cjvt Center za jezikovne vire in tehnologije

REPUBLIKA SLOVENIJA
MINISTRY OF CULTURE

CLARIN.SI

Specifikacije izdelave korpusa Gigafida 2.0 // Gigafida 2.0 Corpus Compilation: Specifications

Simon Krek, Špela Arhar Holdt, Jaka Čibej, Andraž Repar & Nikola Ljubešić

v1.0, 13. 6. 2019

cjvt

1 O projektu

Korpus Gigafida je referenčni korpus slovenščine, tj. zbirka slovenskih besedil najrazličnejših zvrsti, od dnevnih časopisov in revij do knjižnih publikacij vseh vrst (leposlovje, učbeniki, stvarna literatura). Prva različica korpusa Gigafida je nastala v okviru projekta Sporazumevanje v slovenskem jeziku (2008–2013, financerja Evropski socialni sklad in Ministrstvo za izobraževanje, znanost in šport), ki se je v večji meri posvečal gradnji pisnih korpusov, razvoju konkordančnikov za pisne korpuse ter programske opreme za strojno jezikoslovno označevanje korpusov. Poleg korpusa Gigafida so bili v projektu pripravljeni korpusi Kres, iz Gigafide vzorčeni uravnoteženi podkorpus, ter ccGigafida in ccKres, vzorčena odprto dostopna podkorpusa, namenjena razvoju jezikovnih tehnologij za slovenski jezik (Logar et al. 2012).

V sklopu projekta Sporazumevanje v slovenskem jeziku je bil pripravljen tudi uporabniško prijazen konkordančnik, ki raziskovanje jezika približa vsem zainteresiranim uporabniškim skupinam (ne le jezikoslovcem, ki potrebujejo specializirana orodja za delo s korpusnimi podatki). Gigafida je v konkordančniku za širšo rabo prosto na voljo za uporabo od leta 2012 naprej. Do konca leta 2018 je bilo s tem programom opravljenih več kot 2,2 milijona iskanj (pri čemer se štejejo samo iskanja uporabnikov, ki so potrdili sprejem piškotkov). Uporaba skozi leta ostaja stabilna oz. raste, kar nakazuje, da je širša javnost pripoznala korpus kot relevanten vir informacij o sodobni slovenščini. S tega vidika in z upoštevanjem, da je referenčni korpus temeljni vir podatkov za pripravo jezikovnih podatkovnih baz in jezikovnega opisa, je nujno zagotoviti njegovo redno posodabljanje in nadgrajevanje.

Prvo nadgradnjo korpusa Gigafida je omogočil projekt Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres, ki ga je financiralo Ministrstvo za kulturo v letih 2015–2018 v okviru pogodbe št. 33400-15-141007 med ministrstvom in Univerzo v Ljubljani, kjer je izvajalec Center za jezikovne vire in tehnologije. Ker je bil projekt manjšega obsega, je bila nadgradnja usmerjena v tiste segmente, ki lahko največ prispevajo glede na izhodiščno stanje, hkrati pa se opirajo na že izdelano infrastrukturo (konkordančniki, označevalniki itd.) ali uporabljajo orodja za obdelavo besedil, razvita po koncu projekta Sporazumevanje v slovenskem jeziku. V grobem je imel projekt tri cilje: usmerjeno zbiranje novih gradiv; strojna obdelava novih (in obstoječih) gradiv; javna dostopnost in diseminacija nadgrajenih korpusov.

Rezultat nadgradnje korpusa Gigafida je pripravljen v dveh različicah: Gigafida 2.0 je dvojno procesirana, s čimer je zagotovljena boljša reprezentativnost gradiva za raziskave sodobne standardne slovenščine: iz korpusa so odstranjena besedila, pri katerih je mogoče identificirati odstope od trenutnega jezikovnega standarda, obenem pa so odstranjena ponovljena besedila oz. besedilni fragmenti. Za razliko od tega različica Gigafida 2.0 Proto ni deduplicirana in je posledično gradivno obsežnejša. Obe različici korpusa sta na voljo v vrsti konkordančnikov za jezikoslovno rabo, prečiščena Gigafida 2.0 pa tudi v vmesniku za širšo rabo.

Specifikacije natančneje predstavljajo nadgradnjo oz. njene rezultate. Poleg korpusnih statistik in besedilne reprezentiranosti so predstavljene novosti na tehnični ravni: odstranjevanje podvojenih besedil, izboljšano jezikoslovno označevanje in izločitev besedil, ki se iz različnih razlogov odmikajo od jezikovnega standarda. Predstavljene so tudi spremembe v obliki in delovanju konkordančnika, namenjenega širši rabi.

2 Korpusne statistike

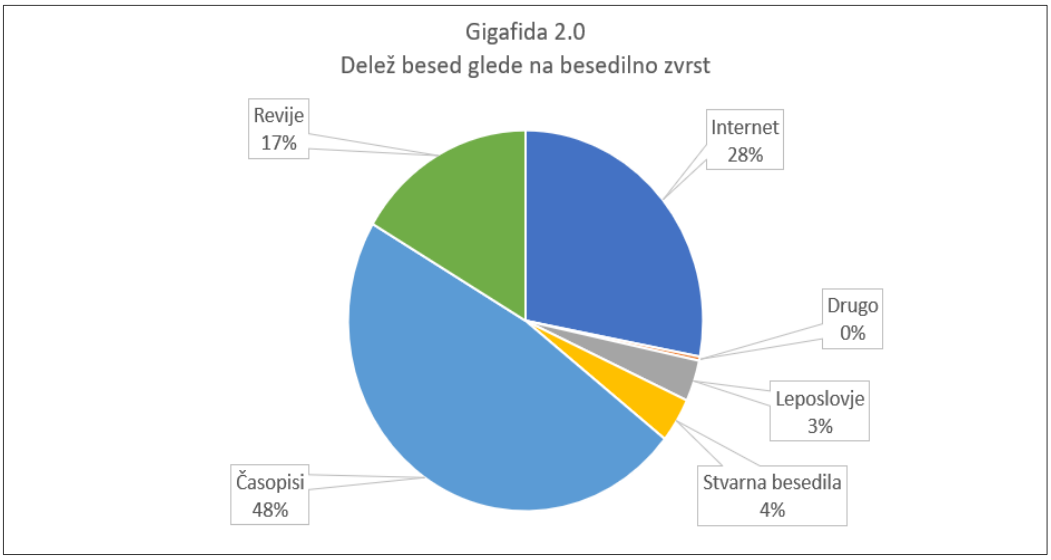
Gigafida 2.0 Proto vsebuje 38.365 besedil oz. 1,8 milijarde besed, kar je približno 29 % več kot Gigafida 1.0 (več o besedilih, ki so bila v korpus dodana na prehodu z različice 1.0 na različico 2.0, v poglavju 3). Korpus Gigafida 2.0 Proto je bil nato še dodatno obdelan: odstranjena so bila podvojena in zelo kratka besedila (glej poglavje 5). V primerjavi z različico 1.0 so bila odstranjena tudi besedila z nestandardnimi jezikovnimi prvinami (glej poglavje 6). Tako je nastala končna različica korpusa - Gigafida 2.0, ki vsebuje 38.310 besedil oz. skupno nekaj več kot 1,1 milijarde besed (1.134.693.933 besed). Velikost Gigafide 2.0 je tako primerljiva z različico 1.0 (1.186.999.699 besed).

Glede na besedilne zvrsti (Slika 1) korpus sestavljajo predvsem časopisi (47,8 % vseh besed), spletna besedila (28,0 %) in revije (16,5 %). Manjše deleže zajemajo še stvarna besedila (3,8 %), leposlovje (3,5 %) in drugo (0,3 %). Kot prikazuje Tabela 1, so deleži pri večini kategorij primerljivi z Gigafido 1.0. Največ odstopanj je pri spletnih besedilih, ki jih je v Gigafidi 2.0 za približno 12 odstotnih točk več. V novi različici je nekoliko manj časopisov (8 odstotnih točk) in revij (5 odstotnih točk manj), nekoliko večji pa je delež leposlovnih besedil (1,5 odstotne točke več). Na tej točki je treba omeniti, da pod spletna besedila štejemo novice, ki so bile v prejšnji različici prisotne le v kategoriji Časopisi, v Gigafidi 2.0 pa so pridobljene v digitalni obliki s pomočjo servisa IJS Newsfeed (glej poglavje 3) in zato

uvrščene pod spletna besedila. Dejanske razlike v razporeditvi besedilnih zvrsti so torej med različicama manjše, kot nakazujejo razporeditve deležev.

Besedilna zvrst	Gigafida 1.0	Gigafida 2.0
Internet	15,6 %	28,0 %
Časopisi	55,9 %	47,8 %
Revije	21,5 %	16,5 %
Stvarna besedila	4,2 %	3,8 %
Leposlovje	2,0 %	3,5 %
Drugo	0,7 %	0,3 %

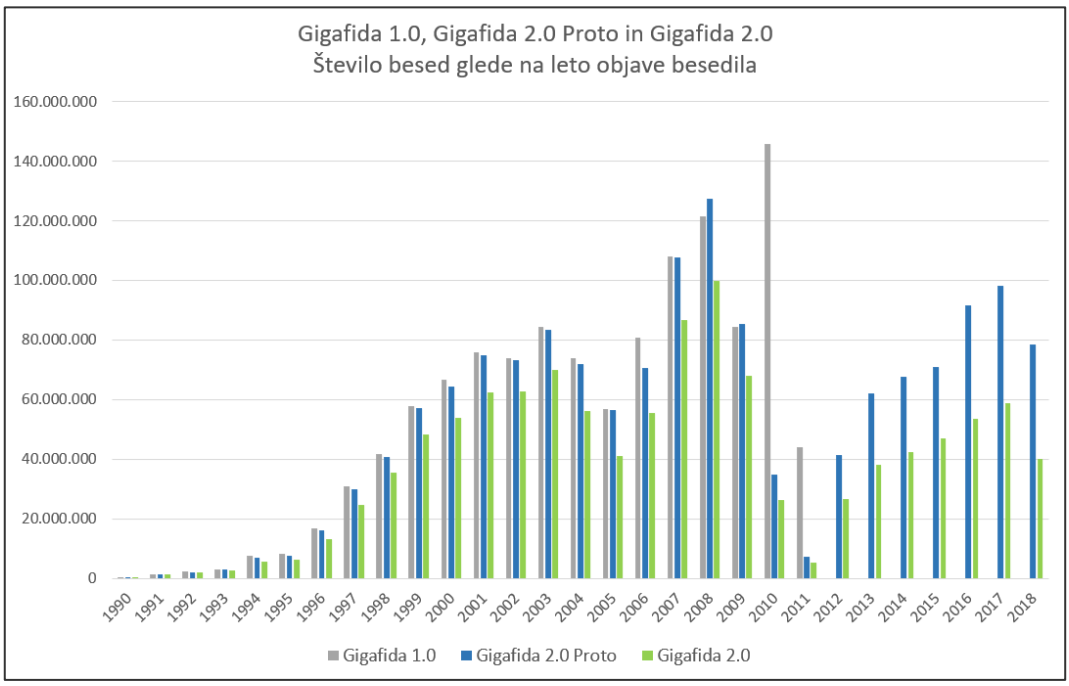
Tabela 1. Delež besed po besedilnih zvrsteh v Gigafidi 1.0 in 2.0.



Slika 1. Razporeditev besed po besedilnih zvrsteh v Gigafidi 2.0.

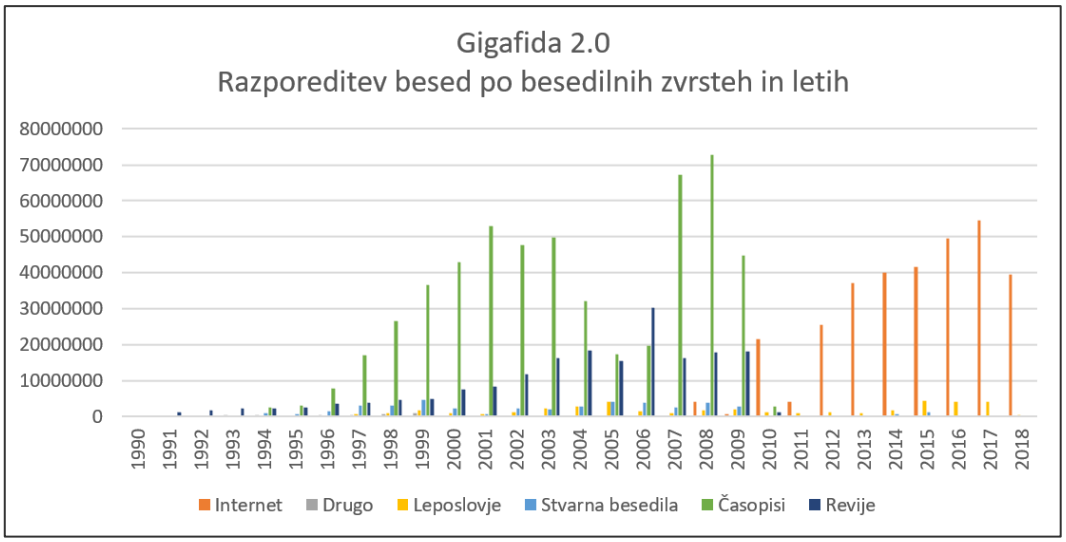
Kot prikazuje Slika 2, so bila v primerjavi z Gigafido 1.0 v različico 2.0 dodana besedila, ki so bila objavljena v letih med 2012 in 2018. Tako dodana besedila zajemajo približno 27 % celotne Gigafide 2.0. Največjo razliko med različicama predstavljajo spletna besedila iz let 2010 in 2011, ki jih je v novi različici zaradi deduplikacije ostalo 17 % oz. 10 %. Iz kategorije spletnih besedil za leti 2010 in 2011 so bila namreč odstranjena besedila z novičarskih portalov (npr. 24ur.com), ki so vsebovala tudi komentarje in forumska sporočila, napisana v nestandardni slovenščini. Ker nestandardnih delov besedila ni bilo mogoče jasno ločiti

od standardnih, so bila ta besedila odstranjena v celoti (več o odstranjevanju nestandardnega gradiva v poglavju 6).



Slika 2. Primerjava korpusov Gigafida 1.0, Gigafida 2.0 Proto in Gigafida 2.0 glede na vsebnost besed po letu objave.

Slika 3 prikazuje razporeditev besed v Gigafidi 2.0 po besedilnih zvrsteh in letih. Časopisna besedila in revije v korpusu zajemajo obdobje med letoma 1990 in 2010. Besedila, objavljena po letu 2010, so predvsem spletna, v primerjavi s prejšnjo različico pa je več tudi sodobnega leposlovja (zlasti iz let 2015, 2016 in 2017).



Slika 3. Razporeditev besed v Gigafidi 2.0 po besedilnih zvrsteh in letih.

3 Zbiranje gradiva in besedilodajalci

Ena od pomembnih značilnosti korpusa Gigafida 2.0 so urejena pravna razmerja z besedilodajalci, ki omogočajo javni dostop do vsebine in nadaljnjo distribucijo korpusov pod pogodbeno dogovorjenimi pogoji. Zbiranje besedil in urejanje pogodbenih razmerij z besedilodajalci predstavlja časovno in finančno velik napor, v projektu nadgradnje pa smo želeli ohraniti podobne možnosti za javno objavo in nadaljnjo distribucijo korpusov, kakršne so se kot dobra praksa uveljavile pri obstoječih korpusih. Zato se je projekt osredotočal na selektivno ciljno zbiranje novih gradiv glede na ugotovljene pomanjkljivosti obstoječih korpusov, ne pa na splošno zbiranje vseh gradiv po načelih, ki so bili uporabljeni pri gradnji korpusa Gigafida in njegovih predhodnikov, korpusov FIDA in FidaPLUS.

Projekt nadgradnje korpusa Gigafida (Krek et al. 2016) je imel na področju zbiranja besedil dva cilja, in sicer:

- besedila, za katera je bilo glede na tip, zvrst ali druge kriterije po analizi korpusov Gigafida in Kres ugotovljeno, da so podreprezentirana in
- besedila izbranih spletnih besedilodajalcev z večjo produkcijo (npr. novičarski portali, dnevni časopisi ipd.), ki zagotavljajo večjo aktualnost korpusnega gradiva.

V prvo kategorijo spadajo šolska gradiva (učbeniki, delovni zvezki ter sorodna gradiva) in leposlovna besedila, zlasti novejše izdaje ter tista besedila, ki imajo visoko branost glede

na podatke o knjižnični izposoji, v drugi kategoriji pa so bila zbrana besedila izbranih besedilodajalcev z največjo besedilno produkcijo.

Za potrebe zbiranja gradiv je bila pripravljena pogodba o prenosu ustreznih pravic z besedilodajalca na Univerzo v Ljubljani kot naročnika pogodbe in nosilca projekta nadgradnje korpusa. Pogodba določa neizključen, neodplačen ter časovno neomejen prenos pravic (pravica reproduciranja, vključno s pravico shranitve v elektronski obliki iz 23. člena Zakona o avtorski in sorodnih pravicah (ZASP) in pravico predelave teh del iz 33. člena ZASP) za celotna besedila za potrebe projekta nadgradnje korpusa Gigafida, poleg tega pa naročniku omogoča, da do 10 % posameznega dela izda pod pogoji licence Creative Commons 4.0 “priznanje avtorstva”, kar predstavlja osnovo za prosto dostopni korpus ccGigafida 2.0.

Na področju šolskih gradiv smo v korpus vključili 15 učbenikov in delovnih zvezkov, ki jih je prispevala založba Rokus-Klett (12 s področja zgodovine, 3 s področja slovenskega jezika), ter 32 e-učbenikov, ki jih je Zavod Republike Slovenije za šolstvo izdal pod licenco “Creative Commons 2.5. (priznanje avtorstva - nekomercialno - deljenje pod enakimi pogoji)” na spletnem mestu http://eucbeniki.sio.si/.

Na področju leposlovnih besedil smo pridobili 310 del, ki jih je prispevalo 9 različnih založb (med zbranimi deli so izvirna slovenska besedila in prevodi tujih del):

- Založba Mladinska knjiga: 121 del
- Založba Didakta: 51 del
- Založba Litera: 50 del
- Založba Beletrina: 30 del
- Kulturno-umetniško društvo Police Dubove: 23 del
- Nadja Dobnik (prevajalka): 16 del
- Založba Emanuel: 7 del
- Cankarjeva založba: 7 del
- Založba Kmečki glas: 1 delo

Na področju spletnih novičarskih besedil smo zbrali spletna besedila z naslednjih portalov:

- delo.si (od maja 2008 do novembra 2018)
- dnevnik.si (od maja 2008 do novembra 2018)
- rtvslo.si (od maja 2008 do novembra 2018)

- 24ur.com (od maja 2008 do novembra 2018)
- siol.net (od maja 2008 do novembra 2018)
- vecer.si (od maja 2008 do novembra 2018)
- zurnal24.si (od septembra 2009 do novembra 2018)
- sta.si (od novembra 2008 do novembra 2018)
- slovenskenovice.si (od novembra 2012 do novembra 2018)
- svet24.si (od februarja 2016 do novembra 2018)

Pogodbe o prenosu pravic smo podpisali s posameznimi portali oziroma matičnimi podjetji, ki so lastniki teh portalov. Besedila smo zbrali s pomočjo storitve IJS Newsfeed (Trampus in Novak, 2012), ki redno spremlja vire RSS novičarskih portalov in prenaša ter pretvarja članke v obliko, primerno za računalniško obdelavo.

4 Označevanje korpusnih besedil

Osnovno jezikoslovno označevanje korpusa Gigafida je bilo opravljeno na naslednjih ravneh: (1) tokenizacija in stavčna segmentacija, (2) oblikoskladenjsko označevanje in (3) lematizacija.

Pri označevanju korpusa smo razpolagali z dvema osnovnima cevovodoma: prvi je bil razvit v okviru projekta SSJ (Obeliks, glej Grčar et al. 2012, od tu dalje »cevovod SSJ«), drugi pa v okviru projekta CLARIN.SI (označevalnik ReLDI, glej Ljubešić in Erjavec 2016, od tu dalje »cevovod CLARIN.SI«). Glede na to, da je tokenizator Obeliks4J (tokenizator cevovoda SSJ) temeljil na pravilih, ki smo jih prav tako upoštevali pri tokenizaciji in stavčni segmentaciji slovenskega učnega podatkovnega niza ssj500k, in da je bil v uporabi tudi pri prejšnji različici korpusa Gigafida, smo se odločili, da ga ponovno uporabimo za tokenizacijo in stavčno segmentacijo.

Pri ostalih ravneh jezikoslovnega procesiranja izbira med obstoječima cevovodoma ni bila do te mere enoznačna: cevovod SSJ je namreč dosegal slabše rezultate na primerjalni platformi babushka-bench (pri označevalniku Obeliks je bila zabeležena točnost oblikoskladenjskih oznak (MSD) 92,67, točnost lematizacije pa 98,19) kot cevovod Clarin.si (pri označevalniku ReLDI je bila zabeležena točnost MSD 94,21 in točnost lematizacije 98,35). Vendar pa kljub temu, da oba cevovoda pri svojih osnovnih nalogah uporabljata algoritme strojnega učenja, cevovod SSJ vsebuje tudi dodaten modul, ki upošteva skrbno premišljena jezikoslovna pravila, medtem ko sledi cevovod Clarin.si paradigmi »vse je

možno«. Slednja pri jezikoslovcih, ki so osnovna ciljna skupina uporabnikov Gigafida korpusa, ni priljubljena. Zato smo razvili metaoznačevalnik, ki uporablja izhoda obeh cevovodov in se nato na podlagi obeh odloča, katero rešitev bo sprejel.

Metaoznačevalnik temelji na paradigmi nadzorovanega strojnega učenja z napovednim modelom, ki se je učil na 10.000 pojavnicah, pri katerih sta se izhoda obeh označevalnikov razlikovala, obenem pa sta bili za posamično pojavnico znani tako pravilna MSD-oznaka kot tudi lema. Na podlagi podatkov o napovedani MSD-oznaki, besedni vrsti (prvi dve črki MSD-oznake) in oblikoskladenjskih značilnosti (pridobljenih iz MSD-oznake) obeh označevalnikov, se metaoznačevalnik odloči, katera MSD-oznaka in lema sta najverjetneje pravilni. Med razvojem metaoznačevalnika smo prav tako preizkusili uporabo različnih površinskih oblik in njihovih pripon kot tudi značilnosti pridobljenih iz potencialnih lem, vendar pa ti podatki niso izboljšali rezultatov metaoznačevalnika. Poleg tega se je metaoznačevalnik, ocenjen na osnovi primerjalne platforme babushka-bench, izkazal za na splošno bolj natančnega kot obe posamični rešitvi, z natančnostjo MSD 94,34 in natančnostjo lematizacije 98,66; gre za 2 % manjšo relativno napako kot pri najboljši posamični rešitvi pri oblikoskladenjskem označevanju in 19 % manjšo pri lematizaciji. Še večji dosežek pa je, da smo z uporabo metaoznačevalnika uspeli združiti jezikoslovne restrikcije cevovoda SSJ, ki so pri uporabnikih bolj priljubljene, z višjo natančnostjo cevovoda Clarin.si.

V Tabeli 2 je podanih 20 najpogostejših napak, ki jih je naredil metaoznačevalnik, od skupno 2.773 napak, ki jih je platforma babushka-bench zabeležila na testnih podatkih. Analiza je pokazala, da gre pri večini napak za zamenjavo tožilnika in imenovalnika pri samostalnikih srednjega in moškega spola (10-odstotni delež vseh napak), in za zamenjavo pridevnikov in prislovov, kar predstavlja 2 % vseh napak. Ti že tako majhni deleži kažejo, da gre za porazdelitev napak z dolgim repom in raznovrstnimi oblikami napak.

	Pravilna oznaka	Strojno označeno	Število napak	Razvezana pravilna oznaka	Razvezana napačna oznaka
1	Sometn	Somei	109	samostalnik vrsta=občno_ime spol=moški število=ednina sklon=tožilnik živost=ne	samostalnik vrsta=občno_ime spol=moški število=ednina sklon=imenovalnik

2	Somei	Sometn	71	samostalnik vrsta=občno_ime spol=moški število=ednina sklon=imenovalnik	samostalnik vrsta=občno_ime spol=moški število=ednina sklon=tožilnik živost=ne
3	Soset	Sosei	61	samostalnik vrsta=občno_ime spol=srednji število=ednina sklon=tožilnik	samostalnik vrsta=občno_ime spol=srednji število=ednina sklon=imenovalnik
4	Ppnsei	Rsn	41	pridevnik vrsta=splošni stopnja=nedoločeno spol=srednji število=ednina sklon=imenovalnik	prislov vrsta=splošni stopnja=nedoločeno
5	Sozmt	Sozer	38	samostalnik vrsta=občno_ime spol=ženski število=množina sklon=tožilnik	samostalnik vrsta=občno_ime spol=ženski število=ednina sklon=rodilnik
6	Sosei	Soset	35	samostalnik vrsta=občno_ime spol=srednji število=ednina sklon=imenovalnik	samostalnik vrsta=občno_ime spol=srednji število=ednina sklon=tožilnik
7	Sozmi	Sozer	32	samostalnik vrsta=občno_ime spol=ženski število=množina sklon=imenovalnik	samostalnik vrsta=občno_ime spol=ženski število=ednina sklon=rodilnik
8	Sozmi	Sozmt	31	samostalnik vrsta=občno_ime spol=ženski število=množina sklon=imenovalnik	samostalnik vrsta=občno_ime spol=ženski število=množina sklon=tožilnik
9	Ppnset	Ppnsei	31	pridevnik vrsta=splošni stopnja=nedoločeno spol=srednji	pridevnik vrsta=splošni stopnja=nedoločeno spol=srednji

				število=ednina sklon=tožilnik	število=ednina sklon=imenovalnik
10	Slmetd	Slmer	27	samostalnik vrsta=lastno_ime spol=moški število=ednina sklon=tožilnik živost=da	samostalnik vrsta=lastno_ime spol=moški število=ednina sklon=rodilnik
11	Rsn	Ppnsei	25	prislov vrsta=splošni stopnja=nedoločeno	pridevnik vrsta=splošni stopnja=nedoločeno spol=srednji število=ednina sklon=imenovalnik
12	Sozer	Sozmt	25	samostalnik vrsta=občno_ime spol=ženski število=ednina sklon=rodilnik	samostalnik vrsta=občno_ime spol=ženski število=množina sklon=tožilnik
13	Sozmt	Sozmi	25	samostalnik vrsta=občno_ime spol=ženski število=množina sklon=tožilnik	samostalnik vrsta=občno_ime spol=ženski število=množina sklon=imenovalnik
14	Ppnmetd	Ppnmeid	25	pridevnik vrsta=splošni stopnja=nedoločeno spol=moški število=ednina sklon=tožilnik določnost=da	pridevnik vrsta=splošni stopnja=nedoločeno spol=moški število=ednina sklon=imenovalnik določnost=da
15	Ppnmeid	Ppnmetd	24	pridevnik vrsta=splošni stopnja=nedoločeno spol=moški število=ednina sklon=imenovalnik določnost=da	pridevnik vrsta=splošni stopnja=nedoločeno spol=moški število=ednina sklon=tožilnik določnost=da
16	Rsn	Vp	23	prislov vrsta=splošni stopnja=nedoločeno	veznik vrsta=priredni
17	Ppnzmi	Ppnzer	23	pridevnik vrsta=splošni	pridevnik vrsta=splošni

				stopnja=nedoločeno spol=ženski število=množina sklon=imenovalnik	stopnja=nedoločeno spol=ženski število=ednina sklon=rodilnik
18	Slmei	Slzei	22	samostalnik vrsta=lastno_ime spol=moški število=ednina sklon=imenovalnik	samostalnik vrsta=lastno_ime spol=ženski število=ednina sklon=imenovalnik
19	Dm	Dt	21	predlog sklon=mestnik	predlog sklon=tožilnik
20	Somei	Slmei	21	samostalnik vrsta=občno_ime spol=moški število=ednina sklon=imenovalnik	samostalnik vrsta=lastno_ime spol=moški število=ednina sklon=imenovalnik

Tabela 2. Najpogostejše označevalne napake glede na opravljeno evalvacijo.

5 Deduplikacija in filtriranje

Raba korpusa Gigafida po objavi je pokazala, da bi bilo smiselno proces odstranjevanja dvojnikov izvesti tudi na obstoječih besedilih iz preteklih različic korpusa, saj se v besedilih, ki izhajajo iz tiskanih medijev, pogosto pojavljajo ponavljajoči se deli besedil, ki izkrivljajo statistične podatke pri poizvedbah po celotnem korpusu (Logar et al., 2015). Tipičen primer takih besedil so radijski in televizijski programi, ki so z isto vsebino objavljeni v različnih virih, podobno tudi identične (časopisne) objave, povzete po istem viru in podobno. V procesu priprave korpusa Gigafida 2.0 je bil proces deduplikacije izveden tudi na vseh besedilih, ki so bila pred tem vključena v korpus FIDA, FidaPLUS in prvo verzijo korpusa Gigafida, v zaporedju Gigafida 1.0 -> na novo zbrana besedila. Deduplikacija je bila izvedena na ravni odstavkov s pomočjo orodja Onion (ONe Instance ONly, Pomikálek 2011), ki omogoča nastavitve dveh parametrov -n in -t. Parameter -n določa dolžino identičnih n-gramov (n-gram length), ki jih program upošteva pri deduplikaciji, parameter -t določa delež identičnih n-gramov v odstavku (duplicate content threshold). Odstavek je opredeljen kot duplikat, če presega prag števila n-gramov, kot je določeno s parametrom -t, ki so bili že zabeleženi v predhodnih besedilih. Parameter -t ima lahko vrednosti med 0 in 1, pri čemer vrednost 1 pomeni, da gre za 100 % identičnost odstavka. Nastavitve, ki smo

jih uporabili, so bile: -n = 9, -t = 0,5, s čimer je bil korpus zmanjšan za 24,9 % (prim. Benko 2013).

Poleg odstavčne je bila deduplikacija izvedena tudi na ravni celotnih besedil. Če je bilo v besedilu več kot 95 % podvojenih odstavkov, je bilo kot duplikat označeno celotno besedilo in izločeno iz dedupliciranega korpusa. Poleg tega je bilo iz celote odstranjenih tudi 6.915 besedil, ki (1) niso vsebovala č, š, in ž, s čimer smo odstranili večino besedil, pri katerih so bile težave s uporabo napačnih kodnih tabel ali avtorji niso uporabljali standardnega zapisa (gl. poglavje 6), ali (2) so vsebovala manj kot 500 znakov. Besedila z manj kot 500 znaki so bila odstranjena z namenom, da se zmanjša število datotek (z metapodatki), ki brez dobrega razloga podaljšujejo postopke procesiranja korpusa.

6 Standardno in nestandardno gradivo

Do različice 2.0 je korpus Gigafida vseboval tudi gradivo, za katerega je bilo znano in predvideno, da so v njem prisotne nestandardne jezikovne prvine, tj. da se v njem kažejo odkloni od standardne slovenščine na ravni zapisa, besedišča, skladnje in sloga. V to kategorijo spadajo zlasti uporabniške spletne vsebine, npr. komentarji, forumska sporočila in druga spletna besedila, v nekaterih primerih pa tudi leposlovje (dober primer je denimo prevod romana Trainspotting) in časopisi (npr. lokalni mediji, (delno) pisani v regionalni jezikovni različici slovenščine; primer je glasilo zamejskih Slovencev v Italiji – Novi Matajur).

Tovrstno gradivo je bilo v korpusu Gigafida v različici 2.0 v čim večji meri minimizirano, in sicer iz več razlogov. Prvič, v času od objave Gigafide 1.0 so bili za raziskovanje nestandardne slovenščine zgrajeni drugi, primernejši viri, v prvi vrsti korpus spletne slovenščine JANES 1.0 (Fišer et al. 2016), ki vsebuje slovenske tvite, forumska sporočila, blogovske zapise in komentarje na novice (skoraj 253 milijonov pojavnic). Drugič, delež tovrstnih besedil je bil v korpusu majhen in zato nereprezentativen za raziskave nestandardnih jezikovnih prvin. In tretjič, korpusna besedila niso bila opremljena z metapodatki o stopnji standardnosti besedil, zato jih je bilo iz poizvedb mogoče izločati le pogojno na podlagi naslovov, podatkov o tipu vira ipd.

Osnovno izhodišče za izločanje besedil s stališča (ne)standardnosti je bilo predvidevanje, ali so njihovi avtorji imeli intenco, ali bili pod družbenim pritiskom, da ustvarijo besedilo v standardni slovenščini, pri čemer standardnost obravnavamo z izhodišč, opisanih v Krek 2015. To v osnovi sociolingvistično izhodišče je bilo preverjeno tudi empirično.

Identifikacijo nestandardnih besedil smo opravili strojno, in sicer z orodjem za avtomatsko določanje stopnje nestandardnosti besedil (Ljubešić et al. 2015), ki je bilo razvito v sklopu projekta JANES. Orodje besedilu pripiše vrednost med 1 (standardno) in 3 (zelo nestandardno), in sicer na dveh nivojih: tehnična nestandardnost (T) upošteva predvsem oblikovno ustreznost besedila (npr. ustrezna raba presledkov, ločil in velikih začetnic), jezikovna nestandardnost (L) pa standardnost besedila glede na besedišče in zapis. Osredotočili smo se na jezikovno nestandardnost besedil.

Orodje je bilo naučeno na ročno označenem naboru kratkih spletnih besedil, kot so tviti, forumska sporočila, in komentarji na novice. Gigafida 2.0 vsebuje daljša besedila, zato smo stopnjo nestandardnosti pripisali vsakemu odstavku posebej. S tem smo dobili natančnejšo sliko o distribuciji nestandardnega znotraj besedila: vsakemu besedilu je bil torej pripisan vektor stopenj nestandardnosti, ki je vseboval vrednosti za vsakega od odstavkov v besedilu.

Razporeditev vrednosti v besedilu smo nato primerjali z razporeditvijo vrednosti v celotnem korpusu s pomočjo statističnega testa Kolmogorova-Smirnova, s katerim je mogoče preveriti, ali se razporeditev vrednosti v vzorcu statistično veljavno razlikuje od razporeditve v celotnem korpusu. Vsa besedila, katerih razporeditev vrednosti se je statistično veljavno razlikovala od razporeditev v korpusu in ki so obenem imela v povprečju višjo vrednost jezikovne nestandardnosti, so bila dodana na seznam problematičnih besedil. Na ta način pridobljeni nabor je bil nato ročno pregledan, izbrana problematična besedila pa so bila iz korpusa izločena.

7 Korpusni konkordančnik

Kot je omenjeno v uvodu, sta tako Gigafida 2.0 kot Gigafida 2.0 Proto na voljo v različnih korpusnih orodjih. Na eni strani sta obe korpusni različici vključeni v NoSketch Engine, ki je nameščen na Odseku za tehnologije znanja na Institutu Jožef Stefan, ter v konkordančniku KonText, ki je na voljo v sklopu infrastrukture Clarin.si. Dogovorjena je tudi umestitev v zmogljivo orodje SketchEngine. Korpus Gigafida 2.0 poleg tega ostaja dostopen tudi v uporabniško prijaznem konkordančniku za širšo rabo.

Gigafida 1.0 je (bila) na voljo v konkordančniku, ki je bil razvit v projektu Sporazumevanje v slovenskem jeziku. Kot je natančneje opisano v Logar et al. (2012: 98–118), je bil ta program zasnovan z namenom, da bi bila korpusna raba dostopna vsem zainteresiranim

uporabniškim skupinam, mdr. so to učitelji, lektorji, prevajalci, pisci besedil za različno rabo itd. Ob razvoju funkcionalnosti in podobe konkordančnika so bila upoštevana spoznanja predhodnih uporabniških študij (Arhar Holdt 2009), novosti pa so bile nekaj let po lansiranju podrobneje ocenjene v posebni uporabniški evalvaciji. Slednja je pokazala, da ciljne uporabniške skupine značilnosti konkordančnika ocenjujejo kot pozitivne, kot glavni prednosti pa sta bila izpostavljena preprosto iskanje po korpusnih podatkih ter izčiščenost vmesnika (Arhar Holdt et al. 2019).

Konkordančnik, v katerem je na voljo Gigafida 2.0, omogoča uporabnikom, da korpus uporabljajo, kot so navajeni od prej. Skladno z rezultati uporabniške evalvacije so ohranjene so vse glavne funkcionalnosti: zavihki za iskanje po konkordancah, kolokacijah in za izdelavo besednih seznamov; postopna vmesniška navigacija, ki jo omogoča medsebojna povezanost podatkov (npr. prehajanje iz seznama kolokatorjev v konkordančni niz); preprosto zasnovano enostavno in napredno iskanje; interaktivni podatkovni filtri; dostop do zgodovine iskanj; možnost izvoza podatkov za nadaljnjo obdelavo v drugih programih.

Čeprav projektni okvir nadgradnje programa ni predvideval, so bile v izvedbo vključene manjše spremembe vmesnika, namenjene lažjemu iskanju in boljšemu pregledu nad rezultati: (a) gumbi za preklop med različnimi načini iskanja so v novem vmesniku bližje iskalnega okenca in zato bolj pri roki, (b) pri preklopu med zavihki *Iskanje, Okolica* in *Seznam* se izvede samodejna priprava podatkov za izbrani iskalni pogoj, kar omogoča hitrejše delo, (c) podatki o pogostosti so v filtrih razporejeni na preglednejši način kot v predhodni različici vmesnika, (d) filtri že v prvem koraku omogočajo izbiro dveh ali več kategorij hkrati, kot tudi preprosto ugašanje izbranih filtrov, (e) navigacija po straneh konkordančnega niza omogoča neposreden skok na izbrano stran, (f) v zavihku Okolica je namesto statistike *MI* na voljo relevantnejša *logdice* in (g) dodana sta gumba za deljenje vsebin na družbenih omrežjih. Čeprav je vmesnik kot jezikovno orodje primarno razvit za uporabo na računalniških zaslonih, je bila pozornost namenjena tudi postavitvi za tablice in pametne telefone, ki mora biti ustrezno prilagojena prostorskim omejitvam.

Pomembnejša novost je oblikovna in podatkovna vključitev korpusa Gigafida 2.0 v portal viri.cjvt.si. Slednji uporabnikom omogoča enostavnejše povezovanje med jezikovnimi podatki različnih vrst. S klikom na poseben gumb ob iskalnem okencu uporabnik dobi možnost, da izbrani iskalni pogoj enostavno in hitro poišče tudi v drugi virih CJVT, npr. Slovarju sopomenk sodobne slovenščine in Kolokacijskem slovarju sodobne slovenščin

8 About the project

Gigafida is a reference corpus of Slovene, i.e. a collection of texts of various genres encompassing daily newspapers, magazines, and numerous types of publications (fiction, textbooks, non-fiction). The first version of Gigafida was completed within the framework of the Communication in Slovene project (2008-2013; financed by European social funds and the Ministry of Education, Science and Sport of the Republic of Slovenia). The project was predominantly concerned with compiling written corpora and developing concordancers for written corpora and software for automatic linguistic annotation of corpora. Apart from Gigafida, the project also delivered the Kres corpus (a balanced sub-corpus sampled from Gigafida), as well as the ccGigafida and ccKres corpora, open-access sampled sub-corpora designed for the further development of language technologies for Slovene (Logar et al. 2012).

The *Communication in Slovene* project also delivered a user-friendly concordancer, which facilitates language research not only for linguists, but for all interested user groups who need specialised tools for processing corpus data. The Gigafida web concordancer is intended for general use and has been openly accessible since 2012. By the end of 2018, more than 2.2 million searches have been conducted with the programme (taking into account only the searches made by users who accepted the use of cookies). Its use has been steadily increasing, which indicates that the general public has recognised the corpus as a relevant source of information on modern Slovene. Because of this and the fact that the reference corpus serves as the basic source of information for language description and the development of language databases, it is vital to ensure that it is regularly updated and upgraded.

The first Gigafida upgrade was made possible by The Upgrade of Corpora Gigafida, Kres, ccGigafida and ccKress, a project financed by the Ministry of Culture of the Republic of Slovenia in the years 2015-2018 under contract no. 33400-15-141007 signed by the Ministry and the University of Ljubljana, with the Centre for Language Resources and Technologies as the main contractor. Due to the relatively small scope of the project, the upgrade focused on segments which would make the greatest possible impact in relation to the initial state while relying on existing infrastructure (concordancers, taggers, etc.) and language processing tools developed after the conclusion of the *Communication in Slovene* project. Broadly speaking, the project had three aims: automatic processing of new (and existing) materials; provision of open access; and the dissemination of upgraded corpora.

The Gigafida upgrade resulted in two versions: Gigafida 2.0 has been processed in two aspects in order to ensure that it is more representative for research on modern standard Slovene. First, the corpus was cleaned of texts in which deviations from current language norms could be identified. Second, duplicate texts and texts fragments were removed from the corpus. Unlike Gigafida 2.0, Gigafida 2.0 Proto has not been deduplicated and is thus slightly larger in size. Both corpus versions are available through a number of concordancers intended for linguistic use; the consolidated version of Gigafida 2.0 is also accessible through a general use interface.

The specifications give a more precise overview of the upgrade and its results. Presented below are corpora statistics and text type distribution as well as the latest technical developments: removal of duplicate texts, improved linguistic annotation, and elimination of texts deviating from the linguistic standard. Also described are the changes in form and functioning of the concordancer for general use.

9 Corpus statistics

Gigafida 2.0 Proto contains 38,364 texts or 1.8 billion words, an increase of 29% compared to Gigafida 1.0 (see Section 10 for more about texts added to the corpus in the transition from version 1.0 to 2.0). The Gigafida 2.0 Proto corpus was further processed by removing duplicate texts and very short texts (see Section 12). In comparison to version 1.0, all texts with non-standard linguistic elements were also eliminated (see Section 13). This resulted in the final version of the corpus, Gigafida 2.0, which contains 38,310 texts or somewhere in excess of 1.1 billion words (1,134,693,333 words). In terms of size, Gigafida 2.0 is thus comparable to Gigafida 1.0 (1,186,999,699 words).

In terms of genre (Figure 1), the corpus is mainly comprised of newspapers (47.8% of all words), online texts (28.0%), and periodicals (16.5%). A smaller percentage of texts belong to the categories of non-fiction (3.8%), fiction (3.5%), and other (0.3%). As seen in Table 1, the percentages are comparable to Gigafida 1.0. The greatest discrepancy occurs with online texts, with Gigafida 2.0 containing about 12 percentage points more. The new version has fewer newspaper texts (a decrease of 8 percentage points) and periodicals (a decrease of 5 percentage points) and a slightly larger percentage of fiction (an increase of 1.5 percentage points). However, it should be mentioned that online texts also include news: in the previous version, these were only included in the *Newspapers* category. In Gigafida 2.0, they were obtained in digital form through the IJS Newsfeed (see Section 10),

and were thus categorised as online texts. The actual differences in the distribution of genres between the two versions are thus smaller than suggested by the percentages.

Text type	Gigafida 1.0	Gigafida 2.0
Internet	15.6 %	28.0 %
Newspapers	55.9 %	47.8 %
Periodicals	21.5 %	16.5 %
Non-fiction	4.2 %	3.8 %
Fiction	2.0 %	3.5 %
Other	0.7 %	0.3 %

Table 1. Text type distribution in Gigafida 1.0 and 2.0.

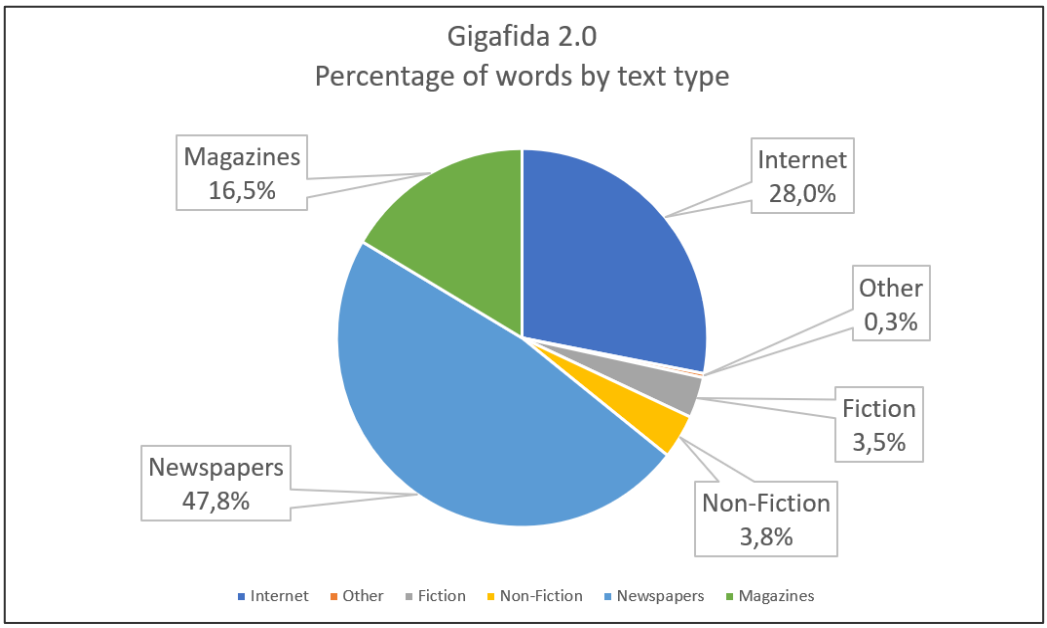


Figure 1. Distribution of words by text type in Gigafida 2.0.

As seen in Figure 2, unlike Gigafida 1.0, Gigafida 2.0 also contains texts published between the years 2012 and 2018. These texts represent about 27% of the Gigafida 2.0 corpus. The biggest discrepancy between the two versions lies in the percentage of online texts from the years 2010 and 2011. Only 17% and 10% of these texts (for 2010 and 2011, respectively) remain in the new version, which is mainly due to the fact that texts published by online news portals (e.g. 24ur.com) were removed from the Internet category because they

included comments and forum messages written in non-standard Slovene. Because non-standard segments could not be clearly distinguished from standard ones, these texts were removed in their entirety (more about elimination of non-standard texts in Section 13).

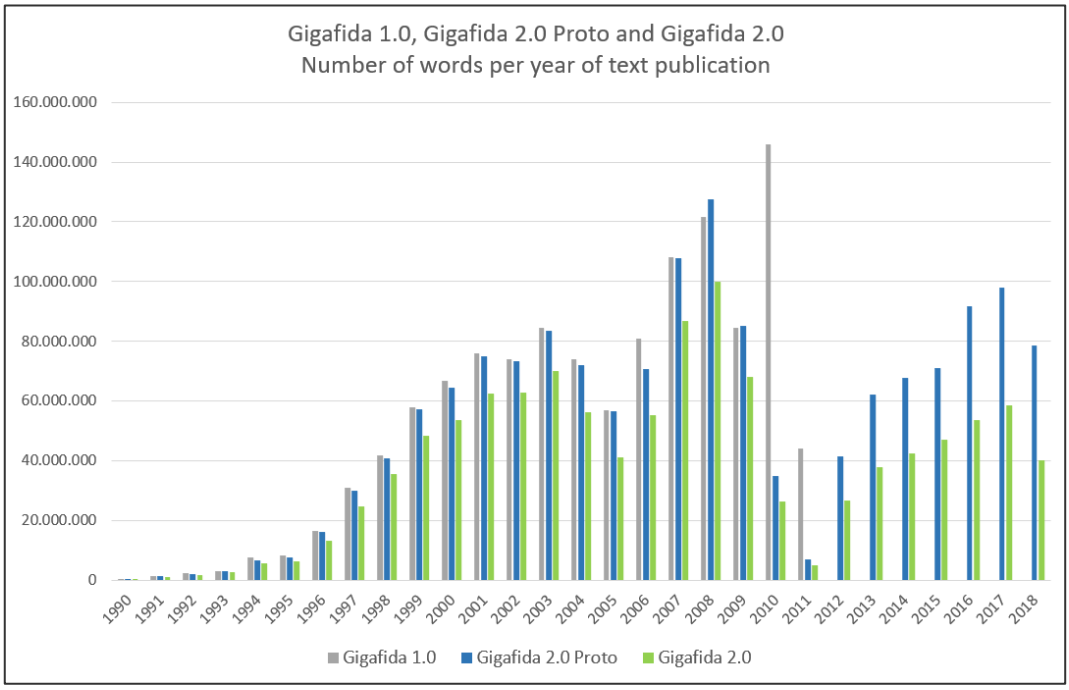


Figure 2. Comparison of Gigafida 1.0, Gigafida 2.0 Proto and Gigafida 2.0 by number of words per year of publication.

Figure 3 represents the distribution of words in Gigafida 2.0 by genre and year of publication. Newspaper and periodical articles in the corpus cover the period from 1990 to 2010. Texts published after 2010 mainly include online texts; compared to the previous version, there is also a greater share of contemporary fiction (particularly for the years 2015, 2016, and 2017).

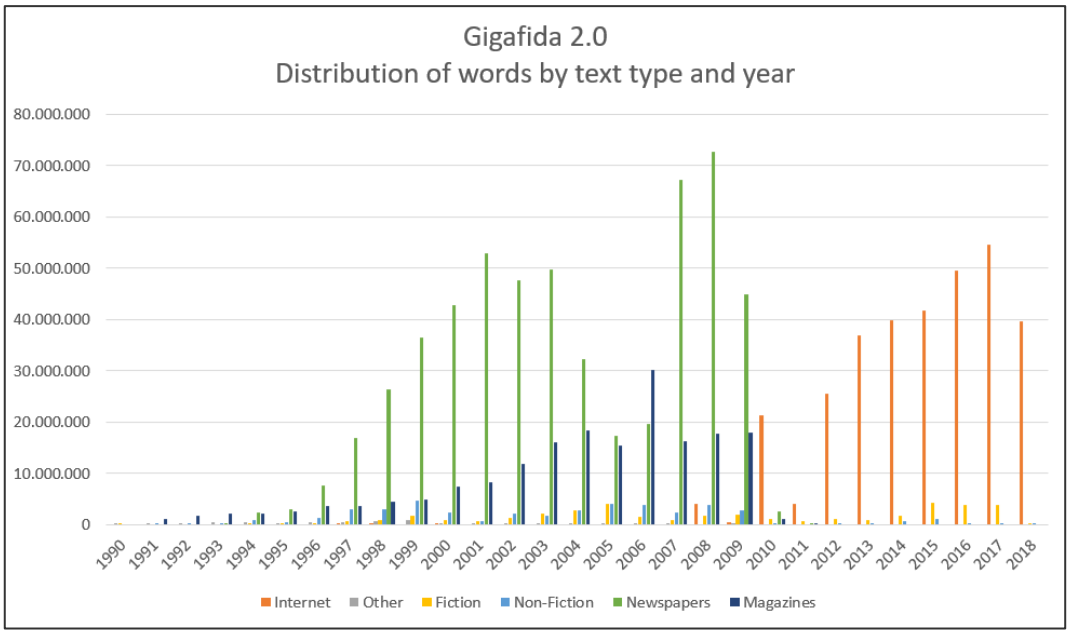


Figure 3. Distribution of words in Gigafida 2.0 by genre and year.

10 Text providers and Material Collection

One of the most important aspects of Gigafida 2.0 is its formalised legal relationship with text providers, which safeguards public access and the further distribution of corpora under the conditions stipulated in the relevant agreements. Collecting materials and formalising contractual relationships with text providers is both financially and time-consuming; however, we wished to retain conditions for publication and further distribution of corpora that had been previously established as good practice in existing corpora. This is why the project focused on selective, targeted collection of new materials chosen on the basis of lacunae in existing corpora, as opposed to general collection of materials according to principles applied in the compilation of the Gigafida corpus and its predecessors, FIDA and FidaPLUS.

The Gigafida upgrade (Krek et al. 2016) had two main goals in collecting material:

- compiling texts under-represented in terms of type, genre or other criteria based on the analysis of Gigafida and Kres;

- compiling texts by selected text providers with a large volume of production (e.g. news portals, daily newspapers and similar), which ensures that the corpus material is up to date.

The first category contains school materials (textbooks, workbooks etc.) and fiction, especially newer editions, and texts popular with readers, as determined on the basis of library lending figures; the second category includes texts by selected text providers with the greatest volume of text production.

For the purposes of collecting the materials, a contract was drafted transferring pertinent rights from text providers to the University of Ljubljana as the contracting authority and the principal project promoter of the corpus upgrade. The contract stipulates a non-exclusive, non-remunerated and perpetual transfer of rights (the right to reproduce, including the right to electronic storage, as stipulated in paragraph 23 of the Copyright and Related Rights Act (ZASP), and the right to adapt, as defined in article 33 of the same Act) to entire texts for the needs of the Gigafida project; it also allows the contracting authority to publish up to 10% of an individual work under the Creative Commons 4.0 Attribution Licence, which is the basis for open access to ccGigafida 2.0.

In terms of educational material, the corpus includes 15 textbooks and workbooks provided by Rokus-Klett Publishing (12 textbooks on history and 3 on the Slovene language) and 32 e-textbooks published by the National Education Institute of the Republic of Slovenia under the Creative Commons Licence 2.5. (attribution-non-commercial-share alike) at http://eucbeniki.sio.si/.

In the field of fiction, 310 works were collected, contributed by 9 different publishers (they include both original works in Slovene and translations of foreign language works into Slovene):

- Mladinska knjiga Publishing: 121 works
- Didakta Publishing House: 51 works
- Litera Publishing House: 50 works
- Beletrina Academic Press: 30 works
- Polica Dubova Publishing House: 23 works
- Nadja Dobnik (translator): 16 works
- Emanuel Publishing: 7 works

- Cankarjeva Publishing House: 7 works
- Kmečki glas Publishing House: 1 work

With online news, online texts were collected from the following portals:

- delo.si (from May 2008 to November 2018)
- dnevnik.si (from May 2008 to November 2018)
- rtvslo.si (from May 2008 to November 2018)
- 24ur.com (from May 2008 to November 2018)
- siol.net (from May 2008 to November 2018)
- vecer.si (from May 2008 to November 2018)
- zurnal24.si (from September 2009 to November 2018)
- sta.si (from November 2008 to November 2018)
- slovenskenovice.si (from November 2012 to November 2018)
- svet24.si (from February 2016 to November 2018)

Copyright transfer agreements were signed with individual portals or their parent companies. The texts were collected with the aid of the IJS Newsfeed (Trampus and Novak, 2012), which regularly crawls RSS feeds of various news portals and converts articles into formats that facilitate digital processing.

11 Annotation of corpus texts

The basic linguistic processing of the Gigafida corpus was performed on the following levels: (1) tokenization and sentence segmentation, (2) morphosyntactic annotation and (3) lemmatization.

At the time of corpus annotation, we had two major pipelines at our disposal: the first was developed within the SSJ project (Obeliks, Grčar et al. 2012, hereafter "SSJ pipeline"), and another developed within CLARIN.SI (reldi-tagger, Ljubešić and Erjavec 2016, hereafter "CLARIN.SI pipeline"). Given that the Obeliks4J tokenizer (the tokenizer of the SSJ project's pipeline) was based on rules that were also followed during the tokenization and sentence segmentation of the Slovene training dataset ssj500k, and that this was the tokenizer that was used in the previous iteration of the Gigafida corpus, we decided to stick with this solution for tokenization and sentence segmentation.

On the remaining levels of linguistic processing, the decision between the two existing pipelines was less obvious. Namely, the SSJ pipeline had lower overall scores on the babushka-bench benchmarking platform (the Obeliks tagger having a measured MSD accuracy of 92.67 and lemmatization accuracy of 98.19) than the CLARIN.SI pipeline (the ReLDI tagger having a measured MSD accuracy of 94.21 and the lemmatization accuracy of 98.35). However, regardless of both pipelines having machine-learning solutions for the two tasks at their core, the SSJ pipeline has one additional rule-based module which incorporates well-thought-through linguistic rules, while the CLARIN.SI pipeline follows the paradigm of "everything is possible", which is not popular with linguists, the primary intended users of the Gigafida corpus. Therefore we developed a meta-tagger - a tagger using the output of both pipelines and deciding, based on the two outputs, which solution it should accept.

The meta-tagger is based on the supervised machine learning paradigm as well, the predictive model being trained on 10,000 instances (tokens) where the outputs of the two basic taggers differ, having the information about what the correct MSD tag and lemma for the specific token are. Using the information about predicted MSD, part-of-speech (first two letters of the MSD tag), and morphosyntactic features (extracted from the MSD tag) from both taggers, the meta-tagger decides which MSD tag and lemma are more likely to be correct. During the development of the meta-tagger we also experimented with using the surface forms and their suffixes, as well as features extracted from potential lemmas, but this information did not improve the meta-tagger's results. Finally, the meta-tagger, evaluated on the babushka-bench platform, proved to be overall more accurate than any of the two solutions by themselves, with the measured MSD accuracy of 94.34 and the lemmatization accuracy of 98.66, a relative error reduction from the best single solution of 2% for MSD tagging and 19% for lemmatization. More importantly, by using the meta-tagger, we managed to combine the preferred linguistic restrictions of the SSJ pipeline and the higher accuracy of the CLARIN.SI pipeline.

The 20 most frequent mistakes made by the meta-tagger, out of all 2,773 mistakes performed on the test set, as calculated on the babushka-bench platform, are given in Table X. The analysis shows that the most mistakes are those of confusing the accusative and nominative case of nouns in masculinum or neutrum, covering 10% of all errors, followed by the confusion of adjectives and adverbs, covering 2% of all mistakes. These

already low percentages show that this error distribution does not have a heavy head, but rather a long tail with quite diverse error types.

	Manually assigned tag	Automatically assigned tag	Error freq.	Manually assigned tag (Extended)	Automatically assigned tag (Extended)
1	Ncmsan	Ncmsn	109	Noun Type=common Gender=masculine Number=singular Case=accusative Animate=no	Noun Type=common Gender=masculine Number=singular Case=nominative
2	Ncmsn	Ncmsan	71	Noun Type=common Gender=masculine Number=singular Case=nominative	Noun Type=common Gender=masculine Number=singular Case=accusative Animate=no
3	Ncnsa	Ncnsn	61	Noun Type=common Gender=neuter Number=singular Case=accusative	Noun Type=common Gender=neuter Number=singular Case=nominative
4	Agpnsn	Rgp	41	Adjective Type=general Degree=positive Gender=neuter Number=singular Case=nominative	Adverb Type=general Degree=positive
5	Ncfpa	Ncfsg	38	Noun Type=common Gender=feminine Number=plural Case=accusative	Noun Type=common Gender=feminine Number=singular Case=genitive
6	Ncnsn	Ncnsa	35	Noun Type=common Gender=neuter Number=singular Case=nominative	Noun Type=common Gender=neuter Number=singular Case=accusative
7	Ncfpn	Ncfsg	32	Noun Type=common Gender=feminine	Noun Type=common Gender=feminine

				Number=plural Case=nominative	Number=singular Case=genitive
8	Ncfpn	Ncfpa	31	Noun Type=common Gender=feminine Number=plural Case=nominative	Noun Type=common Gender=feminine Number=plural Case=accusative
9	Agpnsa	Agpnsn	31	Adjective Type=general Degree=positive Gender=neuter Number=singular Case=accusative	Adjective Type=general Degree=positive Gender=neuter Number=singular Case=nominative
10	Npmsay	Npmsg	27	Noun Type=proper Gender=masculine Number=singular Case=accusative Animate=yes	Noun Type=proper Gender=masculine Number=singular Case=genitive
11	Rgp	Agpnsn	25	Adverb Type=general Degree=positive	Adjective Type=general Degree=positive Gender=neuter Number=singular Case=nominative
12	Ncfsg	Ncfpa	25	Noun Type=common Gender=feminine Number=singular Case=genitive	Noun Type=common Gender=feminine Number=plural Case=accusative
13	Ncfpa	Ncfpn	25	Noun Type=common Gender=feminine Number=plural Case=accusative	Noun Type=common Gender=feminine Number=plural Case=nominative
14	Agpmsay	Agpmsny	25	Adjective Type=general Degree=positive Gender=masculine Number=singular Case=accusative Definiteness=yes	Adjective Type=general Degree=positive Gender=masculine Number=singular Case=nominative Definiteness=yes

15	Agpmsny	Agpmsay	24	Adjective Type=general Degree=positive Gender=masculine Number=singular Case=nominative Definiteness=yes	Adjective Type=general Degree=positive Gender=masculine Number=singular Case=accusative Definiteness=yes
16	Rgp	Cc	23	Adverb Type=general Degree=positive	Conjunction Type=coordinating
17	Agpfpn	Agpfsg	23	Adjective Type=general Degree=positive Gender=feminine Number=plural Case=nominative	Adjective Type=general Degree=positive Gender=feminine Number=singular Case=genitive
18	Npmsn	Npfsn	22	Noun Type=proper Gender=masculine Number=singular Case=nominative	Noun Type=proper Gender=feminine Number=singular Case=nominative
19	Sl	Sa	21	Preposition Case=locative	Preposition Case=accusative
20	Ncmsn	Npmsn	21	Noun Type=common Gender=masculine Number=singular Case=nominative	Noun Type=proper Gender=masculine Number=singular Case=nominative

Table 2. The most frequent tagging errors during evaluation.

12 De-duplication and filtering

After its publication, a more extensive use of Gigafida demonstrated the need to remove duplicates from texts in previous versions of the corpus, since printed media texts often contain duplicate segments that skew statistical data analyses of the corpus as a whole (Logar et al. 2015). A typical example includes radio and TV programmes with identical content published in different sources, as well as identical (newspaper) publications summarising the same source, etc. During the process of compiling Gigafida 2.0, de-duplication was performed on all texts that had been previously included in FIDA, FidaPLUS and the first version of Gigafida, beginning with Gigafida 1.0 and then proceeding

to newly collected texts. De-duplication was performed on the level of paragraphs with the help of Onion (ONe Instance Only, Pomikalek 2011), which allows for the setting of two parameters, -n and -t. Parameter -n defines the length of identical n-grams, which the programme adheres to during de-duplication, whereas parameter -t defines the quantity of identical n-grams within a paragraph (duplicate content threshold). An individual paragraph counts as a duplicate when it exceeds the threshold of the number of n-grams, as defined by parameter -t, which had already been detected in previous paragraphs. Parameter -t may have values between 0 and 1, with value 1 signifying that the paragraph is 100% identical. The following settings were applied: -n=9, -t=0.5, which reduced the corpus by 24.9% (cf. Benko 2013).

Apart from paragraphs, de-duplication was also performed on the level of entire texts. When the text contained more than 95% of duplicate paragraphs, the whole text was marked as a duplicate and removed from the de-duplicated corpus. Furthermore, 6,915 texts were removed from the corpus as whole; this included texts which: 1) did not include letters 'č', 'ž' and 'š' (this resulted in the elimination of texts with issues caused by the use of wrong encoding tables or where the authors failed to use standard spelling (see chapter 6)); and 2) which contained fewer than 500 characters. Texts with fewer than 500 characters were removed with the intention to reduce the number of files (containing metadata) that unnecessarily lengthen corpus processing.

13 Standard and non-standard material

Before version 2.0, corpus included material which was either known to or thought to contain non-standard linguistic features, i.e. deviations from standard Slovene in terms of spelling, vocabulary, syntax, and style. This category was predominantly comprised of online user content, such as comments, board messages and other online texts, and in some cases fiction (a good example being the Slovene translation of the novel *Trainspotting*) and newspapers (e.g. local media, partially written in regional Slovene language variations; e.g. the newspaper of the Slovene minority in Italy, *Novi Matajur*).

In Gigafida 2.0 this type of content was minimised to the greatest degree possible for several reasons. Firstly, following the publication of Gigafida 1.0 other, more reliable sources were created for the analysis of non-standard Slovene, primarily the corpus of online Slovene Janes 1.0 (Fišer et al. 2016; nl.ijs.si/janes), which contains Slovene tweets, board messages, blog entries and news comments (nearly 253 million tokens). Secondly,

the number of these texts in the corpus was relatively small to begin with and thus insufficiently representative for the analysis of non-standard linguistic features. Finally, corpus texts did not include metadata on their degree of standardness, which is why it was impossible to exclude them from queries on the basis of criteria, such as titles, type of source etc.

The main principle for the removal of texts in relation to their level of (non)standardness was the assumption whether the author had the intention or was socially pressured to write a text in standard Slovene, with standardness being defined on the basis of principles described in Krek 2015. This, in effect a socio-linguistic standpoint, was also empirically tested.

Non-standard texts were identified mechanically, with tools for automatic measuring of the degree of non-standardness (Ljubešić et al. 2015) developed within project JANES. The tool ascribes texts with values between 1 (standard) and 3 (highly non-standard), on two levels: the level of technical non-standardness (T) looks at the formal suitability of a text (e.g. the use of spaces, punctuation, and capitalisation), whereas linguistic non-standardness (L) determines the standardness of a text from the perspective of vocabulary and spelling. We decided to focus on the linguistic non-standardness of texts.

The tool was trained on a manually annotated selection of short online texts, such as tweets, board messages, and news comments. Gigafida 2.0 contains longer texts, which is why the level of non-standardness was ascribed to each individual paragraph. This provided us with a more precise overview of the distribution of non-standardness within a text: each text was thus ascribed with a vector of the level of non-standardness, which included values for each individual paragraph in the text.

This distribution of values was then compared with the distribution of values within the entire corpus by applying the Kolmogorov-Smirnov test, which shows whether the distribution of values within a sample differs in a statistically valid manner from the distribution across the entire corpus. Texts with a statistically valid deviation from the distribution in the corpus and with a higher average value for linguistic non-standardness were included on a list of texts with potential issues. This selection was then manually verified and the chosen unsuitable texts removed from the corpus.

14 Corpus concordancer

As already mentioned in the introduction, both Gigafida 2.0 and Gigafida 2.0 Proto are available through various corpus tools. On the one hand, both versions are included in the NoSketch Engine, which can be accessed at the natural language server hosted by the *Department of Knowledge Technologies* at the Jožef Stefan Institute, and in the concordancer Kontext, which is accessible as part of the Clarin.si infrastructure. Its inclusion into the powerful tool SketchEngine has been agreed. On the other hand, Gigafida 2.0 is also available through a user-friendly concordancer for general use.

Gigafida 1.0 is available through a concordancer developed within the framework of the project Communication in Slovene. As described in greater detail in Logar et al. (2012: 98-118), the programme was developed with the intention to facilitate corpus use and access to all interested user groups, such as teachers, editors, translators, writers of texts for various uses etc. The functionality and look of the concordancer were designed on the basis of prior user studies (Arhar Holdt 2009); its newly introduced features were further evaluated a few years after its launch in a separate user evaluation. The target user groups have positively rated the concordancer features, the two main highlights being the simplicity of searching for corpus data and the clean look of the interface (Arhar Holdt et al. 2019).

The concordancer which provides access to Gigafida 2.0 enables the users to use the corpus very much like in the past. In keeping with the results of the user evaluation, all the main functionality has been retained: separate tabs for concordance and collocation searches and for generating word lists; progressive interface navigation, which is enabled by data interconnectedness (e.g. shifting from the list of collocates to concordance sequence); simple basic and advanced searches; interactive data filters; access to search history; the possibility of exporting data to other programmes for further processing.

Even though the upgrade was not planned as part of the original framework of the project, minor interface adjustments were included in the execution, allowing for greater ease of search and a better overview of the results: a) buttons for shifting between different search modes are now placed conveniently close to search window; b) shifting between tabs *Search*, *Collocation* and *List* initiates automatic data processing for the selected query, increasing time efficiency; c) frequency data is now more accessibly categorised within filters; d) the filters allow for an initial, simultaneous choice of two or more categories as

well as for a simple clearing of selected filters; e) navigation through concordance string pages allows for a direct jump to the selected page; f) instead of MI statistics, the tab *Collocations* now provides a more relevant *logDice*; and g) the addition of buttons for sharing content on social media. As a language tool, the interface was primarily developed for computer use; despite this, the design was adjusted for tablet and smartphone use and considered the accompanying space limitations.

One of the more important developments is the inclusion of Gigafida 2.0 in the portal viri.cjvt.si, both in terms of format and data. The portal facilitates a simple interlinking of various types of language data. By clicking a special button next to the search window, the users are given the possibility to search for the selected query simply and quickly in other CJVT sources, e.g. *Thesaurus of Modern Slovene* and the *Collocations Dictionary of Modern Slovene*.

15 Reference // References

ARHAR HOLDT, Špela (2009). Uporabniška evalvacija korpusa FidaPLUS: zasnova vprašalnika, prvi rezultati. V: STABEJ, Marko (ur.). *Infrastruktura slovenščine in slovenistike* (Obdobja 28). Ljubljana: Znanstvena založba Filozofske fakultete. 19–26.

ARHAR HOLDT, Špela, DOBROVOLJC, Kaja, LOGAR, Nataša (2019). Simplicity matters: user evaluation of the Slovene reference corpus. *Language resources and evaluation*, vol. 53, no. 1. 173-190.

BENKO, Vladimír. Data Deduplication in Slovak Corpora (2013). *Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning* . 27-39.

FIŠER, Darja, ERJAVEC, Tomaž, LJUBEŠIĆ, Nikola (2016). JANES v0.4: Corpus of Slovene User-Generated Content. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 4(2). 67-99.

KREK, Simon, GANTAR, Polona, ARHAR HOLDT, Špela, GORJANC, Vojko (2016). Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres. V: ERJAVEC, Tomaž (ur.), FIŠER, Darja (ur.). *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, 29. september do 1. oktober 2016, Ljubljana, Slovenija. Ljubljana: Znanstvena založba Filozofske fakultete. 200–203.

KREK, Simon. Standardni in knjižni jezik - drugi poskus (2015). V: SMOLEJ, Mojca (ur.). *Slovnica in slovar - aktualni jezikovni opis*, (Obdobja 34). Ljubljana: Znanstvena založba Filozofske fakultete. 401–407.

LJUBEŠIĆ, Nikola, FIŠER, Darja, ERJAVEC, Tomaž, ČIBEJ, Jaka, MARKO, Dafne, POLLAK, Senja, ŠKRJANEC, Iza (2015). Predicting the level of text standardness in user-generated content. *1oth International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference*, 7-9 September 2015, Hissar, Bulgaria. 371–378.

LOGAR, 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.

LOGAR, Nataša, DOBROVOLJC, Kaja, ARHAR HOLDT, Špela (2015). Gigafida : interpretacija korpusnih podatkov. V: SMOLEJ, Mojca (ur.). *Slovnica in slovar - aktualni jezikovni opis* (Obdobja 34). Ljubljana: Znanstvena založba Filozofske fakultete. 467–477.

POMIKÁLEK, Jan (2011). *Removing boilerplate and duplicate content from web corpora.* PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech republic.

TRAMPUŠ, Mitja, NOVAK, Erik (2012): *The Internals of an Aggregated Web News Feed*. Proceedings of 15th Multiconference on Information Society 2012 (IS-2012).

16 Priloga // Appendix

Tabela v korpus Gigafida 2.0 vključenih besedil z naborom korpusnih metapodatkov.
Podatke je pripravil dr. Tomaž Erjavec

Table - metadata on the texts included in Gigafida 2.0. The table was made by dr. Tomaž Erjavec.