Slovenska akademska besedila: prototipni korpus in načrt analiz

Tomaž Erjavec,[*] Darja Fišer,[†*] Nikola Ljubešić,[*♣] Nataša Logar,[‡] Milan Ojsteršek[♠]

* Odsek za tehnologije znanja, Institut »Jožef Stefan«, Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

† Oddelek za prevajalstvo, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si

♣ Filozofska fakulteta, Univerza v Zagrebu, Ivana Lučića 3, 10000 Zagreb
nikola.ljubesic@ffzg.hr

‡ Fakulteta za družbene vede, Univerza v Ljubljani, Kardeljeva ploščad 5, 1000 Ljubljana
natasa.logar@fdv.uni-lj.si

♠ Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru, Smetanova ulica 17, 2000 Maribor
milan.ojstersek@um.si

Povzetek

Razvitost jezika, ki se rabi v akademskem okolju, je pomemben kazalnik njegove vitalnosti. V prispevku prikažemo izdelavo sodobnega referenčnega vira za ta del slovenščine in podamo okvirni nabor nadaljnjih na njem osnovanih raziskav. Prototipni korpus KAS vsebuje besedila, zajeta z Nacionalnega portala odprte znanosti, ter vključuje več kot 50.000 znanstvenih in strokovnih besedil s preko milijardo pojavnic. Njegov nastanek je med drugim vključeval postopke zajema, filtriranja, čiščenja in jezikoslovnega označevanja, na njem osnovane raziskave pa bodo dale rezultate na področju klasifikacije besedil, razvoja orodij in zbirk za terminološko delo ter opisa tega dela sodobnega slovenskega jezika.

Slovene Academic Texts: Prototype Corpus and Research Plan

The development of the academic part of any language is an important indicator of its vitality. The paper presents the construction of a contemporary academic language resource for Slovene and provides a framework for further research based on it. The KAS prototype corpus contains texts harvested from the Open Science portal of Slovenia and contains about 50,000 scientific texts with over one billion tokens. Its compilation included the collection, filtering, cleaning and linguistic annotation of its texts, while KAS corpus research will give results in the fields of text classification, terminological tool and database development and in description of contemporary academic Slovene.

1 Uvod

Razvoj in uporaba slovenskega jezika v visokem šolstvu ter znanosti je zadnja leta eno osrednjih vprašanj slovenske jezikovne politike (Kalin Golob et al., 2014; gl. še druge vire v Logar, 2013a: 247). Problem nepripravljenosti slovenskega jezika za digitalno dobo na več mestih izpostavlja tudi *Resolucija o Nacionalnem programu za jezikovno politiko 2014–2018*, še izraziteje pa iz resolucije izhajajoča *Akcijski načrt za jezikovno izobraževanje* in *Akcijski načrt za jezikovno opremljenost* (oba 2015). Namen akcijskih načrtov je bil konkretizirati »izzive, ki so potrebni hitrega in učinkovitega ukrepanja«. V prvem akcijskem načrtu sta tako kot dva od štirih ciljev v zvezi s slovenščino v visokem šolstvu in znanosti navedena prav »razvijanje sporazumevalne zmožnosti v /slovenskem/ strokovnem jeziku« ter »izboljšanje položaja slovenščine kot jezika znanosti«. V drugem akcijskem načrtu je na slovenščino kot jezik znanosti vezanih 8 izmed 47 ciljev.

Leta 2016 se je začel izvajati triletni temeljni raziskovalni projekt »Slovenska znanstvena besedila: viri in opis«. Namen projekta je uresničiti del zgornjih izzivov, podatkovni temelj zanj pa predstavlja obsežen korpus pisnih besedil akademske slovenščine.[1] V prispevku bomo predstavili, kako je potekalo pridobivanje besedil skupaj z metapodatki, kako so bila besedila pretvorjena in kakšna sta korpusov zapis ter trenutna zgradba. Predstavitvi korpusa bo v drugem delu prispevka sledil še kratek oris ključnih korpusnih analiz, ki bodo zajemale klasifikacijo besedil, luščenje terminologij in izgradnjo terminološke baze ter jezikovnoopisne študije.

2 Nacionalni portal odprte znanosti

V zadnjem času so slovenske univerze in druge raziskovalne institucije začele vzpostavljati digitalne zbirke svojih publikacij, ki vsebujejo raznorodna besedila, od diplomskih, magistrskih in doktorskih del do znanstvenih ter strokovnih prispevkov. Pomemben mejnik je pri tem leta 2013 vzpostavljeni Nacionalni portal odprte znanosti, ki agregira vsebine iz repozitorijev slovenskih univerz, slovenskih raziskovalnih organizacij in drugih zbirk (dLib, DKMORS, VideoLectures.NET, repozitorij ScieVie, CLARIN.SI, arhiv ADP) za potrebe skupnega iskalnika, priporočilnega sistema in detektorja podobnih vsebin (Ojsteršek et al., 2014). Repozitoriji omogočajo izvoz metapodatkov v imenike odprtega dostopa (OpenDOAR, ROAR, BASE, DART-Europe itd.), Google Scholar in OpenAIRE. Nacionalni portal in repozitoriji so povezani s slovenskim bibliografskim katalogom COBISS.SI. Če je vir iz nacionalne infrastrukture zaveden v COBISS, se njegovi metapodatki dopolnijo z metapodatki, ki so jih knjižničarji vnesli v COBISS. Na ta način se bistveno izboljša kakovost metapodatkov vstavljenih gradiv. Portal že ponuja dostop do prek 124.000 slovenskih objav s širokega nabora strokovnih

[1] Poimenovanje akademska slovenščina kot nadpomenka za zelo različne žanre, ki nastajajo v akademskem okolju (od izvirnih znanstvenih člankov do študentskih poročil), izhaja iz žanrske teorije (npr. Bhatia, 1993; Swales, 2000). Ker bo več nadaljnjih analiz v projektu teoretično in metodološko izhajalo iz tega pristopa, smo se pri poimenovanju korpusa odločili za ta termin in ga posledično večkrat uporabljamo tudi v tukajšnjem nadaljnjem besedilu, gre pa za – če uporabimo v slovenskem prostoru bolj razširjeno poimenovanje iz zvrstne teorije – strokovno-znanstvena besedila.

področij. Ta dela so izjemno dragocen, a zaenkrat še pomanjkljivo izkoriščen vir podatkov o akademski slovenščini, kot tudi bogat vir terminologije.

3 KAS-proto

Na osnovi gradiv iz podatkovne baze Nacionalnega portala odprte znanosti smo v začetku leta 2016 izdelali prvo različico korpusa slovenskih akademskih besedil, korpus KAS-proto.

3.1 Izvoz podatkov za prototipni korpus

Podatkovna baza, ki smo jo izvozili, je za vsako besedilo obsegala metapodatke, datoteko z izvornim formatom besedila in iz njega izluščeno besedilo. Omejili smo se samo na del metapodatkov (naslov, avtorji, povzetek, univerza, fakulteta, ključne besede, leto nastanka gradiva, vrstilci UDK, COBISS id vira in podatki, ki so potrebni za citiranje vira). Za preslikavo numerično zapisanega UDK v ključne besede področij smo uporabili odprte povezane podatke konzorcija UDK.[2]

3.2 Pretvorba korpusa

Pri pretvorbi zajetih podatkov smo v KAS-proto vključili samo besedila, ki so imela:

- izvorni zapis v PDF, saj je ključno, da je poleg besedila v korpusu raziskovalcem dostopen tudi vpogled v izvirnik, ki poleg besedila ponuja tudi njegovo oblikovanje in slike;
- s tehničnega vidika razmeroma kakovostno besedilo, saj PDF ni idealen format za luščenje besedila in je veliko besedil pokvarjenih do te mere, da so za raziskave neuporabna;
- pripisane vsaj minimalne metapodatke (podatki o avtorjih, naslovu, letnici, univerzi in fakulteti ter zvrsti po tipologiji COBISS), saj je brez teh podatkov nemogoče korektno citiranje, otežene pa so tudi analize korpusa;
- leto izdaje 2000 ali mlajše, saj je bilo starejših besedil zelo malo, zaradi česar korpus ne bi bil reprezentativen za starejša obdobja;
- dovolj velik delež slovenskega besedila, saj so bila v izvozu tudi besedila, ki so v celoti ali pretežno v angleščini.

Po filtriranju smo korpus pretvorili v format XML po shemi, ki smo jo izdelali zanj. Pretvorba je vsebovala naslednje korake:

- popravljanje najpogostsejših napak kodiranja znakov, saj ima veliko avtomatsko izluščenih besedil sistematične napake v kodiranju (*ž* je npr. zapisan kot *ţ*, *µz*, *z˘*, *œ*, *ſ* itd.);
- brisanje slabih znakov, ki bodisi niso veljaven UTF-8 ali pa so v področju zasebne uporabe (PUA);
- odstranjevanje glav in nog strani, ki vsebujejo številko strani, naslov dela, fakulteto itd. in bi sicer zelo izkrivili jezikovno podobo besedil;
- hevristično določanje mej med odstavki, saj so ti osnovna enota diskurza, so pa v neposredno izluščenem besedilu pogosto napačno identificirani;
- odstranjevanje odstavkov, ki so bodisi prazni ali vsebujejo samo ločila;
- avtomatsko določanje jezika posameznega odstavka, bodisi slovenščina ali angleščina, saj za večino analiz potrebujemo samo slovenske dele besedila, obenem pa je koristno ohraniti tudi angleške dele;
- zapis v skladu s shemo XML korpusa.

3.3 Jezikoslovno označevanje

V naslednji fazi je bilo besedilo vsakega dokumenta razčlenjeno na stavke in pojavnice (tokenizirano), oblikoskladenjsko označeno ter lematizirano, za kar smo uporabili nov označevalnik (Ljubešić in Erjavec, 2016), ki deluje na osnovi pogojnih naključnih polj in je bil modela jezika naučen na korpusu ssj500k (Krek et al., 2015) ter leksikonu Sloleks (Dobrovoljc et al., 2015). Označevalnik je sicer počasen, daje pa kakovostne rezultate. Evalvacija je namreč pokazala, da doseže na testni množici iz ssj500k 94,27-odstotno natančnost, kar je signifikantno več v primerjavi z 92,49-odstotno natančnostjo označevalnika Obeliks (Grčar et al., 2012), ki je do sedaj veljal za najboljši oblikoskladenjski označevalnik za slovenščino.

3.4 Zapis korpusa

Kot kaže slika 1, je vsako besedilo v korpusu zapisano kot svoj dokument XML. Korenski element `document` ima pripisane že omenjene metapodatkovne atribute, kamor spadajo tudi URL izvornega dokumenta v repozitoriju svoje univerze in URL (z geslom zaščitene) lokalne kopije PDF celotnega dela.

Dokument je sestavljen iz posameznih strani (`page`) in odstavkov (`p`) znotraj njih; če je bil prelom strani znotraj odstavka, je v XML premaknjen na njegov začetek, posamezna stran pa je lahko tudi prazna. Atributi strani so njena zaporedna številka in kazalka na lokalno kopijo strani v izvirniku, medtem ko ima odstavek pripisano kodo jezika vsebovanega besedila (`sl` ali `en`). Vsi elementi so opremljeni tudi z identifikatorjem (`@xml:id`).

Odstavki nato vsebujejo jezikoslovno označeno besedilo, in sicer povedi (`s`), znotraj njih pa besede (`w`), ločila (`pc`) in presledke (`c`), pri čemer sta prvima dvema pripisani še lema (`@lema`) in oblikoskladenjska oznaka (`@ana`) po priporočilih JOS.

Kot omenjeno, so izvirniki besedil dostopni tudi na strežniku projekta, nanje pa kažemo s kazalci URL iz elementa `document` in na določeno stran iz elementa `page`. Ta pristop je problematičen za javno uporabo korpusa, saj so celotni dokumenti PDF dostopni prek repozitorijev posameznih univerz, ki rade prepovejo njihovo nadaljnje razširjanje, dostop pa omogočijo šele, ko se uporabnik strinja s pogoji uporabe. Zato smo PDF razdelili na strani in te shranili na strežniku kot grafične datoteke PNG, pri čemer je vsaki dodan ključ (gl. `page/@facs_url` na sliki 1). S tem dobimo možnost uporabniku korpusa pokazati nekaj strani besedila, ob tem pa ne omogočamo prevzema celotnega izvirnika, podobno kot to dela spletna storitev Google Books.

3.5 Korpus na spletu

Korpusna besedila smo iz izvornih dokumentov XML pretvorili v t. i. vertikalni format, primeren za uvoz v konkordančnik, in ga vključili v lokalno instalacijo orodja noSketch Engine (Rychlý, 2007; Erjavec, 2013), s čimer dobimo možnost raznovrstnih analiz korpusnih podatkov.

[2] Gl. več na: http://udcdata.info/.

```
<document xml:id="kas-10000" doc_id="10000" text_id="16514" cobiss_id="7078419"
 title="Uravnoteženi sistem kazalnikov v poslovni banki X"
 author="Aver, Goran" supervisor="Bernik, Mojca" year="2012"
 publisher_abbr="UM FOV" publisher="Fakulteta za organizacijske vede" place="Kranj"
 url="http://dkum.uni-mb.si/Dokument.php?id=28143"
 type="Diplomsko delo" udc="005" udc_desc="Menedžment"
 pdf_url="http://nl.ijs.si/project/kas/pdf/000/kas-10000.pdf">
  <page xml:id="pb1" n="1"
   pdf_url="http://nl.ijs.si/project/kas/pdf/000/kas-10000.pdf#page=1"
   facs_url="http://nl.ijs.si/kas/facs/000/kas-10000/p0001-Pr4U.png">
    <p xml:id="pb1.p1" xml:lang="sl">
     <s>
      <w lemma="diplomski" ana="jos:Agpnsn">Diplomsko</w>
      <c> </c>
      <w lemma="delo" ana="jos:Ncnsn">delo</w>
        ...
```

Slika 1: Zapis korpusa v XML.

Zvrst	*besedil*	*%*	*strani*	*%*	*sl. besed*	*%*	*besed*	*%*	*pojavnic*
KAS-proto	50.793	100	3.796.957	100	952.172.179	100	992.429.078	100	1.189.100.226
Diplomska	41.212	81,14	2.819.462	74,26	686.276.048	72,07	711.048.854	71,65	850.937.549
Magistrska	6.401	12,60	704.960	18,57	192.438.734	20,21	200.965.696	20,25	240.145.441
Doktorska	700	1,38	147.049	3,87	38.208.949	4,01	42.872.974	4,32	52.874.876
Specialistična	573	1,13	50.144	1,32	14.153.388	1,49	14.474.521	1,46	17.068.921
Znanstvena	782	1,54	29.635	0,78	9.206.780	0,97	10.475.965	1,06	12.737.433
Strokovna	393	0,77	9.568	0,25	3.730.797	0,39	3.977.127	0,40	4.694.058
Ostalo	732	1,44	36.139	0,95	8.157.483	0,86	8.613.941	0,87	10.641.948

Tabela 1: Velikost korpusa KAS-proto po zvrsteh besedil.

4 Zgradba korpusa

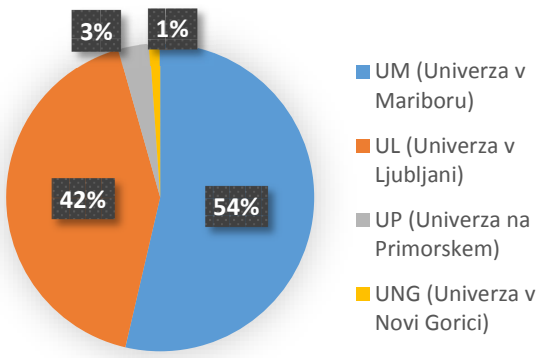
4.1 Zvrst

Tabela 1 vsebuje podatke o velikosti korpusa in njegovih posameznih zvrsteh po besedilih, straneh, številu besed v odstavkih, ki so bili identificirani kot slovenski in v celoti, ter po pojavnicah. Korpus vsebuje skoraj 1,2 milijarde pojavnic, s čimer je po velikosti primerljiv s trenutno največjim korpusom slovenskega jezika Gigafido (Logar Berginc et al., 2012).

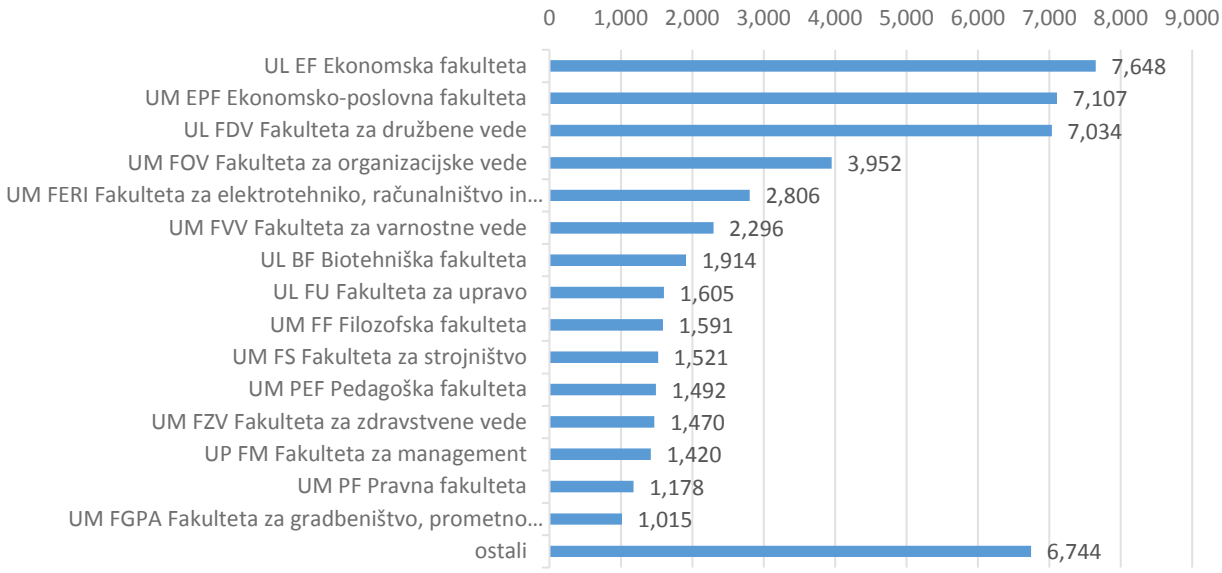
Daleč največji del korpusa predstavljajo diplomska dela, saj zajemajo dobre štiri petine vseh del oz. v korpus prinašajo skoraj tri četrtine strani ali 72 % vseh slovenskih besed. Sledijo magistrska dela in doktorske disertacije, iz katerih je v korpus prišla skoraj četrtina slovenskih besed.

Pri zvrsteh besedil, navedenih v zadnjih treh vrsticah, smo združili več tipov objav: pod znanstvena dela spadajo znanstveni prispevki na konferencah, izvirni znanstveni članki itd., pod strokovna dela strokovni članki, strokovne monografije itd., pod ostala dela pa npr. predgovori, spremna besedila in učna gradiva. Kljub temu, da vsa ta dela skupaj predstavljajo samo nekaj odstotkov korpusa, gre še vedno za razmeroma velike podkorpuse: nezaključna znanstvena dela npr. vsebujejo skoraj deset milijonov besed.

Zanimiv je tudi podatek, kolikšen delež korpusa predstavljajo angleški deli besedil, saj jih je, po eni strani, treba za enojezične raziskave filtrirati, po drugi pa so ti deli dragoceni kot neke vrste vzporedni ali vsaj primerljivi



Slika 2: Število besedil po univerzah.

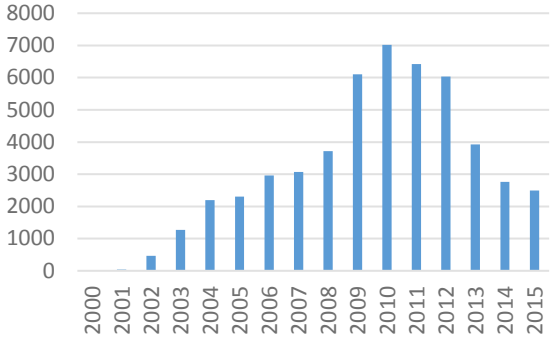


Slika 3: Število besedil po fakultetah.

podkorpus. V korpusu je približno 4 % angleškega besedila: večino prinašajo diplomska dela (3,5 %), po posameznih zvrsteh pa je najvišji delež angleških besedil značilen za doktorske disertacije (skoraj 11 %) in nezaključna znanstvena dela (12 %)

4.2 Vir

Slika 2 kaže, kolikšen delež besedil so v korpus prispevale posamezne slovenske univerze. Tu predvsem preseneča, da prihaja več kot polovica besedil z Univerze v Mariboru (UM), saj ima Univerza v Ljubljani (UL) vsaj dvakrat več študentov. Vzrok je v tem, da je UM repozitorij vzpostavila že leta 2008, UL pa šele leta 2013 (pred tem so na UL obstajale samo podatkovne zbirke posameznih fakultet, npr. Ekonomske fakultete, Fakultete za družbene vede in še nekaterih). Enako velja tudi za ostali dve slovenski univerzi, Univerzo na Primorskem (UP) in Univerzo v Novi Gorici (UNG), ki sta repozitorija prav tako vzpostavili šele leta 2013. Če v korpusu pogledamo število besed po univerzah, je sicer delež obeh največjih (UM in UL) podoben: dela z UM v KAS-proto prinašajo 49 % besed, dela z UL pa 47 % besed.



Slika 5: Število besedil po letih.

Z vzpostavitvijo repozitorijev je povezano tudi to, koliko besedil so v korpus prispevale posamezne fakultete. Na sliki 3, ki prikazuje 15 fakultet z največjim deležem besedil v korpusu KAS-proto (od skupno 55), je razvidno, da kar polovico vseh besedil prispevajo samo štiri visokošolske ustanove, in to vse družboslovne. Tudi sicer je del s tehničnih oz. naravoslovnih fakultet, vključenih v Nacionalni portal odprte znanosti (in posledično v KAS-proto), manj, še posebej pa zaostajajo humanistične vede. Tako se Filozofska fakulteta UM po obsegu del v korpusu še uvrsti na seznam prvih petnajstih, Filozofska fakulteta UL pa je šele na 18. mestu z le 847 deli.

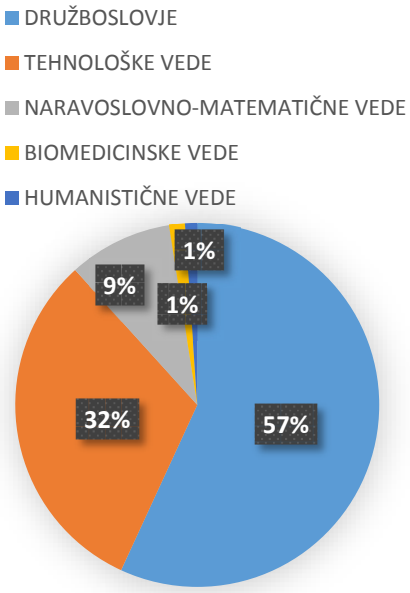
Zanimiva je tudi razporeditev del v korpusu po vedah. Prvi (tipično najpomembnejši) vrstilec UDK smo preslikali v vede taksonomije CERIF (Common European Research Information Format) in vsakemu delu v korpusu pripisali njegovo (glavno) vedo. Slika 4 podaja razmerja med posameznimi vedami.

Vede v veliki meri odslikavajo razdelitev del po fakultetah, saj tudi tu vodijo družboslovne vede, tem pa sledijo tehnološke vede z malo manj kot tretjino vseh del. Če je naravoslovno-matematičnih ved še za slabo desetino, je tako biomedicinskih kot humanističnih ved vsega skupaj 1 % vseh besedil.

4.3 Leto

Kot rečeno, smo iz zajetih besedil izločili dela, starejša od leta 2000, zajem pa je bil narejen v začetku leta 2016. Slika 5 podaja razporeditev števila besedil po letih objave. Iz leta 2000 jih je samo 5, iz naslednjega leta 39, iz leta 2002 pa že 470 in iz vseh kasnejših let po več kot tisoč. Presenetljivo je, da začne po letu 2010, še bolj pa po letu 2012, število besedil v repozitoriju strmo padati. Vzrok je v tem, da so univerze začele po letu 2012 vzpostavljati procese vstavljanja zaključnih del in pravne podlage za podporo tem procesom, kar je upočasnilo dodajanje novih del. Na UL bodo procesi dokončno vzpostavljeni šele do konca septembra 2016. Zaradi tega nekatere fakultete dela svojih študentov samo arhivirajo in jih ne odlagajo v

imenike. Drugi vzrok je v tem, da se število študentov na slovenskih univerzah vsako leto nekoliko zmanjša.



Slika 4: Število besedil po področjih.

5 Načrtovane raziskave in uporaba

Korpus KAS-proto bo omogočil izvedbo več jezikoslovnih in bibliotekarskih (ter v povezavi z obojimi tehnoloških) raziskav slovenskega akademskega jezika.

5.1 Klasifikacija besedil

Na podlagi korpusa KAS-proto in njegovih nadaljnjih različic bomo razvili metode za klasifikacijo besedil in luščenje ključnih besednih zvez, ki bodo izboljšale uporabnost portala odprte znanosti s tem, da bo z njimi omogočeno bolj kompleksno iskanje po vsebinah (Bordea et al., 2015; Fišer et al., 2010; Siddiqi in Aditi, 2015). S priporočili ključnih besednih zvez bo nadgrajen tudi vmesnik za knjižničarje, ki v univerzitetne repozitorije vnašajo nova besedila.

Pri tem bomo uporabili prosto dostopni (vrhnji) del klasifikacije UDK (UDC Linked Data), klasifikaciji ARRS in CERIF, za posamezna področja pa tudi MeSH, Eurovoc in Agrovoc. Zadnje tri klasifikacije so dostopne v XML ali SKOS/RDF, medtem ko bomo ostale za lažje procesiranje in izmenjavo v ta format pretvorili sami.

Kot glavno taksonomijo bomo uporabili UDK, saj je z njo že opremljena večina dokumentov v repozitorijih, poleg tega pa je taksonomiji ARRS in CERIF mogoče enostavno povezati s področji UDK. Taksonomija UDK ni zgolj šifrant, temveč je klasifikacijski sistem s pravili, ki omogočajo relacije med razredi, priredno in zaporedno razširitev področja, enostavne relacije ter podrobne delitve. Razvili bomo razčlenjevalnik za dekompozicijo razredov, ki bodo nato služili kot osnova za večlabelno klasifikacijo dokumentov.

Za učno množico bomo uporabili besedila v korpusu, ki že vsebujejo oznako UDK. Za vsak dokument bomo izluščili besede skupaj z njihovo utežjo TF-IDF. Nato bomo na podatkovni množici preizkusili več metod strojnega učenja. Za razvrščanje dokumentov po podobnosti bomo uporabili rangirno funkcijo BM25, za nadzorovano učenje klasifikacije pa bomo preizkusili različne metode, ki si dostopne v okviru knjižnice Scikit-learn.

V naslednjem koraku bomo izvedli preizkuse s spreminjanjem oz. razširitvijo podatkov ter značilk z:

- uporabo kombinacije določenih delov besedila (naslov, ključne besede, kazalo ipd.) namesto celotnega besedila;
- uporabo lem namesto besednih oblik iz besedila;
- uporabo pogostih n-gramov (enostavnih večbesednih enot) namesto posameznih besed;
- uporabo identificiranih terminov namesto splošnih večbesednih enot;
- dodatno uporabo nadpomenk ali pomenskih relacij pri terminih, ki so povezani z zunanjimi terminološkimi zbirkami;
- dodatno uporabo definicij pri povezanih terminih.

5.2 Razvoj orodij za delo s terminologijo

Luščenje terminologije poteka v treh korakih (Heyle in De Hertog, 2015): zaznavanje jezikovnih prvin, ki sestavljajo večbesedno enoto (zaznavanje enotskosti), razvrščanje po verjetnosti, da so izluščeni termini z določenega področja (zaznavanje terminološkosti) ter združevanje pomensko in konceptualno povezanih terminov (zaznavanje variantnosti), kar je pomemben korak, saj sem sodi kar 15–35 % izluščenih terminoloških kandidatov (Daille, 2005). Četrti, opcijski korak je identifikacija prevodnih ustreznic terminov v drugem jeziku (Daille et al., 1994).

5.2.1 Luščenje terminologije

Osrednji cilj je nadgradnja in evalvacija orodja CollTerm (Pinnis et al., 2012), ki je bil razvit v naših preteklih raziskavah. Trenutna različica terminološke kandidate izlušči s pomočjo oblikoskladenjskih vzorcev in več statistik za sopojavitev besed oz. besednih zvez, kot izhod pa ponudi urejen seznam terminoloških kandidatov, po en seznam za vsako stopnjo n-gramov, pri čemer je n tipično 1 – 6. Nadgradnja bo orodju dodala modul za nadzorovano učenje, ki bo filtriral in rangiral vsak element seznamov, s čimer bomo kot izhod dobili en sam seznam rangiranih terminoloških kandidatov.

Za delovanje modula bomo potrebovali korpus, ročno označen s termini, kar bomo izvedli s pomočjo spletne platforme za označevanje korpusov WebAnno (Eckart de Castilho et al., 2014).

V sklopu projekta bomo razvili tudi sistem za identifikacijo angleških prevodnih ustreznic, ki so na voljo v izvornih dokumentih v obliki dvojezičnih seznamov ključnih besed, tabelaričnih dvojezičnih glosarjev, dvojezičnih izvlečkov in povzetkov ali kot prevodne ustreznice v oklepajih.

5.2.2 Zaznavanje terminoloških variant

Drugi cilj je nadgradnja sistema z identifikacijo različnih poimenovanj istega pojma. Do terminološke variantnosti ne prihaja le zaradi terminološke večpomenskosti, ki je pogost meddisciplinarni pojav, temveč tudi zaradi stilističnih načel tvorjenja besedil in jezikovne gospodarnosti, s katero se izogibamo prekomernemu ponavljanju, zlasti v primeru daljših

terminov. Pričakujemo še, da bodo raznorodne rešitve uporabljali različni avtorji in v različnih časovnih obdobjih na področjih, na katerih se terminologija šele uveljavlja.

5.2.3 Analiza rabe terminov

Na podlagi identificiranih terminoloških kandidatov in njihovih variant bomo izvedli analize, ki bodo prvič omogočile celosten in dragocen vpogled v stanje ter trende terminološke rabe na različnih raziskovalnih področjih v Sloveniji. Za različne vrste akademskih besedil, strokovna področja in časovna obdobja bomo izmerili terminološko gostoto, stopnjo terminološke variantnosti in stopnjo terminološke interdisciplinarnosti.

5.2.4 Izgradnja terminoloških zbirk

Izluščeni terminološki kandidati bodo objavljeni v prosto dostopnem spletnem slovarskem urejevalniku, ki bo slovenskim znanstvenim in strokovnim skupnostim omogočal upravljanje s terminologijo lastnih področij. V izbranih skupnostih bomo pridobili tudi odziv na terminološko zbirko, ki jo bomo zanje pripravili v projektu.

5.3 Korpusni podatki za opis slovenskega akademskega jezika

Podatki iz obsežnega in področno raznolikega korpusa KAS nam bodo omogočili pripravo med vedami ter področji primerjalnega in različnim besedilnim žanrom prilagojenega opisa sodobnega slovenskega jezika, kakršen se rabi v akademskem okolju. Opis bo nastal s sintezo rezultatov treh vrst analiz:

- leksikalne analize,[3]
- besediloslovne in slovnične analize ter
- stilne analize.

5.3.1 Leksikalna analiza

Z metodo frekvenčnega profila (Rayson in Garside, 2000) bomo med drugim analizirali za akademsko pisanje značilno področno nespecializirano leksiko, za katero bi lahko rekli, da je del splošnega strokovnega jezika, npr. *definirati*, *določiti*, *analizirati*, *ključna beseda*, *metoda*, *vzorec*. Funkcija Besedne skice v leksikografskem orodju Sketch Engine (Kilgarriff et al., 2004) nam bo omogočila podrobnejšo analizo tipičnega besedilnega okolja tega besedišča (Logar, 2013b: 115–124), rezultate katere bomo nato ročno pregledali in jih enako kot zgoraj terminološke kandidate vnesli na prosto dostopen spletni portal.

5.3.2 Besediloslovna in slovnična analiza

Besediloslovna in slovnična analiza bo najobširnejša, saj nam bo korpus omogočal npr. analizo skladenjske zapletenosti povedi akademskega pisanja ter značilnih in izstopajočih skladenjskopomenskih kategorij, ki pripomorejo k temu, da je akademsko (znotraj njega zlasti znanstveno) pisanje natančno, jasno ter zgoščeno (Skubic, 1994/95). Posamezne slovnične kategorije (npr. trpnik, gl. Toporišič, 2000: 27–30) bomo opazovali primerjalno med posameznimi področji in primerjalno z leposlovnim delom korpusa ccGigafida (Logar Berginc et al., 2012). Korpusno bomo razčlenili tudi medbesedilnost (Hyland, 2004), ki je blizu področju plagiatorstva (Chandrasoma et al., 2004), in pregledali pojavljanje metabesedilnosti (Williams, 1981) kot dela retoričnih konvencij, ki se nanašajo na besedilo samo ali na odnos med tvorcem in naslovnikom.

5.3.3 Stilna analiza

Izbrani podkorpusi besedil različnih področij nam bodo dali vpogled v prvine (ne)osebnega stila akademskega pisanja in omogočili ugotovitev, ali poudarjena intelektualizacijska vloga (Južnič, 1992; Skubic, 2005) res v celoti izključuje avtorjevo prisotnost v besedilu. Preverili bomo tudi, kakšna je norma slovenskega akademskega pisanja, ter to, ali je znana tipa pisanja – germanski in anglosaksonski (Kalin Golob, 2008: 88–89) – na slovenskih besedilih mogoče (še) prepoznati ter ločiti.[4]

6 Zaključek

V prispevku smo predstavili korpus KAS-proto, njegovo izdelavo in kvantitativni vpogled v njegovo sestavo ter pregled načrtovanih raziskav.

Pred tremi leti smo v zvezi z aktualnimi terminološkimi opisi in njihovo dostopnostjo razmišljali takole: »V času nujnosti internacionalizacije strok in mednarodnega odpiranja njenih nosilcev je za polno funkcionalnost nacionalnega jezika na področju strokovnega jezika mogoče poskrbeti predvsem tako, da ga digitalno celostno podpremo ter si pri tem pomagamo prav z orodji, ki jih je prinesla digitalizacija« (Logar, 2013a: 251). V projektu »Slovenska znanstvena besedila« si bomo prizadevali to podporo še okrepiti in dokazati, da je pravzaprav šele s pomočjo takih orodij ter virov mogoče slovenščino v različnih žanrih akademskega diskurza zares dobro opisati; pri čemer ni zanemarljiva vloga še enega dejavnika: odprtosti tujih – in v prihodnje tudi naših – znanstvenih rezultatov.

Zahvala

Avtorji se zahvaljujejo anonimnima recenzentoma za koristne pripombe. Raziskavo, opisano v prispevku, je podprl projekt ARRS J6-7094 »Slovenska znanstvena besedila: viri in opis«.

7 Literatura

Akcijski načrt za jezikovno izobraževanje. 2015, http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/raziskave-analize/slovenski_jezik/Akcijska_nacrta/ANJI.pdf.

Akcijski načrt za jezikovno opremljenost. 2015, http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/raziskave-analize/slovenski_jezik/Akcijska_nacrta/ANJO.pdf.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy Extraction Evaluation (TexEval). *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 902–910, Denver, Colorado, June 4–5, 2015. ACL.

[3] V tem delu tiste leksike, ki jo lahko (oz. kolikor jo lahko) ločimo od terminologije.

[4] O drugem gl. še npr. Lengar Verovnik, Logar, Kalin Golob (2013: 29–49).

Ranamukalage Chandrasoma, Celia Thompson, Alastair Pennycook. 2004. Beyond Plagiarism: Transgressive and Nontransgressive Intertextuality. *Journal of Language, Identity, and Education* 3, 171–193.

Béatrice Daille. 2005. Variations and application-oriented terminology engineering. *Terminology*, 11/1, 181–197.

Béatrice Daille, Éric Gaussier, Jean-Marc Langé. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. *Proceedings of the 15th conference on Computational linguistics, Volume 1*, Kyoto, Japan, 515–521.

Vijay K. Bhatia. 1993. *Analysing genre: Language use in professional settings*. London, New York: Longman.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih. 2015. Oblikoskladenjski leksikon Sloleks 1.2, *Slovenian Language Resource Repository CLARIN.SI*, http://hdl.handle.net/11356/1039.

Richard Eckart de Castilho, Chris Biemann, Iryna Gurevych in Sied Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands.

Tomaž Erjavec. 2013. Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0* 1/1, 24–49, http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf.

Tomaž Erjavec. 2009. Odprtost jezikovnih virov za slovenščino. V: M. Stabej, ur.: *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: ZIFF.

Tomaž Erjavec, Jan Jona Javoršek, Simon Krek. 2014. Raziskovalna infrastruktura CLARIN.SI. *Zbornik 9. konference Jezikovne tehnologije.* Ljubljana: IJS. 19–24.

Andrej Ermenc Skubic. 2005. *Obrazi jezika*. Ljubljana: Študentska založba.

Darja Fišer, Senja Pollak, Špela Vintar. 2010. Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources. *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*.

Ken Hayland. 2004. *Disciplinary Discourses: Social Interactions in Academic Writing*. Michigan: University of Michigan.

Kris Heylen, Dirk De Hertog. 2015. Automatic Term Extraction. *Handbook of Terminology, Volume 1*. John Benjamins Publishing Company, 203–221.

Stane Južnič. 1992. *Diplomska naloga: napotki za izdelavo*. Ljubljana: Amalietti.

Monika Kalin Golob. 2008. *Jezikovnokulturni pristop h knjižni slovenščini*. Ljubljana: FDV.

Monika Kalin Golob, Marko Stabej, Mojca Stritar Kučuk, Gaja Červ, Samo Kropivnik. 2014. *Jezikovna politika in jeziki visokega šolstva v Sloveniji*. Ljubljana: Založba FDV.

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

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz. 2015. Učni korpus ssj500k 1.4, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1052.

Tina Lengar Verovnik, Nataša, Logar, Monika Kalin Golob. 2013. *Slovenščina kot strokovni jezik na slovenskih univerzah: pregled stanja ter razčlenitev pomena, načina in možnosti njene večje vključitve*, http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/raziskave-analize/slovenski_jezik/Slovenscina_kot_strokovni_jezik_na_slovenskih_univerzah_01.pdf.

Nikola Ljubešić, Tomaž Erjavec. 2016. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. *Proceedings of 10th Language Resources and Evaluation Conference (LREC 2016)*.

Nataša Logar. 2013a. Aktualni terminološki opisi in njihova dostopnost. V: A. Žele, ur.: *Družbene funkcijskost jezika (vidiki, merila, opredelitve)*. Ljubljana: ZIFF. 247–253.

Nataša Logar. 2013b. *Korpusna terminologija: primer odnosov z javnostmi*. Ljubljana: Trojina: zavod za uporabno slovenistiko; Založba FDV.

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

Milan Ojsteršek, Mojca Kotar, Marko Ferme, Goran Hrovat, Mladen Borovič, Albin Bregant, Jan Bezget, Janez Brezovnik. 2014. Vzpostavitev repozitorijev slovenskih univerz in nacionalnega portala odprte znanosti. *Knjižnica* 58/3, 15–39, http://knjiznica.zbds-zveza.si/index.php/knjiznica/article/view/499.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, Tatiana Gornostay. 2012. Term Extraction, Tagging, and Mapping Tools for Under-resourced Languages. *Proceedings of the 10th Conference on Terminology and Knowledge Engineering*, Madrid, Spain, 193–208.

Paul Rayson, Roger Garside. 2000. Comparing Corpora Using Frequency Profiling. *Proceedings of the ACL Workshop on Comparing Corpora*. Hong Kong, 1–6.

Resolucija o Nacionalnem programu za jezikovno politiko 2014–2018. 2013, http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/Zakonodaja/2013/Resolucija_-_sprejeto_besedilo__15.7.2013_.pdf.

Pavel Rychlý. 2007. Manatee/Bonito – A Modular Corpus Manager. *Proceedings of the Workshop on Recent Advances in Slavonic Natural Language Processing.* Brno: Masaryk University, 65–70.

Sifatullah Siddiq, Aditi Sharan. 2015. Keyword and Keyphrase Extraction Techniques: A Literature Review. *International Journal of Computer Applications*, 109. 2.

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

John M. Swales. 2000. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Jože Toporišič. 2000. *Slovenska slovnica*. Maribor: Obzorja.

Joseph M. Williams. 1981. *Style: Ten Lessons in Clarity & Grace*. Glenview, IL: Scott, Foresman and Company.