# Strokovno-znanstvena slovenščina: besednovrstne in oblikoskladenjske značilnosti

**Nataša Logar,* Tomaž Erjavec†**

* Fakulteta za družbene vede, Univerza v Ljubljani
Kardeljeva ploščad 5, 1000 Ljubljana
natasa.logar@fdv.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

**Povzetek**

V prispevku prikazujemo besednovrstne in oblikoskladenjske značilnosti slovenskih strokovno-znanstvenih besedil, do katerih smo prišli z metodo frekvenčnega profila, in sicer na podlagi primerjave štirih (pod)korpusov: celotnega uravnoteženega korpusa slovenščine Kres in njegovega leposlovnega dela ter disertacijskega in diplomskega dela korpusa akademske slovenščine KAS. Rezultati analize so med drugim pokazali, da so za slovensko strokovno-znanstveno pisanje bolj značilni samostalniki, pridevniki in okrajšave (primerjalno s Kresom pa najmanj glagoli, zaimki ter prislovi) in da med MSD-oznakami v disertacijah primerjalno s Kresom najbolj izstopajo občni samostalniki vseh treh spolov v ednini ali množini, ki so v rodilniku ali imenovalniku. Ugotovitve tako opozarjajo na slovnična mesta, ki jim bo treba pri pripravi prihodnjega opisa strokovno-znanstvene slovenščine posvetiti še več pozornosti.

**Academic Slovene: Part-of Speech and Morphosyntactic Characteristics**

The article presents PoS and MSD characteristics of academic Slovene. Using the frequency profiling method, we assembled the data by comparing four (sub)corpora: the balances corpus of Slovene Kres as a whole, its fiction part, the PhD part of the corpus of academic Slovene KAS and its BSc and BA thesis part. Among other findings, the results show that nouns, adjectives and abbreviations are used much more in academic genres (in contrast with verbs, pronouns and adverbs that are typical for non-specialized language). In PhD texts, the following MSDs stand out: common nouns of all three genders in singular or plural and in genitive or nominative. The findings point to grammatical issues that should not be overlooked in new descriptions of academic Slovene.

## 1. Uvod

Z veliko verjetnostjo je mogoče trditi, da je izmed strokovno-znanstvenih podzvrsti vseh jezikov najbolje raziskana akademska angleščina (npr. Biber in Barbieri, 2007; Gardner in Davies, 2013; Hyland, 2008). Enega njenih tehtnejših, obenem pa nejezikoslovnim uporabnikom razumljivejših opisov npr. najdemo v priročniku z naslovom *Academic Writing for Graduate Students: Essential Tasks and Skills* avtorjev J. M. Swalesa in C. B. Feak (2012). Pri pripravi te knjige sta si avtorja pomagala s korpusom študentskih izdelkov *Michigan Corpus of Upper-Level Student Papers*[1] (2004–2009; 2,6 milijona besed; Römer, 2009). To jima je omogočilo, da sta leksikalno-slovnične značilnosti akademskega pisanja ponazorila z realnimi zgledi, pri čemer sta izhajala iz pomena, namena in kohezivnosti besedil. Drugo, sicer izrazito na slovnične značilnosti akademske angleščine osredotočeno delo, ki pa ga izmed več takih prav tako velja izpostaviti, je priročnik *Advanced Grammar: For Academic Writing* avtorja R. Stevensona (2010). Stevensonovo delo sicer ni zasnovano korpusno, vendar pa avtor v njem pojave, kot so zgradba povedi, modalnost, kohezija, ton pisanja, samostalniškost, glagolski časi itn., razlaga na številnih primerih ter na bralcu prijazen način.[2]

Dela, ki bi bilo vsaj podobno prvemu ali drugemu od omenjenih priročnikov, za slovensko strokovno-znanstveno pisanje še nimamo, imamo pa obsežen Korpus akademskih besedil KAS (Erjavec et al., 2016), ki že omogoča analize, na katerih bi lahko bil osnovan tak opis. V prispevku zato na primeru besednovrstnih in oblikoskladenjskih podatkov, ki korpus KAS značilno ločijo od uravnoteženega korpusa slovenščine Kres (Logar Berginc et al., 2012), prikazujemo ugotovitve začetnih tovrstnih analiz in razmišljamo, kako nam podatki iz njih lahko pomagajo pri pripravi prvega, na obsežnem naboru besedil z različnih področij temelječega priročniškega prikaza te podzvrsti slovenskega jezika.

## 2. Metoda in (pod)korpusi

Podatke o besednovrstnih in oblikoskladenjskih oznakah besedil (angl. *part-of-speech*, dalje PoS; *morphosyntactic description*, dalje MSD), ki so zajeta v korpus KAS in korpus Kres, smo pridobili z metodo frekvenčnega profila (angl. *frequency profiling*).[3] Metodo sta zasnovala Rayson in Garside (2000), omogoča pa primerjavo dveh korpusov (ali podkorpusov) po ključnih besedah in slovničnih kategorijah. Rezultat primerjave so seznami, ki kažejo, kateri elementi so bolj značilni za enega oz. drugega od korpusov, če ju primerjamo med seboj. Na slovenskem gradivu je bila metoda prvič uporabljena za ugotavljanje razlik med korpusoma

[1] http://www.helsinki.fi/varieng/CoRD/corpora/MICUSP/

[2] Kratek pregled in primerjavo tujih priročnikov na temo akademskega pisanja gl. v Logar (2017: 25–40).

[3] Pri korpusu KAS smo uporabili njegovo različico v2.0 (2000–2015), https://www.clarin.si/noske/run.cgi/corp_info?corpname=kas (tudi tukajšnje Slike 1–4 in 6 so od tam), pri korpusu Kres pa različico 1.0, http://www.slovenscina.eu/korpusi/kres.

slovenščine Kres in Gigafida (Logar Berginc et al., 2012: 95–97), nato pa še za primerjavo prve različice spletnega korpusa slovenščine $slWaC_1$ z (a) njegovo nadgradnjo $slWaC_2$ ter (b) z obema že imenovanima korpusoma slovenščine, Kresom in Gigafido (Erjavec et al., 2015).

Preizkus je tudi tokrat potekal po enakem postopku: »najprej smo izdelali frekvenčni seznam lem /ter MSD oznak/ obeh korpusov, nato pa za vsako lemo /ter PoS in MSD oznako/ izračunali njeno logaritemsko verjetnost (angl. *log-likelihood*, LL). LL upošteva pogostost elementa, kot tudi velikosti obeh korpusov in večji, kot je, bolj je element značilen za enega od njiju. Elementi z najvišjimi vrednostmi razlik v LL /.../ najočitneje kažejo glavne razlike med korpusoma« (Erjavec et al., 2015: 38).

(Pod)korpusi, ki smo jih primerjali, so bili štirje:

a) Kres,
b) leposlovni del Kresa (dalje Kres-fict),
c) disertacijski del KAS-a (dalje KAS-dr) in
č) diplomski del KAS-a (dalje KAS-dipl).[4]

Iz frekvenčnih seznamov vseh štirih (pod)korpusov smo izločili enote, ki bi šumno obremenile končne sezname LL, tj. besede, zapisane v tujem jeziku, besede, zapisane s številkami, ločila, ki so bila označena kot »besede«, nize ločil ipd. Med seboj smo nato primerjali:

- KAS-dr : Kres,
- KAS-dr : Kres-fict,
- KAS-dipl : KAS-dr,
- KAS-dipl : Kres in
- KAS-dipl : Kres-fict.

Končnih seznamov je bilo skupno 15, torej po pet za vsak opazovani element (PoS, MSD in leme).[5]

| **KAS-dr : Kres** | **KAS-dr : Kres-fict** | **KAS-dipl : Kres** | **KAS-dipl : Kres-fict** |
|---|---|---|---|
| **1. S*** | **1. S** | **1. S** | **1. S** |
| **2. P** | **2. P** | **2. P** | **2. P** |
| **3. O** | **3. O** | **3. O** | **3. O** |
| **4. D** | **4. D** | **4. D** | **4. D** |
| 5. V** | 5. V | 5. V | 5. V |
| 6. M | 6. M | 6. K | 6. M |
| 7. K | 7. L | 7. M | 7. L |
| 8. L | 8. R | 8. L | 8. R |
| 9. R | 9. Z | 9. R | 9. Z |
| 10. Z | 10. G | 10. Z | 10. G |
| 11. G | | 11. G | |

Tabela 1: PoS: rezultati (pod)korpusnih primerjav po metodi frekvenčnega profila.

* Kode pomenijo: S – samostalnik, P – pridevnik, O – okrajšava, D – predlog, V – veznik, M – medmet, K – številnik, L – členek, R – prislov, Z – zaimek, G – glagol.

** Črna barva – značilno za prvi (pod)korpus; siva barva (negativne vrednosti LL) – značilno za drugi (pod)korpus. Enako velja za Tabelo 2. Več o oznakah gl. v Erjavec et al. (2010) in na http://nl.ijs.si/jos/josMSD-sl.html (o označevalniku korpusa Kres gl. Grčar et al. (2012), o orodjih reldi-tokeniser[6] in reldi-tagger,[7] s katerima je bil označen korpus KAS, pa Ljubešić in Erjavec (2016)).

## 3. Analiza rezultatov

Podatke smo razvrstili po vrednosti LL. Zgornji del tabel tako prikazuje elemente, ki so bolj značilnosti za prvega od primerjanih (pod)korpusov, spodnji del z negativnimi vrednostmi LL pa elemente, ki so bolj značilni za drugega od primerjanih (pod)korpusov.

### 3.1. PoS

Ogled LL-vrednosti PoS-oznak je pokazal, da so seznami štirih izmed petih primerjav v vrhnjem delu po zaporedju povsem enaki (Tabela 1 v levem stolpcu besedila) in kažejo, da so za slovensko strokovno-znanstveno pisanje v primerjavi s Kresom ter njegovim leposlovnim delom bolj značilni samostalniki, pridevniki, okrajšave in predlogi (S – P – O – D). (O peti primerjavi, tj. primerjavi KAS-dipl : KAS-dr, gl. razdelek 3.1.1.)

Rezultati v Tabeli 1 potrjujejo izrazito samostalniškost strokovno-znanstvenih besedil, kar je skupaj s pridevnikom mogoče razumeti predvsem kot njihovo gosto terminološkost.[8] Tudi tretje mesto okrajšav ne preseneča, pri čemer je treba dodati, da prisotnost te »besedne vrste« v opazovanih besedilih poleg običajnih *oz.*, *npr.*, *angl.*, *t. i.*, *idr.*, *itd.*, *ipd.*, *tj.* in še nekaterih močno krepijo okrajšave v navedenkah ter seznamih virov in literature (zlasti okrajšave osebnih lastnih imen).

Predlogi so zadnja od štirih besednih vrst, katerih prisotnost je primerjalno večja v diplomah in doktoratih. Kot je razvidno na Sliki 1, se v večjem obsegu (s pojavitvami nad 20.000 v več kot 40-milijonskem podkorpusu doktoratov, če upoštevamo samo njihov slovenski del) pravzaprav rabi le 15 različnih predlogov, s tem, da je v izraziti prednosti predlog *v*. Vendar pa je ta slika – in skoraj enako je pri diplomah – zelo podobna Sliki 2, ki kaže pogostost predlogov v korpusu Kres. To pomeni, da se po *konkretnih* predlogih in njihovem medsebojnem razmerju doktorati ter diplome od splošne slovenščine bistveno ne razlikujejo. Odstopanje, ki ni veliko, a je dovoljšnje, da ga je v prid strokovno-znanstvenemu pisanju zaznala metoda frekvenčnega profila, gre torej pripisati večjemu skupnemu obsegu vseh predlogov (njihova pogostost na milijon pojavnic je npr. v KAS-dr 115.083, v Kresu pa 104.379).

Preostanek Tabele 1 kaže, da je po drugi strani v diplomah in doktoratih – če jih seveda še vedno primerjamo z leposlovjem in celotnim Kresom – izrazito najmanj glagolov, sledijo zaimki in prislovi, takoj za tem pa členki. Ker smo izpustili številnike (K), zapisane s številko, se ti v seznam, ki kaže razlike med diplomami in doktorati na eni strani ter Kresom in njegovim leposlovnim delom na drugi strani, sploh ne uvrstijo (gl. drugi in četrti stolpec Tabele 1) ali pa vsaj ne uvrstijo s pomembno razliko (gl. prvi in tretji stolpec).[9] Manj, vendar še vedno, so za leposlovje in Kres značilni še medmeti, primerjalno najmanj pa vezniki. Predlog in veznik sta torej besedni vrsti, pri katerih je med strokovno-znanstvenim pisanjem na eni strani ter drugimi

[4] Namesto izrazov *diplomsko delo* in *magistrsko delo drugostopenjskega bolonjskega študija* v nadaljnjem besedilu uporabljamo krajši izraz *diploma*, namesto *doktorska disertacija* pa *doktorat*.

[5] Analizo rezultatov razlik v lemah gl. v Logar in Erjavec (2017).

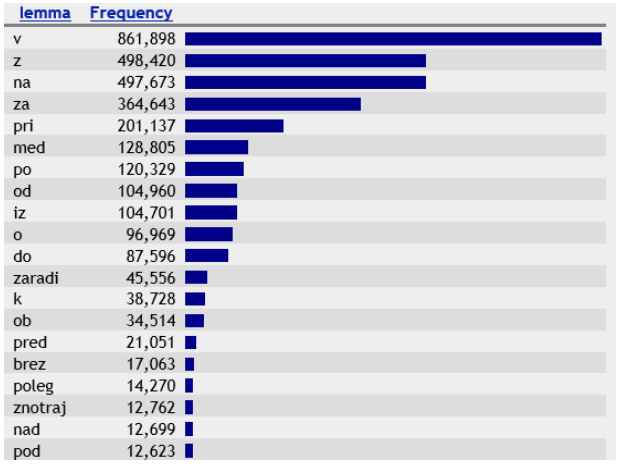
[6] https://github.com/clarinsi/reldi-tokeniser

[7] https://github.com/clarinsi/reldi-tagger

[8] Od vrhnjih 500 lem (tj. lem vseh besednih vrst) pri primerjavi KAS-dr : Kres-fict je bilo glagolnikov (samostalnikov iz glagolov s pomenom (tudi) dejanja, npr. *razumevanje*, *primerjava*, *meritev*) le 10 %, kar še dodatno potrjuje, da je značilnost besedil te zvrsti osredotočenost na statičnost in ne na delovanje oz. početje (o glagolih v zvezi s terminologijo gl. npr. Žele (2004)).

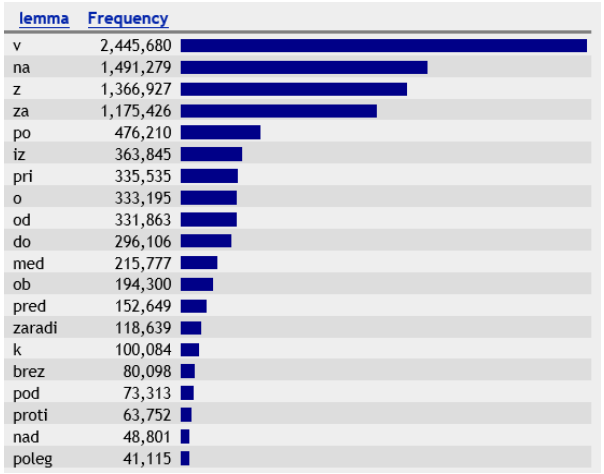
[9] Če bi številke vključili, pa bi se številnik kot besedna vrsta zelo verjetno pojavil v zgornjem, za KAS značilnem delu tabele.

zvrstmi besedil, ki so vse prisotne v Kresu, razlik najmanj (prim. tudi podobnost konkretnih veznikov po pogostosti v KAS-dr in Kresu na Slikah 3 in 4).



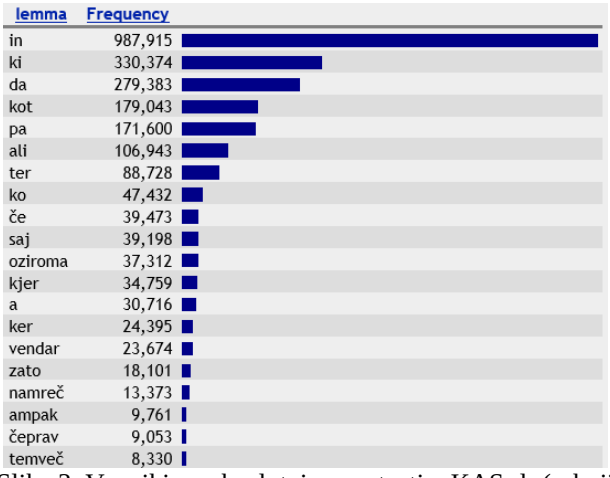
| lemma | Frequency |
|---|---|
| v | 861,898 |
| z | 498,420 |
| na | 497,673 |
| za | 364,643 |
| pri | 201,137 |
| med | 128,805 |
| po | 120,329 |
| od | 104,960 |
| iz | 104,701 |
| o | 96,969 |
| do | 87,596 |
| zaradi | 45,556 |
| k | 38,728 |
| ob | 34,514 |
| pred | 21,051 |
| brez | 17,063 |
| poleg | 14,270 |
| znotraj | 12,762 |
| nad | 12,699 |
| pod | 12,623 |

Slika 1: Predlogi po absolutni pogostosti v KAS-dr (vrhnji del).



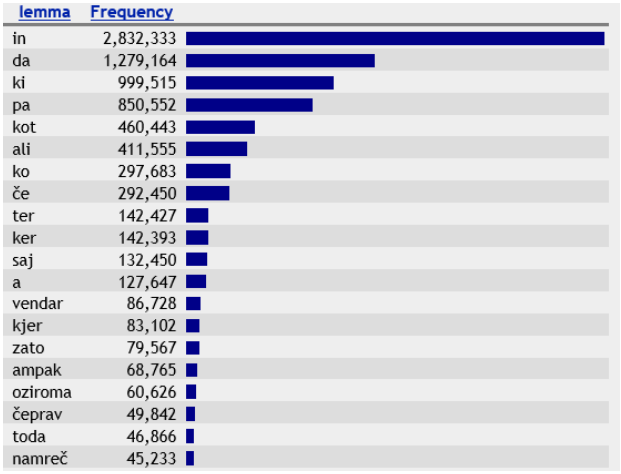
| lemma | Frequency |
|---|---|
| v | 2,445,680 |
| na | 1,491,279 |
| z | 1,366,927 |
| za | 1,175,426 |
| po | 476,210 |
| iz | 363,845 |
| pri | 335,535 |
| o | 333,195 |
| od | 331,863 |
| do | 296,106 |
| med | 215,777 |
| ob | 194,300 |
| pred | 152,649 |
| zaradi | 118,639 |
| k | 100,084 |
| brez | 80,098 |
| pod | 73,313 |
| proti | 63,752 |
| nad | 48,801 |
| poleg | 41,115 |

Slika 2: Predlogi po absolutni pogostosti v Kresu (vrhnji del).



| lemma | Frequency |
|---|---|
| in | 987,915 |
| ki | 330,374 |
| da | 279,383 |
| kot | 179,043 |
| pa | 171,600 |
| ali | 106,943 |
| ter | 88,728 |
| ko | 47,432 |
| če | 39,473 |
| saj | 39,198 |
| oziroma | 37,312 |
| kjer | 34,759 |
| a | 30,716 |
| ker | 24,395 |
| vendar | 23,674 |
| zato | 18,101 |
| namreč | 13,373 |
| ampak | 9,761 |
| čeprav | 9,053 |
| temveč | 8,330 |

Slika 3: Vezniki po absolutni pogostosti v KAS-dr (vrhnji del).



| lemma | Frequency |
|---|---|
| in | 2,832,333 |
| da | 1,279,164 |
| ki | 999,515 |
| pa | 850,552 |
| kot | 460,443 |
| ali | 411,555 |
| ko | 297,683 |
| če | 292,450 |
| ter | 142,427 |
| ker | 142,393 |
| saj | 132,450 |
| a | 127,647 |
| vendar | 86,728 |
| kjer | 83,102 |
| zato | 79,567 |
| ampak | 68,765 |
| oziroma | 60,626 |
| čeprav | 49,842 |
| toda | 46,866 |
| namreč | 45,233 |

Slika 4: Vezniki po absolutni pogostosti v Kresu (vrhnji del).

#### 3.1.1. PoS pri KAS-dipl : KAS-dr

Kot smo že nakazali, smo pri primerjavi KAS-dipl : KAS-dr – pričakovano – dobili drugačne rezultate. Kot kaže Tabela 2, so za diplome bolj značilni glagoli, zaimki,[10] prislovi, členki in vezniki; za doktorate pa bolj samostalniki, sledijo okrajšave, pridevniki in nazadnje predlogi. Enako kot pri ostalih štirih primerjavah sta si torej obe besedilni vrsti najbolj podobni pri predlogih in veznikih.

| PoS | LL | KAS-dipl (na mio pojavnic) | KAS-dr (na mio pojavnic) |
|---|---|---|---|
| **1. G** | 44790 | 129.765 | 116.670 |
| **2. Z** | 39120 | 53.301 | 45.524 |
| **3. R** | 9186 | 43.705 | 40.249 |
| **4. L** | 5535 | 21.651 | 19.766 |
| **5. V** | 1746 | 87.997 | 85.838 |
| **6. D** | –2838 | 111.949 | 115.083 |
| **7. P** | –6511 | 149.127 | 154.613 |
| **8. O** | –12345 | 8.220 | 10.048 |
| **9. S** | –26472 | 384.693 | 402.489 |

Tabela 2: KAS-dipl : KAS-dr, PoS: rezultati primerjave po metodi frekvenčnega profila.

Rezultati te primerjave torej ožijo zgornjo ugotovitev: v primerjavi s splošnim jezikom je celotno strokovno-znanstveno pisanje (KAS-dr in KAS-dipl) bolj samostalniško-pridevniško (in najmanj glagolsko), obenem pa se v notranji primerjavi med doktorati in diplomami ta lastnost izkazuje za izrazitejšo pri prvih; z drugimi besedami: če diplome »zoperstavimo« doktoratom, se pokaže, da so diplome besednovrstno bližje splošnemu jeziku, kakršnega z besedili izkazujeta Kres in njegov leposlovni del.

### 3.2. MSD

Primerjava MSD-oznak je ob potrditvi zgornjih ugotovitev o besednovrstnih značilnostih besedil v KAS-u

[10] Pri tem je treba opozoriti, da je glagolski morfem *se* označen kot zaimek (gl. o tem tudi dalje).

in Kresu podala še uvid v razlike pri podrobnejših oblikoskladenjskih kategorijah. V nadaljevanju se bomo osredotočili predvsem na rezultate primerjave med KAS-dr in Kresom, katerih vrhnji del prikazuje Slika 5.[11]

Na Sliki 5 vidimo, da so na prvih petnajstih mestih po vrednosti LL, torej pri enotah, ki so najbolj značilne za KAS-dr v primerjavi s korpusom splošne slovenščine, predvsem samostalniške oblike (8/15), na celotnem seznamu 118 enot pa je sicer največ oblik pridevnika (64/118; samostalniških je 31/118). Rezultati nadalje pokažejo, da je na prvem mestu oznaka *Sozer*, torej 'samostalnik, občni, ženskega spola, edninski in v rodilniku' (v KAS-dr so najpogostejše tri take pojavnice: *vrednosti*, *uporabe*, *analize*); na drugem mestu je enak samostalnik, le da tokrat množinski (*storitev*, *informacij*, *sprememb*); na tretjem mestu pa 'občni samostalnik srednjega spola ednine in v rodilniku' (*leta*, *podjetja*, *dela*; več takih samostalnikov gl. na Sliki 6). Med pridevniki je na najvišjem, 7. mestu 'splošni (tj. nesvojilni in nedeležniški) pridevnik, ki ni stopnjevan, je ženskega spola, v množini in rodilniku' (daleč največji obseg ima pojavnica *človekovih*, sledijo *otrokovih*, *dušikovih* ipd.). Naslednji na seznamu ima enake lastnosti, le da ima namesto množinske edninsko obliko (*človekove*, *posameznikove*, *otrokove*) itd. Prvi glagol se pojavi na 11. mestu in ima oznako *Gp-spm-n*, ki označuje pomožnik *smo*. Zopet je visoko, tj. na 6. mestu, enorodna skupina okrajšav (*str.*, *oz.*, *npr.* ipd.).

Nadaljnje značilnosti MSD-profila doktoratov so še: pri samostalniških oblikah so najbolj značilni skloni imenovalnik, rodilnik, mestnik in orodnik, sicer pa z ženskim spolom kombinacije ednina + rodilnik, množina + rodilnik in ednina + imenovalnik; z moškim spolom ednina + imenovalnik in množina + rodilnik; s srednjim spolom pa ednina + rodilnik. Značilne pridevniške oblike so večinoma v osnovniku, v presežniku ni nobene, izmed vrst splošni/svojilni/deležniški pa ni nobenega svojilnega, deležniških je tretjina. Med zaimki sta najvišjo vrednost LL dobili oznaki *Zz-sei* (31. mesto) in *Zz-sem* (36. mesto), ki pomenita 'oziralni zaimek srednjega spola ednine v imenovalniku' oz. 'mestniku' (močno prevladujeta pojavnici *kar* in *čemer*). Od predlogov so na najvišjem, 16. mestu tisti, ki se vežejo z mestnikom (po pogostosti prevladuje predlog *v*, kar smo videli že pri analizi PoS-oznak, sledita *na*, *pri* idr.), na 29. mestu je še predlog, ki se veže z orodnikom (v veliki večini *s/z*, sledita *med* in *pred*), pri glagolih pa ob pomožniku višjo vrednost LL (25. mesto) kaže še 'nedovršnik v sedanjiku tretje osebe množine' (*vplivajo*, *predstavljajo*, *kažejo* itd.). Le nekoliko izstopajo še priredni vezniki (s skoraj milijonom pojavitev je v KAS-dr prevladujoč veznik *in*, sledi mu veznik *pa* s 170 tisoč pojavitvami, nato veznik *ali* s 100 tisoč pojavitvami idr.). Prvi števniki se uvrstijo na 70. mesto, in sicer gre za številke, zapisane z rimsko številko.[12]

Ostale tri primerjave MSD-oznak so dale zelo podobne rezultate.

### 3.2.1. MSD pri KAS-dipl : KAS-dr

Primerjava MSD-oznak med KAS-dipl : KAS-dr je dala najkrajši seznam enot (115 v primerjavi z npr. KAS-dr : Kres, ki jih ima 357), kar kaže na oblikoskladenjsko podobnost teh dveh besedilnih vrst. Rezultati tu pričakovano ponavljajo večjo glagolskost diplom (med njimi pojavnic, kot so: *pride*, *začne*, *postane*; *doseči*, *zagotoviti*, *ugotoviti*; *postala*, *začela*, *pokazala*; *imeti*, *uporabljati*; *pojavijo*, *dobijo*, *postanejo*). V tem smislu izstopata tudi pomožnik za prvo osebo ednine (*sem*) in tretjo osebo ednine v prihodnjiku (*bo*), seveda pa tudi zaimek (dejansko pa glagolski morfem) *se*. Na drugi strani so doktorati, kot že rečeno, bolj samostalniško-pridevniški. Tudi tu nobena v primerjavi izstopajoča pridevniška oblika ni v presežniku. Izmed redkih za doktorate značilnih glagolskih oblik so na seznamu 'pomožni glagol v prvi osebi množine v trdilni' in 'nikalni obliki' (*smo*, *nismo*) in 'pomožni glagol v dvojini' (*sta*), dalje pa še 'dovršni glagoli v obliki deležnika, v množini in moškega spola' (npr. *uporabili*, *ugotovili*, *izvedli*), 'dovršni glagoli v sedanjiku, prvi osebi in množini' *(dobimo*, *uporabimo*, *najdemo*), 'glagoli v obliki deležnika, v množini in moškega spola' (*upoštevali*, *analizirali*, *podali*) ter še 'nedovršni glagoli v sedanjiku, tretji osebi in dvojini' (*predstavljata*, *ugotavljata*, *navajata*).

Oblikoskladenjske kategorije skupaj z njihovimi najpogostejšimi zapolnitvami torej izkazujejo tipičnost določenih lem v določenih oblikah, pri čemer gre zlasti za občne samostalnike vseh treh spolov v ednini ali množini ter v rodilniku in imenovalniku.

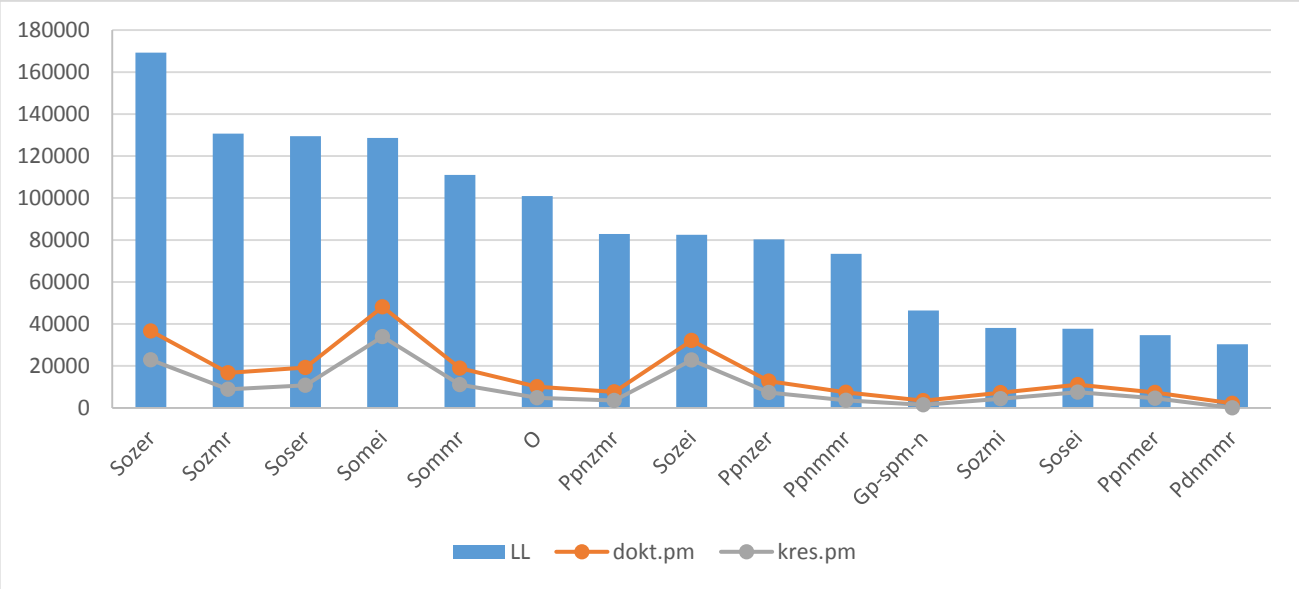
## 4. Iz podatkov v opis

Glavne ugotovitve naše besednovrstne analize so potrdile, da je za opis slovenskih strokovno-znanstvenih besedil še vedno pomembna njihova »samostalniškost« (Žagar Karer, 2011: 145) oz. »neglagolsko izražanje« (Toporišič, 1991: 23). A kot je pokazal frekvenčni profil lem, ki je bil prav tako pridobljen v tej raziskavi (Logar in Erjavec, 2017), se predmetnopoimenovalna zgoščenost strokovno-znanstvenih besedil, ki gre očitno »na škodo« glagolov, ne dosega le s termini, ki so večinoma pridevniško-samostalniški (Logar et al., 2013), temveč tudi s samostalniki, ki so splošnostrokovni (ter seveda njihovih tipičnim – predvsem spet pridevniškim – besedilnim okoljem). Taki splošnostrokovni samostalniki so npr. povezani z zgradbo strokovno-znanstvenih besedil (*slika*, *tabela*, *graf*, *primer*, *literatura*, *priloga*, *podpoglavje* itd.) ali predstavitvijo in interpretacijo rezultatov (npr. *podatek*, *število*, *izračun*, *ugotovitev*, *posledica*, *interpretacija*, *mnenje*).[13] Prav ustrezna raba takih izrazov je poleg poznavanja specializiranih področnih poimenovanj ključna za natančno, jasno in razumljivo pisanje besedil, katerih osrednja funkcija je informativno-spoznavna (Skubic, 1994/95).

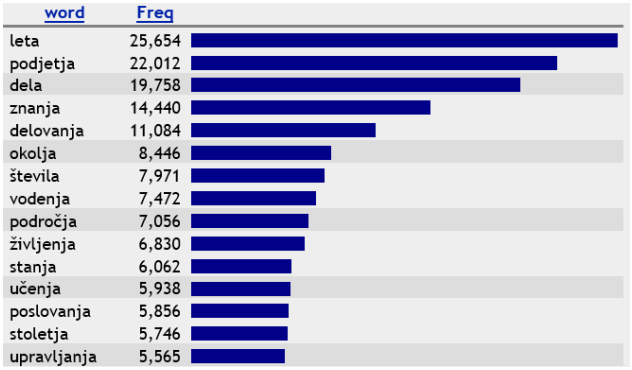
[11] Za lažjo predstavo o metodi smo tu k oznakam dodali še podatke o njihovi relativni pogostosti v obeh korpusih. Vidi se, da so razmerja med višinami stolpcev (LL) drugačna, kot so – medsebojno sicer zelo podobna – razmerja med relativnim številom pojavnic z določeno MSD-oznako (torej višino pik).

[12] A kot smo že opozorili, rezultatov pri številniku zaradi izpusta te besedne vrste v primerih, ko je šlo za zapis z arabsko številko, ne moremo upoštevati.

[13] Izmed lem (samostalniki, pridevniki, glagoli, prislovi, členki in okrajšave), ki so se pokazale kot značilne za KAS-dr, ko smo ta podkorpus primerjali s Kres-fict, smo jih kar 30 % prepoznali kot splošnostrokovnih.




Slika 5: KAS-dr : Kres, MDS: rezultati primerjave po metodi frekvenčnega profila (vrhnji del) in relativne vrednosti njihovih pojavitev (na milijon pojavnic).




Slika 6: KAS-dr: pojavnice z oznako *Soser* (vrhnji del).

Frekvenčni profil MSD-oznak doktoratov in diplom je še podrobneje pokazal tipičnost določenih sklonov ter števil, pri glagolih pa npr. izstopanje sedanjika nedovršnih glagolov in množinskega dovršnega deležnika moškega spola. Dalje smo pri obeh besednih vrstah, ki sicer kažeta največjo sorodnost strokovno-znanstvenega pisanja s splošnim jezikom, ugotovili, da so v doktoratih izrazito pogosti predlogi, ki se vežejo z mestnikom (zlasti *v*), in priredni vezniki (zlasti *in*). Tudi okrajšave (prim. Kompara, 2011) so primerjalno izstopale, zaradi česar bi bilo smiselno še podrobneje pogledati njihovo raznovrstnost, razlog za nastanek in položaj v besedilu.

Raba glagolov, ki so se na drugi strani v celoti izkazali kot značilni za splošni in leposlovni jezik ter bolj za diplome kot za doktorate, je v znanstvenem pisanju omejena, kar pomeni, da ne izkazuje bogate obsegovne in pomenske razpršenosti, značilne npr. za nekatere novinarske žanre ali umetnostno prozo. Prav to oženje v specifično in natančno poimenovanje dejanj (*analizirati*, *meriti*, *ugotavljati* itd.) ter v prevladujočo rabo določenih oblik (gl. npr. podatke o rabi trpnika v Logar et al., 2016), je obenem razlog, da je tudi opis glagolske rabe v strokovno-znanstvenem pisanju nujen, še zlasti če bo tak opis namenjen bralcem, ki se v tej podzvrsti šele opismenjujejo.[14]

Metoda frekvenčnega profila nam je torej dala podatke, katerih prvi izsledki že izpostavljajo besednovrstna in oblikoskladenjska mesta, ki bi bila pri pripravi prihodnjega celovitejšega opisa strokovno-znanstvene slovenščine vredna pozornosti. Na ta način naši rezultati z vidika izrazito slovničnih kategorij relevantno dopolnjujejo druge podatke, ki jih lahko prav tako pridobimo iz korpusa KAS (npr. podatke o kolokacijskem okolju za strokovno-znanstveno pisanje značilne neterminološke leksike v orodju Sketch Engine (Kilgarriff et al., 2004)). Metoda frekvenčnega profila je sicer primerna tudi za primerjavo manjših (pod)korpusov, zato bi jo bilo mogoče uporabiti tudi na drugačnih podkorpusih KAS-a, pri čemer imamo v mislih predvsem vzorčene področne podkorpuse (npr. podkorpus humanističnih in tehničnih ved, lahko pa tudi posameznih strok znotraj njih). Tak pristop bi pokazal tudi, kolikšen vpliv je imelo na naše tukajšnje rezultate dejstvo, da je KAS tako v diplomskem kot v doktorskem delu večinoma družbosloven in zelo malo humanističen (prim. Erjavec et al., 2016). Brez dodatnih analiz lahko ta hip ocenimo le, da je bil vpliv pomemben.

## 5. Sklep

Metodo frekvenčnega profila smo na slovenskih korpusih tokrat uporabili že tretjič, zato smo okvirno vedeli, kakšne rezultate lahko pričakujemo. Primerjave smo zastavili na štirih (pod)korpusih, kar nam je omogočilo, da smo (a) ocenili, kateri rezultati so za naš cilj najboljši, ter (b) da smo lahko prečno preverjali, koliko so rezultati prekrivni in torej relevantni.

[14] Podrobneje bi veljalo pogledati še prislove in člene (prim. Mikolič, 2005), pa tudi zaimke (prim. Gorjanc, 1998) in števnike, ki se jim tu nismo posvetili.

Rezultati primerjav PoS- in MSD-oznak v podkorpusih diplom in doktoratov med seboj – ter s Kresom in leposlovjem na drugi strani – so nam omogočili prvi tako celovit uvid v izrazito oblikoslovno-skladenjske značilnosti strokovno-znanstvenega pisanja pri nas. V nadaljevanju pa bo vsekakor treba tukajšnje dokaj kratko interpretiranje rezultatov dopolniti še z leksikalnimi zapolnitvami (kjer se izkazujejo za tipične) in vpogledom v razloge za večja odstopanja; načrtujemo pa tudi nadgradnjo z leksikalno-skladenjskimi informacijami iz izluščeni n-gramov (Dobrovoljc, 2016).

## Zahvala

Avtorja se zahvaljujeta anonimnim recenzentom za koristne pripombe. Raziskavo, opisano v prispevku, je podprl projekt ARRS J6-7094 *Slovenska znanstvena besedila: viri in opis*.

## 6. Literatura

Douglas Biber in Federica Barbieri. 2007. Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes*, 26(3): 263–286.

Kaja Dobrovoljc. 2016. *Korpus KAS: n-grami (interno gradivo)*. Ljubljana, Trojina, zavod za uporabno slovenistiko; Filozofska fakulteta UL; Fakulteta za družbene vede UL.

Tomaž Erjavec, Darja Fišer, Simon Krek in Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. V: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Malta.

Tomaž Erjavec, Nikola Ljubešić in Nataša Logar. 2015. The slWaC Corpus of the Slovene Web. *Informatica*, 39(1): 35–42.

Dee Gardner in Mark Davies. 2013. A New Academic Vocabulary List. *Applied Linguistics*, 35(3): 305–327.

Vojko Gorjanc. 1998. Konektorji v slovničnem opisu znanstvenega besedila. *Slavistična revija*, 46(4): 367–388.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: T. Erjavec in J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*, str. 89–94. Ljubljana, Institut »Jožef Stefan«.

Ken Hyland. 2008. As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes*, 27(1): 4–21.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. V: *Proceedings of the 11th EURALEX international congress*, str. 105–116. Lorient, Universite de Bretagne-Sud.

Mojca Kompara. 2011. *Slovar krajšav*. Kamnik: Amebis.

Nikola Ljubešić in Tomaž Erjavec. 2016. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA).

Nataša Logar Berginc, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana, Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Nataša Logar, Špela Arhar Holdt in Tomaž Erjavec. 2016. Slovenski strokovni jezik: korpusni opis trpnika. V: E. Kržišnik in M. Hladnik (ur.): *Toporišičeva obdobja*, str. 237–245. Ljubljana, Znanstvena založba Filozofske fakultete.

Nataša Logar in Tomaž Erjavec. 2017. Slovene Academic Writing: A Corpus Approach to Lexical Analysis. V: *Interdisciplinary Knowledge-making, Challenges for LSP Research: Book of Abstracts*, str. 44. Bergen, Norwegian School of Economics. (Celotni prispevek je oddan v objavo.)

Nataša Logar, Špela Vintar in Špela Arhar Holdt. 2013. Terminologija odnosov z javnostmi: korpus – luščenje – terminološka podatkovna zbirka. *Slovenščina 2.0*, 1(2): 113–138.

Nataša Logar. 2017. *Strokovno-znanstveni jezik: študijska literatura in priročniki v Sloveniji ter kratek pregled tujih praks*. Ljubljana, Fakulteta za družbene vede, http://nl.ijs.si/kas/wp-content/uploads/2018/03/KAS-pregled-prirocnikov-navodil-in-predmetov-Logar.pdf.

Vesna Mikolič. 2005. Izrazi moči argumenta v znanstvenih besedilih. V: M. Jesenšek (ur.): *Knjižno in narečno besedoslovje slovenskega jezika*, str. 278–291. Maribor, Slavistično društvo Maribor.

Paul Rayson in Roger Garside. 2000. Comparing Corpora Using Frequency Profiling. V: *Proceedings of the Workshop on Comparing Corpora*, str. 1–6. Hong Kong, Association for Computational Linguistics.

Ute Römer. 2009. English in Academia: Does Nativeness Matter? *Anglistik: International Journal of English Studies*, 20(2): 89–100.

Andrej Skubic. 1994/95. Klasifikacija funkcijske zvrstnosti in pragmatična definicija funkcije. *Jezik in slovstvo*, 40/5: 155–168.

Richard Stevenson. 2010. *Advanced Grammar: For Academic Writing*. Morisville, Academic English Publications.

John M. Swales in Christine B. Feak. 2012 *Academic Writing for Graduate Students: Essential Tasks and Skills*. Michigan, The University of Michigan.

Jože Toporišič. 1991. *Slovenska slovnica*. Maribor, Obzorja.

Mojca Žagar Karer. 2011. *Terminologija med slovarjem in besedilom: analiza elektrotehniške terminologije.* Ljubljana, Založba ZRC, ZRC SAZU.

Andreja Žele. 2004. Stopnje terminologizacije v leksiki (na primerih glagolov). V: M. Humar: *Terminologija v času globalizacije*, str. 77–91. Ljubljana, Založba ZRC, ZRC SAZU.