Slovenski besedilni korpus bibliotekarstva – najsodobnejša slovaropisna podpora bibliotekarski terminologiji[1]

Slovene text corpus of library science – An advanced lexicographic tool for library terminology

Ivan Kanič
Centralna ekonomska knjižnica Ekonomske fakultete Univerze v Ljubljani

Povzetek
Bibliotekarstvo je stroka, ki ima v slovenskem prostoru bogato tradicijo, danes pa uspešno sledi najnaprednejšim tokovom razvoja v svetu. Zato je presenetljivo, da v preteklosti bibliotekarski terminologiji ni bila namenjena sistematična skrb ali organiziran poskus njene kodifikacije in normiranja v slovarju. Vrzel sta v zadnjem desetletju zapolnila prevajalni in razlagalni slovar bibliotekarske terminologije, ki sta nastala po sodobnih načelih leksikografije in temeljita še na ročnem izpisovanju bibliotekarskih izrazov iz obsežnega nabora slovenskih strokovnih besedil in po njih izdelanem geslovniku. Slovaropisna skupina sledi sodobnim leksikografskim tokovom, zato bo uporabila za dopolnjevanje slovarjev učinkovita orodja korpusnega jezikoslovja. Zasnovala in vzpostavila je slovenski besedilni korpus bibliotekarstva, ki je v prvi fazi že presegel 1,8 milijona besed, prevzetih iz 234 slovenskih strokovnih in znanstvenih besedil s področja bibliotekarstva. Korpus omogoča različne oblike iskanja in prikaza besed in besednih zvez v ožjem ali širšem sobesedilu ter izdelavo seznamov in kazal po meri za besedne analize in primerjave. Spletna aplikacija je javno dostopna.

Ključne besede
bibliotekarstvo, informacijska znanost, terminologija, korpusno jezikoslovje, besedilni korpusi, Korpus bibliotekarstva

UDK 02(038)=163.6

Abstract
In Slovenia librarianship has had a rich and successful tradition, and today it is well in line with the most advanced developments and trends in the world. Surprisingly no systematic concern had been paid to library terminology in the past and no organized attempt to attain its codification and authority in dictionaries. Recently two dictionaries have filled the gap, a translation dictionary and an explanatory dictionary of library terminology. They were both concieved and built in accordance with up-to-date principles of modern lexicography, thus they were based on the excerption of a set of selected technical and scientific texts in librarianship and library science, the excerption being manual at the time, of course. A list of selected and evaluated headings was produced from excerpted texts to initiate the building of the dictionary. The lexicographers keep following closely current modern trends in lexicography, consequently they have designed and set up the Slovene text corpus in the field of library and information science, a powerful and efficient tool for editing and amending of the existing Dictionary of Library Terminology. The corpus has reached 1.8 million words extracted from 234 Slovene technical and scientific texts. It supports a varisty of specialized search methods, display of search results – words in close or wider context, and building of customized word lists for text and word analysis. The web based application is in open access.

Keywords
Librarianship, library science, LIS, terminology, corpus linguistics, text corpus, Korpus bibliotekarstva

UDC 02(038)=163.6

---

[1] Objavljeno v: *Knjižnica : odprt prostor za dialog in znanje : zbornik referatov = Library : open space for dialogue and knowledge : proceedings* / Strokovno posvetovanje Zveze bibliotekarskih društev Slovenije, Maribor, 20.-22. oktober 2011 = Professional Conference of Slovenian Library Association, Maribor, October 20-22, 2011 ; [urednici Melita Ambrožič in Damjana Vovk]

1 Uvod

Natančno in enoznačno sporazumevanje med ljudmi je *conditio sine qua non* organizirane družbe, vzdrževanja že doseženih vrednot in vzpostavljanja novih. Vse bolj zapletene oblike sporazumevanja in sodelovanja med posamezniki in skupinami to zahtevo še bolj postavljajo v ospredje. Obsežni enojezični razlagalni slovarji omogočajo tako poglobljeno in enoznačno razumevanje pripadnikov istega jezika, večjezični, pretežno prevajalni slovarji, pa sporazumevanje med pripadniki različnih jezikov. Razvoj družbe in njena stratifikacija na razvejena in ozko specializirana strokovna področja se odraža tudi v razvoju jezika stroke. Strokovna terminologija je tako dokaz moči, pestrosti, razvitosti in samostojnosti nekega jezika, istočasno pa tudi stroke, v kateri se uporablja, zato je jasno, da je treba strokovno terminologijo ustrezno gojiti in negovati, po drugi strani pa tudi normirati in kodificirati v enojezičnih terminoloških slovarjih.

2 Slovenska bibliotekarska terminologija

Bibliotekarstvo je stroka, ki ima med Slovenci in v slovenskem prostoru dokajšnjo tradicijo, bibliotekar v Licejski knjižnici, današnji Narodni in univerzitetni knjižnici, je bil na primer v 19. stoletju tudi genialni jezikoslovec, znalec devetnajstih jezikov Matija Čop. Slovenski strokovni termini so se kljub močnemu in stalnemu vplivu nemškega jezika v 19. in 20. stoletju, v sedanjosti pa vplivu angleške strokovne literature in s tem tudi terminologije, razvili, utrdili in tudi uveljavili. Ves ta čas bibliotekarskim terminom ni bila namenjena sistematična skrb ali organiziran poskus njihove kodifikacije in normiranja, čeprav so bili v vrstah bibliotekarjev tudi najvidnejša imena slovenskega jezikoslovja, dr.Mirko Rupel je bil ravnatelj Narodne in univerzitetne knjižnice. Raba je zato slonela predvsem na jezikovni praksi in normi posameznih “šol” ali na pomembnih in vplivnih posameznikih, npr. Avgustu Pirjevcu, Pavlu Kalanu in drugih priznanih bibliotekarjih, kar pogosto ni bilo niti oblikovno niti semantično usklajeno. Sodobni strokovni stiki v svetu, hiter razvoj računalniške opreme in s tem vnos sprememb v tehnologijo in metodologijo obdelave ter pojav množice novih nekonvencionalnih nosilcev zapisov, ki kot knjižnično gradivo prihajajo v knjižnice, so povzročili ne samo vdor tujih poimenovanj ampak tudi napačno rabo, ki jo je povzročil vir, iz katerega je črpana informacija.

2.1 Slovenski bibliotekarski terminološki slovarji

Zaradi spoznanja, da bibliotekarska stroka v slovenščini nima niti enojezičnega razlagalnega niti večjezičnega prevajalnega slovarja, je postala izdelava slovarja nujna. V ožjem krogu strokovnjakov se je porodila misel o projektu, ki bi zapolnil to vrzel v bibliotekarski stroki. Ideja je postala resničnost, ko sta Ivan Kanič in Mirko Popovič pripravila organizacijska in kasneje tudi strokovna izhodišča za delovanje projektne skupine, ki je začela z delom leta 1987. Začela se je priprava večletnega projekta za sestavo slovenskih bibliotekarskih terminoloških slovarjev. Cilji projekta za ureditev bibliotekarske terminologije in s tem tudi slovenske Bibliotekarske terminološke komisije so bili:

1. Kodificiranje slovenske bibliotekarske terminologije, ki temelji na
   - evidentiranju strokovnih izrazov, ki se v bibliotekarstvu uporabljajo danes ali so se uporabljali v preteklosti,
   - pomenski analizi posameznega termina in ugotavljanju pomenskih povezav ter sinonimnih ali antonimnih odnosov med njimi,
   - normiranju glede na knjižno normo in zahteve urejene strokovne terminologije.
2. Sestava in izdaja terminoloških slovarjev bibliotekarstva in informacijske znanosti, tj.
   - enojezičnega razlagalnega terminološkega slovarja in
   - večjezičnega prevajalnega slovarja.

3. Jezikovno svetovanje in presoja ob tekočih terminoloških vprašanjih rabe strokovnih izrazov v bibliotekarstvu in informacijski znanosti, ki se opira na za slovar opravljenih pomenskih analizah terminov.
4. Objavljanje rezultatov v strokovni literaturi, predvsem v strokovni reviji Knjižnica, in predstavitev delovanja v drugih strokovnih krogih ter na mednarodni ravni.

Komisija se je ravnala po sodobnih načelih leksikografije, zato je delo temeljilo na **izpisih bibliotekarskih izrazov** iz obsežnega nabora slovenskih strokovnih besedil in po njih izdelanem geslovniku. Ta inventarizacija strokovnega besedišča, ki se je v preteklosti uporabljalo v slovenski bibliotekarski strokovni literaturi, je bila osnova za izbor v slovarju obdelanih strokovnih pojmov in primerjanje z izborom v tujejezičnih strokovnih slovarjih. Inventarizacija izrazja je bila z metodo takrat še "ročnega" izpisovanja strokovnih besedil na tak način opravljena kot izhodiščna faza projekta in je trajala več let, zahteven segment selekcije in vrednotenja terminov, pomenska analiza z vzpostavitvijo pomenskih zvez med izbranimi termini in oblikovanje razlag je predstavljalo osrednje delo terminološke skupine, katere ciljni rezultati so bili objavljeni leta 2002 (*Angleško-slovenski slovar bibliotekarske terminologije*) in leta 2009 še *Bibliotekarski terminološki slovar*, oba v tiskani in spletni izdaji.

3 Besedilni korpusi

V jezikoslovju je *korpus* oziroma *besedilni korpus* velika in strukturirana zbirka besedil, navadno grajena, hranjena in obdelana računalniško. Korpuse se uporablja za statistične analize pisanega in/ali govorjenega jezika, za preverjanje pojavitev besed in besednih zvez ali pa za potrditev lingvističnih pravil v določenem jeziku. So tudi nepogrešljivo in nadvse koristno orodje ob pripravi splošnih in tudi *terminoloških slovarjev*. Korpusi so seveda začeli nastajati tam čez veliko lužo in predvsem v "velikih" jezikih, vendar imamo že nekaj časa tudi za slovenski jezik več po nastanku in namenu različnih korpusov.

Referenčni korpusi so temeljna vrsta korpusov, ki naj bi predstavili celovito podobo nekega jezika. So večjega obsega, zanje je glede na tradicijo tudi najnatančneje izdelana metodologija gradnje, predstavljajo pa izhodišče za temeljne jezikoslovne raziskave predvsem s področja slovnice in slovarjev. Pri nas sodita v to kategorijo dva besedilna korpusa: *Nova beseda*, ki vsebuje okrog 240 milijonov besed iz 5.700 leposlovnih, strokovnih in uradnih besedil do leta 2004, in referenčni korpus slovenskega jezika *FidaPLUS*, ki vsebuje okrog 621 milijonov besed iz slovenskih besedil najrazličnejših zvrsti, objavljenih v letih od 1979 do 2006.

Specializirani korpusi predstavljajo jezik v natančno določeni rabi, med njimi so najpomembnejši korpusi strokovnih jezikov, predvsem v okviru terminoloških raziskav in gradnje terminoloških slovarjev. Taki so pri nas na primer *Korpus DSI* (Korpus zbornikov posvetovanja Dnevi slovenske informatike od 2003 do 2010 in revije Uporabna informatika), ki vsebuje 2 milijona besed, in na istem naslovu *korpus iFpX*, ki zajema okrog 14 milijonov izbranih izrazov iz korpusov FidaPLUS in DSI. Zanimiva sta tudi jezikoslovno označeni korpus *Jos* in *Evrokorpus*, ki je zbirka vzporednih dvojezičnih korpusov prevodov in obsega besedila v 22 jezikih držav, ki so bile leta 2007 članice EU, vsebuje pa 98 milijonov besed oz. 600 tisoč prevodnih enot. Nekaj prav posebnega pa je *Vayna* s četrt milijona besed iz 360 časopisnih člankov, ki so v času od aprila do avgusta 1998 obravnavali ti. "*verbalne napade na JLA*" (oz. ozadje in potek procesa JBTZ).

Besedilni korpusi se v terminologiji uporabljajo za inventarizacijo in preverjanje pojavitev besed in besednih zvez v strokovnih besedilih, torej za ugotavljanje, kateri izrazi in na kakšen način se pojavljajo v jeziku neke stroke. Iz njih je mogoče pridobiti zelo različne sezname besed in besednih zvez, z lematizacijo in besednovrstnim označevanjem pa posegati celo v analizo uporabe posameznih besednih vrst. Zato so nepogrešljivo in nadvse koristno orodje ob pripravi sodobnih terminoloških

slovarjev. V ta namen smo si že dalj časa prizadevali vzpostaviti tak korpus tudi na področju slovenskega bibliotekarstva.

3.1 Predhodnik besedilnega korpusa bibliotekarstva

Strokovno delo pri pripravi *Bibliotekarskega terminološkega slovarja* , ki je potem izšel leta 2009, se je ravnalo po sodobnih načelih leksikografije, zato je temeljilo na ugotavljanju rabe izrazja v strokovnem jeziku in evidentiranju terminov s tradicionalnim *"ročnim"* izpisovanjem bibliotekarskih izrazov iz obsežnega seznama tiskanih slovenskih strokovnih besedil. Izpisovanje izbrane slovenske bibliotekarske strokovne literature v letih 1988 do 1999 je zajelo 291 sistematično in v celoti izpisanih besedil na skupno 6575 straneh, ob tem pa še naključne izpiske iz večjega števila ob delu uporabljenih besedil. Upoštevana so bila dela okrog 140 slovenskih avtorjev in tudi nekaj prevodov, npr. standardi ISBD. Na tej osnovi je bil nato izdelan alfabetarij, ki je predstavljal osnovni nabor izrazov za pripravo slovarja. "Ročno" pripravljeni izpiski so bili že tedaj računalniško obdelani, najprej s "tablico" ZX Spectrum, nato z Atarijem in končno z osebnimi računalniki na različnih operacijskih sistemih. Računalniška besedilna zbirka je vsebovala 10.300 ekscerptov in je že bila dostopna na spletu. Besedilnega korpusa in njegovih funkcij s takratno računalniško opremo, sredstvi in znanjem še ni bilo mogoče vzpostaviti. Danes je to že mogoče in po enoletnih pripravah je julija 2011 "shodila" testna zasnova *slovenskega Korpusa bibliotekarstva*, v začetku septembra 2011 pa so delovale že vse predvidene funkcije.

4 Slovenski Korpus bibliotekarstva

Natančno dva meseca je trajalo, da se je *Korpus bibliotekarstva* iz svojih zametkov razvil v povsem delujoč sistem z vsemi funkcijami, kot je bil načrtovan. Namenjen je analizi slovenskih bibliotekarskih strokovnih in znanstvenih besedil in v njih uporabljnih terminov, predvsem pa kot učinkovito orodje za dopolnjevanje *Bibliotekarskega terminološkega slovarja*. Korpus omogoča različne oblike iskanja in prikaza besed in besednih zvez v ožjem ali širšem sobesedilu ter izdelavo seznamov in kazal po meri za analize in primerjave.

Korpus je javno in brezplačno dostopen na spletu pod pogoji licence CC, zato upamo, da bo koristil tudi strokovnim kolegom bibliotekarjem in tudi študentom bibliotekarstva ter morebiti kakšnemu jezikoslovcu, terminologu ali slovaropiscu iz drugih logov.

4.1 Nabor besedil

Korpus zajema skoraj izključno elektronsko objavljena besedila, bodisi izvorno digitalna ali digitalizirana, izjema je le nekaj pomembnejših besedil, ki so bila objavljena samo v tiskani obliki in smo pridobili od avtorjev njihove postprinte, predvsem magistrska in doktorska dela. Ob zasnovi korpusa so bila pripravljena tudi strokovna izhodišča za črpanje besedil, na katerih temelji obširen seznam potencialnih kandidatov za vključitev. To so predvsem zaključna dela s širšega področja bibliotekarstva (diplomska, magistrska in doktorska dela ljubljanskega *Oddelka za bibliotekarstvo, informacijsko znanost in knjigarstvo*, pa tudi nekatera izbrana dela iz drugih univerz; dela, ki niso v slovenskem jeziku, seveda niso bila upoštevana), članki iz novejših letnikov strokovne revije Knjižnica, članki iz revij Organizacija znanja in COBISS obvestila, nekateri izbrani članki iz Knjižničarskih novic, prispevki iz večjega števila zbornikov, nekaj strokovnoinformativnih člankov iz drugih virov in monografske publikacije različnih vrst, predvsem tiste, ki sta jih objavili Narodna in univerzitetna knjižnica in Zveza bibliotekarskih društev Slovenije. Nekaterih del žal ni bilo mogoče vključiti zaradi zaščite objavljenih datotek v pdf formatu, ki ne dovoljuje branja posameznih besed v besedilu.

4.2 Obseg korpusa

Že v začetni fazi vzpostavitve je dosegel korpus občudovanja vreden obseg, predvsem pa zadovoljivo delujejo tudi že vse predvidene funkcije. Vsebuje več kot **1,8 milijona besed**, črpanih iz **234 krajših ali daljših besedil**. Vsa navedena dela so bila objavljena v elektronski obliki, večina izvorno digitalnih oz. vzporednih tiskani izdaji, nekaj pa tudi digitaliziranih. Poudarek je na zajemu besedil, objavljenih v zadnjem desetletju, glede na možnosti pa kdaj tudi starejša. Selektivni seznam potencialno zanimivih besedil obsega še okrog 400 enot, s katerimi bomo v prihodnje dopolnjevali korpus glede na časovne možnosti.

Vrsta objave	**Število besedil**	**Število besed**
Doktorske disertacije	4	215.223
Magistrska dela	19	525.696
Diplomska dela	8	203.805
Monografske publikacije	10	*207.837*
Članki:		
Revija Knjižnica	79	330.518
Organizacija znanja	31	102.491
Knjižničarske novice	21	40.049
Prispevki v zbornikih	59	212.377
Drugi članki in sestavki	3	8.148
Celotni korpus	**234**	**1.811.981**

Prikaz 1: Analizirana besedila po tipu in število iz njih zajetih besed.

Seznam 234 vključenih besedil s hipertekstnimi povezavami na celotna besedila je objavljen v dokumentaciji na spletu.

4.3 Programska oprema in postavitev korpusa

Pri pripravljanju besedil in za postavitev ter javno spletno uporabo korpusa je bila uporabljena domača programska oprema - urejevalnik besedil EVA in njegova internetna različica *NEVA* s specifičnimi funkcijami, ki že nekaj let omogočajo delovanje splošnega referenčnega korpusa slovenskega jezika *Nova beseda*, spletne različice *Slovarja slovenskega knjižnega jezika* in nekaterih drugih slovarskih in jezikoslovnih orodij. Vse priprave in obdelave besedil potekajo na osebnem računalniku, prav tako izdelava številnih indeksov za konkordančno in besedno iskanje ter iskanje po besednih parih, trojčkih, četverčkih in peterčkih.

4.4 Avtorske pravice

Upoštevana besedila niso v korpusu dostopna niti v izvirni obliki niti v celoti, za uporabnike so izdelane le hipertekstne povezave na izvirno objavo (npr. dLib.si, arhiv revije Knjižnica ipd.). Besedila so uporabljena samo za izdelavo kumulativnih statističnih kazalcev jezika, npr. za sezname besed ali besednih zvez s pogostnostjo, in v konkordančnih seznamih, vendar tudi tam le v obliki ožjega citata ne več kot treh povedi - tekoče povedi, povedi pred njo in povedi za njo. Korpus torej ne posega v avtorske pravice piscev besedil ali založnikov. Kjer prispevki niso prosto dostopni, je zato povezava narejena samo na naslovno stran časopisa (npr. Knjižničarske novice), kadar je zbornik objavljen v eni sami datoteki, je za vsakega od prispevkov povezava na celoten zbornik.

Izključni nosilec avtorskih pravic za uporabljene programske rešitve je dr. Primož Jakopin. Korpus bibliotekarstva je zasnoval in pripravil Ivan Kanič, korpus je javno in brezplačno dostopen na spletu pod pogoji licence CC.

4.5 Uporaba in funkcije korpusa

Besedilni korpus je v celoti spletna aplikacija in ne potrebuje nalaganja nobenih komponent na uporabnikov računalnik, prav tako ni nobenih omejitev glede na vrsto in/ali verzijo spletnega brskalnika. Nameščen je kot posebna stran bloga *Bibliotekarska terminologija*, kjer so objavljene tudi vse novosti, dokumentacija za pomoč pri uporabi, nekatere zanimivosti in analize. Uporabniški vmesnik je enostaven in pregleden, omogoča nekaj osnovnih uporabniških nastavitev in izbor načina poizvedovanja ter možnosti omejevanja iskanja na določen tip dokumentov.

3.5.1 Uporabniške nastavitve omogočajo:

- Omejevanje iskanja konkordanc glede na veliko/malo začetnico besede, saj iskanje sicer ne razlikuje med velikimi in malimi črkami.
- Besedno iskanje po celih besedah ali krnih; standardno iskanje poteka samo natančno po vpisanem nizu znakov kot zaključeni celoti – besedi. Nastavitev velja za besedno iskanje, iskanje po parih, trojčkih, četverčkih in peterčkih.
- Omejitev iskalnega razpona: Standardno poteka iskanje po vseh besedilih (označena je izbira *"celotni korpus"*), z izbiro pa lahko uporabnik omeji iskanje po besedilih samo enega tipa ali več tipov dokumentov hkrati (npr. samo po člankih iz revije Knjižnica, samo po doktorskih disertacijah ipd.).

Slika 1: Uporabniški vmesnik (blog *Bibliotekarska terminologija*)

4.5.2 Iskalni postopki

Uporabniški vmesnik omogoča šest različnih vrst iskanja, za vsako je izdelan tudi poseben indeks. V navodilih in pomoči za uporabo korpusa so navedeni tudi ilustrativni prikazi posameznih iskanj in možnosti, ki jih ima uporabnik na razpolago. Iskalni niz je mogoče kombinirati tudi z drugimi podatki, npr. z dolžino besed, njihovo pogostostjo, pojavljanju z določeno drugo besedo ipd.

Konkordance - iskanje in prikaz besed(e) v sobesedilu z navedbo vira.
Kot rezultat poizvedbe se izpiše konkordančni seznam iskanega zaporedja znakov v ožjem sobesedilu tako, da je pred iskano besedo še 45 znakov in enako število znakov za njo. Tradicionalno se takšno kazalo imenuje tudi KWIC indeks ali *ključne besede v besedilu*.

- Standardno poteka iskanje po vpisanem nizu znakov vključno s presledki in upoštevaje vrstni red besed, ki jih je lahko več.
- Maskiranje posameznih znakov ni mogoče.
- Dovoljeno je desno krajšanje vsake od besed z znakom *****.
- Standardno se izpiše do 100 zadetkov na stran, nastavitev je mogoče spremeniti. Pri velikem številu zadetkov je omogočeno listanje po straneh ali začetek prikaza na določenem mestu v seznamu.
- Na desni strani vsakega zadetka se izpiše kodirana oznaka dokumenta, ki s klikom pokaže sobesedilo iskane besede. Izpišejo se do tri povedi, tekoča poved z iskano besedo, poved pred njo in poved za njo. Nad besedilom je skrajšan bibliografski opis dokumenta s hipertekstno povezavo do celotnega izvirnega besedila na strežniku, kjer je objavljeno.

Besedno iskanje
- Iskanje ene same besede, dovoljen je levi in/ali desni odrez z znakom *.
- V rezultatih se izpiše abecedni seznam zadetkov z navedbo pogostosti pojavljanja (brez sobesedila).
- V naslednjem koraku je mogoče prikazati vsakega od zadetkov tudi v sobesedilu in z navedbo vira, ki je prav tako hipertekstno povezan.

Iskanje po pogostih besednih parih
- Iskanje ene ali obeh besed v besednem paru besede, dovoljen je levi in/ali desni odrez z znakom *.
- Znak * lahko nadomešča tudi eno od besed v celoti.
- V rezultatih se izpiše seznam besednih parov, v katerih se iskana beseda pojavlja.
- Seznam je urejen padajoče po pogostosti pojavljanja.
- S hipertekstno povezavo je omogočen prehod na sobesedilo z navedbo vira.

Iskanje po pogostih besednih trojčkih, četverčkih ali peterčkih
- Iskanje ene ali več besed v besednem trojčku (četverčku, peterčku), dovoljen je levi in/ali desni odrez z znakom *.
- Znak * lahko nadomešča tudi eno ali več besed v celoti.
- V rezultatih se izpiše seznam besednih trojčkov (četverčkov, peterčkov), v katerih se iskana beseda pojavlja.
- Seznam je urejen padajoče po pogostosti pojavljanja.
- S hipertekstno povezavo je omogočen prehod na sobesedilo z navedbo vira.

Iskanje po trojčkih, četverčkih in peterčkih je lahko dolgotrajno, posebno z uporabo zvezdic.

Slika 2: Rezultati poizvedbe: konkordančni prikaz iskane besede v ožjem sobesedilu s hiperpovezavo na širše sobesedilo.

Slika 3: Prikaz iskane besede v širšem sobesedilu z navedbo vira in hipertekstno povezavo nanj.

Slika 4: Rezultati poizvedbe: Prikaz trojčkov s pogostostjo in možnostjo prikaza sobesedila.

4. 6 Vpogled v korpus

V korpusu izpričan besedni zaklad in njegova pestrost ter pogostost pojavljanja besed so odraz nabora analiziranih besedil, zato dosti pričakujemo od nadaljnje rasti korpusa, ki ga bomo po najboljših močeh dopolnjevali. Največje bogastvo in pestrost izrazja pričakujemo v številnih člankih, objavljenih v reviji Knjižnica v zadnjem desetletju, ter magistrskih delih. Žal diplomska dela s področja bibliotekarstva praviloma niso dostopna v elektronski obliki.

Analiza *Korpusa bibliotekarstva* z nekaj več kot 200 slovenskimi strokovnimi bibliotekarskimi besedili oz. 1,6 milijona besedami (kolikor je obsegal v času priprave podatkov za ta prispevek) je povsem potrdila osnovne teoretične predpostavke o besedilnih korpusih. V korpusu zajete besede se lahko razvrsti v tri značilne skupine:

- *zelo pogoste besede*, ki pa ne prispevajo k predstavitvi vsebine dokumentov, mednje sodijo tudi funkcijske besede, ki sodijo v sam vrh pogostosti; v tej skupini je razmeroma malo različnih besed, vendar izrazito izstopajo s svojo veliko pogostostjo (absolutni prvak je pomožni glagol **biti** s 93.896 pojavljanji, sledi mu veznik **in** (61.115), predlog **v** (46.315) itn.; opaziti je zelo strm padec pogostosti povsem v skladu z Zipfovom zakonom[2], zato je petnajsta najpogostejša beseda še zadnja s frkvenco nad deset tisoč ,
- *zelo redke besede*, ki prav tako ne predstavljajo vsebine dokumentov (med njimi so tudi imena), ki se iztekajo v dolg rep besed s pogostostjo ena,
- relativno ozek pas besed v sredini, ki so najpomembnejše nosilke vsebine in v našem primeru morebitni kandidati za vključitev v terminološki slovar.

Med tistimi najpogostejšimi so praviloma besede, ki bi jih pri indeksiranju podatkovne zbirke lahko uvrstili med blokirane besede. Korpus te kategorije ne pozna, ker so lahko za analize besedil, primerjavo avtorjev ipd. zanimive vse besede. V ilustracijo navajam dva primera. Beseda d**okaj** je bila evidentirana sedemkrat, vendar samo v enem besedilu! Beseda **namreč** se pojavlja v določeni vrsti znanstvenih publikacij dvakrat pogosteje od povprečja, v prispevkih za zbornike pogosteje kot v člankih revije Knjižnica in najredkeje v monografskih publikacijah . . . Pogostost v tem primeru ne pomeni absolutne frekvence, temveč delež v odnosu do vseh besed, izražen v promilih.

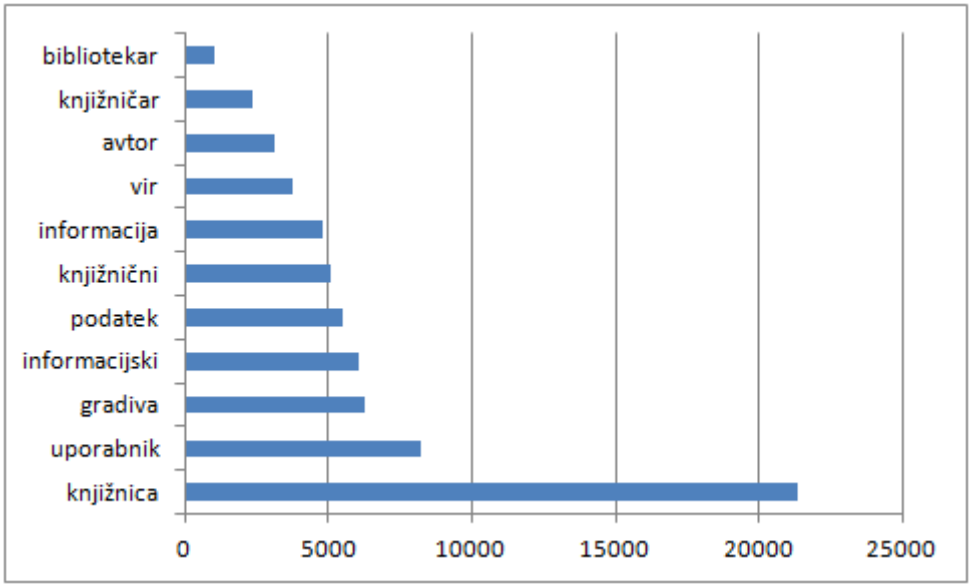
V prvi stotniji najpogostejših besed so na primer:

biti	ona	ter	ves
in	kot	do	le
v	ta	iz	več
za	o	imeti	že
na	pri	med	saj
ki	ali	še	oziroma
da	lahko	svoj	si
tudi	ne	drugi	naj
pa	po	tako	bolj
z	v	kateri	vse
s	od	kar	ko

Med 100 najpogosteje evidentiranimi besedami pa so tudi nekatere za bibliotekarstvo pomembne ključne besede, **knjižnica** je že na 7. mestu, nato si sledijo še gradivo, informacijski, delo, podatek, sistem, uporabnik, tema, področje, knjižničen, informacija, knjiga, vir, zbirka, visokošolski, uporabnikov, znanstven, storitev, program, avtor, revija, raziskovalen, študent, razvoj, rezultat, analiza, dejavnost, vsebina, vprašanje, stran, članek, primer, iskanje, organizacija, dokument, **knjižničar** pa še ravno zaključuje prvo stotnijo najpogostejših. Pri vseh teh je bilo izračunano absolutno pojavljanje vseh oblik besede, ker je bil izveden postopek lematizacije, to je je proces pripisovanja osnovne (slovarske) oblike besednim oblikam v korpusu. Beseda **knjižnica** se npr.

[2] Zipfov zakon temelji na trditvi, da je majhno število besed uporabljeno zelo pogosto, mnogo drugih ali skoraj vse ostale pa zelo poredko.

pojavlja v 21 različnih oblikah (glede na sklon in število, vendar tudi z razlikovanjem velikih in malih črk).



Slika 5: Nekaj najpogostejših bibliotekarskih terminov.

5 Zaključek

Bibliotekarstvo je ena redkih strok, ki lahko za pripravo in dopolnjevanje svojih terminoloških slovarjev že uporablja besedilni korpus, to najsodobnejše in pomembno jezikoslovno in slovaropisno orodje. Slovenski terminološki slovarji, pa tudi večina tujih, še vedno nastajajo brez podpore korpusov, izjeme so le redke in samo potrjujejo pravilo. Čeprav korpus že sedaj zajema širok spekter različnih tipov dokumentov od doktorskih disertacij in znanstvenih člankov do strokovnoinformativnih člankov, ostaja dinamična rast korpusa z dopolnjevanjem že objavljenih in evidentiranih besedil (takšnih je še nad štiristo), predvsem pa najnovejših tekočih objav s širšega področja bibliotekarstva, najpomembnejša naloga ustvarjalcev korpusa. Le tako bo lahko postal korpus za stroko reprezentativen in celovit prikaz rabe strokovnega izrazja v slovenskem bibliotekarstvu. Ker je korpus na spletu javno dostopen, upamo, da bo koristil tudi strokovnim kolegom bibliotekarjem in tudi študentom bibliotekarstva ter morebiti kakšnemu jezikoslovcu, terminologu ali slovaropiscu iz drugih logov. Največ koristi si od njega vsekakor obetamo pri dopolnjevanju *Bibliotekarskega terminološkega slovarja*.

6 Citirani viri in literatura

1. *Bibliotekarska terminologija: Blog*. Dostopno na naslovu: http://terminologija.blogspot.com/
2. *Evrokorpus.* Dostopno na naslovu: http://evrokorpus.gov.si/
3. *FidaPLUS.* Dostopno na naslovu: http://www.fidaplus.net/
4. *Jos.* Dostopno na naslovu: http://nl.ijs.si/jos/
5. *Korpus bibliotekarstva.* Dostopno na naslovu: http://terminologija.blogspot.com/p/korpus.html
6. *Korpus bibliotekarstva: Vključena besedila.* Dostopno na naslovu: http://www.cek.ef.uni-lj.si/terminologija/Korpus/datoteke/seznam_besedil_si.html
7. *Korpus DSI.* Dostopno na naslovu: http://nl2.ijs.si/dsi.html

8. *NEVA - interNet version of EVA*. Dostopno na naslovu: http://www.laze.org/neva/
9. *Nova beseda.* Dostopno na naslovu: http://bos.zrc-sazu.si/s_beseda.html
10. *Oxford English dictionary*. Dostopno na naslovu: http://www.oed.com/
11. *Slovar slovenskega knjižneg ajezika.* Dostopno na naslovu: http://bos.zrc-sazu.si/sskj.html
12. *Vayna.* Dostopno na naslovu: http://nl.ijs.si/et/talks/korpus/vayna-hdr.html
13. *Wikipedia.* Dostopno na naslovu: http://en.wikipedia.org
14. *Wikipedija.* Dostopno na naslovu: http://sl.wikipedia.org