GIGAFIDA IN SLWAC: TEMATSKA PRIMERJAVA

Nataša LOGAR BERGINC
Univerza v Ljubljani, Fakulteta za družbene vede

Nikola LJUBEŠIĆ
Univerza v Zagrebu, Filozofska fakulteta, Odsek za informacijske in komunikacijske znanosti

Logar Berginc, N., Ljubešić, N. (2013): Gigafida in slWaC: tematska primerjava. Slovenščina 2.0, 1 (1): 78–110.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_05.pdf.

V prispevku analiziramo dvoje: (a) vključevanje besedil z interneta v obstoječe referenčne korpuse, ki ga soočamo z obstojem spletnih korpusov, ter (b) dva najnovejša korpusa slovenščine: korpus Gigafida, ki ga pretežno sestavljajo tiskana besedila, v manjši meri pa tudi spletna, in korpus slWaC, ki je v celoti sestavljen iz spletnih besedil. Najprej ugotavljamo podobnosti in razlike med njima z metodo tematskega modeliranja, nato pa isto metodo apliciramo še na posamezne taksonomske kategorije Gigafide. Prvi del analize je pokazal, da je ravnanje sestavljalcev referenčnih korpusov v zvezi z vključevanjem internetnih besedil v korpuse, ki naj bi kazali celovito podobo nekega jezika, trenutno še neenotno, če pa se zanj že odločijo, je nabor vključenih žanrov praviloma širok. Drugi del analize je pokazal dokajšnjo tematsko različnost Gigafide in slWaCa ter izpostavil najznačilnejše teme, ki jih pokriva vsak od šestih Gigafidinih delov.

Ključne besede: slovenščina, referenčni korpus, spletni korpus, tematsko modeliranje

1 UVOD

Vsaki gradnji korpusa sledi analiza, ki šele zares pokaže, kaj korpus vsebuje in kje je pomanjkljiv v svoji "težnji po reprezentativnosti" (Biber 1993: 256). Četudi se oblikovalci referenčnih korpusov že nekaj časa zavedajo, da je reprezentiranje jezika – ali le dela jezika – problematična naloga (prim. npr.

Kilgarriff, Grefenstette 2003: 340–343; Kupietz in dr. 2010: 1849), je natančnejše razpoznavanje vsebine vsakega korpusa takojšnja naslednja naloga po zaključku njegove (trenutne) gradnje. Med glavna merila, ki uravnavajo sestavo korpusov, sodijo besedilne zvrsti in vrste, značilnosti tvorca ter naslovnika, (ne)fikcijskost vsebine ipd., pa tudi prenosnik in besedilna tema oz. predmetno področje, ki se jima bomo posvetili v tem prispevku.

1.1 Prenosnik

Pred množično uporabo interneta je bil prenosnik (angl. *channel*, *medium*) besedil, vključenih v referenčne korpuse, dveh vrst: pisni (angl. *written*) in govorni (angl. *spoken*), pri čemer je bilo pod "pisno" (ki nas bo tu edino zanimalo) razumljeno predvsem tiskano. V zadnjem desetletju je prišlo do preobrata: "tradicionalnemu" pisnemu prenosniku – tisku – se je v javnih sporočanjskih položajih kot vsakodnevni način prenosa sporočil pridružil še elektronski, pri čemer raziskave potrjujejo, da postaja podajanje pisnega jezika v javni rabi (pa tudi zasebni) celo vse manj domena tiska in vse bolj domena elektronskih medijev.

- V Ameriki je po podatkih raziskave Pew Research Centra, ki so jo izvedli leta 2010, na vprašanje, kje vse so včeraj dostopali do novic, največ respondentov navedlo televizijo (58 %), digitalni viri, kot so splet, e-pošta, mobilni telefoni in socialna omrežja, so prišli na drugo mesto (44 %), prek radia je novice spremljalo 34 % vprašanih, na zadnjem mestu pa so bili tiskani časopisi (26 %).
- V Sloveniji je po podatkih Statističnega urada RS (5. 10. 2012) v prvem četrtletju leta 2012 internet redno uporabljalo 70 % oseb v starosti 10–74 let. Pri tem je največ uporabnikov (58 %) internet uporabljalo za pošiljanje in prejemanje e-pošte ter v enakem obsegu za iskanje informacij o blagu in storitvah, 50 % za iskanje informacij, povezanih z zdravjem, 46 % za branje in prenašanje spletnih novic, časopisov ali revij, 46 % za branje spletnih forumov, 45 % za pridobivanje znanja s pomočjo spletnih enciklopedij, 30 % za storitve, povezane s potovanji in nastanitvijo, ter 29 % za prodajo blaga ali storitev.

1.2 Besedilna tema oz. predmetno področje

Členjenost besedilnih tem oz. predmetnih področij (angl. *topic*, *domain*, *subject area*, *subject field*) je v korpusih zelo različna. Tako je bilo npr. v korpusu *Brown* (1964) reportažno časopisje členjeno na politiko, šport, družbo, pomembne novice, finance in kulturo, imaginarna proza pa dalje še na detektivsko, znanstvenofantastično, pustolovsko, ljubezensko in humoristično (prim. Gorjanc 2005: 16–17); v *Českem nacionalnem korpusu SYN2010*[1] so strokovna besedila členjena na religijo, pravo, umetnost, ekonomijo, tehnologijo, naravoslovje, humanistiko in življenjske stile; v *Hrvaškem nacionalnem korpusu*[2] je ista vrsta besedil (strokovna besedila) ločena na šport, politiko, ekologijo, bioetiko itd.; v *Britanskem nacionalnem korpusu*[3] pod *informativno* najdemo med drugim svetovno politiko, trgovino in finance, umetnost, religijo in filozofijo ter prosti čas.

Tematska oz. področna opredelitev besedil je sicer lahko vodilo zgolj pri zbiranju besedil, v taksonomijo korpusa oz. v kolofon korpusnih dokumentov pa nato ni vključena ali pa – nasprotno – jo najdemo tako med merili za zbiranje besedil kot med metapodatki. Izhodiščna, čeprav ne v celoti uresničena tematska oz. področna členitev je tako npr. značilna za referenčni korpus *Oxford English Corpus*,[4] ki ga sestavlja dvajset delov, pretežno poimenovanih po temi oz. področju (npr. računalništvo, okolje, prosti čas, vojska, transport).[5] Ti deli so nadalje razdeljeni še na podteme oz. podpodročja (tako jih ima npr. šport kar okrog štirideset).

1.3 Raziskovalna vprašanja in vrste analiz

Najprej nas bo zanimalo, kako so se glede vključevanja besedil z interneta

[1] http://ucnk.ff.cuni.cz/english/syn2010.php

[2] http://hnk.ffzg.hr/struktura.html

[3] http://www.natcorp.ox.ac.uk/

[4] http://oxforddictionaries.com/words/the-oec-composition-and-structure

[5] Vendarle pa zunaj tematske členitve ostaja polovica korpusa, in sicer nerazvrščena besedila (angl. *unclassified*, 17,1 %), blogi (angl. *blogs*, 8,2 %) in novice (angl. *news*, 24,4 %).

odločali sestavljalci aktualnih referenčnih korpusov. Pred pregledom predpostavljamo, da so tej vključitvi večinsko naklonjeni, da pa delež besedil z interneta ohranjajo pod 50 %. Temu prikazu bomo dodali kratek opis stanja na področju spletnih korpusov. Nato bomo prešli k osrednji analizi, v kateri nas bosta zanimala dva korpusa sodobne slovenščine, s poudarkom na prvem: korpus Gigafida,[6] ki ga v 84 % sestavljajo tiskana besedila, preostalih 16 % pojavnic pa vanj prinašajo spletna besedila, in korpus slWaC,[7] ki je v celoti sestavljen iz spletnih besedil – ogledali si ju bomo prek rezultatov metode tematskega modeliranja (angl. *topic modeling method*; Blei in dr. 2003; Sharoff 2010). Metodo bomo aplicirali na dva načina: najprej bomo naredili tematsko primerjavo med Gigafido in slWaCom, nato pa bomo metodo uporabili še na posameznih Gigafidinih taksonomskih kategorijah. Zanimalo nas bo, katere so tematske podobnosti in razlike med Gigafido ter slWaCom in katere teme so pretežno značilne za vsak posamezni del Gigafide. Analizi rezultatov bo sledil sklep.

2 BESEDILA Z INTERNETA V KORPUSIH

2.1 Besedila z interneta v obstoječih referenčnih korpusih

Sestavljalci najnovejših referenčnih korpusov tujih jezikov so se o vključevanju besedil z interneta odločali različno (Tabela 1).

Korpus	**Besedila z interneta (da/ne; opis)**	**Obseg besedil z interneta**
a) angleščina		
Oxford English Corpus[8] Obseg: 2 milijardi pojavnic Leto: 2010	DA Korpus je skoraj v celoti sestavljen iz besedil z interneta, le nekaj tiskanih	Skoraj v celoti; od tega npr. blogi 8,2 %.

[6] http://www.gigafida.net

[7] http://www.nljubesic.net/resources/corpora/slwac/

[8] Spletne strani korpusov navajamo v razdelku Spletne strani. Ogled smo opravili februarja 2013.

	besedil, kot npr. znanstvene revije, je bilo dodanih zaradi dopolnitve posameznih predmetnih področij. Med dvajsetimi predmetnimi področji so tudi spletne strani podjetij, osebne spletne strani, blogi, forumi ipd.	
Cambridge English Corpus Obseg: milijarda pojavnic Leto: 2012	DA Z interneta so vključene spletne strani podjetij in osebne spletne strani, blogi, tviti, e-pošta,[9] spletne diskusijske skupine in forumi.	Ni podatka.
COCA: Corpus of Contemporary American English Obseg: 450 milijonov pojavnic Leto: 2012	NE	0
b) nemščina Das Deutsche Referenzkorpus – DeReKo Obseg: 5,4 milijarde pojavnic Leto: 2012	NE	0
c) nizozemščina SoNaR: STEVIN Nederlandstalig Referentiecorpus Načrtovani obseg: 500 milijonov pojavnic Leto: potekajoči projekt, podatki iz Reynaert in dr. (2010)	DA Korpus bo vseboval naslednja elektronska besedila (pisna, namenjena branju): forume, e-knjige, e-revije, e-pošto, glasila, sporočila za javnost, podnapise, teletekst, spletna besedila, Wikipedijo, klepetalnice in bloge (Reynaert 2010: 2695)	Načrtovani delež: 55 %.
č) danščina KorpusDK Obseg: 56 milijonov pojavnic Leto: 2000	DA Korpus vsebuje predstavitvene spletne strani in en spletni časopis.	Ni podatka.
d) finščina CSC: Suomen kielen tekstikokoelma Obseg: 180 milijonov pojavnic Leto: besedila iz 90. let 20. st.	NE	0
e) italijanščina CORIS/CODIS: CORpus di Italiano Scritto Obseg: 120 milijonov pojavnic	NE	0

[9] Ni razvidno, ali je šlo za zbiranje zasebne e-pošte ali (in morda hkrati) za zbiranje sporočil, poslanih prek dopisnih seznamov. Enako velja za korpus nizozemščine – gl. točko (c).

Leto: besedila iz 80. in 90. let 20. st.		
f) španščina CREA: Corpus de Referencia del Español Actual CREA Obseg: 154 milijonov pojavnic Leto: 2012	NE	0
g) portugalščina CRPC: Corpus de Referência do Português Contemporâneo Obseg: pisni del: 309 milijonov pojavnic Leto: 2010	NE	0
h) estonščina Reference Corpus of Estonian Obseg: 245 milijonov pojavnic Leto: 2009	DA Korpus vsebuje naslednja besedila z interneta: klepetalnice, forume, novičarske skupine in komentarje na novičarskih portalih.	9 %.
i) češčina SYN2010: Český národní korpus Obseg: 100 milijonov pojavnic Leto: 2010	NE	0
j) poljščina Narodowy korpus języka polskiego – NKJP Načrtovani obseg: 1,500 milijonov pojavnic Leto: potekajoči projekt, podatki iz Górski, Łazinski (2012)	DA Korpus bo vseboval: bloge, forume, klepetalnice, dopisne sezname ipd. ter predstavitvene spletne strani (strani ustanov in osebne spletne strani) (Górski, Łazinski 2012).	Načrtovani delež: 7 %.
k) slovaščina SNK: Slovenský národný korpus Obseg: 719 milijonov pojavnic Leto: 2011	NE	0
l) hrvaščina HNK: Hrvatski nacionalni korpus Trenutni obseg: 101,3 milijona pojavnic Leto: potekajoči projekt	DA V taksonomiji je predvidena kategorija e-besedila.	Ni podatka.

Tabela 1: Besedila z interneta v nekaterih tujih referenčnih korpusih.

Tabela kaže, da je med petnajstimi korpusi dvanajstih jezikov sedem takih, ki

vsebujejo – ali se načrtuje, da bodo vsebovali – besedila z interneta. Gre za dva korpusa angleščine ter korpuse nizozemščine, danščine, estonščine, poljščine in hrvaščine. Od preostalih osmih korpusov, ki ne vsebujejo besedil z interneta, dva ne presenečata, saj vsebujeta besedila iz 80. in 90. let 20. st. (korpus finščine in italijanščine), medtem ko so korpusi ameriške angleščine, nemščine, španščine, portugalščine, češčine in slovaščine iz let 2010–2012, tako da leto nastanka oz. leto izdaje vključenih besedil ni moglo vplivati na nevključitev besedil z interneta.

Obsegi internetnih besedil v korpusih, naštetih v Tabeli 1, so zelo različni: pri dveh korpusih je obseg manjši od 10 %, pri korpusu nizozemščine je načrtovani obseg 55%, pri treh korpusih ta podatek ni razviden, medtem ko korpus *Oxford English Corpus* v tem pogledu izstopa, saj gre za bolj spletni kot pa "tradicionalni" pisni korpus. Pri besedilnih žanrih se zdi, da je izhodišče sestavljalcev korpusov široko: večinoma želijo zajeti vse, od predstavitvenih spletnih strani, prek klepetalnic do tvitov ipd.

2.1.1 BESEDILA Z INTERNETA V NAJNOVEJŠEM REFERENČNEM KORPUSU SLOVENŠČINE

Kot je podrobneje pojasnjeno v Logar Berginc in dr. (2012: 45–67), so besedila z interneta postala tudi del najnovejšega korpusa slovenščine Gigafida, ki ga bomo analizirali v nadaljevanju. Tako odločitev je vodilo zavedanje, da postaja internet vse vplivnejše mesto, na katerem se srečujeta besedilna recepcija in produkcija.[10] Ker je šlo v metodološkem smislu za prvi večji poskus pridobivanja spletnih besedil za referenčni korpus pri nas,[11] so se sestavljalci tega korpusa – dokaj poskusno – pri izbiri spletnih besedil omejili na strani z informativnimi vsebinami (deset strani novičarskih portalov, npr. *24ur.com*, *siol.net*, *pozareport.si*, ter dvainšestdeset predstavitvenih strani ustanov, npr.

[10] Pravzaprav tokrat niti ni šlo za prvo vključitev internetnih besedil v referenčni korpus slovenščine, saj je že FidaPLUS vsebovala 1,24 % takega gradiva, enajst dokumentov z besedili s spletnih strani pa je postalo del referenčnega korpusa FIDA že pred petnajstimi leti.

[11] Pajkanje spletnih besedil je izvedel Miha Grčar (Institut Jožef Stefan), ki je celotni postopek skupaj z Markom Brakusom opisal v Logar Berginc in dr. (2012: 51–67).

up-rs.si, *mirovni-institut.si*, *spasteater.si*, in devetindvajset podjetij, npr. *revoz.si*, *sportina.si*, *kompas.si*), tako da v zajemu ni blogov, forumov, klepetalnic ipd., edino, kar je sorodno tovrstnim žanrom in je bilo vključeno v pajkanje za Gigafido, so komentarji na novičarskih portalih. Pred začetkom zbiranja besedil za Gigafido je bil načrtovan zelo okviren, od 10- do 50-odstotni obseg internetnega dela, ki se je na koncu uresničil v že omenjenih 16 % ali 185.758.467 pojavnicah.

2.2 Spletni korpusi

Ravno nasprotni zgornjim so v tem pogledu korpusi, ki so sestavljeni le iz besedil s spletnih strani: spletni korpusi (angl. *web corpora*), katerih namen je širok, saj želijo biti uporabni kot splošen vir podatkov o nekem jeziku, kot se ta kaže na svetovnem spletu (Baroni in dr. 2009: 1).

Uporaba interneta kot izjemno velikega, prosto in takoj dostopnega vira podatkov za jezikovnotehnološke ter jezikoslovne namene se je močno povečala z nastankom iniciative *WaCky* (Baroni in dr. 2009 ter tam navedena literatura), natančneje: ta iniciativa je v ospredje gradenj spletnih korpusov postavila njihovo uporabnost za jezikoslovne namene, s tem da je kot nujen del sestavljanja korpusov vključila detekcijo jezika, čiščenje, brisanje dvojnikov ter označitev besedil. Do danes je nastalo že več spletnih korpusov različnih jezikov, npr. korpus angleščine *ukWaC*, nemščine *deWaC*, italijanščine *itWaC*, francoščine *frWaC*,[12] leta 2011 pa tudi spletni korpus hrvaščine *hrWaC* in slovenščine *slWaC* (Ljubešić, Erjavec 2011; več o slednjem v nadaljevanju).

Prednosti gradnje spletnih korpusov je več, najočitnejše so avtomatizacija postopka, umik potrebe po urejanju avtorskopravnih razmerij[13] in precejšnja hitrost pridobitve velike količine besedil, je pa v primerjavi z gradnjo "tradicionalnih" pisnih korpusov ta gradnja precej manj izbirajoča oz.

[12] Prim. http://wacky.sslmit.unibo.it/doku.php?id=corpora.

[13] Vprašanje urejanja avtorskih pravic pri spletnih besedilih je različno od države do države, večinoma pa se pri gradnji spletnih korpusov zanemarja.

kontrolirana, zato je analiza njihove vsebine še toliko bolj pomembna.

3 GIGAFIDA IN SLWAC: PREDSTAVITEV, PRIMERJALNA METODA IN ANALIZA

3.1 Gigafida: gradnja in vsebina

Korpus Gigafida vsebuje 1.187.002.502 pojavnic in je nadgradnja referenčnega korpusa slovenskega jezika FidaPLUS, ki je v obsegu več kot 621 milijonov pojavnic na spletu prosto dostopen od leta 2006 ter že vključuje (oz. nadgrajuje) prvi tak korpus za slovenščino, tj. v letih 1997–2000 nastali korpus FIDA. Zbiranje novih besedil za Gigafido je potekalo od januarja 2009 do maja 2010 (tisk) oz. od aprila 2010 do aprila 2011 (internet). Gigafida vsebuje javno dostopna objavljena pisna besedila različnih zvrsti, ki so ločena v šest taksonomskih kategorij, kot prikazuje Tabela 2. Časovno obdobje, ki ga zajemajo besedila, vključena v Gigafido, je 1990–2011, s tem da prihaja pretežni del pojavnic iz besedil, objavljenih po letu 2000.

Taksonomija	**Oznaka**	**Število pojavnic**	**Delež v %**
tisk	T	1.001.244.035	84,35
knjižno	T.K	74.356.531	6,26
leposlovje	T.K.L	23.969.196	2,02
stvarna besedila	T.K.S	50.387.335	4,24
periodično	T.P	918.936.054	77,42
časopisi	T.P.C	663.664.965	55,91
revije	T.P.R	255.271.089	21,51
drugo	T.D	7.951.450	0,67
internet	I	185.758.467	15,65
SKUPAJ		**1.187.002.502**	**100,00**

Tabela 2: Delež pojavnic po taksonomiji v Gigafidi.

Gigafida je označena s statističnim označevalnikom Obeliks (Grčar in dr. 2012) po tabeli oznak JOS (Erjavec in dr. 2010). Obeliks vključuje tri module, povezane v en program: tokenizator, ki deluje na podlagi pravil, ter statistična modula za lematizacijo in označevanje. Korpus je dostopen za široko javno uporabo v spletnem vmesniku, kot baza podatkov pa je prosto dostopen v obsegu 9 % (100 milijonov pojavnic) pod imenom ccGigafida[14] (Arhar Holdt in dr. 2012; Erjavec, Logar Berginc 2012; Kosem 2012; Logar Berginc in dr. 2012: 98–118, 77–97).

3.2 slWaC: gradnja in vsebina

Korpus slWaC vsebuje 380 milijonov pojavnic,[15] ki prihajajo z 11.493 spletnih strani z domene .si. Je oblikoskladenjsko označen in lematiziran z označevalnikom ToTaLe z oznakami iz specifikacij JOS (Erjavec in dr. 2010). Pajkanje za slWaC je potekalo od januarja do marca 2011. Gradnja je vključevala naslednje faze:

a) izbor izhodiščnih URL-naslovov,

b) pajkanje,

c) brisanje dvojnikov,

č) luščenje vsebine (angl. *content extraction*)

d) detekcijo jezika,

e) filtriranje in

f) jezikoslovno označevanje.

Izbor izhodiščnih URL-naslovov je bil izveden s pomočjo API-ja Yahoo BOSS, ki omogoča strojno izvajanje poizvedb na spletnih straneh Yahoo indeksa, poizvedbe so bile sestavljene iz naključnega nabora pojavnic s srednjo pogostostjo (tj. pogostostjo od 1.000 do 10.000), pridobljenih iz časopisnega

[14] http://www.slovenscina.eu/korpusi/proste-zbirke

[15] Trenutno poteka gradnja nove različice v obsegu 500 milijonov pojavnic.

dela korpusa FidaPLUS, ki vsebuje 100 milijonov pojavnic. Na ta način smo dobili okoli 50.000 URL-naslovov z 11.493 spletnih domen (gl. Tabelo 3).

Pridobljeni URL-naslovi so bili izhodišče za pajkanje vrhnje domene .si, ki smo ga izvedli z lastnim algoritmom, ki uporablja iskanje v širino na večnitni način. Zajemali smo dokumente text/html z velikostjo 50 do 500 kilobajtov. S pajkanjem smo tako dobili 9,2 milijona dokumentov.

Odstranjevanje dvojnikov na ravni odstavka smo izvedli z razpršilnim (angl. *hash*) algoritmom SHA224, s čimer smo odstranili 2,3 % pridobljenih dokumentov.

Pridobivanje vsebine smo izvedli z lastnim algoritmom, ki iz HTML-dokumenta pridobi največji obseg vsebine, ki je videti jezikovno pravilen (odstavki se začenjajo z veliko začetnico in zaključujejo s končnim ločilom) ter je hkrati na istem nivoju znotraj hierarhične strukture HTML-dokumenta. Eksperimentalno je bilo potrjeno, da je na ta način pridobljena vsebina opazno "čistejša" od tiste, ki jo dobimo z algoritmom BTE ali zelo priljubljenim orodjem BoilerPipe,[16] ki ima sicer višji priklic kot naša metoda. S to metodo smo vsebino uspešno pridobili iz 17,8 % zbranih dokumentov.

Detekcija jezika je bila izvedena na ravni odstavka z Markovovim algoritmom drugega reda, ki se je že predhodno izkazal za uspešnega pri nadzirani detekciji jezika, in to tudi pri slovenščini sorodnih jezikih, kot sta hrvaščina in srbščina (Ljubešić in dr. 2007). Algoritem je izvedel razlikovanje med 22 jeziki in je 22 % dokumentov označil kot neslovenske.

Sledilo je končno filtriranje vsebine, v katerem so bili izločeni dokumenti, ki so bili prekratki, ki so vsebovali napake v kodiranju ali so imeli visok odstotek interpunkcijskih znakov. V tej fazi je bilo iz korpusa umaknjenih 3,7 % vsebine. V zaključku gradnje korpusa slWaC so bila besedila še oblikoskladenjsko označena in lematizirana z orodjem ToTaLe (Erjavec in dr. 2005).

[16] http://code.google.com/p/boilerpipe/

Rezultat posamezne faze gradnje	**Število domen / dokumentov / pojavnic**
izhodiščne domene	11.493
pajkane domene	18.418
pridobljeni dokumenti	9.247.341
dokumenti po odstranitvi dvojnikov	9.022.716
dokumenti, iz katerih je bila pridobljena vsebina	1.598.011
dokumenti v slovenščini	1.337.286
dokumenti po končnem filtriranju	1.287.895
pojavnice	380.299.844

Tabela 3: Število domen, dokumentov oz. pojavnic po posameznih fazah gradnje slWaCa.

3.3 Metoda tematskega modeliranja

V zadnjih letih je metoda tematskega modeliranja (Blei in dr. 2003) vse bolj priljubljen način proučevanja velikih zbirk besedilnih podatkov, zlasti na področju analize vsebine in digitalne humanistike, pa tudi v korpusnem jezikoslovju za analizo ter primerjavo različnih korpusov (Sharoff 2010).

Metoda temelji na predpostavki, da je vsak dokument v zbirki nastal iz vsebin z več temami. Vsako temo predstavlja verjetnostna distribucija besed – povedano drugače: za vsako besedo obstaja določena verjetnost, da pripada določeni temi. Primer v Tabeli 4, ki je vzet iz Steyvers in Griffiths (2007), prikazuje štiri teme in po pet besed, ki najverjetneje pripadajo vsaki od tem. Teme so izračunane na korpusu TASA (*Touchstone Applied Science Associates*),[17] ki vsebuje 37.000 besedil s področja izobraževanja. Že iz prvih petih besed, ki najverjetneje pripadajo določeni temi, je razvidno, da gre za teme, povezane z zdravili, barvami, spominom in obiskom pri zdravniku. Lahko predvidevamo, da različna besedila vsebujejo različne kombinacije

[17] http://lsa.colorado.edu/spaces.html

posameznih tem. Tako bo npr. besedilo o osebi, ki je zaradi zlorabe zdravila utrpela spremembo v percepciji barv, sestavljeno iz kombinacije prvih treh tem, medtem ko bo besedilo, ki govori o izgubi spomina in obisku pri zdravniku, sestavljeno iz zadnjih dveh tem.

tema 247	**tema 5**	**tema 43**	**tema 56**
drugs	red	mind	doctor
drug	blue	thought	dr.
medicine	green	remember	patient
effects	yellow	memory	hospital
body	white	thinking	care

Tabela 4: Po metodi tematskega modeliranja pridobljene teme in besede, ki najverjetneje pripadajo vsaki od tem, v korpusu TASA (vir: Steyvers in Griffiths 2007).

Vhodni podatki za metodo tematskega modeliranja so zbirke dokumentov in vnaprej predvideno oz. določeno število tem. Rezultat metode sta dve verjetnostni distribuciji:

a) verjetnostna distribucija tem za vsak dokument oz. verjetnost, da nek dokument vsebuje določene teme, in

b) pogojna verjetnostna distribucija besede pri določeni temi oz. verjetnost posamezne besede, da pripada določeni temi.

Tematski model je generativni, ker to, kako je nastala vsebina dokumenta, kaže na osnovi latentnih spremenljivk, tj. tem. Naloga modeliranja je najti tiste latentne spremenljivke, ki najbolje pojasnjujejo zbirko dokumentov kot rezultat povezovanja vsebin, sestavljenih iz istih spremenljivk. Model je zasnovan na latentni Dirichletovi alokaciji (LDA), sklepanje pa se najpogosteje izvede po Gibbsovem vzorčenju.

Ob tem, da se modeliranje tem uporablja za opis vsebine zbirk besedilnih podatkov, se ta metoda vse pogosteje uporablja tudi za iskanje večpomenskosti in izračun podobnosti dokumentov, napisanih v istem jeziku, pa tudi za iskanje jezikovno neodvisnih konceptov v večjezičnih zbirkah

besedil, povezanih na ravni dokumentov.

Kot bo podrobneje pojasnjeno v nadaljevanju, smo tematsko modeliranje uporabili za izgradnjo N tem na vsakem od korpusov (Gigafida, slWaC) oz. podkorpusov, tj. taksonomskih kategorij Gigafide; vsebino (pod)korpusov pa prikazujemo kot skupek tem v obliki vrstic v tabeli. Vsaka tema je prikazana z besedami (samostalniki), ki najverjetneje pripadajo eni temi, pri vsaki temi pa podajamo tudi podatek o njenem obsegu, ki ga ima najverjetneje v celem (pod)korpusu. Najverjetnejše besede torej dajejo uvid v najverjetnejšo temo, tj. nam omogočajo, da temo poimenujemo, obseg te teme pa nam pove, v kolikšni meri je ta zastopana v vsebini celotnega (pod)korpusa.

3.4 Gigafida in slWaC: rezultati metode tematskega modeliranja

Oba korpusa (oz. namesto Gigafide ccGigafida, ki smo jo zaradi dostopnosti v obliki baze podatkov in manjše velikosti vzeli za analizo) smo primerjali z metodo, predstavljeno zgoraj. Tako pri slWaCu kot pri Gigafidi so nas zanimale le samostalniške leme. Najprej smo primerjavo naredili med celotno ccGigafido in slWaCom (razdelek 3.4.1), pri čemer smo število tem omejili na dvajset, nato pa smo metodo aplicirali še na posamezne taksonomske kategorije znotraj ccGigafide, pri čemer smo zaradi celostnega prikaza dobljenih rezultatov v prispevku število tem omejili na pet (razdelek 3.4.2).

Teme smo poimenovali ročno. Pri tem smo skušali čim bolje zajeti skupno vsebino po metodi tematskega modeliranja pridobljenih samostalniških lem. Izkazalo se je, da pri več skupinah zgolj eno poimenovanje (npr. *finance*) ne bo pokrilo vseh lem, da je treba torej uporabiti tudi kombinirana poimenovanja (npr. *izobraževanje + razvoj + gospodarstvo*), in to v različnih kombinacijah (npr. enkrat *naselje + cestni promet + potovanje*, drugič *potovanje + turizem*).

3.4.1 CELOTNA GIGAFIDA PROTI SLWACU

Pri zbiranju besedil za Gigafido (Logar Berginc in dr. 2012: 13–44) so si sestavljalci zastavili cilj pridobiti gradivo različnih tem oz. področij. Določen je

bil naslednji nabor, pri čemer ni šlo za zaprto množico: *aktualni dogodki*; *gospodarstvo, politika*; *vzgoja in izobraževanje*; *narava, dom, hišni ljubljenci*; *ljudje, družina, moški, ženske, otroci, mladina*; *zdravje, hrana*; *posel, finance*; *prosti čas, glasba, film, razvedrilo, moda*; *šport, turizem*; *kultura, umetnost*; *religija, duhovnost* ter *računalništvo* in *avtomobilizem*.

Prva primerjava po metodi tematskega modeliranja je pokazala, da so razlike med Gigafido in slWaCom dokajšnje. V Tabelah 5 in 6, v katerih za oba korpusa prikazujemo delež (%) vsake teme od dvajsetih v korpusu (prvi stolpec) in samostalniške leme, ki z najverjetneje pripadajo eni temi (tretji stolpec), je razvidno naslednje:

- **Osem tem je skupnih** (v spodnjih tabelah krepki tisk): Gre za notranjo politiko, finance, ekipni šport, vojno in terorizem (po svetu), publikacije in kulturo, lokalno (prostorsko) politiko, zdravje ter pravo.
- **Pet tem je različnih** (v spodnjih tabelah ležeči tisk): V Gigafidi so opaznejše teme naselje in cestni promet (zlasti z vidika prometnih nesreč), prireditve (zlasti z vidika njihove najave, opisa), televizijski in radijski program, neekipni športi ter zaposlitev. V slWaCu izstopajo film, glasba, potovanja in turizem, zunanja politika (zlasti EU, Hrvaška) ter mali oglasi.
- **Sedem tem je deloma skupnih** (spodaj podčrtano): Tu gre predvsem za različna tematska druženja, npr.: avtomobilizem in informacijsko-komunikacijska tehnologija sta v Gigafidi združena, v slWaCu sta ločena; hrana je v Gigafidi izrazito enorodna kategorija, medtem ko je v slWaCu zgolj manjši del življenjskega stila; družina je v Gigafidi skupaj z moškim, žensko in domom (*otrok*, *leto*, *družina*, *dan*, *ženska*, *življenje*, *starš*, *čas*, *oče*, *prijatelj*, *človek*, *moški*, *žena*, *sin*, *mama*, *mož*, *mati*, *pomoč*, *dom*), v slWaCu pa skupaj z religijo (*otrok*, *leto*, *cerkev*, *dan*, *oče*, *bog*, *človek*, *čas*, *mati*, *roka*, *družina*, *sin*, *starš*, *gospod*, *beseda*, *svet*, *življenje*, *ime*, *pot*). Razlog za različno druženje

tem je lahko bodisi ta, da je vsako področje zastopano s premalo podatki za lastno temo, ali ta, da si področji delita ključno izrazje in se to pogosto pojavlja pri obeh (razloga pa sta seveda lahko tudi oba hkrati).

Med obsegi skupnih, različnih ter deloma skupnih tem v enem in drugem korpusu ni velikih razlik. Skupne teme (*notranja politika* itd.) v Gigafidi obsegajo 41 %, v slWaCu 36 %; različne teme (npr. v Gigafidi *naselje, cestni promet*, v slWaCu *film*) v Gigafidi obsegajo 22 %, v slWaCu 23 %; preostali, deloma skupni del ima v Gigafidi 37% del, v slWaCu pa 41%. Do enake ugotovitve pridemo, če pogledamo obseg posameznih skupnih tem: razlike so majhne, še največja je pri *notranji politiki*, ki ji je v Gigafidi pripisan 7,31% obseg, v slWaCu pa 4,87% obseg.

Obseg v %	**Tema**	**Samostalniške leme**
8,68	človek, življenje, religija	človek življenje svet čas leto beseda način stvar ljubezen stoletje odnos zgodovina država bog delo resnica moč zgodba cerkev
7,31	**notranja politika**	predsednik vlada svet stranka država minister leto član volitev zakon poslanec odbor komisija zbor uprava predlog seja vprašanje predstavnik
6,42	telo, okolje	voda roka barva glava tla noga meter del vrsta morje čas prostor oči zrak oblika zemlja dan sonce stran
6,16	**finance**	milijon odstotek evro leto tolar cena banka milijarda družba podjetje dolar delnica vrednost delež prodaja trg sit država denar
6,06	razvoj, gospodarstvo	razvoj podjetje država področje trg delo sistem družba okolje človek program gospodarstvo politika projekt cilj možnost znanje sodelovanje storitev
5,76	**šport (ekipni)**	tekma točka liga igralec ekipa zmaga igra sezona klub minuta prvak mesto trener prvenstvo konec krog leto reprezentanca moštvo
5,73	**vojna, terorizem**	leto država vojna človek vojska policija dejanje orožje napad oblast žrtev dan vojak čas policist sila zapor meja kazen
5,48	*naselje, cestni promet, potovanje*	cesta mesto hiša pot leto ura ulica vozilo meter prostor avtomobil vas promet kilometer voznik nesreča del kraj hotel
5,40	moški, ženska, družina, dom	otrok leto družina dan ženska življenje starš čas oče prijatelj človek moški žena sin mama mož mati pomoč dom
4,79	*prireditve*	ura dan leto društvo prireditev dom sobota dvorana občina skupina nedelja mesto šola praznik petek član razstava teden koncert

4,74	informacijsko-komunikacijska tehnologija, avtomobilizem	sistem motor računalnik avtomobil podatek vozilo model uporabnik oprema naprava hitrost slika stran telefon program moč uporaba kartica zaslon
4,46	*televizijski in radijski program*	film oddaja glasba leto poročilo serija dan program čas del novica skupina dnevnik pesem ponovitev svet festival radio teden
4,26	**publikacije, kultura**	leto knjiga naslov delo revija ime razstava stran avtor fotografija številka nagrada zbirka slika beseda muzej časopis članek jezik
4,24	**lokalna (prostorska) politika**	občina leto prostor gradnja objekt območje podjetje okolje voda zemljišče projekt odpadek delo energija načrt ministrstvo cesta sredstvo stanovanje
3,75	**zdravje**	bolezen telo zdravilo koža zdravljenje bolnik težava človek rak kri celica zdravnik bolečina voda snov primer zdravje srce dan
3,68	**pravo**	zakon člen postopek sodišče pravica podatek dan odstavek organ podlaga oseba odločba stranka primer sklad pogodba zadeva določba list
3,50	*šport*	leto mesto prvenstvo pokal dirka tekmovanje tekma sezona dan čas ekipa zmaga nastop tek prvak meter finale kategorija šport
3,48	*zaposlitev*	leto delo plača zakon država delavec zavarovanje pravica sredstvo pogodba čas zavod strošek dejavnost proračun denar mesec oseba pokojnina
3,20	izobraževanje	šola leto delo program fakulteta otrok univerza študent področje učenec izobraževanje znanje študij šport učitelj jezik šolstvo zavod dijak
2,89	hrana	vino olje mleko meso voda sol žlica sladkor zelenjava hrana sadje jed kruh izdelek rastlina krompir jajce kilogram sok

Tabela 5: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po obsegu v Gigafidi.

Obseg v %	**Tema**	**Samostalniške leme**
10,77	človek, moški, ženska, življenje	človek čas življenje stvar svet ženska otrok dan način moški primer odnos vprašanje leto beseda trenutek konec problem resnica
8,02	izobraževanje, razvoj, gospodarstvo	delo področje program projekt leto šola razvoj organizacija znanje podjetje slovenija študent sistem skupina sodelovanje okolje otrok dejavnost cilj
6,13	informacijsko-komunikacijska tehnologija	stran uporabnik podatek sistem računalnik slika program uporaba telefon naprava podjetje internet vsebina tehnologija omrežje volja zaslon oprema storitev
5,98	bivalno okolje	voda energija prostor barva material sistem uporaba del naprava izdelek površina zrak temperatura oblika primer okolje plin stroj

		stena
5,48	*film*	film leto vloga igralec režiser zgodba nagrada svet igralka serija čas dan življenje new york snemanje oskar dekle john
5,40	*glasba*	leto skupina glasba pesem koncert festival album dan slovenija oddaja nastop ura nagrada oder skladba prireditev večer ljubljana čas
5,38	**lokalna (prostorska) politika**	občina leto cesta mesto prostor ljubljana članek območje slovenija objekt hiša del gradnja stanovanje vas dom župan wikipedija naselje
5,34	*potovanje, turizem*	mesto dan pot ura leto hotel morje otok čas soba potovanje vrh ogled del meter letalo obala gora voda
5,18	**finance**	leto evro odstotek podjetje milijon družba banka cena trg država plača delnica denar slovenija delež prodaja delavec rast vrednost
4,87	**notranja politika**	vlada predsednik zakon stranka slovenija minister sodišče predlog svet ministrstvo član komisija poslanec mnenje leto delo zadeva vprašanje seja
4,70	**vojna, terorizem**	leto država vojna človek predsednik oblast zda napad vojska dan sila mesto vojak policija žrtev vlada svet orožje obama
4,43	**publikacije, kultura**	leto knjiga delo razstava ljubljana avtor umetnost jezik del zbirka nagrada ime kultura stoletje zgodovina muzej gledališče roman besedilo
4,42	**pravo**	zakon podatek primer pravica člen oseba plačilo storitev dan postopek pogodba pogoj strošek sklad slovenija podlaga stran račun cena
4,04	**šport (ekipni)**	tekma minuta igra leto točka prvenstvo igralec ekipa mesto zmaga sezona klub konec liga reprezentanca prvak trener finale gol
3,98	življenjski stil, hrana	koža hrana voda olje žival pes rastlina izdelek vrsta las barva dan meso okus čas minuta mleko sestavina zelenjava
3,65	**zdravje**	telo bolezen zdravilo zdravnik leto otrok bolnik zdravljenje težava človek ženska dan primer bolečina zdravje raziskava učinek rak kri
3,60	*zunanja politika*	država slovenija eu leto članica minister predsednik hrvaška vlada unija evropa vprašanje sporazum svet politika komisija meja republika sodelovanje
3,17	religija, družina	otrok leto cerkev dan oče bog človek čas mati roka družina sin starš gospod beseda svet življenje ime pot
2,88	*mali oglasi*	oglas iskanje seznam stran znamka stroj možnost vrh kvadrat cena država model ce traktor mascus leto ukaz zožitev prikolica
2,59	avtomobilizem	vozilo avtomobil motor dirka vožnja leto voznik avto kolo mesto cesta hitrost nesreča model sedež kilometer čas sezona razred

Tabela 6: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po obsegu v slWaCu.

Primerjava rezultatov metode tematskega modeliranja in predhodnega nabora tem, ki je bil vodilo pri zbiranju besedil za Gigafido, pokaže, da med prvimi dvajsetimi temami v Gigafidi ni (vsaj ne dovolj opazno) narave, hišnih ljubljencev, mladine, mode in kulture; če pa se še enkrat osredotočimo le na teme, ki so v Gigafidi in slWaCu različne, lahko okvirno posplošimo: v trenutno največjem referenčnem korpusu slovenščine je opazen delež kronike ter napovednikov prireditev in predvsem televizijskega programa (ki so najpomembnejši vzrok korpusnega šuma že od FidePLUS dalje), kaže pa tudi, da zlasti tiskano časopisje (56 % korpusa) več pozornosti kot spletni viri namenja neekipnim športom. Zaposlitev je tema, ki je v Gigafidinih besedilih uokvirjena v gospodarsko krizo – ta se v slWaCu v naboru 20 tem skoraj ni pojavila, zgolj nakazana je pri financah. Po drugi strani v Gigafidi v naboru dvajsetih tem ni prostega časa in zabave (film, glasba, potovanja, turizem), ki ju najdemo v slWaCu, splet pa se zdi tudi prva izbira za objavo malih oglasov.[18] Preseneča samostojnost teme zunanja politika v slWaCu, ki je v Gigafidi zgolj nakazana v temi vojna in terorizem; konkordančnik *NoSketch Engine* glede tega pokaže, da so med osmimi domenami, na katerih se najpogosteje pojavljajo samostalniki iz zunanje politike, naslednje: dnevnik.si, rtvslo.si, delo.si, mladina.si, radiokoper.si, rtvslovenija.si, eu2008.si, rsi.si in radiomaribor.si. Pretežni vir "zunanjepolitičnih" samostalnikov so torej spletne strani dveh časopisov in ene revije (Dnevnik, Delo, Mladina), kar kaže na to, da je v tiskanih izdajah istoimenskih publikacij zunanje politike najbrž manj kot na spletu, del izmed naštetih virov pa v Gigafido ni bil vključen (radijske postaje, tvslovenija.si, eu2008.si).

Iz celotne primerjave je razvidno, da velja pri pripravah na nadgradnjo Gigafide posebno pozornost nameniti publikacijam s temami, ki so umanjkale, čeprav se bo najverjetneje pokazalo, da je pri kateri od njih svetovni splet pač prva izbira tako za tiste, ki tam objavljajo besedila, kot za bralce. Če povežemo rezultate tematske analize in podatke o informacijah, ki jih uporabniki iščejo

[18] Seveda pa je teh v Gigafidi morda manj tudi zato, ker ta npr. ne vključuje *Salomonovega oglasnika*.

na internetu iz uvoda, sta npr. takšni temi potovanja in mali oglasi.

3.4.2 TAKSONOMSKE KATEGORIJE GIGAFIDE

Kot je bilo razvidno v točki 2.1.1, so besedila v Gigafidi razdeljena v šest taksonomskih kategorij: leposlovje, stvarna besedila, časopisi, revije, drugo in internet. Zanimalo nas je, kaj lahko na podlagi značilnih tem izvemo o vsebini vsake od njih. Analizo smo, kot že rečeno, omejili na pet tem in dobili Tabele 7–12.

3.4.2.1 Leposlovje

Obseg v %	**Tema**	**Samostalniške leme**
28,30	človek	človek leto življenje roka otrok svet beseda oče delo glas stvar knjiga mama smrt prijatelj misel mož gospod mati
23,70	telo, moški, ženska, prostor	oči glava obraz stran noga ženska trenutek pogled okno las telo moški stena usta zrak prst barva morje nebo
22,70	čas, kraj	dan čas hiša mesto konec pot ime noč ulica del cesta prostor moč ura bog šola kraj teden dom
18,10	bivanjski prostor, telo	vrata soba voda miza tla roka postelja papir kri hrbet stol uho hodnik vrh pisarna vino številka kuhinja kozarec
7,20	telo, religija, predmeti	oblika gora les točka teža maščoba gibanje tiger postopek maša križ prepir jed bistvo kamen motor obiskovalec avgust nož

Tabela 7: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po obsegu v leposlovnem delu Gigafide.

Za leposlovni del Gigafide se kot ključna kaže tematizacija človeka v razmerjih do drugega človeka (*partner*, *starš*, *prijatelj* itd.), z vidika telesa (*roka*, *glas*, *oči*, *glava*, *obraz*, *noga* itd.), prostora in časa (npr. *dan*, *noč*, *ura*, *hiša*, *mesto*, *pot*, *dom*, *kuhinja*) ter predmetov, ki ga obdajajo (*postelja*, *stol*, *kozarec*, *križ*, *kamen* ipd.).

3.4.2.2 Stvarna besedila

Obseg v %	**Tema**	**Samostalniške leme**
27,33	človek, RAZNO	življenje človek svet otrok čas delo ženska bog moč način beseda

		odnos ljubezen zgodovina jezik stoletje stvar oblika leto
23,56	izobraževanje	delo leto država šola znanje podjetje skupina družba učenec področje razvoj proces pravica organizacija cilj primer delavec sistem učitelj
18,45	narava, hrana, zdravje	voda rastlina telo vrsta minuta snov hrana roka bolezen list olje tla noga sol zdravilo barva zrak del glava
17,70	čas, prostor, religija	leto mesto stoletje dan pot čas hiša vojna cesta ura del stran vrh cerkev svet konec morje kraj dolina
12,96	besedilo, računalništvo	slika besedilo stran podatek beseda datoteka ime točka okno polje jezik uporaba oblika program del vrstica primer sistem knjiga

Tabela 8: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po obsegu v Gigafidini kategoriji stvarna besedila.

Med stvarnimi besedili Gigafide so predvsem besedila, ki poučujejo o človeškem življenju, družbi, naravi, prehrani, zdravju, religiji in računalništvu.

3.4.2.3 Časopisi

Obseg v %	**Tema**	**Samostalniške leme**
29,47	notranja politika	leto država predsednik delo svet vlada občina zakon človek stranka čas ministrstvo področje minister sodišče zveza vprašanje pravica član
21,57	prireditve, televizijski in radijski program	leto ura film dan čas otrok šola delo življenje svet človek program oddaja del knjiga glasba razstava skupina dom
19,93	RAZNO	cesta leto voda prostor mesto človek dan meter ura vozilo hiša avtomobil čas del območje bolezen vino pot delo
15,72	finance, gospodarstvo	leto odstotek milijon tolar evro podjetje cena družba banka trg milijarda država delnica vrednost denar dolar mesec prodaja delež
13,30	šport	tekma mesto leto točka prvenstvo zmaga sezona ekipa liga igra klub pokal prvak igralec konec minuta trener krog tekmovanje

Tabela 9: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po obsegu v časopisnem delu Gigafide.

V časopisnem delu Gigafide so značilno tematizirani notranja politika, finance, gospodarstvo in šport, opazen pa je tudi napovednik dogodkov oz. televizijskega in radijskega programa.

3.4.2.4 Revije

Obseg v %	Tema	Samostalniške leme
32,10	RAZNO	leto človek življenje čas svet otrok dan mesto delo ženska družina pot film hiša odnos knjiga konec stran vojna
20,20	finance, gospodarstvo, notranja politika	leto podjetje država delo milijon odstotek tolar trg področje zakon cena razvoj družba dejavnost program stranka vlada predsednik pravica
17,19	računalništvo, avtomobilizem	sistem stran računalnik motor podatek model avtomobil slika program hitrost uporabnik oprema naprava prostor vozilo različica zaslon del uporaba
15,42	televizijski in radijski program, šport	leto dan ura film skupina mesto naslov revija minuta tekma igralec čas klub sezona igra glasba ekipa konec oddaja
15,09	zdravje	voda telo koža bolezen barva zdravljenje sit zdravilo težava hrana dan bolnik rak rastlina vrsta oblika snov čas eur

Tabela 10: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po obsegu v revijalnem delu Gigafide.

Pri revijah je na prvem mestu raznorodna skupina, ki je v sedmih primerih od devetnajstih sicer enaka skupini RAZNO pri časopisih, vseeno pa jo od nje ločijo razmeroma nepovezani *življenje*, *svet*, *otrok*, *ženska*, *družina*, *film*, *hiša*, *odnos*, *knjiga*, *konec*, *stran* in *vojna*. Še tri skupine so podobne časopisnim: finance, gospodarstvo in notranja politika, ima pa slednja v časopisih veliko večji obseg, tj. sama zase kar 29%. Pri revijah se pojavijo tudi teme televizijski oz. radijski programi in šport, vendar pa so pri revijah za razliko od časopisov te združene v eno skupino, pa tudi njihov obseg je manjši (15 % proti 35 % pri časopisih). Nove so pri revijah teme računalništvo, avtomobilizem in zdravje.

3.4.2.5 Drugo

Obseg v %	Tema	Samostalniške leme
22,81	gospodarsko pravo	člen zakon odstavek republika oseba leto družba postopek dan pravica sklad organ banka sredstvo rok podatek odločba podjetje podlaga
22,59	pravo v državnem zboru	zakon vlada zbor republika predlog gospod svet zadeva amandma odbor poslanec vprašanje sodišče obravnava sklep ministrstvo

		komisija stranka predsednik
21,23	RAZNO	leto človek čas življenje oddaja beta ura film vojna program prispevek svet otrok minuta insert dan delo mesto vas
20,79	evropsko pravo	država pogodbenica delo člen sklad zakon sporazum varstvo promet program republika dejavnost uporaba sredstvo pogoj objekt organ odstavek področje
12,58	izdelki	izdelek material izdelava cena tovarna voda tara vrednost snov vrsta del energija oblika leto vlakno zrak stroj slika primer

Tabela 11: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po obsegu v Gigafidini kategoriji drugo.

Tabela 11 kaže izrazito pravnost kategorije drugo (66 %), kar ne preseneča, saj 46 % besed v ta del Gigafide prinašajo besedila Državnega zbora RS. Če izločimo leme, ki so skupne vsaj dvema od treh s pravom označenih skupin, dobimo podskupino, ki bi jo morda lahko označili kot gospodarskopravna (*oseba, leto, družba, postopek, dan, pravica, banka, rok, podatek, odločba, podjetje, podlaga*), podskupino, ki je bolj povezana z zadevami državnega zbora (*vlada, zbor, predlog, gospod, svet, zadeva, amandma, odbor, poslanec, vprašanje, sodišče, obravnava, sklep, ministrstvo, komisija, stranka, predsednik*), in splošnejšo podskupino, ki zgolj v svojem začetku nakazuje večjo povezanost z evropskim pravom (*država, pogodbenica, delo, sporazum, varstvo, promet, program, dejavnost, uporaba, pogoj, objekt, področje*).

3.4.2.6 Internet

Obseg v %	**Tema**	**Samostalniške leme**
25,55	RAZNO	človek leto otrok država življenje čas svet dan stran denar ženska delo oblast vojna stvar narod družina beseda predsednik
24,58	notranja politika, gospodarstvo, razvoj	leto delo država evro vlada področje zakon podjetje milijon svet program sredstvo ministrstvo družba predsednik odstotek občina republika razvoj
18,49	šport	leto film tekma mesto sezona igralec igra ekipa prvenstvo konec liga prvak klub dan točka zmaga čas naslov skupina
15,96	promet, avtomobilizem	cesta voda vozilo ura dan mesto avtomobil promet leto prostor cena čas voznik del avto nesreča izdelek vožnja meter

15,43	pravo	zakon člen postopek sodišče dan odstavek pravica podatek organ podlaga stranka oseba odločba pogodba primer določba delo sklad predlog

Tabela 12: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po obsegu v internetnem delu Gigafide.

V Tabeli 12 je predvsem razvidna tematska sorodnost kategorije internet s časopisi in revijami (notranja politika, gospodarstvo, šport, avtomobilizem). Na petem mestu je pravo z lemami, ki smo jih videli že pri kategoriji drugo. Glede na spletne strani, s katerih so prišla besedila v Gigafido z interneta – novičarski portali 66 %, ostalo predstavitvene spletne strani podjetij in ustanov (od tega le 4 % s strani podjetij, medtem ko je ostalih 40 % s strani ustanov, od tega 25 % s strani državnih ustanov tipa *dz-rs.si*, *sodisce.si*, *ip-rs.si* ipd.) – je rezultat pričakovan. Opazno je, da med petimi najznačilnejšimi temami svoje "predstavnice" nimajo verjetno preveč raznorodna, pa hkrati premalo obsežna besedila s predstavitvenih strani podjetij.

Tematska analiza taksonomskih kategorij Gigafide je dala glede na predhodno poznavanje besedil, ki so bila vključena v Gigafido (oz. so bila že predhodno vključena v FidoPLUS; prim. Logar Berginc in dr. 2012), razmeroma pričakovane rezultate. Ob tem velja poudariti, da je omejitev na pet tem minimalna (posledično smo pri kar štirih in pol od šestih kategorij dobili temo RAZNO) in je na spodnji meji povednosti. Pri vseh kategorijah smo naredili tudi analizo z desetimi temami, a je za tukajšnji celovit prikaz preobsežna – v Tabeli 13 jo podajamo samo za internetni del korpusa.

Obseg v %	**Tema**	**Samostalniške leme**
17,12	RAZNO	človek leto država otrok življenje čas dan denar svet vojna narod oblast ženska stran družina stvar žrtev roka primer
15,02	prireditve, kultura, zabava	leto film dan svet čas glasba ura knjiga življenje skupina nagrada delo zgodba pesem festival fotografija mesto vloga človek
12,05	promet	cesta voda ura vozilo mesto avtomobil promet dan prostor avto voznik čas vožnja motor nesreča meter morje del smer
11,86	notranja politika	vlada predsednik zakon svet zbor minister predlog seja država republika zadeva leto komisija poslanec odbor skupina član

		stranka delo
10,67	razvoj, izobraževanje	področje delo program leto razvoj država projekt sredstvo okolje študent sistem univerza šola fakulteta ministrstvo mesto organizacija prostor sodelovanje
8,84	šport	tekma leto sezona mesto prvenstvo liga prvak ekipa točka zmaga klub igra igralec konec minuta finale krog pokal gol
8,38	gospodarstvo, finance	leto evro milijon odstotek plača podjetje banka država denar družba sredstvo mesec delo cena višina delavec eur vrednost proračun
6,76	pravo	podatek oseba zakon dovoljenje delo storitev član država pogoj pravica podlaga pogodba organ sklad zemljišče republika dejavnost družba dan
4,71	pravo	zakon člen odstavek pravica organ postopek informacija določba ustava list podatek sklad podlaga sodišče sprememba predlog odločba značaj točka
4,59	pravo	sodišče postopek dan stranka član odločba stopnja pogodba sklep pravica sodba odstavek razlog pritožba podlaga rok zahtevek odločitev primer

Tabela 13: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po obsegu v internetnem delu Gigafide: deset tem.

V seznamu desetih tem kategorije internet so glede na Tabelo 12 v celoti nove le teme prireditve, kultura, zabava, bolj opazna pa je tema pravo, ki je zdaj zastopana v kar treh skupinah (v tabeli smo pri vseh pustili kar enorodno poimenovanje *pravo*). Če seštejemo njihov delež, dobimo 16,06 %, kar to temo na lestvici desetih tem uvršča na drugo mesto. Tolikšna pravnost internetnega dela Gigafide je bila že opažena (gl. Erjavec, Logar Berginc 2012: 61–62), vendar pa tokrat uporabljena metoda tematskega modeliranja vendarle kaže ugodnejšo sliko tematske razpršenosti Gigafidine kategorije internet kot takrat izvedena metoda frekvenčnega profila (angl. *frequency profiling*; Rayson, Garside 2000).[19]

[19] Metoda frekvenčnega profila temelji na logaritemski verjetnosti (angl. *log-likelihood*, LL), izvedli pa smo jo tako, da smo najprej izdelali frekvenčni seznam lem vsakega od podkorpusov Gigafide (tj. njenih taksonomskih kategorij) ter preostalega dela Gigafide, nato pa za vsako lemo izračunali njeno logaritemsko verjetnost, se pravi, da smo vsako posamezno taksonomsko kategorijo Gigafide primerjali s celotnim preostalim delom istega korpusa. V tem prispevku prikazana metoda tematskega modeliranja kot način primerjave *(opomba se nadaljuje na naslednji strani)*

4 SKLEP

Korpusno jezikoslovje je v iskanju načinov, kako sestaviti referenčni korpus, ki bi lahko veljal za trdno empirično osnovo, ki bi jo raziskovalci jezika proučili in nato na njeni podlagi posplošili svoje ugotovitve na celotni jezik, oblikovalo mrežo različnih meril, med katerimi sta tudi prenosnik ter besedilna tema oz. predmetno področje. V zvezi s prvim je nov izziv prinesel razvoj informacijsko-komunikacijskih tehnologij, posledica katerega je npr. ta, da tiskano časopisje kot vir novic že prehitevajo digitalni viri (splet, e-pošta, mobilni telefoni in socialna omrežja). Oblikovalci referenčnih korpusov to okoliščino upoštevajo različno. Kot je pokazal kratek pregled, skupna težnja, da bi se v referenčne korpuse vključevalo besedila z interneta in v kolikšnem obsegu bi to bilo, še ni jasno razvidna, če pa korpus že vsebuje ali bo vseboval besedila z interneta, se vanj v glavnem zajemajo besedila različnih žanrov.

Celovito poznavanje vsebine korpusa je mogoče pridobiti šele po zaključku njegove gradnje. Dober uvid v teme, ki jih pokrivajo korpusni dokumenti, daje metoda tematskega modeliranja. Na Gigafido smo jo aplicirali dvakrat, enkrat na njene taksonomske kategorije, drugič na celoto in hkrati primerjalno s slWaCom. V obeh primerih smo dobili podatke o tem, kaj v njej (in slWaCu) je, le omejeno (kolikor jih pač daje primerjava dveh entitet) pa podatke o tem, kaj v njej (oz. njem) manjka. Bolje torej razumemo, kakšen je vzorec, in imamo izhodišče za premislek, kaj bi v njem s tematskega vidika še lahko bilo. Vsekakor pa smo potrdili, da sta korpusa Gigafida in slWaC dokaj različna, kar navaja na sklepanje, da je v prihodnjih gradnjah referenčnih korpusov

dveh korpusov in metoda frekvenčnega profila, in smo jo uporabili v Erjavec, Logar Berginc (2012), se razlikujeta v tem, da je izhodišče pri tematskem modeliranju korpusa neodvisno, najprej namreč izračunamo teme v vsakem korpusu posebej, šele nato jih primerjamo med seboj. Na drugi strani pri primerjavi korpusov na osnovi razlike v pogostosti posameznih besed dobimo besedišče, ki je bolj značilno za en korpus, vendar to velja le v primerjavi s konkretnim drugim korpusom oz. podkorpusom, nato pa na tej osnovi sklepamo o značilnejših temah (oz. značilnejši vsebini) enega in drugega. Velja pa še povedati, da se v analizi Erjavec, Logar Berginc (2012) nismo omejili na samostalniške leme, temveč smo upoštevali vse leme ne glede na besedno vrsto.

smiselno združiti (čim več) tako besedil iz tiska kot besedil z interneta. V prihodnje bomo tematsko klasifikacijo, ki smo jo izdelali v raziskavi, uporabili še za klasifikacijo datotek v obeh obravnavanih korpusih (tj. izdelali bomo popis, kateri temi oz. temam pripada vsaka od vključenih datotek).

V zagovor nujnosti gradnje korpusov – takrat sicer korpusov *govorjenih* besedil – sta Stabej in Vitez leta 2000 zapisala: "dejstvo je, da je analitična slika nekega jezika, ki elemente zajema samo iz pisnih besedil, izrazito delna in nepopolna" (79). In dalje še: "če je idealni cilj korpusno podprtega jezikoslovja spoznavanje jezika, kot je izpričan v vseh razsežnostih sporazumevanja, je samo pisni korpus premalo" (80). Navedeno je mogoče oz. celo nujno prenesti na besedila, ki jih desetletje pozneje pišemo za "nove medije" in beremo na njih. Njihova vnaprejšnja opustitev iz korpusov, ki so osnova za jeziko(slo)vne opise jezika v vseh razsežnostih sporazumevanja in utemeljitve zanje, bi pomenila diskvalifikacijo pomembnega dela jezika. Še celo več – rastoči podatki o obsegu in dometu besedil, objavljenih na spletu (ali v digitalnem formatu, vendar na zaprtih platformah), zastavljajo obratno vprašanje: katera in kakšna so sploh še besedila, ki so dostopna le v tiskani obliki ter kakšen vpliv in ugled imajo? Kaže, da bodo tudi prihodnji "tradicionalni" korpusi postali (zlasti) korpusi digitalno dostopnih besedil, to pa je prihodnost, ki terja predvsem povsem nov premislek o njihovem uravnoteževanju, medtem ko bodo spletni korpusi nastajali še naprej in bodo brez težav postajali vse večji, a je za to, da bi bili podlaga jezikovnim opisom ter predpisom, njihova nestrukturiranost in precej manjša kontrola ter uvid nad tem, kaj smo vanje dobili izmed vsega, kar "je tam zunaj" (Atkins in dr. 2005: 96), trenutno vendarle še ovira.[20]

ZAHVALA

Avtorja se zahvaljujeta anonimnima recenzentoma za izredno koristne pripombe in predloge.

[20] A ne ovira, ki se je ne bi dalo odstraniti – prim. npr. poskus žanrske identifikacije za gradnjo referenčnega korpusa spletnih žanrov v Rehm in dr. 2008.

VIRI

Gigafida. Dostopno prek: http://www.gigafida.net.

slWaC. Dostopno prek: http://www.nljubesic.net/resources/corpora/slwac/; http://nl.ijs.si/.

LITERATURA

Arhar Holdt, Š., Kosem, I., in Logar Berginc, N. (2012): Izdelava korpusa Gigafida in njegovega spletnega vmesnika. V T. Erjavec, J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*: 16–21. Ljubljana: Institut Jožef Stefan.

Atkins, S., Kilgarriff, A., in Rundell, M. (2005): *Lexicom*. Brno: Masaryk University.

Baroni, M., Bernardini, S., Ferraresi, A., in Zanchetta, E. (2009): The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation,* 43 (3): 209–226.

Biber, D. (1993): Representativeness in Corpus Design. *Literary and Linguistic Computing,* 8 (4): 243–257.

Blei, D. M., Ng, A. Y., Jordan, M. I., in Lafferty, J. (2003): Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022.

Erjavec, T., Ignat, C., Pouliquen, B., in Steinberger, R. (2005): Massive Multi-Lingual Corpus Compilation: Acquis Communautaire and ToTaLe. V Z. Vetulani (ur.): *Proceedings of the 2nd Language & Technology Conference*: 32–36. Poznan.

Erjavec, T., Fišer, D., Krek, S., in Ledinek, N. (2010): The JOS Linguistically Tagged Corpus of Slovene. V N. Calzolari in dr. (ur.): *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*: 1806–1809. Valletta: European Language Resources Association (ELRA).

Erjavec, T., in Logar Berginc, N. (2012): Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES. V T. Erjavec, J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*: 57–62. Ljubljana: Institut Jožef Stefan.

Gorjanc, V. (2005): *Uvod v korpusno jezikoslovje*. Domžale: Založba Izolit.

Górski, R. L., in Łazinski, M. (2012): Typologia tekstów w NKJP. V A. Przepiórkowski, M. Bańko, R. L. Górski, B. Lewandowska - Tomaszczyk (ur.): *Narodowy Korpus Języka Polskiego*: 13–23. Warsaw: Wydawnictwo Naukowe PWN.

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

Kilgarriff, A., in Grefenstette, G. (2003): Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29 (3): 333–347.

Kosem, I. (2012): User-Friendly Concordancers for Corpora of Slovene. *Prace Filologiczne*, 63: 167–180.

Kupietz, M., Belica, C., Keibel, H., in Witt, A. (2010): The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. V N. Calzolari in dr. (ur.): *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*: 1848–1854. Valletta: European Language Resources Association (ELRA).

Ljubešić, N., in Erjavec, T. (2011): hrWac in slWaC: Compiling Web Corpora for Croatian and Slovene. V I. Habernal, V. Matoušek (ur.): *Text, Speech and Dialog: Proceedings of the 14th International Conference, TSD*: 395–402. Pilsen: Springer Berlin Heidelberg.

Ljubešić, N., Mikelić, N., in Boras, D. (2007): Language Identification: How to Distinguish Similar Languages. V: *Proceedings of the 29th*

International Conference on Information Technology Interfaces: 541–546. Zagreb: SRCE.

Logar Berginc, N., in dr. (2012): *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Pew Research Center (2010): Americans Spending More Time Following the News – *Ideological News Sources: Who Watches and Why*. Dostopno prek: http://www.people-press.org/.

Reynaert, M., in dr. (2010): Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus. V N. Calzolari in dr. (ur.): *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*: 2693–2698. Valletta: European Language Resources Association (ELRA).

Rayson, P., in Garside, R. (2000): Comparing Corpora Using Frequency Profiling. *Proceedings of the ACL Workshop on Comparing Corpora*: 1–6. Hong Kong.

Sharoff, S. (2010): Analysing Similarities and Differences between Corpora. V T. Erjavec, J. Žganec Gros (ur.): *Zbornik Sedme konference Jezikovne tehnologije*: 5–11. Ljubljana: Institut Jožef Stefan.

Stabej, M., in Vitez, P. (2000): KGB (korpus govorjenih besedil) v slovenščini. V T. Erjavec, J. Žganec Gros (ur.): *Zbornik konference Jezikovne tehnologije*: 79–81. Ljubljana: Institut Jožef Stefan.

Statistični urad RS (5. 10. 2012): *Uporaba informacijsko-komunikacijske tehnologije v gospodinjstvih in pri posameznikih, Slovenija, 2012: končni podatki*. Dostopno prek: http://www.stat.si/novica_prikazi.aspx?id=5037.

Steyvers, M., in Griffiths, T. (2007): Probabilistic Topic Models. V T. Landauer, D. S. McNamara, S. Dennis in W. Kintsch (ur.): *Handbook of Latent Semantic Analysis: A Road to Meaning*: 1–15. Hillsdale, NJ: Laurence Erlbaum.

SPLETNE STRANI

Tuji referenčni korpusi:

Cambridge English Corpus. Dostopno prek: http://www.cambridge.org/gb/elt/catalogue/subject/item2701617/Cambridge-International-Corpus/?site_locale=en_GB.

COCA: Corpus of Contemporary American English. Dostopno prek: http://corpus.byu.edu/coca/.

CORIS/CODIS: CORpus di Italiano Scritto. Dostopno prek: http://dslo.unibo.it/coris_eng.html.

CREA: Corpus de Referencia del Español Actual. Dostopno prek: http://ntlle.rae.es/nomina/jsp/NominaFor.jsp.

CRPC: Corpus de Referência do Português Contemporâneo. Dostopno prek: http://www.clul.ul.pt/pt/recursos/183-reference-corpus-of-contemporary-portuguese-crpc.

CSC: Suomen kielen tekstikokoelma. Dostopno prek: http://www.csc.fi/english/research/software/ftc.

Das Deutsche Referenzkorpus – DeReKo. Dostopno prek: http://www.ids-mannheim.de/kl/projekte/korpora/.

HNK: Hrvatski nacionalni korpus. Dostopno prek: http://www.hnk.ffzg.hr/.

KorpusDK. Dostopno prek: http://ordnet.dk/korpusdk/.

Narodowy korpus języka polskiego – NKJP. Dostopno prek: http://nkjp.pl/.

Oxford English Corpus. Dostopno prek: http://oxforddictionaries.com/words/the-oxford-english-corpus.

Reference Corpus of Estonian. Dostopno prek:
http://www.keeletehnoloogia.ee/projects-1/the-reference-corpus-of-the-estonian-language/comprehensive-corpus-of-estonian?set_language=et.

SNK: Slovenský národný korpus. Dostopno prek: http://korpus.juls.savba.sk/.

SoNaR: STEVIN Nederlandstalig Referentiecorpus. Dostopno prek:
http://lands.let.ru.nl/projects/SoNaR/.

SYN2010: Český národní korpus. Dostopno prek: http://ucnk.ff.cuni.cz/.

GIGAFIDA AND SLWAC: TOPIC COMPARISON

In the article, the following two issues are analyzed: (a) incorporation of texts from the Internet into existing reference corpora and comparison with the existence of web corpora, and (b) the latest two corpora of Slovenian language texts: the Gigafida corpus consisting mainly of printed texts and to a lesser extent also web texts, and the slWaC corpus which is entirely compiled from web texts. First, similarities and differences between the two corpora are identified using the topic modelling method, and then the same method is applied to the individual taxonomic categories of the Gigafida corpus. The first part of the analysis showed that the work of reference corpus compilers is currently still incoherent with regard to the incorporation of Internet texts into corpora which should reveal the overall picture of a certain language. In case compilers decide to incorporate web texts, the range of included genres is generally broad. The second part of the analysis showed a significant thematic variation between the Gigafida and slWaC corpora, and pointed out the most typical themes covered by each of the six Gigafida corpus parts.

Keywords: Slovenian language, reference corpus, Web corpus, topic modeling