

GIGAFIDA: INTERPRETACIJA KORPUSNIH PODATKOV

¹Nataša Logar, ²Kaja Dobrovoljc, ^{2,3}Špela Arhar Holdt

¹Fakulteta za družbene vede, Ljubljana; ²Trojina, zavod za uporabno slovenistiko, Ljubljana;

³Filozofska fakulteta, Ljubljana

UDK 81'322'324:004.5

V prispevku obravnavamo interpretacijo podatkov, ki jih uporabnik dobi, ko odgovore na jezikovna vprašanja išče prek spletnega konkordančnika korpusa Gigafida. S pomočjo štirih zgledov opozarjamo na nevarnosti prehitrega posploševanja, zlasti pri podatku o pogostosti, filtrih *Vrsta besedila*, *Vir* in *Leto* (osnovno iskanje), ter na nujnost poznavanja značilnosti in natančnosti strojnega pripisovanja oznak (napredno iskanje). V diskusiji izpostavljamo, da je treba – ne glede na vrsto iskanja in iskalni pogoj – pri interpretaciji rezultatov korpusnih poizvedb imeti v mislih tudi vsebino ter zgradbo korpusa.

korpus, interpretacija, konkordančnik, vmesnik, korpusno označevanje

The paper discusses the interpretation of data that users of the Gigafida corpus obtain when trying to solve their language dilemmas through Gigafida's web concordancer. Illustrated by four examples, we emphasize the perils of making hasty generalizations, especially when dealing with frequency information and data presented in the *Text type*, *Source* and *Year* filters (within basic search), as well the need for understanding the characteristics and accuracy of automated corpus annotation (within advanced search). In the discussion, we emphasize that the interpretation of query results should be made by taking the corpus content and structure into account, regardless of the search query or search type.

corpus, interpretation, concordancer, interface, corpus tagging

1 Uvod

1.1

Interpretacija podatkov, naj bo preprosta (npr. primerjava pogostosti) ali kompleksna (npr. primerjava pojavnosti v posameznih delih korpusa v kombinaciji z naslovom in avtorjem ter upoštevajoč korpusno zgradbo), je sestavni del vsake korpusne poizvedbe. Čeprav interpretacije korpusnih podatkov v tem prispevku ne bomo razumeli v statističnometodološkem smislu, je razumevanje korpusa kot vzorca, tj. kot »podmnožice celotne populacije, ki je bila na predviden način (i)zbrana za raziskavo« (Peck, Olsen, Delore 2009: 8), relevantno tudi za nas, še zlasti s poudarkom, da morajo raziskovalci pri posploševanju podatkov na podlagi vzorca upoštevati način, na katerega je bil ta narejen (prav tam: 33).

Konkordančnik SSJ je bil skupaj z vmesnikom, v katerem je dostopen korpus Gigafida, razvit z »mislijo na najširšo možno korpusno rabo« (Logar Berginc idr. 2012: 98; Arhar Holdt, Kosem, Logar Berginc 2012: 20–21; Kosem 2012). S to

mislijo so bili tudi korpusovi metapodatki odmerjeni in v vmesnik umeščeni tako, da uporabnikom karseda olajšajo interpretacijo rezultatov poizvedb.

1.2

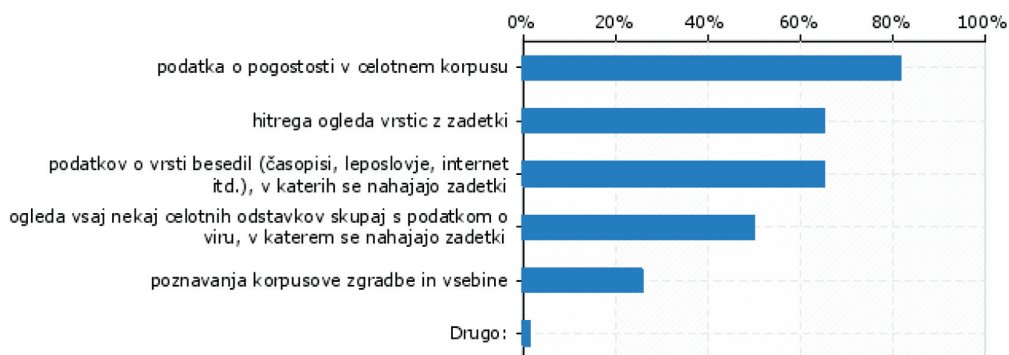
V korpusu Gigafida so za interpretacijo rezultatov pomembni dvoji podatki:

- a) *predočeni podatki* so podatki, ki so prikazani hkrati z rezultati poizvedb, npr. pri osnovnem iskanju pogostost v celotnem korpusu in njegovih posameznih taksonomskih delih (filtri) ter podatki, ki so dostopni s klikom na konkordančno vrstico (sobesedilo, bibliografski podatki in korpusne oznake);
- b) *zaledni podatki* so podatki, ki so povezani z izdelavo korpusa, npr. značilnosti in natančnost označevanja na eni strani ter korpusna zgradba na drugi.

V prispevku bomo opozorili na nekatere izmed tistih delov korpusa, na katere so uporabniki morda premalo pozorni ali v zvezi s katerimi so pri posploševanju rezultatov poizvedb premalo previdni. Izkušnje z jezikovnotehnoloških izobraževanj učiteljev slovenščine (Stritar, Dobrovoljc 2013) npr. kažejo, da uporabniki pri interpretaciji upoštevajo predvsem predočene številske podatke, bistveno manj pa zaledne podatke o korpusu.

2 Anketno izhodišče

Avgusta 2014 smo začeli izvajati anketno raziskavo o tem, kako in zakaj se uporablja korpus Gigafida ter kako ga uporabniki ocenjujejo. K izpolnitvi spletne ankete (izdelali smo jo z orodjem <http://www.lka.si>) smo povabili vse Gigafidine uporabnike, in to bodisi z osebnim stikom (npr. na jezikovnotehnoloških izpopolnjevanjih učiteljev, ki jih je od septembra do novembra 2014 izvajalo Slovensko društvo za jezikovne tehnologije; <http://ucitelji.sdjt.si>), bodisi prek stanovskega foruma (SlovLit in imenik članov že omenjenega društva za jezikovne tehnologije), bodisi z nagovorom na korpusni začetni spletni strani. Do začetka aprila 2015 je na vprašanja odgovorilo 82 anketiranih (od tega 62 v celoti in 20 delno). V anketo smo vključili tudi vprašanje o interpretaciji podatkov, ki je imelo naslednje izhodišče: *Pri interpretaciji podatkov iz korpusa Gigafida večinoma izhajam iz ..., izbrati pa je bilo mogoče enega ali več zaključkov te povedi* (Slika 1).



Slika 1: Deleži odgovorov na anketno vprašanje, iz česa uporabnik izhaja pri interpretaciji podatkov iz korpusa Gigafida

Vseh klikov na prvih pet zaključkov povedi je bilo 190 (pri čemer smo odšteli odgovor »drugo«, ki je bil izbran le enkrat), kar pomeni, da je največji delež posameznih odgovorov sicer res pripadal »pogostosti«, dejansko pa je bila zmagovalna kombinacija »pogostost (82 %) – konkordance (65 %) – filter *Vrsta besedila* (65 %)«. Na četrtem mestu je bil odgovor »ogled odstavkov in bibliografskih podatkov«, ki ga je v primerjavi z odgovoroma »konkordance« oz. »filter« izbralo 15 % manj anketiranih. Samo 17 anketiranih (26 %) je povedalo, da pri interpretaciji podatkov izhajajo tudi iz poznavanja korpusove vsebine in zgradbe.¹ Ker so ključne informacije o vsebini in zgradbi Gigafide dostopne tudi na korpusni spletni strani, je mogoče sklepati, da jih vsaj nekateri uporabniki poznajo, očitno pa imajo pri interpretaciji rezultatov omejen vpliv.

3 Podatki, pri katerih sta potrebni pozornost in previdnost

V tem poglavju bomo na štirih tipološko različnih primerih prikazali, kako podatke, ki spremljajo rezultate poizvedb po Gigafidi oz. nanje vplivajo, ustrezno uporabiti pri interpretaciji rezultatov.

3.1 Analiza primerov

3.1.1 Osnovno iskanje: pogostost, filtri *Vrsta besedila*, *Vir*, *Leto* ter odstavek in bibliografski podatki

Za premislek o interpretaciji podatkov, vidnih ob konkordančnih vrsticah v osnovnem iskanju, smo izbrali primera *po navadi* : *ponavadi* in *prekaren* : *prekeren*.

¹ Ob tem je pomemben podatek, da za 59 % anketiranih velja, da se pri svojem delu ukvarjajo z jezikom.

a) *Po navadi* : *ponavadi*

1 2 3 4 5 6 7 8 9 10

Prikazujem 1-20 od 16.550 konkordanc (0.0 sekund).

Osnovne oblike

► [po navada \(16.546\)](#)

► [po navaditi \(4\)](#)

Vrsta besedila

► [Magazines \(6.452\)](#)

► [Newspapers \(4.324\)](#)

► [Internet \(2.885\)](#)

► [Non-fiction \(2.239\)](#)

► [Fiction \(585\)](#)

► [Več](#)

dojki. Bruha tja do 4. meseca, nato se **po navadi** slabost umiri, seveda pa se

Potem **po navadi** rečem: 'No ja, res sem že

Sila moderni so tudi animatorji na rojstnih dnevih, ki **po navadi** uro ali dve zabavajo celo gr

pa odkrita. Največ je nevretenčarjev, pri katerih lahko **po navadi** novo vrsto ugotovijo samo z

. Če obstajajo ali jih človek ustvari, bodo možnosti **po navadi** tudi uporabljene.

potem zavijemo nekam drugam, ampak so nam ti razlogi **po navadi** zelo znani. Izziv in tisto, kar

že dalj časa, pri mednarodni skupnosti pa tovrstno početje **po navadi** ne pade na plodna tla.

Slika 2: Del konkordančnega niza prislova *po navadi* s pogostostjo in filtrom *Vrsta besedila* v korpusu Gigafida

1 2 3 4 5 6 7 8 9 10

Prikazujem 1-20 od 76.353 konkordanc (0.421 sekund).

Osnovne oblike

► [ponavadi \(76.353\)](#)

Vrsta besedila

► [Časopisi \(35.186\)](#)

► [Revije \(23.440\)](#)

► [Internet \(9.213\)](#)

► [Stvarna besedila \(5.521\)](#)

► [Leposlovje \(2.514\)](#)

► [Več](#)

spodbudilo in šparglji so se pojavili mesec dni prej kot **ponavadi**. Loredana Ivančič jih je na tržnici

jih ni bilo. Zlorabo so opazili seveda prepozno, **ponavadi** čez nekaj tednov, ko niso prejeli p

nekaj zveni predobro, da bi bilo res, potem **ponavadi** tudi ni. Niti en normalen in pošten

jih bomo priklopili. Namenski projektorji za domači kino imajo **ponavadi** več video vvhodov in posebna vide

programa prepuščena trenutni inovativnosti nekakšnega kadrovskega konglomerata, k **ponavadi** predstavlja kot novinarski sindika

po zaslugi fundacije Lilly bogatejša za 105 milijonov dolarjev. **Ponavadi** so posamezniki darežljivejši kot p

Slika 3: Del konkordančnega niza prislova *ponavadi* s pogostostjo in filtrom *Vrsta besedila* v korpusu Gigafida

Slik 2 in 3 je razvidno, da se je prislov *po()**navadi* v rabi uveljavil zapisan skupaj, saj je v celotni Gigafidi skoraj petkrat pogostejši od svoje dvojnice.² Ogled filtra *Vrsta besedila* pove, da je največ pojavitev zapisa *po navadi* v revijah, medtem ko je zapis *ponavadi* najbolj pogost v časopisih, v revijah pa ga je približno 11.000 zadetkov manj. Iz tega lahko sledi sklep, da so novinarji in lektorji periodičnega tiska tisti, ki ne poznajo ali ne upoštevajo pravopisnega določila o edino pravilnem zapisu narazen (SP 2001: § 571; gl. tudi Dobrovoljc, Jakop 2011: 113), vendar pa ta sklep drži le delno. Če upoštevamo zgradbo korpusa in primerjamo število enega in drugega zapisa na milijon besed v posameznem delu Gigafide, je namreč največji razmik v razmerju med zapisom narazen in zapisom skupaj res pri časopisih (1 : 8), že na drugem mestu pa sledi leposlovje (1 : 4), kjer je tudi sicer ta prislov enkrat pogostejši kot v dnevnem tisku (še nekoliko pogostejši pa je v stvarnih besedilih). Zgolj na podlagi absolutnega števila pojavitev v filtru *Vrsta besedila* lahko torej uporabnik, ki pri interpretaciji ne upošteva zgradbe korpusa, pride do napačnega sklepa.

² Za večino uporabnikov je to verjetno že zadosten podatek za interpretacijo v smislu »prednost ima zapis skupaj, zato bom uporabil tega«.

b) *Prekaren* : *prekeren*

Pridevnik *prekaren/prekeren* se je v rabi razširil v času zadnje gospodarske krize (del slovenščine pa je bil sicer že prej). Uporabniku Gigafide, ki želi pomoč pri izbiri zapisa, sam podatek o pogostosti odločitve ne bo olajšal; *prekaren* se namreč v korpusu pojavi 66-krat, *prekeren* pa 62-krat. Pri izbiri se moramo torej opreti na druge razpoložljive informacije. Iz konkordančnih vrstic je mogoče razbrati, da imata *prekaren* in *prekeren* enak pomen ter enako besedilno okolje (*delavec, zaposlitev, delo* ipd.). Iz filtrov izvemo (Slika 4), da se po različnih besedilih pojavljata v zelo podobnih obsegih in pogosteje po letu 2008. Pri primerjavi filtrov se razkrije upad v letu 2011, pri razumevanju katerega je zopet treba upoštevati zgradbo korpusa. Upad je namreč pričakovan, saj Gigafida iz tega obdobja vsebuje le besedila iz tiska, zbrana do maja 2010, in besedila z interneta, zbrana od aprila 2010 do aprila 2011 (Logar Berginc idr. 2012: 26, 46, 47).

Osnovne oblike ▶ prekaren (66) Vrsta besedila ▶ Internet (35) ▶ Časopisi (23) ▶ Stvarna besedila (5) ▶ Revije (3) ▶ Več Vir ▶ Internet, ustanove (16) ▶ Dnevnik (13) ▶ drugo (11) ▶ rtvslo.si (7) ▶ siol.net (5) ▶ Več Leto ▶ 2010 (33) ▶ 2009 (12) ▶ 2011 (8) ▶ 2008 (7) ▶ 2007 (2) ▶ Več	Osnovne oblike ▶ prekeren (61) ▶ prekerena (1) Vrsta besedila ▶ Internet (31) ▶ Časopisi (25) ▶ Revije (4) ▶ Stvarna besedila (2) ▶ Več Vir ▶ Dnevnik (15) ▶ drugo (12) ▶ dz-rs.si (7) ▶ rtvslo.si (7) ▶ Internet, novice (6) ▶ Več Leto ▶ 2010 (27) ▶ 2008 (13) ▶ 2011 (9) ▶ 2009 (4) ▶ 2007 (4)
---	--

Slika 4: Podatki v filtrih *Vrsta besedila*, *Vir*, *Leto* ob konkordančnemu nizu pridevnikov *prekaren* in *prekeren* v korpusu Gigafida

Tudi ogled filtrov torej ni izkazal nobene besedilne prvine, ki bi dala prednost enemu od zapisov. Ker pa gre za vrstni pridevnik in kot tak del strokovnih poimenovanj, si je smiselno ogledati še odstavke ter njihove vire v podkorpusu Stvarna besedila. Tak pogled lahko pokaže dodatne relevantne informacije, podane s strani strokovno podkovanega avtorja (Slika 5).

Prikaz odstavka	Korpusne oznake
Izraz "prekarizacija" se s stališča delavk in delavcev nanaša na tisto, čemur se iz perspektive kapitala in njegovih zastopnikov pravi "fleksibilizacija trga delovne sile, fleksibilnost ali prožnost dela". "Prekarizacija" je novejša časnikarska izpeljanka iz pridevnika "prekaren, prekeren"; ta je prišel iz latinskega pridevnika <i>precarius</i> (prvi pomen je specifičen: "izprošen, priberačen, po milosti dobljen"; izpeljani pomen je posplošen: "negotov, nestalen"); iz ablativa je nastal prislov <i>precario</i>, ki ima tudi dva pomena specifičnega "po milosti, na prošnjo" in posplošenega v tehničnem pravnem besedišču "do preklica". Izraz je v moderne jezike najprej prišel prav v tehničnem pravnem pomenu (1336) in je šele v 16. stoletju pridobil splošni pomen, ki je zdaj v rabi. Verbinčev Slovar tujk pod geslom "prekeren" daje tako pravno tehnični kakor posplošeni pomen: "izprošen, na prošnjo ali do preklica dovoljen; odvisen od tuje volje ali okoliščin". (Verbinc prinaša tudi pravni izraz "prekarij: začasna uporaba ali užitek tuje lastnine".) Samostalniki "prekarnost" ("preklicnost; negotovost, nestalnost, tveganost, dvomljivost, kočljivost") je nastal šele v 19. stoletju. Latinski pridevnik je v zvezi z	vir: drugo leto nastanka: 2006 vrsta besedila: Stvarna besedila naslov: Svetovno gospodarstvo in revolucionarna politika avtor: Rastko Močnik

Slika 5: Del odstavka ene od konkordančnih vrstic pridevnika *prekeren* s podatki o viru v korpusu Gigafida

Predstavitev tega primera lahko zaključimo z ugotovitvijo, da tudi poglobljena analiza potrjuje izhodiščno stanje: uporabnik se lahko o varianti obravnavanega pridevnika odloči sam – oz. se mora odločiti sam. Korpusna analiza pri odločitvi torej ni pomagala neposredno, je pa omogočila uporabniku pridobiti dovolj jezikovnega in vsebinskega znanja, da lahko svojo izbiro interpretira in utemelji kot ustrezno.

3.1.2 Napredno iskanje: podatek o osnovni obliki, besedni vrsti in oblikoskladenjskih lastnostih

Pri naprednem iskanju po korpusnih besedilih uporabniki oblikujejo kompleksnejše iskalne pogoje, v katerih lahko poleg oblike določijo tudi besedno vrsto in druge oblikoslovne lastnosti iskane besede ali besede v njeni okolici. Za oblikovanje naprednih poizvedb in njihovo ustrezno interpretacijo je zato pomembno poznavanje značilnosti in natančnosti strojnega pripisovanja tovrstnih zalednih podatkov, kar v nadaljevanju ponazarjamo na primerih iskanja pretvorb frazema *dati besedo* in primerniških oblik prislova *težko*.

a) *Dati besedo: dana beseda: dajanje besede*

Slika 6 prikazuje del konkordančnega niza za frazem *dati besedo*, ki smo ga v naprednem iskanju opredelili kot kombinacijo glagola *dati*, ob katerem se do tri besede levo ali desno pojavlja samostalnik *beseda* v tožilniku ednine. Vidimo lahko, da je poizvedba vrnila nize različnih glagolskih oblik z različnimi oblikoskladenjskimi lastnostmi, ki jim je bila v skladu s specifikacijami, razvitimi v okviru projekta Jezikoslovno označevanje slovenščine (JOS; Erjavec, Krek 2008), vsem pripisana besedna vrsta *glagol* in nedoločniška lema *dati*.

, ki se od drugih zliva v vaš sistem. Dati in držati besedo , ne le drugim, temveč tudi
k njegovemu zadnjemu podvigu in tako preusmeri pozornost nanj. Dala si mu besedo in mu priskrbela poslušalce, torej si
najbolj ugodno vplivati na rezultate. Upoštevajte sogovornika enakovredno, dajte mu prednostno besedo , vi pa potem dodajte svoje mnenje
kratka replika? Potem izvolite, ker bi gospodu Selihoviču dala potem še besedo . Izvolite.
: V redu. Najlepša hvala. Saj vam bom dal še preostali besedo . Mogoče samo neko pojasnilo vas,
VIZJAK: Hvala lepa. Seveda bom, ampak bi dal še gospe besedo pa potem. Ampak bi res prosil

Slika 6: Del konkordančnega niza za napredno iskanje sopojavitev lem *dati* in *beseda* v korpusu Gigafida

Ena izmed značilnosti glagolskih frazemov je njihova možnost t. i. celovitih pretvorb (Kržišnik 1994: 95–100), v katerih se skladenjska zgradba frazema spremeni zaradi prehoda glagola v drugo oblikoskladenjsko kategorijo, denimo samostalnik (posamostaljenje, npr. *dajanje besede*) ali samostojni stavek (potrpnjenje, npr. *beseda je dana*), pri čemer korpusne analize potrjujejo, da ta možnost ni sistemska, temveč odvisna od posamezne frazeološke enote (Gantar 2007: 126–127, 237–240). Pri priklicu tovrstnih pretvorb v korpusu Gigafida je treba upoštevati, da so v označevalnem sistemu JOS trpni deležniki na -n/-t označeni kot deležniški pridevniki, glagolniki pa kot samostalniki, kar se ne sklada nujno s kategorizacijami v drugih priročnikih (prim. npr. *Slovenski pravopis* 2001) ali uporabnikovo intuicijo.

To pomeni, da moramo za celostno analizo pretvorbenih možnosti frazema *dati besedo* iskanje ponoviti še za kombinacijo samostalnika *beseda* s pridevnikom *dan* in samostalnikom *dajanje*. Analiza rezultatov razkrije, da besedna zveza v obeh pomenih ('zagotoviti' in 'dati možnost govorjenja') zelo redko izkazuje možnost potrpnjenja (*premierjeva beseda ni bila dana iskreno; končna beseda bo dana volilcem*), medtem ko je pretvorba v pridevniško zvezo (*dana beseda*) precej pogosta, a značilna samo za prvi pomen, posamostaljenje (*dajanje besede*) pa samo za drugega (Slika 7).

glede vodenja seje, glede sklicevanja seje, glede načina dajanja besede in podobno, polno avtoriteto in je poslanec neprijetno
V razpravi je opozicija možnost »arbitriranja« pri dajanju besede poslancem označila za pokop demokracije v Sloveniji, predstavniki
naj bi iz predlaganih sprememb poslovnika črtali sporno določbo o dajanju besede poslancem za replike. V obrazložitev so zapisali,
odgovarjanje prevladuje informativni metarazgovor. Usmerjanje k cilju razgovora, dajanje besede , priklic k razgovoru so govorna dejanja, ki
je prosila, naj opazujem, koliko je dosledna pri dajanju besede samo tistim učencem, ki so dvigovali roko.
slabe. Kar v revijalnem tisku manjka, je več dajanja besede praktikom filma in njihovim izkušnjam, njihovim znanjem in
, spoštovani predsednik. Vi veste kakšen je vrsti red dajanja besede - razprava, replika.

Slika 7: Del konkordančnega niza za napredno iskanje sopojavitev lem *dajanje* in *beseda* v korpusu Gigafida

Pomembno je torej zavedanje, da so bile metajezikovne lastnosti pojavnica v korpusu Gigafida pripisane v skladu s specifičnim (in podrobno dokumentiranim – gl. Holozan idr. 2008; <http://nl.ijs.si/jos>; <http://www.slovenscina.eu/sloleks>) sistemom jezikoslovnega označevanja, ki lahko zaradi omejitev strojnih postopkov na nekaterih

mestih odstopa od bolj ali manj poenotenih kategorizacijskih in lematizacijskih načel drugih obstoječih jezikoslovnih opisov slovenščine.³

b) *Teže* : *težje*

Poleg raziskav pomenskih, skladenjskih ali pravopisnih lastnosti leksike korpusi omogočajo tudi analize njegovih oblikoslovnih značilnosti, kot so vzorci pregibanja ali tvorjenja besed. Ena izmed takih oblikoslovnih zadreg je tudi variantnost v obrazilnem stopnjevanju lastnostnih prislovov, ki besedotvorno izvirajo iz pridevnika (Toporišič 2004: 409).

Variantnost obrazilnega stopnjevanja, do katere se obstoječi referenčni priročniki opredeljujejo precej različno,⁴ se v jezikovni rabi npr. pojavlja pri primerniških oblikah prislova *težko* (*teže* in *težje*). Ločeni poizvedbi z naprednim iskanjem po tistih pojavitvah obeh oblik, ki so v korpusu označene kot prislov, kažeta, da je pogostnostno razmerje med njima 1 : 4 v prid obliki z *j*. Toda podrobnejši pregled konkordanc razkriva, da se med primeri rabe prislovne oblike *teže* prikazujejo tudi oblike samostalnika *teža* (Slika 8), med prvimi 20 primeri rabe prislovne oblike *težje* pa je kar 6 primernikov pridevnika *težko* (Slika 9). Ugotovitev, da sta bili v procesu avtomatskega označevanja tem oblikam pripisani napačna besedna vrsta in osnovna oblika, torej narekuje, da pri morebitnih kvantitativnih analizah upoštevamo tudi oceno zanesljivosti strojno pripisanih podatkov. Te pa spet ni mogoče posploševati na celoten korpus, saj je natančnost označevanja močno odvisna od vrste leksike (več gl. v Grčar idr. 2012).

◀ prejšnja stran	1 2 3 4 5 6 7 8 9 10 11 12	naslednja stran ▶
Prikazujem 41-60 od 7.119 konkordanc (0.437 sekund).		
spet slovesno. V času, ko je športnikom vse teže dobiti pomoč za treninge in tekmovanja, je mladi radovljiški		
na prizorišče prav nič miroljubno, zato je bilo toliko teže vsem drugim demonstrantom, ki niso prišli uničevati mesta.		
izumrl, a se še vedno dogaja, da ženske teže dobijo službo kot moški, ker bodo mogoče izkoristile porodniški		
državnih interesov. Sodobna evropska regionalna politika daje nekaj več teže neposredno ljudem, kolikor se jih pač lahko zajame skozi		
Ashdown vse teže kroti nasprotja		

Slika 8: Del konkordančnega niza za oblike *teže*, označene kot primernik prislova *težko*, v korpusu Gigafida

³ Posebnosti se denimo pojavljajo pri označevanju posamostaljenih pridevnikov (pridevnik), deležij (prislov), elativov (samostojna lema s stopnjo nedoločeno), nekaterih skupin slovnčnih besednih vrst ipd. (več gl., kot že omenjeno, v Holozan idr. 2008).

⁴ Medtem ko SSKJ, v katerem o priporočeni stopnjevanju obliki izpridevniških prislovov sklepamo na podlagi zgledov, daje prednost oblikam z *j* (*nižje*, *višje*, *tišje*), *Slovenska slovnica* (Toporišič 2004: 409) v zbornem jeziku priporoča rabo oblik brez *š* oz. *j*, v knjižnem in neknjižnem pogovornem jeziku pa obratno. Ta ločnica je precej bolj zrelativizirana v *Slovenskem pravopisu* 2001 (§ 990), ki opisuje, da so prislovi iz primernikov enaki obliki za srednji spol; če pa se ti končajo na *-jše*, *-šje*, *-žje* in *-čje*, se zadnji soglasnik navadno lahko tudi izpušča (npr. *ljubeznivejšelljubezniveje*). Kljub načelni strpnosti pravopisnih pravil do obeh obrazilnih variant pa spremljajoči pravopisni slovar pri nekaterih iztočnicah priporoča zgolj eno izmed možnosti, npr. *jaje* pri iztočnici *jako*, *ože* pri iztočnici *ozko* in *teže* pri iztočnici *težko*.

1	2	3	4	5	6	7	8	9	10	naslednja stran ►
Prikazujem 1-20 od 29.456 konkordanc (0.156 sekund).										
krize. Vzrokov za njen padec je več: čedalje težje je pridobivala svež denar na posojilnih trgih, zaskrbljeni komitenti										
je še odprto. V zadnjih sedmih dirkah bo vse težje za vsakega dirkača,« meni Häkkinen.Sporno enoto										
tudi njegovo življenje ni (bilo) lahko. Še težje se zdijo preizkušnje drugih filmskih likov: vidimo jih lahko										
na položaju in omogoča dovolj izkušenj in moči tudi za težje izzive.« Ne spomnimo se, da bi kdaj										
programu imagemap. S starejšimi različicami programa imagemap je bilo težje delati in so zahtevale posebno konfiguracijsko datoteko. Program,										
bom takrat privoščil Potrčev golaž, mlajše primerke bo namreč težje ujet..										
. A veter me ne skrbi, saj sodim med težje kolesarje, lažje bo še precej bolj oviral,"										

Slika 9: Del konkordančnega niza za oblike *težje*, označene kot primernik prislova *težko*, v korpusu Gigafida

4 Razprava

Konkordančnik SSJ in vmesnik Gigafida sta v primerjavi s svojim predhodnikom ASP32 uporabniško prijaznejša (Kosem 2012; Logar Berginc idr. 2012: 98–118). Med njunimi novostmi je opazna zlasti uvedba filtrov, ki omogočajo hkraten prikaz dveh vrst informacij: zadetkov v konkordančnem nizu, besednem seznamu ali seznamu kolokatorjev ter obenem informacijo o razpršenosti po delih Gigafide.⁵ Čeprav je bil primarni namen filtra *Vrsta besedila* pravzaprav omogočanje enostavnega selekcioniranja zadetkov za nadaljnji podrobnejši pregled, se je njegova druga vloga, tj. prikaz pogostosti po podkorpusih, pri interpretaciji izkazala celo za bolj pomembno. V nadaljevanju bi bilo zato smiselno razmisliti o nadgradnji tega filtra s podatkom o relativni pogostosti, torej pogostosti glede na število vseh besed v določenem podkorpusu.

Za pravilno interpretacijo korpusnih podatkov je zelo pomembno tudi to, da se ne zadržimo le pri številskih delih rezultatov. Že preprost primer *prekaren/prekeren* pokaže, da lahko pomembne podatke (morda za odločitev v jezikovni zadregi celo najbolj pomembne podatke) o leksiki dobimo šele s pregledom širšega sobesedila. V zvezi s sklepanjem zgolj iz podatka o pogostosti v celotnem korpusu pa je treba dodati še opozorilo o podvojenih konkordancah, t. i. dvojnikih: detekcija dvojnikov in približnih dvojnikov je bila pri gradnji Gigafide temeljito izvedena le pri spletnih besedilih (Logar Berginc idr. 2012: 64–65), šele naknadne analize po zaključku gradnje pa so pokazale, da bi morali isti postopek uporabiti tudi drugje. Podvojeni rezultati (enako besedilo, čeprav iz različnih virov) zlasti specifičnih iskanj so v konkordančnih vrsticah sicer hitro opazni, vseeno pa bi bilo treba v prihodnje, kolikor je mogoče, izločiti ta moteči in za interpretacijo potencialno škodljivi korpusni šum.

Če navedena opozorila veljajo za veliko večino uporabnikov Gigafide (po podatkih ankete skoraj 90 % anketiranih uporablja osnovno iskanje), pa morajo biti na pozna-

⁵ Drugih filtrov, ki ne izhajajo iz podatkov v kolofonu korpusnih dokumentov, npr. filter *Zapis oblike*, v prispevku nismo obravnavali.

vanje značilnosti ter natančnosti strojnega pripisovanja zalednih podatkov, kot so besedna vrsta in oblikoslovne lastnosti, pozorni predvsem tisti, ki uporabljajo napredno iskanje (med anketiranimi je takih skoraj 60 %). Natančnost označevalnika Obeliks (Grčar idr. 2012) je sicer visoka, v povprečju na vsakih sto besed le dvema pripiše napačno besedno vrsto ali osnovno obliko (prav tam: 92), vendar je ob tem treba upoštevati, da so strojno pripisane osnovne oblike, besednovrstne kategorije in oblikoskladenjske lastnosti pri nekaterih skupinah besed manj zanesljive kot pri drugih. To sicer ne pomeni, da njihova kvantitativna korpusna analiza ni smiselna ali mogoča, mora pa uporabnik to korpusno inherentno pomanjkljivost pri krajših rezultatih poizvedb zaobiti z ročnim pregledom konkordanc, pri daljših pa z dodatnimi kontekstualnimi omejitvami iskanj ali poizvedbami brez opiranja na strojno pripisane podatke.

V korpusnem jezikoslovju se pogosto opozarja na nevarnosti prehitrega sklepanja o jeziku zgolj na podlagi bežnih poizvedb po korpusu (npr. O’Keeffe, McCarthy 2010: 382), na drugi strani korpusi kot orodje prinašajo določene omejitve, pri čemer ni nepomembno, da so referenčni korpusi primarno leksikografski vir in zato za iskanje odgovorov na jezikovne zadrege, ki jih sicer rešujejo ali pa bi ji morali reševati drugi jezikovi viri, niso posebej prilagojeni. A kot velja za podatkovne zbirke, statistične vzorce ali jezikovne vire na splošno, je omejitve mogoče preseči, le dobro je treba vedeti, katere so in od kod izvirajo.

Literatura

- ARHAR HOLDT, Špela, KOSEM, Iztok, LOGAR, Nataša, 2012: Izdelava korpusa Gigafida in njegovega spletnega vmesnika. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 16–21.
- DOBROVOLJC, Helena, JAKOP, Nataša, 2011: *Sodobni pravopisni priročnik: med normo in predpisom*. Ljubljana: Založba ZRC, ZRC SAZU.
- ERJAVEC, Tomaž, KREK, Simon, 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 49–53.
- GANTAR, Polona, 2007: *Stalne besedne zveze v slovenščini: korpusni pristop*. Ljubljana: Založba ZRC, ZRC SAZU.
- Gigafida: <http://www.gigafida.net>
- GRČAR, Miha, KREK, Simon, DOBROVOLJC, Kaja, 2012: Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 89–94.
- HOLOZAN, Peter, KREK, Simon, PIVEC, Matej, RIGAČ, Simon, ROZMAN, Simon, VELUŠČEK, Aleš, 2008: *Specifikacije za učni korpus – projekt »Sporazumevanje v slovenskem jeziku«: kazalnik 2*. http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik2/SSJ_Kazalnik_2_Specifikacije-ucni-korpus_v1.pdf
- KOSEM, Iztok, 2012: User-Friendly Interfaces for Corpora of Slovene. *Prace Filologiczne* 63. 167–180.
- KRŽIŠNIK, Erika, 1994: *Slovenski glagolski frazemi (ob primeru frazemov govorjenja)*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.
- LOGAR BERGINČ, Nataša, GRČAR, Miha, BRAKUS, Marko, ERJAVEC, Tomaž, ARHAR HOLDT, Špela, KREK, Simon, 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko, Fakulteta za družbene vede.

- O'KEEFFE, Anne, MCCARTHY, Michael (ur.), 2010: *The Routledge Handbook of Corpus Linguistics*. New York: Routledge.
- PECK, Roxy, OLSEN, Chris, DEVORE, Jay L., 2009: *Introduction to Statistics and Data Analysis*. Belmont: Cengage Learning.
- Slovar slovenskega knjižnega jezika*, 1970–1991. Ljubljana: DZS, SAZU.
- Slovenski pravopis*, 2001. Ljubljana: Založba ZRC, ZRC SAZU.
- STRITAR, Mojca, DOBROVOLJC, Kaja, 2013: Korpusi na poti v šole: jezikovnotehnološko izpopolnjevanje učiteljev. *Slovenščina 2.0* 1/1. 181–194.
- TOPORIŠIČ, Jože, 2004: *Slovenska slovnica*. Maribor: Založba Obzorja.