še ni popolnoma izoblikovala in zato vsebuje veliko besed, ki so nemške izposojenke.

V poeziji pa se naraščanje preneha že pri sedmih (7) črkah. Besed, daljših od osemnajst (18) črk, praktično ni. K temu pripomorejo krajše osnovne enote poezije. Vendar se pojavlja ena beseda, ki je združek ponavljanja besede *pada*: "padapadapadapadapadapadapadapadapadapadapadapadapadapadapadapadapadapadapadapadapadapadapada …"

Pri člankih smo se pogosto srečali z dolgimi besedami, predvsem iz razloga, ker se med pisanjem sestavljenih besed niso uporabljali vezaji, sklici na slike pa so vsebovali opis le nje združeno v eno besedo. Primeri so našteti spodaj:

- "zahodnovirginijskauniverza" – program bi to besedo ločil na tri različne, če bi se v njej uporabili vezaji: "zahodno", "virginijska" in "univerza";
- "znanstvenoraziskovalnemu" bi postalo: "znanstveno" in "raziskovalnemu";
- "evangeličansko-country-glasbo" bi postalo: "evangeličasnko", "country" in "glasbo".

Tu imamo še znanstveno-kemijsko ime "polytetrafluoroethylene" in nekakšna matematična funkcija, ki vsebuje *arccos*. Zadnja se je očitno med pretvorbo v .txt datoteko zapisala v eni vrstici, da smo dobili sledeč nesmisel: "arccosxxrixririxririair".

Pri blogih gre za skoraj popolno enakomerno naraščanje in upadanje. Na žalost pa se srečamo z besedami na kakršne naletimo, ko se pisci polastijo forumske pisave:

- "hahahahahahahahahahahah",
- "dgfdgfgjhgfhfghdfresidenca",
- "vsivenglasafnagmailpikacom" (če bi tu bila pred *com* pika, bi ga program ločil od preostale besede),
- "klikklikklikklikklikklikklikklikklikklikklikklikklikklikklikklik",
- "bruhhhhhhhhhhhhhhhhhhh".

Pri Wikipediji imamo veliko besed nad 25 črk, ki jih je naš program za prepisovanje z interneta pobral namesto slik in tabel. Gre za njihove naslove v ozadju, ki sicer vizualno na strani niso vidni, nahajajo pa se v html obliki strani. Primeri teh naslovov so:

- "meteorologijimeteorologijavremepsihrometrijavode",
- "unijaportugalskaportugalsko",
- "matematikalogikaaksiomi",
- "elementovklasifikacijski",
- ipd..

# 6. ANALIZA ČRK

Ko smo opravili z besedami, smo se še bolj poglobili v naše raziskovanje in izluščili uporabo samih črk. Naša analiza črk sestoji iz štetja posameznih črk in pa statistično analizo njihovega položaja v besedi.

Za boljši pregled nad vrednostmi, smo spodaj naredili tabelo števila vseh črk v posamezni kategoriji korpusov:

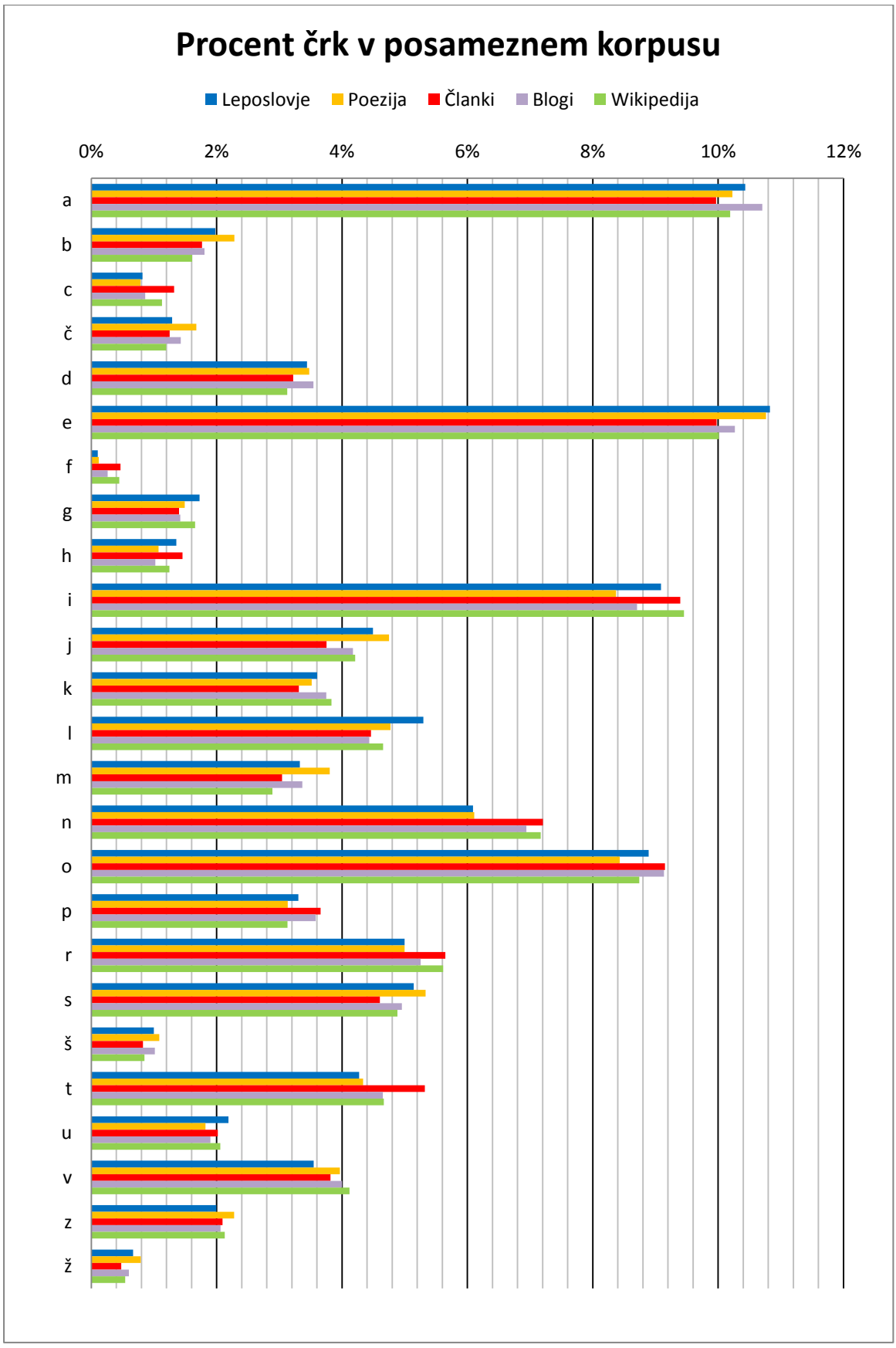
| Leposlovje | Poezija | Članki | Blogi | Wikipedija |
|---|---|---|---|---|
| 6.007.498 | 1.239.560 | 1.268.010 | 2.192.853 | 2.157.981 |

**Tabela 8: Število črk v posameznem korpusu**

## 6.1. Število posameznih črk

Ponovno smo se polastili programerskega jezika *Python* in sprogramirali kratek program, ki je prebral vsa besedila, naredil tabelo črk in štel kolikokrat se posamezna črka pojavi v njih. Štetja v že narejenih korpusih smo se izognili, saj je postopek soroden tistemu pri štetju uporabe posameznih besed. Iz teh rezultatov smo naredili tabelo in iz nje graf, ki prikazuje uporabo posameznih črk v vsaki kategoriji posebej (Graf 3).

Vnaprej smo predvidevali, da se samoglasniki pojavijo največkrat, medtem ko se naj bi šumniki prikazali najmanjkrat. Graf smo prikazali v procentnih vrednostih, da bi rezultati bili enakovredni, ne glede na število črk v posameznem korpusu.




Graf 3: Procent črk v posameznem korpusu

Graf je urejen po abecednem redu in prikazuje delež uporabe posamezne črke v procentih. Opazimo da se vsi samoglasniki, razen črke *u*, pojavljajo za 2% večkrat kot ostale črke. Prvi soglasnik se pojavi šele na 5. mestu. Čeprav so šumniki v spodnji četrtini uporabe, pa je najmanj uporabljena črka *f*. Razen pri člankih, kjer se mnogokrat zgodi, da šumnikov sploh ne pišejo (ostajajo na črkah *c, s* in *z*) in na Wikipediji, kjer količina vseh črk zasenči malo število šumnikov.

## 6.2. Samoglasniki

Samoglasniki so seveda zelo popularni, vendar ali so v prednosti pred ostalimi črkami? Seveda bomo v skriti besedi verjetno odkrili kakšen samoglasnik ali dva, vendar, če imamo odkritih že tretjino črk – vsi so samoglasniki – se nam še splača klicati ostale, neodkrite, samoglasnike? Ali bi bilo bolje, da bi se lotili ugibanja soglasnikov? Statistično frekvenco samoglasnikov smo primerjali s statistično frekvenco soglasnikov, da bi dobili nekakšen pregled, kako velik delež črk v besedi samoglasniki pravzaprav zasedajo. Spodnja tabela (Tabela 9) je rezultat tega kratkega izračuna.

| GLAS | Frekvenca (v %) | | | | | |
|---|---|---|---|---|---|---|
| | Leposlovje | Poezija | Članki | Blogi | Wikipedija | POVPREČJE |
| soglasniki | 41,426 | 39,617 | 40,507 | 40,715 | 40,458 | 40,544 |
| samoglasniki (a, e, i, o, u) | 58,574 | 60,383 | 59,493 | 59,285 | 59,542 | 59,456 |
| **SKUPAJ** | **100** | **100** | **100** | **100** | **100** | **100** |

**Tabela 9: Frekvenčna primerjava med soglasniki in samoglasniki**

Samoglasniki zasedajo približno 40% besede, kot je vidno iz izračuna povprečja, kar pomeni , da besedo dolgo pet (5) črk sestavljata dva (2) samoglasnika. Če sta ta dva že odkrita, imamo večjo možnost, da bo naslednja odkrita črka soglasnik.

Vidno je tudi, da se v leposlovju in blogih, kjer gre za daljše tekste, pojavlja več samoglasnikov kot v drugih korpusih.

## 6.3. Tujke in naglašene črke

Do sedaj smo govorili le o črkah, kot jih poznamo iz naše abecede. Zelo malokrat, pa vseeno, se v besedilih pojavljajo črke z naglasi in tuje črke. Ker se tovrstni znaki navadno ne uporabljajo, smo jih ločili od zgornje kategorije in jih združili v popolnoma novo, svojo kategorijo. Naglasi se mnogokrat pojavljajo v besedah, katerih izgovor bi avtor rad poudaril, saj bi jih v nasprotnem primeru pogosto naglasili narobe. Prav tako so naglasi včasih uporabljeni tudi v pesmih, saj je tam izgovor besede pomembnejši zaradi rime.

Tujke pa se v slovenščini pojavljajo v imenih, ki izhajajo iz drugih jezikov, in pa frazah, ki izvirajo iz angleščine. Naslednja tabela (Tabela 10) prikazuje te posebne črke in naglase, ki smo jih iz zgornjih grafov izpustili, ter število njihovih pojavitev v korpusih.

| ČRKA | ŠTEVILO POJAVITEV | | | | |
|---|---|---|---|---|---|
| | Leposlovje | Poezija | Članki | Blogi | Wikipedija |
| á | 452 | 1284 | 5 | 8 | 244 |
| â | 2 | 109 | 0 | 0 | 14 |
| ä | 62 | 4 | 11 | 5 | 84 |
| ç | 5 | 0 | 1 | 1 | 16 |
| ć | 7 | 7 | 128 | 169 | 117 |
| é | 777 | 2321 | 19 | 18 | 272 |
| ę | 1 | 0 | 0 | 0 | 7 |
| í | 180 | 1102 | 6 | 5 | 153 |
| ó | 539 | 1715 | 1 | 11 | 128 |
| ô | 33 | 386 | 0 | 1 | 26 |
| ö | 34 | 11 | 26 | 22 | 208 |
| q | 247 | 13 | 133 | 61 | 357 |
| ú | 115 | 244 | 5 | 5 | 45 |
| ü | 21 | 17 | 9 | 11 | 260 |
| w | 730 | 88 | 2028 | 1558 | 3609 |
| x | 451 | 99 | 665 | 396 | 1016 |
| y | 3622 | 125 | 2012 | 1401 | 2821 |
| ź | 2 | 0 | 0 | 0 | 10 |
| **SKUPNO ŠTEVILO VSEH ČRK** | **6.007.498** | **1.239.560** | **1.268.010** | **2.192.853** | **2.157.981** |

**Tabela 10: Število tujke in naglašenih črk v korpusih**

## 6.4. Položaj črk

Položaj črk v besedi je pomemben element analize besed. Na način statistično prikaže samo 'obliko' našega jezika. Ko prvič vidimo pisavo nekega – nam tujega – jezika, poleg dolžine besed najprej opazimo, katere črke se največkrat pojavijo in kje. Če je položaj črk podoben našemu maternemu jeziku, se nam tudi sam jezik zdi bolj domač in pravilna izgovarjava lažja. Ko ima tujec pred seboj slovensko besedilo, zazna s katero črko se besede največkrat začnejo in končajo. Z hitrim pogledom poskuša prepoznati in povezati tujo besedo z neko, ki jo že pozna [19].

Program nam je iz našega korpusa prebral vse besede – tako se le te niso ponavljale. Nato je za vsako posamezno črko prištel, kolikokrat se ta pojavi kot prva, druga, predzadnja ali zadnja črka besede. Pri dvočrkovnih besedah je program prebral le prvo in zadnjo črko. Pri tričrkovnih smo sredinski znak vzeli kot črko na drugi poziciji. Ostale, daljše besede nam seveda v tej analizi ne povzročajo problema kot je neuporabljena pozicija. Enočrkovne besede smo spustili iz analize.

Na koncu smo za lažji pregled vse skupaj prikazali v dveh različnih zapisih.

V prvem zapisu smo kot celoto vzeli posamezno pozicijo (Tabela 12 in Tabela 13). 100% črk se pojavi na poziciji ena (1). Kar pomeni, da smo bili osredotočeni na prvo črko besede. Posamezna črka je dobila le del te celote glede na to, kolikokrat se je pojavila na tem mestu. S takšnim razmišljanjem smo dobili spodnji graf, ki predstavlja vse štiri pozicije, ki smo jih raziskovali (prva [1], druga [2], predzadnja [-2] in zadnja [-1] pozicija v besedi). Vsaka pozicija je razdeljena na dele, ki seveda predstavljajo naše črke. Da se grafi ne bi ponavljali z različnimi korpusi, smo naredili tabelo, ki je bila povprečje vseh korpusov.

Za razumevanje celotne slike, smo naredili tudi tabelo števila črk v posamezni poziciji in posameznem korpusu (Tabela 11).

| | [1] | [2] | [-2] | [-1] |
|---|---|---|---|---|
| **Leposlovje** | 102.346 | 101.943 | 100.890 | 101.869 |
| **Poezija** | 43.914 | 42.849 | 41.863 | 43.322 |
| **Članki** | 40.352 | 40.019 | 39.036 | 40.113 |
| **Blogi** | 59.763 | 59.639 | 58.645 | 59.662 |
| **Wikipedija** | 63.048 | 62.550 | 61.574 | 62.923 |
| **SKUPAJ** | **309.423** | **307.000** | **302.008** | **307.889** |

**Tabela 11: Število črk na posamezni poziciji v različnih korpusih**

V nadaljevanju smo naredili tabele procenta črk na določeni poziciji za vsak korpus posebej (Tabela 12 in Tabela 13).

| ČRKA | Leposlovje | | | | Poezija | | | | Članki | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | [1] | [2] | [-2] | [-1] | [1] | [2] | [-2] | [-1] | [1] | [2] | [-2] | [-1] |
| a | 1,1 | 17,4 | 6,5 | 18,7 | 1,1 | 16,5 | 7,9 | 18,2 | 2,9 | 16,4 | 4,8 | 17,3 |
| b | 3,6 | 2,4 | 0,7 | 0,1 | 4,1 | 2,0 | 1,1 | 0,2 | 2,9 | 2,0 | 0,9 | 0,3 |
| c | 1,3 | 0,3 | 4,0 | 0,9 | 1,1 | 0,1 | 3,5 | 1,0 | 2,2 | 0,6 | 2,9 | 1,1 |
| č | 1,1 | 0,3 | 1,7 | 0,6 | 1,6 | 0,9 | 2,4 | 1,0 | 0,9 | 0,3 | 1,1 | 0,5 |
| d | 4,2 | 2,5 | 1,4 | 0,5 | 4,5 | 2,3 | 2,3 | 0,8 | 5,2 | 2,3 | 1,5 | 1,4 |
| e | 0,6 | 11,6 | 8,4 | 13,9 | 0,5 | 12,1 | 9,1 | 14,2 | 1,8 | 12,9 | 9,2 | 14,7 |
| f | 0,7 | 0,1 | 0,1 | 0,1 | 0,7 | 0,1 | 0,1 | 0,1 | 1,7 | 0,2 | 0,2 | 0,2 |
| g | 2,9 | 1,0 | 3,2 | 0,2 | 3,3 | 1,1 | 2,1 | 0,3 | 2,3 | 0,9 | 3,5 | 1,0 |
| h | 1,6 | 1,7 | 1,2 | 4,3 | 1,7 | 0,3 | 0,5 | 3,4 | 1,8 | 0,9 | 0,4 | 5,3 |
| i | 3,3 | 5,3 | 10,0 | 18,4 | 2,1 | 6,1 | 8,7 | 16,4 | 4,0 | 7,6 | 9,8 | 15,8 |
| j | 1,2 | 0,5 | 8,1 | 1,1 | 1,5 | 1,0 | 10,2 | 2,3 | 1,2 | 0,6 | 11,0 | 1,0 |
| k | 5,2 | 1,7 | 5,6 | 1,3 | 5,5 | 1,8 | 5,9 | 1,8 | 5,3 | 1,9 | 6,4 | 1,5 |
| l | 2,3 | 4,6 | 11,9 | 5,4 | 2,6 | 5,5 | 9,0 | 5,1 | 2,6 | 4,1 | 7,0 | 3,2 |
| m | 4,0 | 1,7 | 4,1 | 7,1 | 4,7 | 1,8 | 3,1 | 6,2 | 5,0 | 1,6 | 4,3 | 5,6 |
| n | 6,7 | 1,8 | 11,5 | 2,5 | 6,0 | 1,7 | 11,5 | 3,5 | 6,3 | 3,0 | 14,7 | 4,0 |
| o | 7,3 | 16,7 | 3,6 | 12,6 | 6,1 | 16,3 | 3,8 | 12,3 | 6,6 | 16,6 | 4,9 | 12,2 |
| p | 17,3 | 2,4 | 0,4 | 0,1 | 15,7 | 2,4 | 0,7 | 0,2 | 15,7 | 3,1 | 0,5 | 0,4 |
| r | 4,3 | 14,7 | 1,8 | 1,0 | 4,3 | 14,5 | 2,6 | 1,4 | 4,4 | 13,0 | 2,8 | 2,2 |
| s | 10,1 | 1,9 | 1,5 | 1,0 | 10,2 | 1,8 | 2,0 | 0,7 | 8,9 | 1,9 | 2,1 | 2,7 |
| š | 1,4 | 0,1 | 1,1 | 0,9 | 1,5 | 0,2 | 1,0 | 1,8 | 1,0 | 0,1 | 0,9 | 0,2 |
| t | 3,9 | 2,7 | 7,9 | 2,2 | 4,1 | 3,1 | 6,9 | 3,0 | 4,8 | 2,8 | 6,3 | 2,7 |
| u | 2,6 | 3,9 | 1,1 | 4,8 | 2,3 | 3,5 | 0,6 | 3,2 | 2,6 | 3,2 | 0,7 | 3,5 |
| v | 5,3 | 1,9 | 3,4 | 1,9 | 6,0 | 2,5 | 3,7 | 2,2 | 4,4 | 1,6 | 3,1 | 2,6 |
| z | 7,0 | 2,5 | 0,5 | 0,1 | 7,3 | 2,2 | 0,6 | 0,2 | 4,8 | 2,6 | 0,6 | 0,3 |
| ž | 1,0 | 0,2 | 0,5 | 0,1 | 1,4 | 0,3 | 0,7 | 0,3 | 0,7 | 0,1 | 0,4 | 0,1 |
| SKUPAJ | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

**Tabela 12: Frekvenca pojavitve črk na posamezni poziciji v leposlovju, poeziji in člankih (v %)**

| ČRKA | Blogi | | | | Wikipedija | | | |
|---|---|---|---|---|---|---|---|---|
| | [1] | [2] | [-2] | [-1] | [1] | [2] | [-2] | [-1] |
| a | 2,1 | 17,6 | 5,3 | 19,8 | 4,0 | 16,8 | 4,9 | 20,1 |
| b | 3,5 | 2,0 | 0,8 | 0,2 | 3,9 | 1,8 | 0,7 | 0,2 |
| c | 1,5 | 0,3 | 3,5 | 0,9 | 2,4 | 0,5 | 3,1 | 0,9 |
| č | 1,0 | 0,3 | 1,4 | 0,5 | 0,8 | 0,3 | 1,0 | 0,4 |
| d | 4,9 | 2,4 | 1,5 | 0,7 | 4,7 | 1,8 | 1,7 | 1,1 |
| e | 1,3 | 12,1 | 9,2 | 13,8 | 2,4 | 13,7 | 9,2 | 14,2 |
| f | 1,7 | 0,2 | 0,2 | 0,1 | 1,9 | 0,2 | 0,2 | 0,2 |
| g | 2,5 | 1,0 | 3,4 | 0,4 | 3,2 | 0,8 | 3,1 | 0,8 |
| h | 1,5 | 0,6 | 0,4 | 4,7 | 2,2 | 0,9 | 0,6 | 5,1 |
| i | 3,5 | 6,2 | 9,3 | 16,5 | 3,6 | 7,9 | 9,9 | 15,9 |
| j | 1,4 | 0,5 | 10,9 | 1,0 | 1,4 | 0,4 | 10,4 | 1,2 |
| k | 5,7 | 1,9 | 7,1 | 1,6 | 6,2 | 2,0 | 9,1 | 1,7 |
| l | 2,2 | 4,4 | 8,3 | 3,5 | 2,8 | 4,6 | 6,3 | 2,7 |
| m | 4,7 | 1,7 | 4,0 | 7,2 | 5,5 | 1,4 | 3,7 | 5,7 |
| n | 6,8 | 2,2 | 13,5 | 3,2 | 5,5 | 3,0 | 14,2 | 4,0 |
| o | 6,5 | 17,2 | 4,3 | 12,9 | 5,5 | 15,4 | 5,2 | 10,6 |
| p | 16,7 | 2,9 | 0,5 | 0,3 | 13,7 | 2,4 | 0,4 | 0,3 |
| r | 4,4 | 13,9 | 2,1 | 1,4 | 4,4 | 12,9 | 3,0 | 2,1 |
| s | 9,2 | 1,9 | 2,0 | 1,3 | 8,8 | 1,9 | 1,9 | 2,9 |
| š | 1,3 | 0,1 | 0,8 | 0,6 | 1,2 | 0,1 | 0,7 | 0,1 |
| t | 3,9 | 2,6 | 6,9 | 2,2 | 4,4 | 2,7 | 5,4 | 2,5 |
| u | 2,8 | 3,4 | 0,5 | 4,1 | 2,4 | 3,9 | 1,1 | 3,6 |
| v | 4,3 | 1,8 | 3,2 | 2,7 | 4,3 | 2,0 | 3,3 | 3,0 |
| z | 5,9 | 2,6 | 0,5 | 0,2 | 4,0 | 2,3 | 0,6 | 0,3 |
| ž | 0,8 | 0,1 | 0,5 | 0,1 | 0,6 | 0,1 | 0,3 | 0,1 |
| SKUPAJ | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

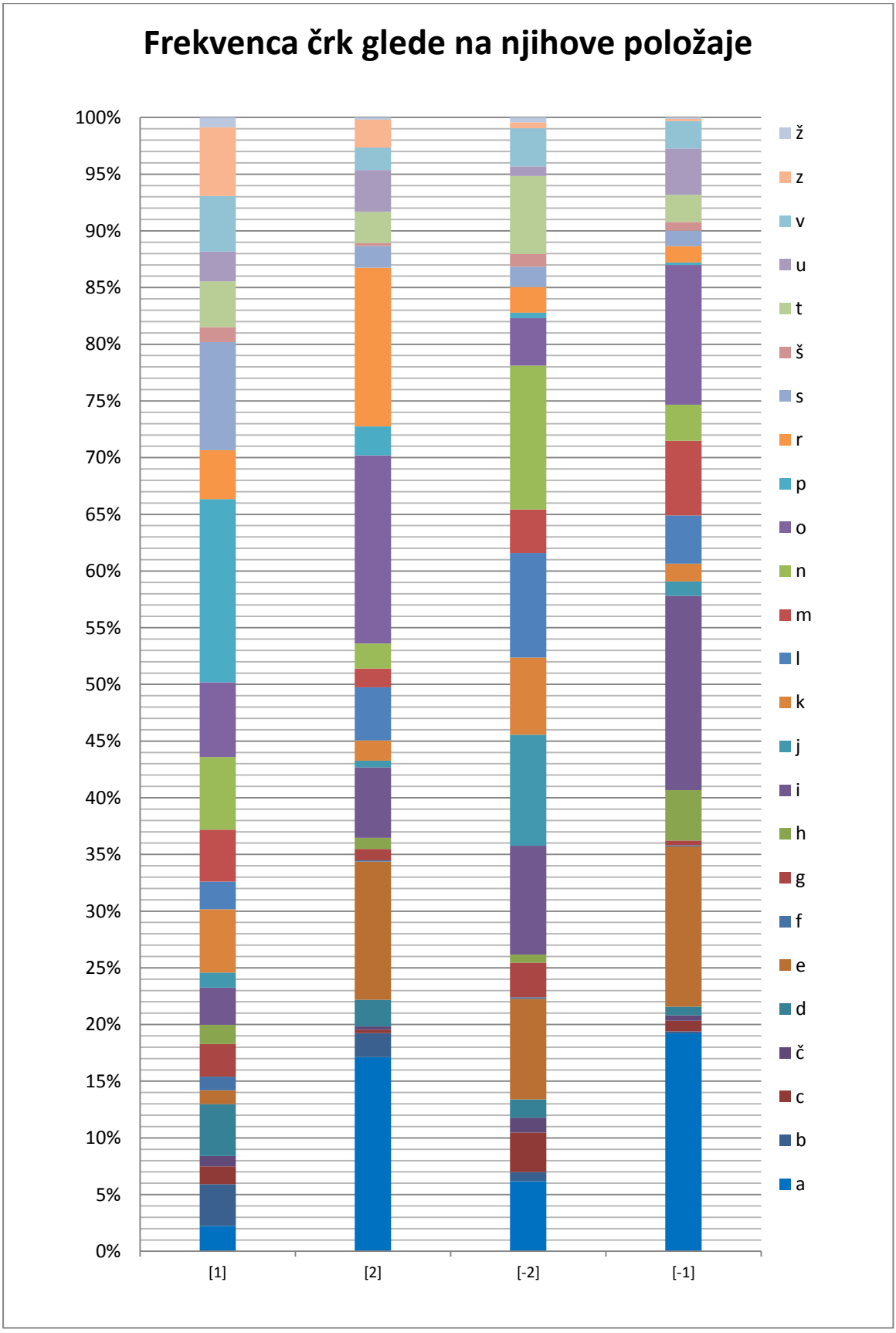
**Tabela 13: Frekvenca pojavitve črk na posamezni poziciji v blogih in korpus Wikipedije (v %)**

Nadaljevali smo z zgradbo tabele povprečja za vsak položaj (Tabela 14). To smo storili s seštetjem vseh korpusov in deljenjem z že izračunanim številom vseh črk za posamezen položaj (glej Tabelo 11). Da bi se izognili na videz nesmiselnim in dolgim številkam, smo tabelo nato prikazali v procentih.

| ČRKA | [1] | [2] | [-2] | [-1] |
|---|---|---|---|---|
| a | 2,1 | 17,0 | 5,9 | 19,0 |
| b | 3,6 | 2,1 | 0,8 | 0,2 |
| c | 1,7 | 0,3 | 3,5 | 1,0 |
| č | 1,1 | 0,4 | 1,5 | 0,6 |
| d | 4,6 | 2,3 | 1,6 | 0,8 |
| e | 1,3 | 12,3 | 8,9 | 14,1 |
| f | 1,3 | 0,1 | 0,1 | 0,1 |
| g | 2,9 | 1,0 | 3,1 | 0,5 |
| h | 1,7 | 1,0 | 0,7 | 4,5 |
| i | 3,3 | 6,4 | 9,6 | 16,9 |
| j | 1,3 | 0,6 | 9,8 | 1,2 |
| k | 5,5 | 1,8 | 6,7 | 1,6 |
| l | 2,5 | 4,6 | 9,0 | 4,2 |
| m | 4,7 | 1,7 | 3,9 | 6,5 |
| n | 6,3 | 2,3 | 12,8 | 3,3 |
| o | 6,5 | 16,5 | 4,3 | 12,2 |
| p | 16,0 | 2,6 | 0,5 | 0,2 |
| r | 4,4 | 13,9 | 2,3 | 1,5 |
| s | 9,5 | 1,9 | 1,8 | 1,6 |
| š | 1,3 | 0,1 | 0,9 | 0,7 |
| t | 4,1 | 2,8 | 6,9 | 2,4 |
| u | 2,6 | 3,7 | 0,8 | 4,0 |
| v | 4,9 | 1,9 | 3,3 | 2,4 |
| z | 5,9 | 2,5 | 0,5 | 0,2 |
| ž | 0,9 | 0,2 | 0,4 | 0,1 |
| **SKUPAJ** | **100** | **100** | **100** | **100** |

**Tabela 14: Povprečna frekvenca pojavitve črk na posamezni poziciji (v %)**

S pomočjo nove tabele smo lahko zgradili graf (Graf 4), ki vizualno prikaže delež vsake črke na pozicijah. Na desni strani smo dodali legendo barv in črk, ter stolpce uredili po abecednem redu, da ne bi prihajalo do zmede.




Graf 4: Povprečna frekvenca črk glede na mesto pojavitve

Iz vrhnjega grafa opazimo, da se največ besed začne s črko *p* in nadaljuje s samoglasniki ali črko *r*. Predzadnje črke imajo še najbolj raznoliko razdelitev; prevladujejo samoglasniki in pa *j, k, l, n* in *t*. Samoglasniki ravno tako predstavljajo večino zadnjih črk, kar je zaradi oblike slovenskih končnic pravzaprav bilo pričakovano.

Pri drugem zapisu pa smo se osredotočili na posamezne črke, ter prikazali kolikšen procent zaseda katera od štirih pozicij pri posamezni črki. Črka *a* se 100% pojavlja v nekem položaju naših besed. Če primerjamo število ponovitev pri posamezni poziciji, koliko velik delež celote zaseda ena pozicija. Glede na naše prejšnje ugotovitve, predvidevamo, da se bodo samoglasniki največkrat pojavljali na drugem in zadnjem mestu besede. Ponovno smo najprej naredili tabelo za vsak korpus posebej (Tabela 15 in Tabela 16).

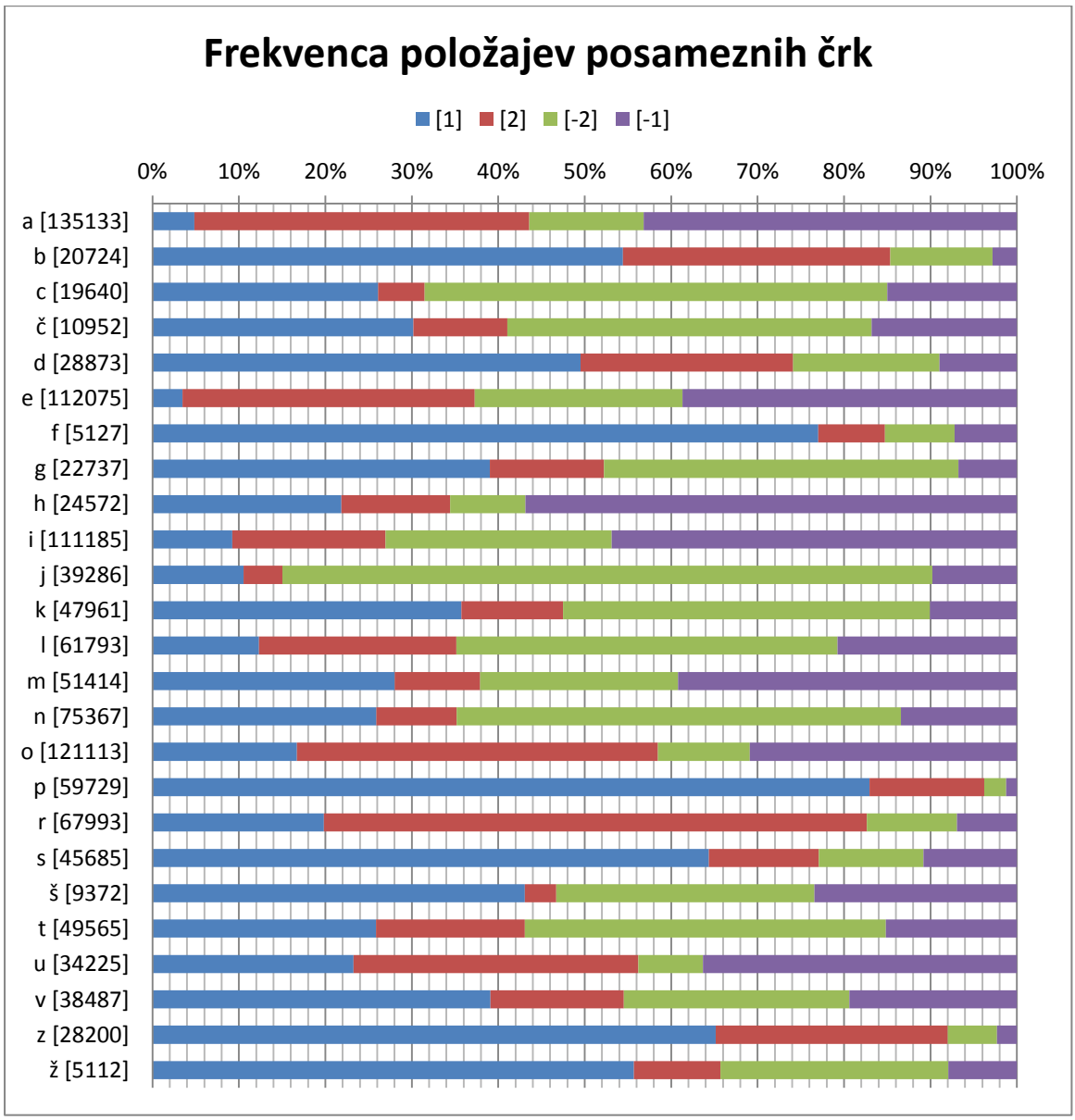
| ČRKA | Leposlovje | | | | Poezija | | | | Članki | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | [1] | [2] | [-2] | [-1] | [1] | [2] | [-2] | [-1] | [1] | [2] | [-2] | [-1] | SKUPAJ |
| a | 2,5 | 39,9 | 14,8 | 42,9 | 2,6 | 37,7 | 17,7 | 42,0 | 7,1 | 39,6 | 11,3 | 42,0 | 100 |
| b | 52,7 | 35,0 | 10,6 | 1,8 | 56,0 | 26,7 | 14,7 | 2,7 | 48,4 | 32,7 | 14,1 | 4,8 | 100 |
| c | 20,8 | 4,0 | 60,5 | 14,6 | 19,7 | 2,1 | 60,7 | 17,6 | 32,8 | 8,5 | 42,7 | 16,1 | 100 |
| č | 28,8 | 9,2 | 45,7 | 16,4 | 28,1 | 14,7 | 40,1 | 17,2 | 32,6 | 9,6 | 38,9 | 18,9 | 100 |
| d | 48,9 | 29,0 | 16,1 | 6,0 | 46,3 | 22,8 | 22,6 | 8,4 | 49,8 | 22,3 | 14,4 | 13,5 | 100 |
| e | 1,8 | 33,6 | 24,2 | 40,4 | 1,5 | 33,8 | 24,7 | 40,0 | 4,8 | 33,5 | 23,5 | 38,3 | 100 |
| f | 77,4 | 8,4 | 8,5 | 5,8 | 73,7 | 6,3 | 10,6 | 9,4 | 76,7 | 8,7 | 6,5 | 8,1 | 100 |
| g | 40,3 | 13,6 | 43,0 | 3,1 | 48,9 | 16,2 | 30,2 | 4,7 | 30,7 | 11,9 | 43,8 | 13,6 | 100 |
| h | 18,2 | 19,3 | 13,3 | 49,2 | 28,9 | 4,8 | 8,8 | 57,5 | 21,1 | 10,4 | 4,7 | 63,8 | 100 |
| i | 8,9 | 14,3 | 26,8 | 50,0 | 6,6 | 18,3 | 25,5 | 49,6 | 11,0 | 20,4 | 25,7 | 42,9 | 100 |
| j | 11,5 | 5,0 | 73,9 | 9,7 | 10,5 | 6,6 | 67,0 | 15,8 | 8,9 | 4,3 | 79,3 | 7,5 | 100 |
| k | 37,6 | 12,4 | 40,3 | 9,7 | 37,4 | 11,9 | 38,3 | 12,4 | 35,7 | 12,4 | 41,6 | 10,3 | 100 |
| l | 9,5 | 19,0 | 48,9 | 22,5 | 12,1 | 24,8 | 39,6 | 23,5 | 15,6 | 24,5 | 40,8 | 19,2 | 100 |
| m | 23,7 | 10,2 | 24,1 | 42,0 | 30,1 | 11,3 | 18,9 | 39,8 | 30,5 | 9,7 | 25,6 | 34,2 | 100 |
| n | 29,9 | 8,2 | 50,7 | 11,2 | 27,1 | 7,3 | 49,7 | 15,9 | 22,9 | 10,9 | 51,9 | 14,3 | 100 |
| o | 18,3 | 41,6 | 8,8 | 31,3 | 16,2 | 42,2 | 9,6 | 32,1 | 16,5 | 41,3 | 11,9 | 30,3 | 100 |
| p | 85,5 | 11,8 | 2,1 | 0,6 | 83,2 | 12,2 | 3,4 | 1,2 | 79,8 | 15,4 | 2,6 | 2,2 | 100 |
| r | 19,7 | 67,5 | 8,1 | 4,8 | 19,3 | 63,3 | 11,1 | 6,3 | 20,0 | 57,9 | 12,4 | 9,7 | 100 |
| s | 69,9 | 13,2 | 10,3 | 6,7 | 70,0 | 12,0 | 13,2 | 4,8 | 57,2 | 12,0 | 13,3 | 17,4 | 100 |
| š | 40,5 | 3,7 | 30,6 | 25,2 | 33,6 | 3,9 | 22,2 | 40,3 | 47,8 | 2,9 | 39,6 | 9,8 | 100 |
| t | 23,4 | 16,4 | 47,1 | 13,0 | 24,5 | 18,3 | 39,4 | 17,9 | 29,4 | 17,2 | 37,2 | 16,2 | 100 |
| u | 21,3 | 31,5 | 8,6 | 38,5 | 24,5 | 36,2 | 5,7 | 33,5 | 26,2 | 32,0 | 6,5 | 35,3 | 100 |
| v | 42,6 | 15,1 | 26,9 | 15,5 | 42,6 | 17,5 | 24,8 | 15,1 | 38,3 | 13,4 | 25,7 | 22,6 | 100 |
| z | 69,4 | 25,1 | 4,4 | 1,2 | 71,5 | 21,0 | 5,6 | 1,8 | 57,8 | 31,7 | 6,8 | 3,8 | 100 |
| ž | 57,2 | 10,1 | 26,2 | 6,4 | 53,8 | 11,8 | 25,0 | 9,4 | 55,7 | 7,0 | 28,5 | 8,8 | 100 |

**Tabela 15: Frekvenca posamezne pozicije pri vsaki črki v leposlovju, poeziji in člankih (v %)**

| ČRKA | Blogi | | | | Wikipedija | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | [1] | [2] | [-2] | [-1] | [1] | [2] | [-2] | [-1] | SKUPAJ |
| a | 4,8 | 39,3 | 11,6 | 44,3 | 8,8 | 36,6 | 10,6 | 44,0 | 100 |
| b | 54,3 | 30,4 | 12,2 | 3,1 | 59,9 | 26,8 | 10,0 | 3,3 | 100 |
| c | 24,4 | 4,7 | 56,0 | 14,9 | 34,8 | 7,9 | 43,7 | 13,5 | 100 |
| č | 32,5 | 9,7 | 42,5 | 15,3 | 32,5 | 11,1 | 39,0 | 17,4 | 100 |
| d | 51,5 | 25,8 | 15,3 | 7,4 | 50,9 | 19,6 | 17,5 | 12,1 | 100 |
| e | 3,6 | 33,3 | 25,0 | 38,1 | 6,1 | 34,7 | 23,0 | 36,2 | 100 |
| f | 78,1 | 8,0 | 7,5 | 6,5 | 77,0 | 6,9 | 8,4 | 7,7 | 100 |
| g | 33,8 | 14,3 | 45,8 | 6,1 | 41,1 | 10,5 | 38,6 | 9,8 | 100 |
| h | 21,5 | 7,8 | 5,5 | 65,1 | 25,0 | 10,6 | 6,3 | 58,1 | 100 |
| i | 9,9 | 17,6 | 25,9 | 46,6 | 9,7 | 21,3 | 26,2 | 42,9 | 100 |
| j | 10,2 | 3,6 | 78,6 | 7,7 | 10,6 | 3,3 | 77,2 | 8,8 | 100 |
| k | 35,3 | 11,7 | 43,0 | 9,9 | 33,0 | 10,6 | 47,3 | 9,2 | 100 |
| l | 12,1 | 24,1 | 44,7 | 19,0 | 17,2 | 27,9 | 38,0 | 16,9 | 100 |
| m | 26,7 | 9,4 | 22,6 | 41,2 | 33,7 | 8,9 | 22,2 | 35,2 | 100 |
| n | 26,8 | 8,7 | 51,9 | 12,5 | 20,8 | 11,4 | 52,5 | 15,2 | 100 |
| o | 15,9 | 42,2 | 10,5 | 31,5 | 15,2 | 41,8 | 13,8 | 29,1 | 100 |
| p | 81,7 | 14,4 | 2,6 | 1,4 | 81,5 | 14,2 | 2,6 | 1,7 | 100 |
| r | 20,3 | 63,7 | 9,6 | 6,4 | 19,9 | 57,4 | 13,2 | 9,5 | 100 |
| s | 64,1 | 13,5 | 13,4 | 8,9 | 57,2 | 12,2 | 12,2 | 18,4 | 100 |
| š | 46,2 | 4,1 | 28,5 | 21,2 | 57,1 | 3,0 | 34,8 | 5,2 | 100 |
| t | 25,2 | 16,8 | 43,6 | 14,4 | 29,5 | 18,2 | 35,4 | 17,0 | 100 |
| u | 25,9 | 31,7 | 4,3 | 38,2 | 21,8 | 35,3 | 10,0 | 32,9 | 100 |
| v | 35,8 | 15,4 | 26,2 | 22,7 | 34,2 | 15,7 | 26,0 | 24,1 | 100 |
| z | 64,7 | 27,9 | 5,3 | 2,1 | 55,4 | 31,8 | 8,3 | 4,5 | 100 |
| ž | 52,4 | 9,3 | 29,9 | 8,4 | 58,9 | 9,8 | 23,1 | 8,2 | 100 |

**Tabela 16: Frekvenca posamezne pozicije pri vsaki črki v blogih in korpusu Wikipedije (v %)**

Ponovno smo iz dobljenih tabel zgradili tabelo povprečja, iz nje pa smo izpeljali naslednji graf (Graf 5). V grafu smo poleg črk tudi zapisali število vseh pojavitev posamezne črke – sešteli smo le števila pojavitev, ko se je črka pojavila na prvem, drugem, predzadnjem in zadnjem mestu besede.




**Graf 5: Povprečna razdelitev črk glede na pojavitev v posameznih položajih**

Zgornji graf potrdi naše prejšnje ugotovitve, da se na drugem in zadnjem mestu največkrat pojavljajo samoglasniki. Posledično lahko trdimo, da so to mesta, ki zavzamejo največji procent samoglasnikov.

Če se osredotočimo na črko *f*, ki je najmanj uporabljena črka v slovenščini, opazimo, da največji del njene linije predstavlja prvo mesto besede. Isto velja za črko *p,* ki smo jo pri prejšnjem grafu (glej Graf 4), določili kot za najbolj pogosto pojavitev na prvem mestu. Ravno tako opazimo, da največji delež črke *r* sestavlja drugo mesto besede – kar smo ravno tako že razbrali iz prejšnjega grafa.

Črke *b*, *g*, *p* in *z* se skoraj sploh ne pojavljajo na zadnjemu mestu besede. Črka *i* se največkrat pojavlja na zadnjemu mestu, presenetljivo manjkrat pa na ostalih treh mestih – očitno gre za črko, ki je največkrat uporabljena v končnicah besed. Vendar poleg tega samoglasnika lahko ostale razberemo iz grafa tudi brez skale na oseh, saj imajo vsi zelo podoben vzorec pojavitve.

# 7. N-GRAMI

N-gram je zaporedje *n* črk, ki se pojavlja v besedilu. Ker ima slovenščina v nasprotju z angleščino veliko enočrkovnih besed, je tudi n-gramov v naših rezultatih primerno manj, saj se te besede niso vključile v naše štetje. V programu *Python* smo spisali program, ki je kot vir besed uporabil naše narejene tabele, ki jih je sprva združil v eno samo. Tako smo preprečili ponavljanje besed in zato neprimerno veliko razliko med n-grami. Pri računanju možnih rezultatov smo uporabljali 43 kot število črk (25 osnovnih črk + 18 tujih in naglašenih).

## 7.1. 2-gram

2-gram vključuje vsa zaporedja dveh črk, ki se nahajajo v našemu korpusu, oziroma na splošno v slovenščini. Program je izpustil enočrkovne besede, nato pa iz ostalih besed izluščil vse mogoče kombinacije dveh črk. Postopek je bil enostaven: iz besede (primer: *sosed*) je razbral kombinacijo prvih dveh črk (primer*: so*), nato pa prvo črko odstranil iz besede (primer: *osed*) in ponovil postopek, dokler ni odstranil vseh črk razen zadnje. Za vsako kombinacijo posebej je seštel njihovo uporabo v korpusu.

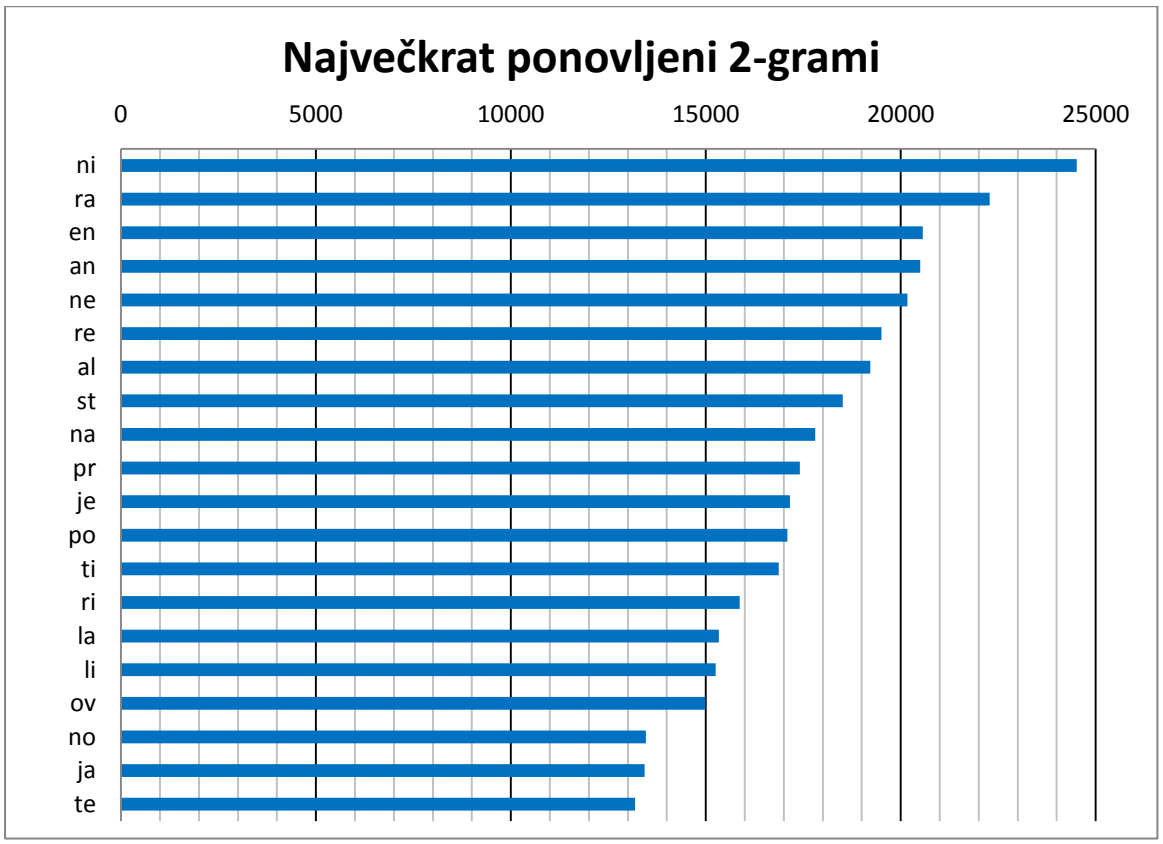
Račun za izračun vseh možnih kombinacij (6.1):

$r$ – velikost posamezne kombinacije
$n$ – velikost množice iz katere kombiniramo

$$n^r = 43^2 = 1849 \tag{6.1}$$

Možnih kombinacij je 1849, naša tabela pa je vsebovala 1318 različnih 2-gramov ter število njihovih ponovitev. Iz našega korpusa je bilo mogoče prebrati 1.485.133 vseh 2-gramov.

Namesto dolgih grafov, smo se odločili, da je bolj primerno, če naredimo krajšo tabelo rezultatov. Tako smo v grafu (Graf 6) prikazali dvajset (20) največkrat ponovljenih 2-gramov ter število njihovih ponovitev.




Graf 6: Največkrat ponovljeni 2-grami

Opazimo, da gre večinoma za kombinacije črk, ki se uporabljajo v končnicah besed. V isti smeri bomo nadaljevali še v nadaljnjih n-gramih in primerjali rezultate.

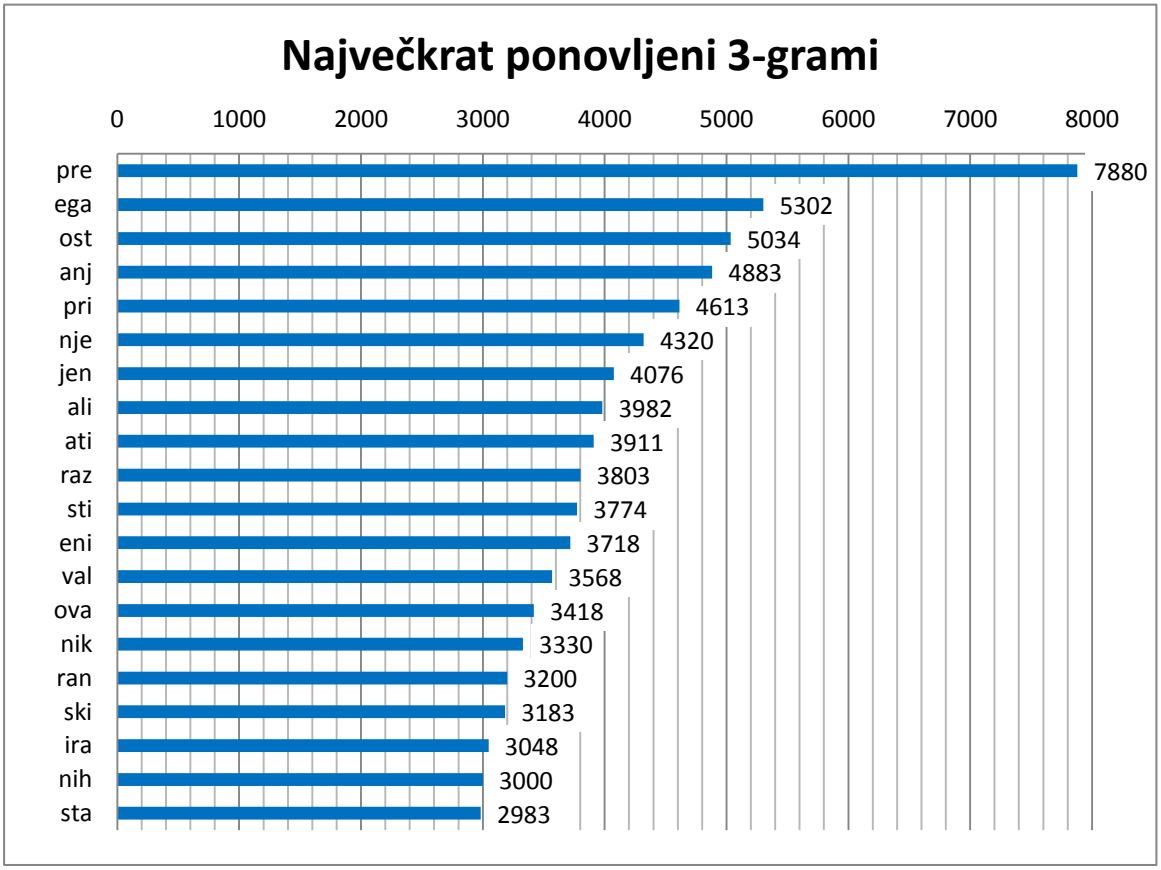
## 7.2. 3-gram

3-gram je zaporedje treh črk, ki se pojavljajo v besedilu. V tem primeru smo za besedilo vzeli naš korpus, da se pregledane besede ne bi ponavljale. Postopek pridobitve informacij je bil enak tistemu, ki smo ga uporabili pri iskanju 2-gramov, le da smo tokrat izpustili vse besede, ki so bile krajše od treh znakov, nato pa prebirali zaporedje do predzadnje črke posamezne besede.

Račun za izračun vseh možnih rezultatov (6.2):

$$n^r = 43^3 = 15.625 \tag{6.2}$$

Mogočih je 15.625 različnih kombinacij črk. Naša tabela jih vsebuje 14.953, kar je velika večina kombinacij. Iz naših korpusov smo razbrali 1.276.909 vseh 3-gramov. Sledeči graf (Graf 7) predstavlja tiste z največ ponovitvami.




**Graf 7: Največkrat ponovljeni 3-grami**

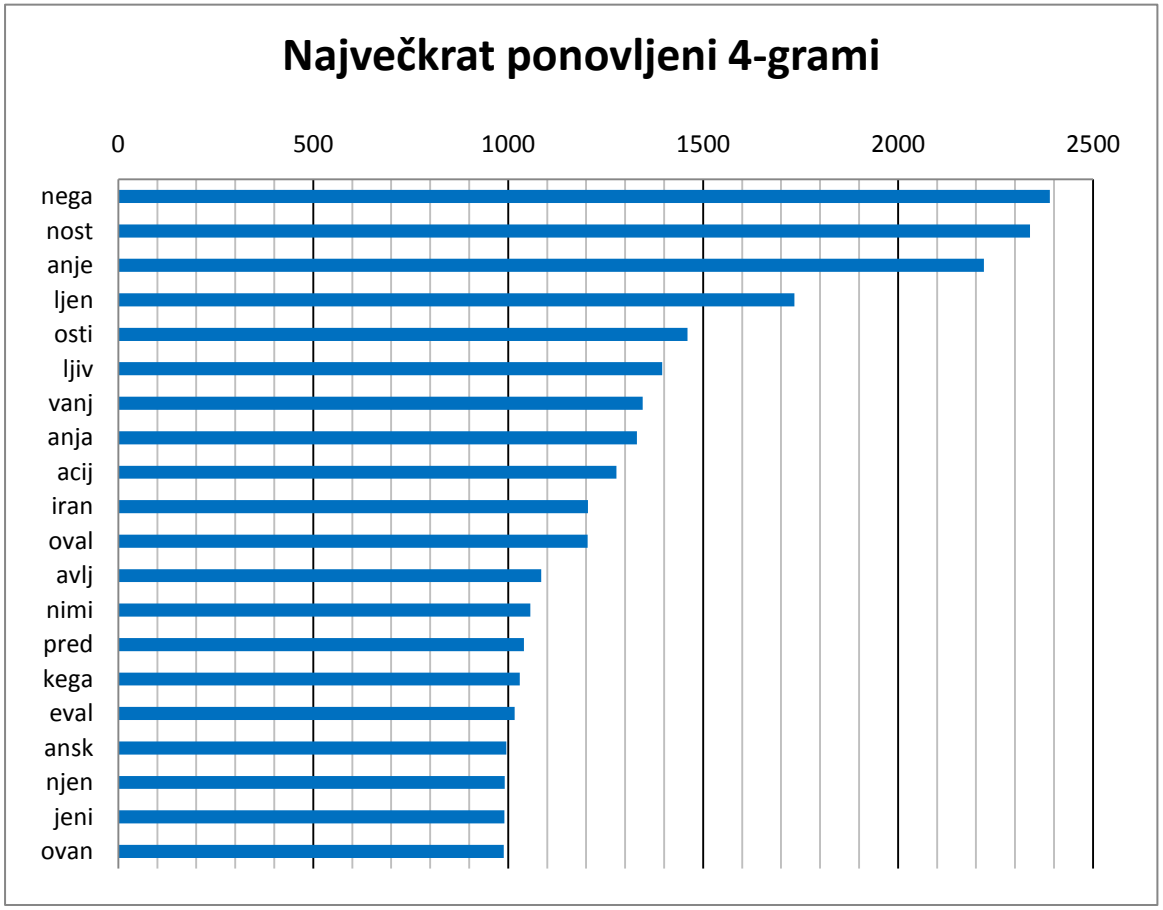
3-gram *pre* se pojavi skoraj dvakrat več kot ostali. Gre za predpono besedam. Ponovno je na grafu največ končnic in predpon.

## 7.3. 4-gram

Kakor prejšnja dva, je 4-gram zaporedje – tokrat štirih črk. Izluščili smo besede krajše od štirih znakov in naše rezultate primerjali z izračunom možnih 4-gramov, ki je prikazan spodaj (6.3).

$$n^r = 43^4 = 3.418.801 \qquad (6.3)$$

Program je v združenemu korpusu našel 71.038 različnih 4-gramov. Število vseh najdenih je bilo 1.069.367. Opazimo, da se število najdenih n-gramov manjša z večjim *n*-jem, kot se tudi mora.  Iz naših rezultatov smo naredili graf najbolj uporabljenih (Graf 8).



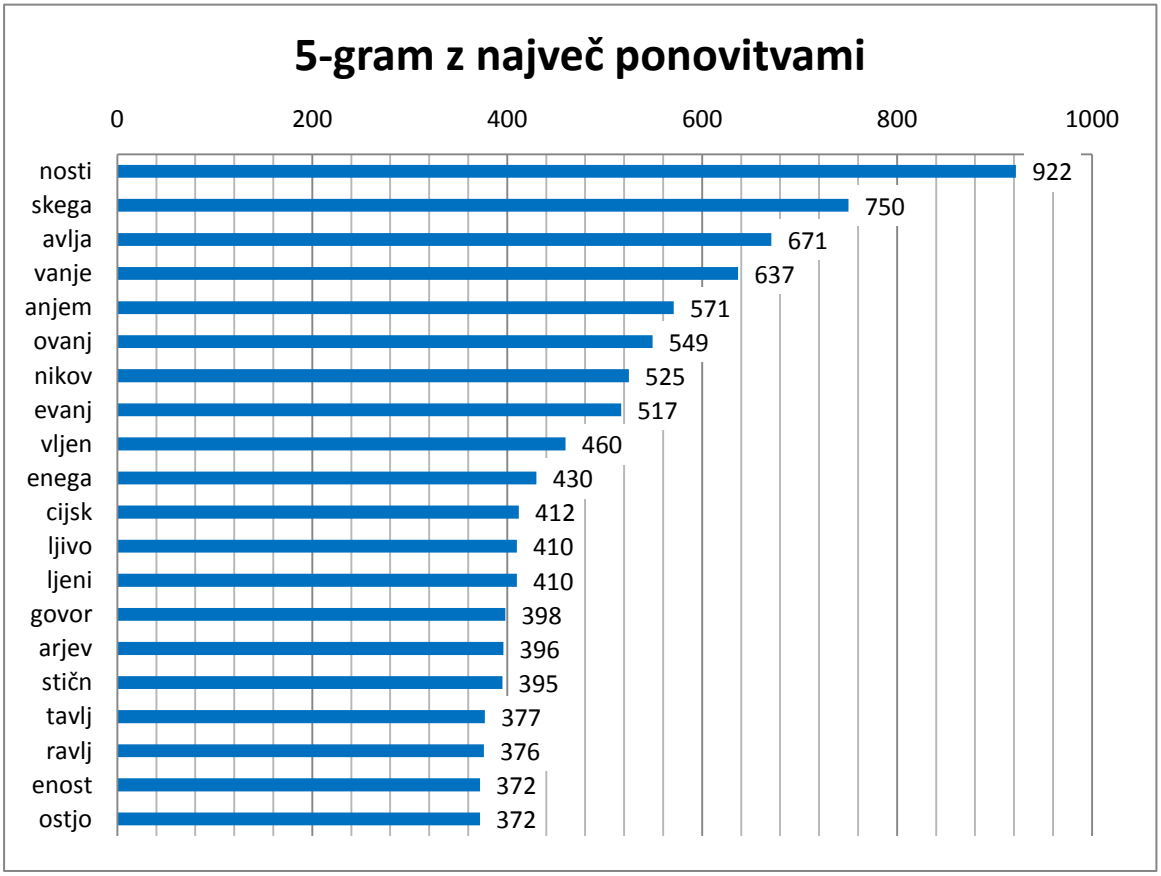

Graf 8: Največkrat ponovljeni 4-grami

## 7.4. 5-gram

Gre za zaporedje petih črk. Tokrat je v poštev prišlo še manj besed, saj smo izključili vse, ki so bile krajše od petih znakov. Vendar je iz grafa dolžin besed (Graf 2) razvidno, da je še vedno velika večina besed prišla v poštev, saj se krajše besede ponovijo večkrat, kar pomeni, da v našem korpusu zasedajo manj prostora kot bi, če bi prebirali vsa besedila znova (6.4).

$$n^r = 43^5 = 147.008.443 \qquad (6.4)$$

Možnih kombinacij črk je 147.008.443. V našem korpusu se nahaja le 165.431 različnih 5-gramov in 865.019 ponovitev. Iz grafa najbolj ponovljenih 5-gramov (Graf 9) lahko vidimo, da se je tudi abscisna os zelo skrajšala.




**Graf 9: Največkrat ponovljeni 5-grami**

## 7.5. Ugotovitve

Največkrat ponovljen 2-gram je zaporedje ni – končnica slovenščine, ki je očitno najbolj pogosto uporabljena. Pri ostalih n-gramih se ravno tako najbolj pogosto pojavljajo kombinacije črk, ki sestavljajo slovenske predpone in končnice. Iz tega razloga smo v nadaljevanju naredili še n-grame lematiziranih besed, da bi primerjali koliko ponovitev se pravzaprav nahaja v takšnih, skrajšanih, besedah.

Seveda je bilo predvideno, da se bodo z večjimi možnostmi kombinacij ponovitve zmanjšale. Primerjajmo največkrat uporabljeni 2-gram *ni* z skoraj 25.000 ponovitvami. Nato 3-gram *pre* s približno 8.000 ponovitvami, ki pa ima kar dva tisoč ponovitev več kot

3-gram, ki se nahaja drugi v vrsti. 4-gram *nega* se pojavi manj kot 2.500-krat. 5-gram *nosti* pa skoraj 1000-krat.

# 8. LEMATIZACIJA BESED

Lematizacija besed je v sorodu z, ni pa enako kot, krnjenjem. Krnenje dodatek osnovni besedi le odseka, medtem pa lematizacija spremeni isto besedo v njeno osnovno obliko (*govorjenje* postane *govoriti*) [1].

Ker je program, ki smo ga sprogramirali v Pythonu vzel preveč časa in dajal nepravilne rezultate, smo se zatekli k programu *LemmaGen* (*http://lemmatise.ijs.si/*). S pomočjo le tega, smo naš združeni korpus lematizirali in združili enake osnovne besede. Tako smo dobili tabelo 'golih' besed in njihovih ponovitev. V nadaljevanju smo izvedli podobno analizo kot pri vseh ostalih korpusih.

## 8.1. Najbolj uporabljene besede

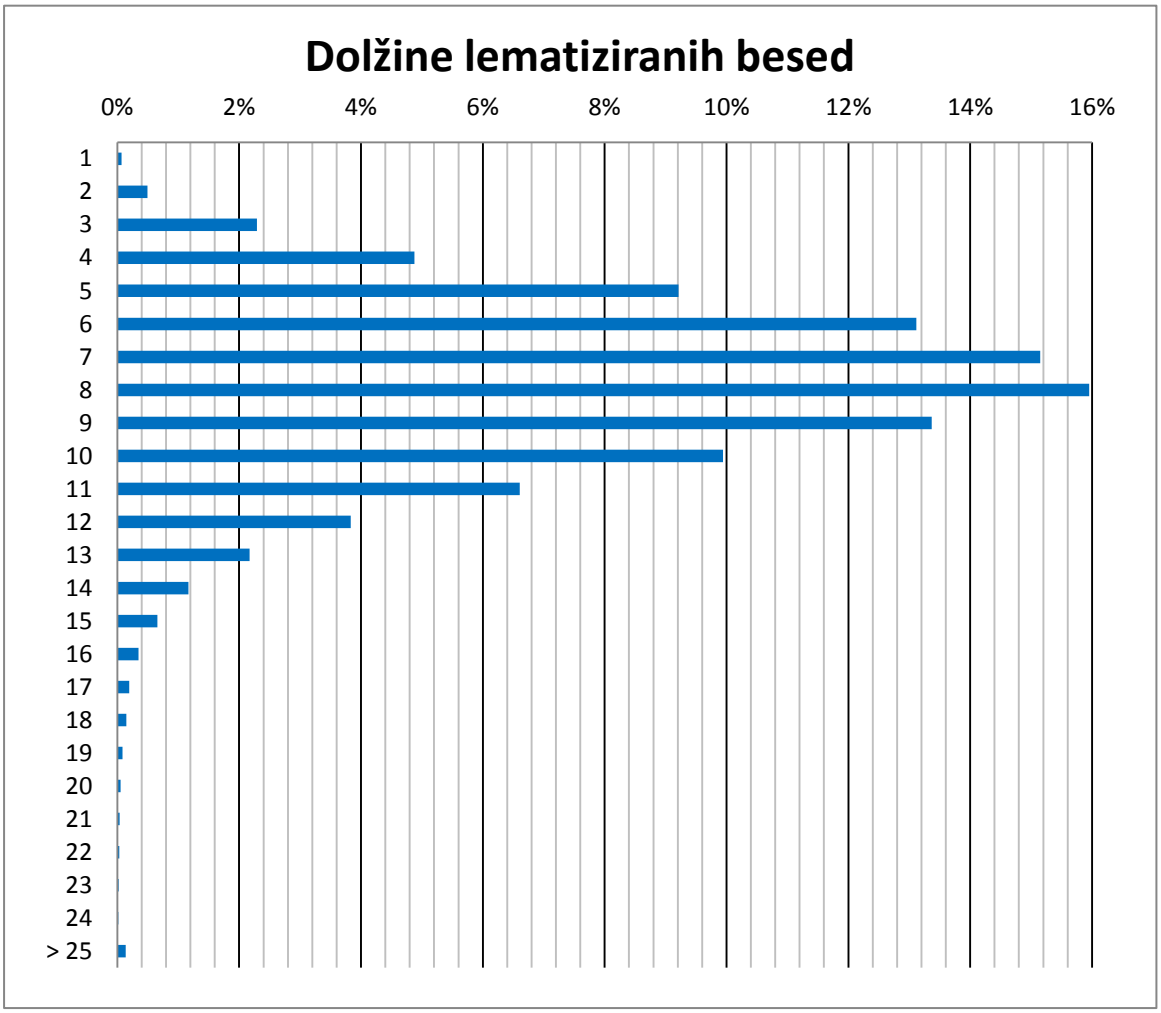
Iz našega združenega korpusa smo najprej odstranili polnila, nato pa ga poslali skozi zgoraj opisani program, in dobili rezultate. Primerjali smo dvajset največkrat ponovljenih besed med lematizirano verzijo in ostalimi korpusi (Tabela 17). Z rdečo smo označili besede, ki so enake ali vidno izpeljane iz lematizirane verzije. Popolnoma enake smo še odebelili.

| Lematizirano | Leposlovje | Poezija | Članki | Blogi | Wikipedija |
|---|---|---|---|---|---|
| **dan** | sam | **dan** | slika | janša | leta |
| dejati | oči | svet | let | danes | glej |
| **čas** | dobro | sam | leta | sds | del |
| **srce** | **dan** | **srce** | dela | **dan** | povezave |
| **noč** | gospod | **noč** | bolnikov | gre | zunanje |
| **nazaj** | bog | oči | **glede** | leta | ima |
| misel | rekel | **čas** | hoje | tek | org |
| takoj | oče | bog | št | dobro | ljubljana |
| **najbolj** | imel | dni | otrok | ima | sl |
| kmalu | roko | gre | rehabilitacija | **čas** | št |
| naprej | človek | glej | test | the | vzpostavljeno |
| glas | ima | življenje | letn | sam | km |
| delo | glavo | pesem | stran | skupaj | države |
| daleč | mati | pesmi | **najbolj** | **najbolj** | cerkev |
| enkrat | prišel | **nazaj** | dne | pot | članek |
| vrata | šel | vem | ima | slovenija | pridobljeno |
| tukaj | dolgo | dekle | vrednosti | strani | let |
| **ime** | obraz | en | amputaciji | poti | mesto |
| dovolj | videl | lepo | protezo | časa | **ime** |
| **glede** | janez | dva | objavil | del | snovi |

**Tabela 17: Primerjava lematiziranega korpusa z ostalimi**

## 8.2. Dolžine lematiziranih besed

Poleg primerjanja besed, smo primerjali še njihove dolžine. Naslednji graf predstavlja dolžine lematiziranih besed (Graf 10).



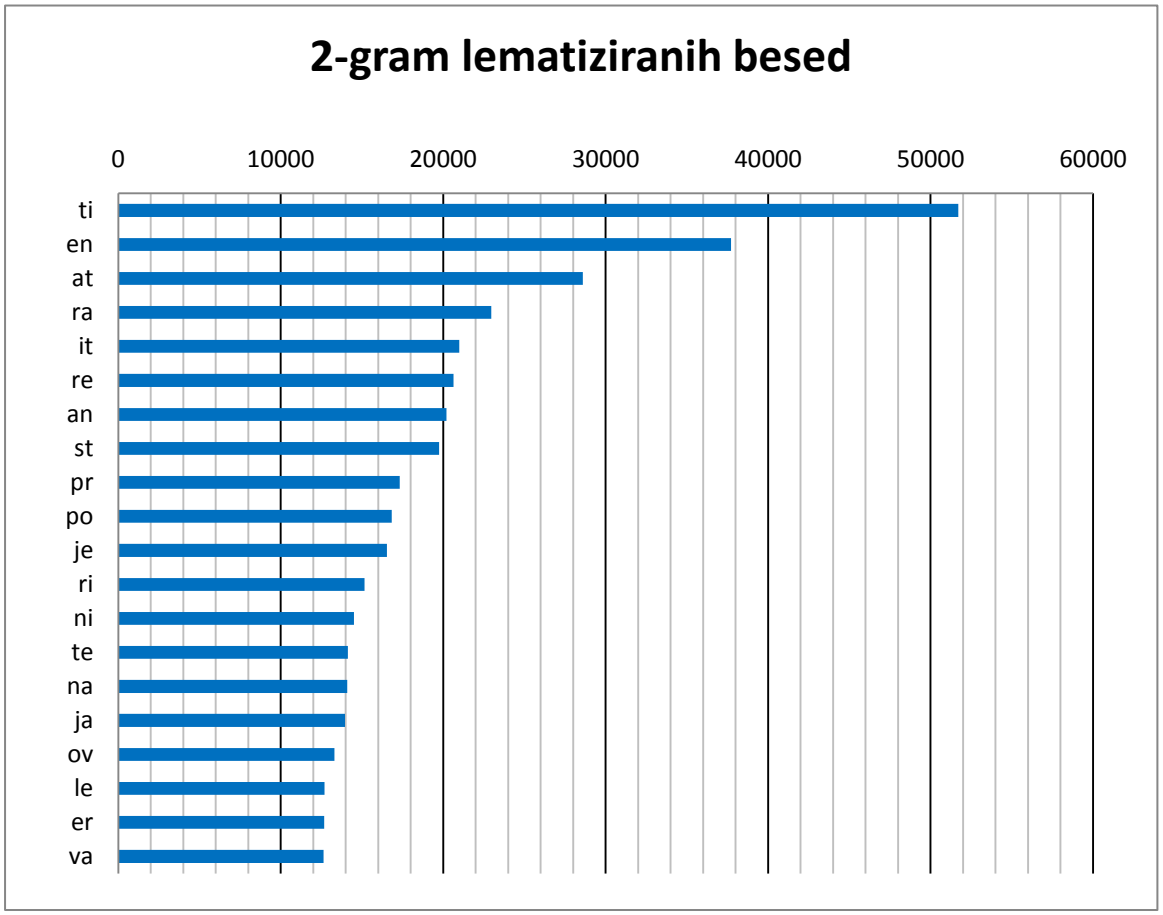

**Graf 10: Dolžine nelematiziranih besed**

Če zadnji graf primerjamo z grafom dolžin besed nelematiziranih korpusov (Graf 2), opazimo, da tudi tu graf neha naraščati pri osmem (8) znaku. Naraščanje pa je hitrje, kakor tudi upadanje besed. To povzroča lematizacija glagolov, saj pri nekaterih besedah vzame nekaj črk (primer: *prišel* postane *šel)* ter jih pri drugih nekaj doda (primer: *ima* postane *imeti*).

## 8.3. N-grami lematiziranih besed

Za popolno analiziranje lematiziranih besed smo iz njih sestavili naslednje n-grame.

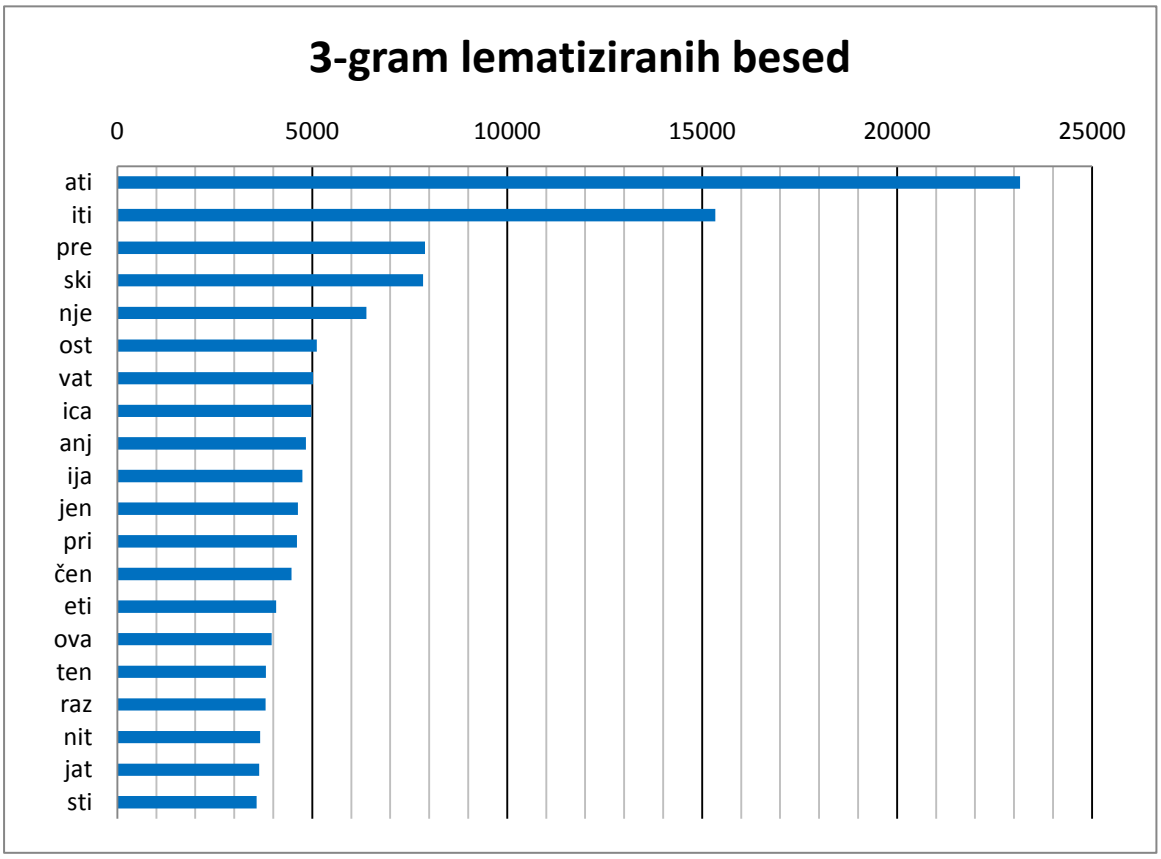
2-gram je štel 1.355 različnih kombinacij, ter 1.405.151 vseh. Naredili smo tudi graf dvajsetih (20) najbolj ponovljenih (Graf 11).




**Graf 11: 2-gram lematiziranih besed**

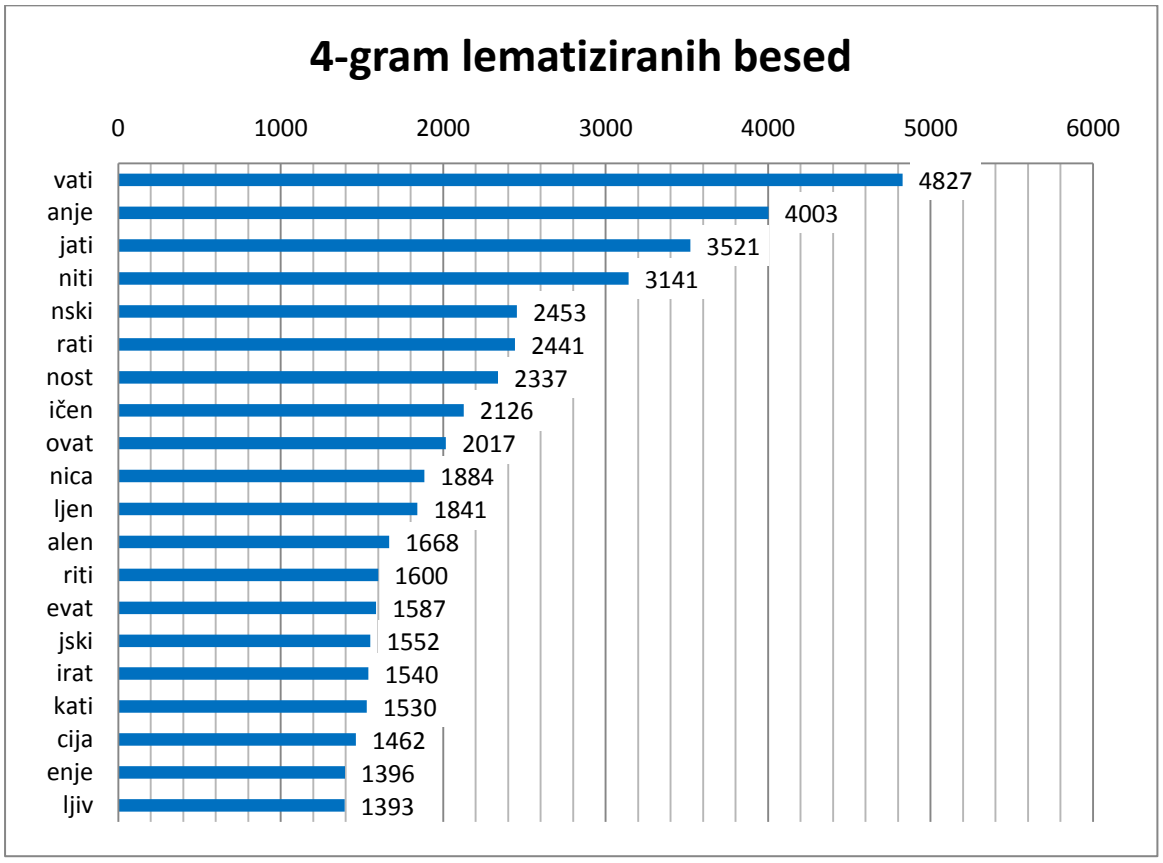
Iz rezultatov opazimo, da je kombinacija *ni* upadla za 10.000 ponovitev. Medtem se je kombinacija *ti* tako povečala, da je kar več kot dvakrat bolj pogosta od večine ostalih. K temu pripomorejo osnovne oblike glagolov, ki smo jih lematizirali (primer: *imeti*).

3-gram nam je medtem prinesel 1.197.081 kombinacij, od tega 14.430 različnih. Iz grafa dvajsetih (20) največkrat ponovljenih (Graf 12) vidimo, da je upadla večina predpon in končnic, ne pa vse – kar pomeni, da so kombinacije očitno del besede, ki se ne nahaja na začetku ali koncu. Število končnic osnovnih oblik besed se povečalo. Kombinacija *ati* se pojavlja trikrat več kot večina drugih, kombinacija *iti* pa dvakrat.



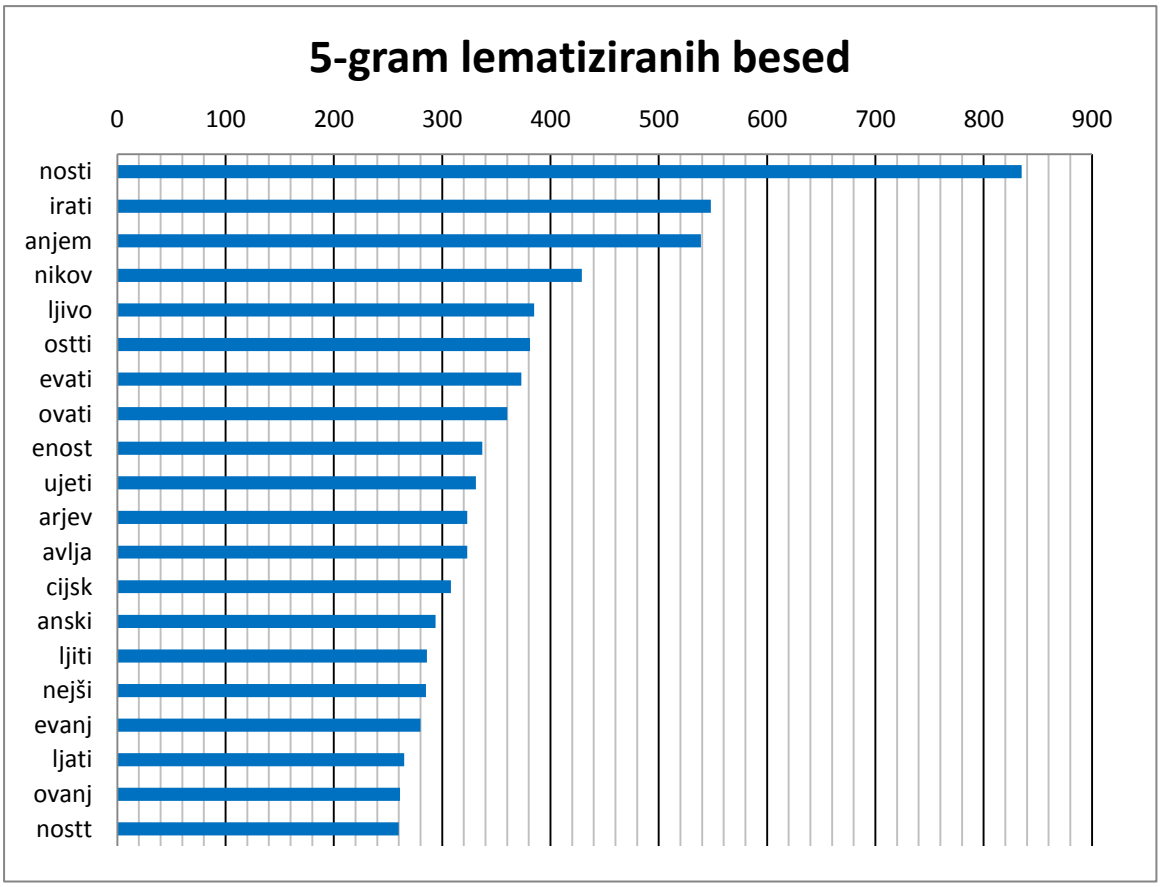

Graf 12: 3-gram lematiziranih besed

V grafu 4-gramov (Graf 13) se količina končnic osnovnih oblik besed še veča, s tem da dobimo več kombinacij tistih, ki so že prikazane v prejšnjih grafih (primer: *ati* se razdeli na *bati* in *jati*). Vseh ponovitev 64.964 kombinacij je 990.159. Opazimo, da se število vseh kombinacij primerno manjša, saj krajših besed ne moremo več vključevati v naš program.




**Graf 13: 4-gram lematiziranih besed**

Graf 5-gramov ima 589.295 kombinacij, od tega 148.463 različnih. Vidimo, da je število različnih kombinacij poskočilo za več kot dvakrat, kar nam potrdi prejšnje ugotovitve, da gre za končnice, ki se začenjajo razlikovati z več dodanimi črkami. Tu tudi naletimo na prve napake našega preprostega lematizatorja in sicer, zaporedne črke. To tovrstnih napak pride, če beseda ustreza kriterijem za lematizacijo, vendar pa gre za enega od nepredvidenih in zato nepredpisanih možnosti.




Graf 14: 5-gram lematiziranih besed

# 9. UPORABA STATISTIČNE ANALIZE

Poleg vrednosti znanja o sestavi slovenskih besedil, smo s primerjanjem korpusov dobili večji vpogled o tem, kako izvor besedila vpliva na obliko slovenščine. Očitno je, da se je z elektronskim razvojem tudi uporaba domačega jezika nekoliko spremenila, potujila, oziroma za uporabnike interneta verjetno postala bolj praktična. Še vedno pa imamo daljša besedila (blogi), ki uporabljajo čisto slovensko besedišče, brez velikih sprememb.

Znanje sestav besedila nam lahko pride prav pri kriptografiji. Kriptografija je, ko ima neko sporočilo skriti pomen, ki pa ga lahko dešifriramo samo s pomočjo nekega ključa, več o tem smo opisali v nadaljevanju.

## 9.1. Skrita sporočila

Prvo skrivanje podatkov se je imenovalo *steganografija* [20]. Gre za fizično prekritje samega sporočila. To vključuje njihovo tihotapljenje, medtem ko so bile same besede zalite z voskom in tako niso vzbujale pozornosti dokler niso dosegle cilja. Kitajci so sporočilo napisano na svili zmečkali v kroglice in jih potopili v vosek – te kroglice je sel nato pogoltnil.

Naslednja stopnja je bila *kriptografija* [20], za uporabo katere dandanes večina sporočil potrebuje nek ključ, ki je poslan na destinacijo ločeno od sporočila. Za daljša sporočila se je uporabljala *transpozicija* – mešanje črk v besedi (anagrami), za krajša pa tudi *substitucija* – zamenjava črk z drugimi črkami ali simboli. Substitucija se deli na *šifriranje* – zamenjava črk v besedilu in *kodiranje* – zamenjava celotnih besed.

Naš ključ v in sporočilo v tem času lahko pošljemo skozi alogaritem, iz katerega nato dobimo kodirano besedilo. Na cilju nato s pomočjo svojega ključa in alogaritma to besedilo razberejo.

S pomočjo naše analize lahko z nekaj truda tudi razberemo manj komplicirano kodo. Za to uporabimo frekvenco besed ali črk (odvisno ali gre za šifriranje ali kodiranje). Nad sporočilom naredimo podobno analizo kot smo jo prikazovali čez celotno nalogo in primerjamo največkrat ponovljene besede (ali črke) s temi, ki jih že poznamo. Seveda je možnosti, da bo največkrat ponovljena beseda v sporočilu sovpadala z največkrat ponovljeno besedo v poznanem jeziku zelo malo, lahko pa sklepamo, da so najvišje tri besede v urejenem grafu jezika enake kot najvišje tri besede v urejenem grafu sporočila.

Če nismo prepričani lahko to domnevo še nekoliko razširimo na pet največkrat ponovljenih besed [20].

V elektronski dobi obstajajo celotne funkcije, ki z uporabo entropije prikažejo težavnost ugotavljanja neke besede [21]. Pri tem bi lahko prav prišla naša raziskava in ideja o številčnejših korpusih. Seveda imamo lahko tudi en združeni/skupni korpus. Vendar če približno poznamo temo kodiranega besedila, si lahko množico možnosti zmanjšamo in tako porabimo manj časa in prostora. Seveda pa je potrebno vključiti tudi naše rezultate o poziciji posamezne črke. Primer: pri iskanju drugega znaka besede je pametno upoštevati naše tabele in grafe ter naše iskanje pričeti pri črkah z največjo možno frekvenco, ne pa po abecednem vrstnem redu. Takšno iskanje je seveda veliko bolj uspešno.

## 9.2. Primerjava z ostalimi jeziki

Za ljudi, ki ne poznajo drugega jezika, je analiza besed v pomoč pri razumevanju besedila. Naše ugotovitve o slovenščini smo primerjali z analizo angleščine na internetni spletni strani [22]. Sicer se slovenščina in angleščina močno razlikujeta ena od druge, saj ima angleščina besede, ki jih slovenščina ne pozna in obratno. Besede, ki pa le imajo enak oziroma podoben pomen, pa se nahajajo na podobnih položajih v grafih. Iz našega združenega korpusa smo ponovno razbrali dvajset najbolj uporabljenih besed in procent njihovih ponovitev. S primerjanjem teh sorodno-pomenskih besed med grafoma, smo dobili naslednjo tabelo (Tabela 18):

| Angleščina | | Slovenščina | |
|---|---|---|---|
| **Beseda** | **Procent ponovitev (v %)** | **Beseda** | **Procent ponovitev (v %)** |
| and | 3,04 | in | 2,94 |
| in | 2,27 | v | 2,19 |
| is | 1,13 | se | 1,74 |
| for | 0,88 | za | 0,78 |
| with | 0,70 | z | 0,67 |
| on | 0,62 | na | 1,32 |
| not | 0,61 | ne | 0,64 |
| was | 0,74 | je | 3,85 |
| that | 1,08 | da | 0,99 |
| by | 0,63 | s | 0,57 |

**Tabela 18: Primerjava podobno-pomenskih besed v angleškem in slovenskem jeziku**

Opazimo, da ima kar nekaj besed – posebno tiste, ki so dobesedni prevod – zelo podoben procent uporabe v posameznem jeziku. Tak primer so besede: *and – in, in – v, for – za, with – z* in *not – ne*. Takšne besede pridejo zelo prav pri učenju jezika, saj so nekakšen temelj na katere se lahko učenec opira dokler mu je slovnica jezika še tuja.

## 9.3. Črkovne igre

Pri znani igri *Scrabble* imamo na voljo le nekaj črk – ploščic. Na igralni plošči imamo že sestavljenih nekaj besed. S pomočjo frekvence, ki nam pove, kje se neka črka največkrat pojavi, lahko hitro ugotovimo, če imamo na voljo primerne črke za sestavo neke besede. Seveda, bi nam te ugotovitve prišle prav na veliko večjem primeru takšne igre. Lahko bi tudi sprogramirali program, ki ponovno glede na naše rezultate začne z najboljšo pozicijo črke. Pri ravno tako znani igri *Boggle,* vidimo, da se črka *a* nahaja b kotu mreže – naša raziskava nam pove, da je neprimerno zapravljati čas z iskanjem besede, ki se prične na to črko.

## 9.4. Preiskovanje podatkovnih baz

Zelo prav bi prišli korpusi, ki bi se sproti povečevali, tako da jih ne bi povozil čas. Recimo: z objavo na spletni strani *24ur.com*, bi povečali korpus člankov; z objavo nove leposlovne knjige pa korpus leposlovja. Tako bi lahko študentje, ki ne vedo, kje bi najlažje našli vir za svojo nalogo, vpisali ključne besede v iskalnik, ter za rezultat dobili procentno vsebnost te besede v posameznem korpusu. Če se v korpusu člankov ta beseda pokaže 57% ter v korpusu raziskovalnih knjig 43%, bi bilo iskanje dobro začeti pri različnih člankih. Z drevesno razdelitvijo korpusov na pod-korpuse, bi bil rezultat še bolj natančen.

# 10. SKLEPNE UGOTOVITVE

Slovenščina ima kot vsak jezik različne načine uporabljanja glede na temo in način pisanja. V nasprotju z marsikaterimi drugimi jeziki, je slovensko besedišče težje lematizirati zaradi velikega števila končnic in predpon – kako veliko je to število smo lahko razbrali iz n-gramov – poleg tega pa je potrebno paziti, da ne odsekamo preveč posamezne besede. Če bi lahko slovenščino spreminjali v osnovno obliko samo po enem ključu, bi bili rezultati še boljši. Lahko pa bi rezultate tudi primerjali iz samega slovarja in tako popravili napačno lematizirane besede.

Zaradi oblike slovenščine je njen korpus daljši, saj namesto, da bi šteli ponovljeno besedo v edini, osnovni obliki njenega pomena, seštevamo vsako obliko besede posebej. Če bi namesto lematizacije besede krnili, bi dobili preveč enakih besed, čeprav bi bil pomen nekrnjenih besed popolnoma drugačen med seboj. Vendar ne glede na to, ko odstranimo polnila, iz preostalih besed ni težko odkriti vrsto besedila.

V igri skritih besed bi z našimi statističnimi ugotovitvami hitreje prišli do rezultatov. Dolžine besed pa bi nam pomagale pri ugotovitvi, koliko različnih možnosti sploh je, če vemo, da gre za besedo z desetimi znaki; te pa se še zmanjšajo, če vemo, da je beseda prišla iz korpusa leposlovja.

Pri primerjavi z ostalimi jeziki, smo ugotovili, da čeprav sta jezika daleč od tega, da bi bila v sorodu, lahko s frekvenco uporabe besed nekatere od teh kar prezrcalimo v drugi jezik. Če bi šlo za jezika, ki imata enak izvor (recimo *hrvaščina*), bi bilo razbiranje besedila še toliko lažje.

S pomočjo naših rezultatov bi s tem znanjem kaj hitro razbrali šifrirano sporočilo Julija Cezarja, ki je za pošiljanje svojih skrivnih načrtov uporabljal substitucijo [20].

Med pisanjem diplomske naloge smo odkrili nekaj zanimivih informacij, ki so nam bile poprej neznane ali pa se nanje nismo nikoli osredotočili, kot je bila ugotovitev o količini polnil v posameznih besedilih; o očitnem razlikovanju najbolj uporabljenih besed med korpusi; da je *f* ena najmanj uporabljenih črk in podobno. Predvsem smo našli kar nekaj koristnih spletnih strani za iskanje besedil. Prebrali smo nekaj drugih raziskovalnih nalog (na katere smo se sklicevali v našem delu) in še dodatno razširili svoje znanje, predvsem programiranja v programerskem jeziku *Python* in zgradbe tabel v *Microsoft Excel 2010*.

Rezultate naše naloge je mogoče uporabiti kot nekakšen smerokaz ali pa referenco za druge podobne ali enake raziskave. Pri korpusih, ki so manjši od naših, se lahko prepričamo o dovolj veliki raznolikosti, če se rezultati drugih podobnih raziskav približno

sovpadajo z našo. To je uporabno predvsem pri frekvenci črk in besed. Lahko pa so uporabljene tudi ugotovitve za namene, kot smo jih opisali v prejšnjem poglavju.

Korpusi in izvorne kode naših programov v *Pythonu* smo objavili na GitHubu za lažji doseg.

Povezava: https://github.com/SuhadolcBarbara/Statisticna-Analiza-Slovenskih-Besedil

# LITERATURA


[1] (2013, Aug.) Wikipedija. [Online]. http://sl.wikipedia.org/

[2] Prof. def. Martina Ozbič. (2013, Sep.) Akustična spektralna fft analiza samoglasniškega sistema slovenskega jezika. [Online]. http://nl.ijs.si/isjt98/zbornik/sdjt98-Ozbic.pdf

[3] (2013, Sep.) Fran Ramovš Institute of Slovenian Language. [Online]. http://bos.zrc-sazu.si/nova_beseda.html

[4] (2013, Aug.) SloLyrics.com. [Online]. http:/slolyrics.com

[5] (2013, Aug.) Pesmi.si. [Online]. http://pesmi.si

[6] (2013, Sep.) Digitalna Knjižnica Slovenije. [Online]. http://www.dlib.si/

[7] (2013, July) 24ur. [Online]. http://www.24ur.com

[8] (2013, Aug.) Žurnal24. [Online]. http://www.zurnal24.si

[9] (2013, Aug.) Delo. [Online]. http://www.delo.si

[10] Danuta Reah, *The Language of Newspapers;* New York: 29 West 35th Street, 1998, pp. 4, 5, 25.

[11] (2013, Aug.) Clanki.net. [Online]. http://www.clanki.net/

[12] (2013, Aug.) Revija Ventil. [Online]. http://www.revija-ventil.si/domov/

[13] (2013, Aug.) Univerzitetni rehabilitacijski inštitut Republike Slovenije - Soča. [Online]. http://www.ir-rs.si/

[14] (2013, Aug.) National Geographic Slovenija. [Online]. http://www.nationalgeographic.si/

[15] (2013, Aug.) Dictionary.com. [Online]. http://dictionary.reference.com

[16] (2013, Aug.) Slovenski Blogi. [Online]. http://sloblogi.drugisvet.com

[17] (2013, Sep.) Beautiful Soup Documentation. [Online]. http://www.crummy.com/software/BeautifulSoup/bs4/doc/

[18] (2013, Aug.) Slovar slovenskega knjižnega jezika. [Online]. http://bos.zrc-sazu.si/sskj.html

[19] Ed.D. Beatrice S. Mikulecky, "Teaching Reading in a Second Language," 2008, http://www.longmanhomeusa.com/content/FINAL-LO%20RES-Mikulecky-Reading%20Monograph%20.pdf.

[20] Simon Singh, *The Code Book;* England: Four Estate Ltd, 2006.

[21] Pavla Lah. (2013, Aug.) Uporaba kriptografije v internetu. [Online]. http://www.si-ca.si/kripto/index.htm

[22] (2013, Aug.) English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU. [Online]. http://norvig.com/mayzner.html

[23] "Mladi za napredek Maribora 2013". (2013, Sep.) Analiza besedil v slovenski popularni glasbi. [Online]. http://www.zpm-mb.si/attachments/sl/1185/SS_Slovensk_jezik_Analiza_besedil_v_slovenski_popularni_glasbi.pdf

[24] Mate Goznik. (2013, Aug.) Analiza novinarskih vsebin spletnih izdaj slovenskih časopisov in portala Dostop.si. [Online]. http://dkum.uni-mb.si/IzpisGradiva.php?id=22588

[25] Ministrstvo za kulturo. (2013, Sep.) Raziskave na področju slovenskega jezika. [Online]. http://www.mk.gov.si/si/delovna_podrocja/sluzba_za_slovenski_jezik/raziskave_na_podrocju_slovenskega_jezika/