REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI SKLAD ZA REGIONALNI RAZVOJ
NALOŽBA V VAŠO PRIHODNOST

Korpus metaFida 1.0

Poročilo projekta Razvoj slovenščine v digitalnem okolju

Aktivnost DS1.8

Avtor: Tomaž Erjavec

Institut "Jožef Stefan", Odsek za tehnologije znanja

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Vsebina

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

1 Projektni okvir

Poročilo oz. kazalnik *Korpus metaFida 1.0* je nastalo pod okriljem projekta Razvoj slovenščine v digitalnem okolju, ki sta ga med leti 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (Operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020). Cilj projekta je bil zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, podjetja in širšo javnost. Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.

Kazalnik se umešča v prvi projektni delovni sklop z naslovom *Jezikovni viri*. Namen delovnega sklopa je bila nadgraditi slovenske besedilne korpuse in leksikon besednih oblik. Prenovili smo učne množice in postopke za strojno označevanje sodobne slovenščine. Rezultat so osveženi in povečani jezikovni viri, ki so na voljo tako uporabniški skupnosti kot za strojno rabo. Z razvitimi postopki in orodji bo posodabljanje slovenskih korpusov v prihodnosti hitrejše in preprostejše.

Med področja, ki smo se jim na projektu posvetili, je sodila tudi izdelava korpusa metaFida, ki ga opisujemo v tem kazalniku. Korpus sestavlja izbor javno dostopnih obstoječih korpusov slovenskega jezika (tudi tistih, ki so bili narejeni ali nadgrajeni v projektu RSDO), s čimer bi omogočal poizvedbe po več korpusih naenkrat. Korpus je prosto dostopen za rabo prek konkordančnikov CLARIN.SI. Ciljna velikost korpusa je bila tri milijarde besed.

Korpus metaFida je izšel v dveh različicah, in sicer metaFida 0.1 (Erjavec, 2021) kmalu po začetku projekta RSDO in metaFida 1.0 (Erjavec, 2023) povsem na koncu projekta. Različica 0.1 je bila narejena iz obstoječih korpusov ob izdaji, različica 1.0 pa je nekatere korpuse nadomestila oz. dodala s korpusi narejenimi v okviru projekta RSDO. Poleg tega je bil nadgrajen postopek gradnje korpusa.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

2 Opis projektnih aktivnosti

Cilj aktivnosti 1.8 *Metakorpus vseh večjih slovenskih korpusov* je bil pripraviti načrt in izdelati korpus metaFida, ki naj bi združeval vse relevantne slovenske javno dostopne korpuse v enoten korpus, ki naj bi bil dostopen prek konkordančnikov CLARIN.SI.

Slovenščina ima namreč za spletno analizo prek konkordančnikov CLARIN.SI dostopnih večje število raznovrstnih korpusov, v okviru projekta RSDO pa so bili nekateri tudi nadgrajeni ali na novo ustvarjeni. Vendar morajo uporabniki, v primeru, da jih zanimajo iste poizvedbe po različnih korpusih, iskati relevantne informacije po vsakem korpusu posebej, nato pa te informacije ročno združevati, kar je zamudno, vodi pa lahko tudi do napak pri analizi. Dodatni problem je, da imajo korpusi tipično različne metapodatke in da so lahko tudi označeni po različnih jezikoslovnih ravneh, kar enaka iskanja po različnih korpusih dodatno otežuje.

V sklopu projekta smo naredili pregled javno dostopnih obstoječih korpusov slovenskega jezika in tistih, ki so bili nadgrajeni v okviru projekta, in izbor vključili v združeni korpus, imenovan metaFida. Pri tem je bilo najprej treba poenotiti metapodatke in uskladiti jezikoslovne in strukturne oznake med korpusi ter izdelati pretvorbe posameznih korpusov iz njihovih vertikalnih formatov, ki jih kot vhod uporabljajo konkordančniki CLARIN.SI, v vertikalni format metaFide. Korpus smo dali na voljo raziskovalni skupnosti prek konkordančnikov CLARIN.SI v sklopu aktivnosti 6.1.

Na projektni aktivnosti je sodeloval raziskovalec z Instituta “Jožef Stefan”.

Delo je potekalo po naslednjih korakih:

- Pregled obstoječih in v projektu nadgrajenih korpusov in utemeljen izbor za metaFido.
- Izdelava skupnega formata za vertikalno datoteko metaFida, podatkovnih struktur in algoritma za pretvorbo.
- Pretvorbe obstoječih korpusov za prvo različico korpusa metaFida 0.1, deduplikacija, postavitev pod konkordančnike (v okviru sklopa 6.1).
- Pretvorbe za celoten nabor korpusov, vključno s korpusi RSDO, deduplikacija, postavitev pod konkordančnike (v okviru sklopa 6.1).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

REZULTATA AKTIVNOSTI

- Korpus metaFida 0.1 v velikosti 3,6 milijarde besed kot baza na repozitoriju CLARIN.SI: http://hdl.handle.net/11356/1746 in vključen v konkordančnike na CLARIN.SI:
  - KonText: https://www.clarin.si/kontext/query?corpname=mfida01
  - NoSketch Engine Bonito: https://www.clarin.si/noske/sl.cgi/first?corpname=mfida01&reload=1&iquery=
  - NoSketch Engine Crystal: https://www.clarin.si/ske/#dashboard?corpname=mfida01
- Korpus metaFida 1.0 v velikosti 4,7 milijarde besed kot baza na repozitoriju CLARIN.SI: http://hdl.handle.net/11356/1775 in vključen v konkordančnike na CLARIN.SI:
  - KonText: https://www.clarin.si/kontext/query?corpname=mfida10
  - NoSketch Engine Bonito: https://www.clarin.si/noske/sl.cgi/first?corpname=mfida10&reload=1&iquery=
  - NoSketch Engine Crystal: https://www.clarin.si/ske/#dashboard?corpname=mfida10

3 Postopek gradnje

Postopek izgradnje metaFide je bil sledeč:

1. Izdelava seznama korpusov, ki sestavljajo korpus metaFida (cf. Tabela 1, 2). Identifikatorju vsakega korpusa je dopisana tudi letnica najmlajšega besedila v korpusu, za korpuse, ki nimajo pripisanega datuma izdaje besedil, pa letnica formiranja korpusa. Ta podatek služi za določitev atributa besedil year_max, ki je podrobneje razložen v nadaljevanju.
2. Izdelava programa, ki s seznama in iz konfiguracijskih datotek vključenih korpusov (registry files) izdela tabelo z informacijami o korpusih, ki so pomembne za nadaljnjo obdelavo: identifikator korpusa, ime korpusa, lokacija vertikalne datoteke, ime elementa, ki ustreza posameznemu besedilu v korpusu, seznam strukturnih oznak z njihovimi atributi, seznam pozicijskih atributov. V primeru, da je atribut lahko večvrednosten, je k njegovemu imenu pripisan znak, ki služi za ločevanje posameznih vrednosti.
3. Na osnovi tabele iz 2. točke smo ročno izdelali tabelo s preslikavo od korpusa neodvisnih strukturnih elementov in njihovih atributov v strukturne elemente in atribute metaFide.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

4. Izdelava programa, ki na osnovi tabel iz 2. in 3. točke za vsak vključeni korpus:
    a. odpre ustrezno vertikalno datoteko,
    b. glede na tabelo preslikave iz 3. preslika (ali ignorira) strukturne elemente in njihove atribute,
    c. glede na tabelo iz 2. z informacijami o korpusih preslika izvorne pozicijske atribute v pozicijske atribute metaFide; po potrebi nadomesti ločilo za večvrednostne atribute in določi slovensko oz. angleško oblikoskladenjsko oznako glede na tisto, ki obstaja v izvornem korpusu,
    d. vsakemu odstavku v metaFidi doda identifikator, bodisi izvornega bodisi na novo tvorjeni v primerih, ko v izvornem korpusu struktura, ki ustreza odstavku v metaFidi, identifikatorja nima.

Po izvajanju programa v 4. točki dobimo različico metaFide, ki vsebuje vso besedilo izvornih korpusov – vrstni red korpusov je tak, kot je podan v seznamu iz 1. točke, vrstni red besedil pa tak, kot je v izvornih korpusih. Za izdelavo končne verzije smo dodali še dva koraka:

5. Odstranjevanje podvojenih odstavkov: napisali smo program, ki za niz, ki je sestavljen iz besed odstavka, izračuna kodo MD5 in jo izpiše, skupaj z identifikatorjem odstavka. Nato za vse večkrat ponovljene kode MD5, razen prve, odstranimo odstavke z ustreznim identifikatorjem iz korpusa metaFida. V primeru, da je ta postopek izbrisal vse odstavke nekega besedila, odstranimo celotno besedilo. Odstranjeni odstavki, glede na površen pregled, vsebujejo tipično en razmeroma kratek stavek, dostikrat kar iz istega besedila (predvsem premi govor), medtem ko se odstranjena besedila tipično pojavljajo v več korpusih, predvsem tudi v korpusu slWaC, npr. korpus IMP je kot digitalna knjižnica v celoti dostopen na strežniku https://nl.ijs.si/. Pri takih besedilih je pomemben vrsti red korpusov v seznamu iz 1. točke, saj hočemo ohraniti besedila, ki imajo bolj kvalitetne oznake.
6. Sortiranje po letnici besedila: ker je korpus metaFida še najbližje diahronemu korpusu, smo se odločili, da bomo besedila v korpusu sortirali po letnici nastanka. V različici 0.1 so bila besedila brez letnice nastanka na koncu korpusa, v 1.0 pa smo izboljšali sortiranje tako, da smo vsakemu besedilu pripisali letnico, po kateri besedilo ni moglo nastati (year_max), in sortirali po tej (več o tem atributu v razdelku Strukturne oznake in atributi).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

4 Vključeni korpusi

Osnovno vodilo za vključitev korpusov v metaFido je bilo, da vključimo čim večje število čim bolj raznovrstnih korpusov. Konkordančniki CLARIN.SI ponujajo prek 160 korpusov oz. prek 100, če ne štejemo takih korpusov, ki so (tipično dodatno označena) podmnožica drugih.

V korpus metaFida smo vključili vse obstoječe slovenske korpuse, ki:

1. niso podmnožica drugih korpusov, s čimer so mdr. izključeni skoraj vsi korpusi, ki so ročno označeni;
2. imajo pojavnice, označene vsaj z lemo in oblikoskladenjsko oznako MULTEXT-East (bodisi njeno angleško bodisi slovensko različico), s čimer je izključen npr. korpus DGT-UD-sl, ki nima oznak MULTEXT-East;
3. v primeru, ko korpus obstaja v dveh različicah (izvorna besedila, popravljena besedila), smo v korpus metaFida vključili samo izvorno različico.

Pri dveh vključenih korpusih obstajajo dodatne posebnosti:

- v nabor smo raje kot celotno vključili deduplicirano različico korpusa Gigafida 2.0, saj ne bi imelo smisla vključevati korpus, ki ima veliko število (skoraj) duplikatov;
- korpus Janes je dostopen kot celoten korpus, pa tudi kot štirje (pod)korpusi, ki ločijo izvor besedil (tviti, blogi, novice, Wikipedija komentarji); v metaFido smo vključili posamezne podkorpuse, saj s tem damo uporabnikom boljši vpogled v izvor posameznih konkordanc, pa tudi bolj precizno možnost izbire podkorpusov metaFide .

Vsi zajeti korpusi raje kot vzorce besedil vsebujejo integralna besedila (z izjemo vrzeli zaradi deduplikacije v Gigafidi) in skoraj vsi so – z delno izjemo Janes-Norm 3.0 – avtomatsko označeni.

V inicialno različico metaFide 0.1 je bilo vključenih 34 korpusov (Tabela 1); število pojavnic in besed je takšno, kot jih javi konkordančnik noSketch Engine.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 1: Pregled korpusov, vključenih v različico 0.1 korpusa metaFida. Korpusi, ki so bili v različici 1.0 nadomeščeni z drugimi, so v tabeli prečrtani. Število pojavnic in besed je takšno, kot jih javi konkordančnik noSketch Engine.

Identifikator	Ime	Pojavnic	Besed
classlawiki_sl	CLASSLAWiki-sl (Slovenian Wikipedia)	54.608.642	41.543.793
dgt15_sl	EU DGT 2015: Slovene	62.303.744	49.145.476
dsi	DSI (informatika)	5.245.073	4.254.177
eltec_slv	ELTeC-slv (100 romanov)	6.901.534	5.596.656
filmi	FILMI (filmske kritike)	936.446	764.764
gfida20_dedup	Gigafida v2.0 (referenčni, dedupliciran)	1.333.360.653	1.105.200.611
~~gos_vl42~~	GosVL 4.2 (govorni, VideoLectures)	179.063	178.748
~~gos11~~	Gos 1.1.1 (referenčni, govorni)	1.063.861	1.033.697
imp	IMP (starejša besedila)	17.723.874	14.348.452
ispac_sl	ISPAC: slovensko	1.432.798	1.169.486
janes_blog	Janes-Blog (blogi s komentarji)	34.534.431	27.596.463
janes_forum	Janes-Forum (spletni forumi)	47.066.575	37.654.809
janes_news	Janes-News (komentarji na novice)	14.838.074	11.908.481
janes_tweet	Janes-Tweet (tviti 2013-2017)	151.457.091	108.769.902
janes_wiki	Janes-Wiki (Wikipedija komentarji)	5.008.067	3.917.428
jaslo_sl	jaSlo: slovensko	532.395	425.434
~~kas_dipl~~	KAS Dipl (diplome)	1.101.796.659	867.652.073
~~kas_dr~~	KAS Dr (doktorati)	101.473.395	76.460.363
~~kas_mag~~	KAS Mag (magisteriji)	495.827.656	389.493.726
konji	Konji (konjeništvo)	469.894	395.718
korp	KoRP (odnosi z javnostmi)	2.194.130	1.756.731

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

lemonde_sl	LeMonde: slovensko	615.617	506.358
~~maj68~~	Maj68 (Maj 1968 v literaturi)	794.382	645.002
maks	MAKS (mladinska književnost)	12.072.273	9.881.294
prilit	PriLit (starejša pripovedna proza)	1.275.209	1.060.538
rsdo5	RSDO5 (s termini označena besedila)	310.588	241.797
sbsj	SBSJ (šolska besedila)	1.836.810	1.424.887
~~siparl20~~	siParl 2.0 (parlament 1990-2018)	239.749.733	202.461.367
slwac	slWaC (Slovene Web)	895.903.321	749.372.269
~~solar~~	Šolar v2 Clear (šolski spisi)	1.907.731	1.621.671
suss	ŠUSS (jezikovna vprašanja)	365.371	272.541
trans5_sl	TRANS5: slovensko	1.594.120	1.297.269
tweet_sl	Tweet-sl (stari tviti)	6.291.820	4.854.229
vayna	VAYNA (napadi na JNA)	300.666	256.429
Σ	*34*	*4.601.971.696*	*3.723.162.639*

Tudi v različico metaFide 1.0 je bilo vključenih 34 korpusov, pri čemer jih je bilo 8 iz različice 0.1 izločenih, ravno toliko pa dodanih na novo (Tabela 2).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 2: Pregled korpusov, vključenih v različico 1.0 korpusa metaFida. Korpusi, ki so bili nadgrajeni oz. dodani, so v tabeli podčrtani. Število pojavnic in besed je takšno, kot jih javi konkordančnik noSketch Engine.

Identifikator	Ime	Pojavnic	Besed
classlawiki_sl	CLASSLAWiki-sl (Slovenian Wikipedia)	54.608.642	41.543.793
dgt15_sl	EU DGT 2015: Slovene	62.303.744	49.145.476
dsi	DSI (informatika)	5.245.073	4.254.177
eltec_slv	ELTeC-slv (100 romanov)	6.901.534	5.596.656
filmi	FILMI (filmske kritike)	936.446	764.764
gfida20_dedup	Gigafida v2.0 (referenčni, dedupliciran)	1.333.360.653	1.105.200.611
gos20	Gos 2.0 (referenčni, govorni)	2.496.110	2.436.386
imp	IMP (starejša besedila)	17.723.874	14.348.452
ispac_sl	ISPAC: slovensko	1.432.798	1.169.486
janes_norm30	Janes-Norm 3.0 (ročno normaliziran)	336.497	253.651
janes_blog	Janes-Blog (blogi s komentarji)	34.534.431	27.596.463
janes_forum	Janes-Forum (spletni forumi)	47.066.575	37.654.809
janes_news	Janes-News (komentarji na novice)	14.838.074	11.908.481
janes_tweet	Janes-Tweet (tviti 2013-2017)	151.457.091	108.769.902
janes_wiki	Janes-Wiki (Wikipedija komentarji)	5.008.067	3.917.428
jaslo_sl	jaSlo: slovensko	532.395	425.434
jezkor	JezKor (jezikoslovje)	9.272.854	6.243.898
kost10_orig	KOST: izvorni (L2)	1.209.129	1.020.509
oss10	OSS (znanstvena besedila)	3.261.941.663	2.342.855.598
konji	Konji (konjeništvo)	469.894	395.718
korp	KoRP (odnosi z javnostmi)	2.194.130	1.756.731

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

lemonde_sl	LeMonde: slovensko	615.617	506.358
maj68	Maj68 (Maj 1968 v literaturi)	1.271.144	1.033.971
maks	MAKS (mladinska književnost)	12.072.273	9.881.294
prilit	PriLit (starejša pripovedna proza)	1.275.209	1.060.538
rsdo5	RSDO5 (s termini označena besedila)	310.588	241.797
sbsj	SBSJ (šolska besedila)	1.836.810	1.424.887
siparl30	siParl 3.0 (parlament 1990-2022)	239.749.733	202.461.367
slwac	slWaC (Slovene Web)	895.903.321	749.372.269
solar30_orig	Šolar: učenci (razvojni)	1.907.562	1.621.527
suss	ŠUSS (jezikovna vprašanja)	365.371	272.541
trans5_sl	TRANS5: slovensko	1.594.120	1.297.269
tweet_sl	Tweet-sl (stari tviti)	6.291.820	4.854.229
vayna	VAYNA (napadi na JNA)	300.666	256.429
Σ	*34*	*6.177.363.908*	*4.741.542.899*

5 Oznake korpusa

V korpusu metaFida smo obdržali samo tiste informacije, ki so skupne večini izbranih korpusov ali pa so zadosti pomembne, da jih pripišemo vsaj tistim besedilom, ki jih imajo (kot npr. avtor). Struktura je plitva (<text> z metapodatki, nato <p> in <s>), saj je tako laže ustvarjati podkorpuse oz. zamejiti iskanje na posamezne vrste besedil.

5.1 Strukturne oznake in atributi

metaFida ima naslednje strukturne oznake:

1. **text**: posamezno besedilo, ki se pretvori iz ustreznega elementa izvornega korpusa (text, session). Element ima naslednje atribute:
    a. **corpus_id**: identifikator korpusa (kot v Tabelah 1 in 2), iz katerega je besedilo.
    b. **corpus**: ime tega korpusa (Tabeli 1, 2).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

   c. **info**: URL do korpusa v CLARIN.SI noSketch Engine, s čimer uporabniku omogočimo enostaven prehod v možnost bolj podrobnega iskanja po tem korpusu.
   d. **id**: identifikator besedila. Kjer obstaja, je prevzet iz izvornega korpusa, kjer ne, je izdelan iz imena korpusa in zaporedne številke besedila (pred deduplikacijo, gl. spodaj).
   e. **year**: letnica izdaje besedila, pri čemer podatek manjka pri 19,4 % besedil, tipično zato, ker ni identificiran v besedilih izvornega korpusa, kar zadeva predvsem spletna besedila, saj datum objave tam tipično ni znan.
   f. **year_max**: letnica, po kateri besedilo ni moglo biti objavljeno. Pri besedil z znanim letom izdaje je "year_max" enak "year", pri besedilih brez te letnice pa je enak letnici izdelave oz. zajema besedil izvornega korpusa. S tem je tudi približno petini besedil brez znanega datuma izida pripisan vsaj približek tega datuma (dodano v različici 1.0).
   g. **publisher**: izdajatelj besedila, pri čemer je to za spletna besedila lahko tudi domena.
   h. **title**: naslov besedila, za spletna besedila brez naslova pa vsebuje URL. Večina besedil naslova nima, bodisi zaradi tega, ker ga ni oz. ni identificiran (npr. tviti), bodisi je anonimiziran (npr. korpus Šolar).
   i. **author**: avtor oz. avtorji besedila (atribut ima lahko več vrednosti, ločene so s podpičjem); podobno oz. še bolj kot pri naslovu ta podatek večinoma manjka.
   j. **wordcount**: število besed v besedilu.
2. **p**: odstavek, ki se pretvori iz ustreznega elementa izvornega korpusa (ab, align, note, p, speech). Element ima naslednji atribut:
   a. **id**: identifikator odstavka. Kjer obstaja, je prevzet iz izvornega korpusa, kjer ne, je izdelan iz identifikatorja besedila in zaporedne številke odstavka (pred deduplikacijo).
3. **s**: stavek oz. poved. Brez atributov.
4. **gap**: prazna oznaka za vrzel, ki pomeni, da del besedila manjka, bodisi že v izvirnem korpusu (npr. v Gigafidi zaradi deduplikacije) bodisi zaradi dedupliciranja metaFide.
5. **g**: (interna) prazna oznaka za "lepilo", ki imogoči konkordančniku, da pravilno izpiše stičnost besed v kontekstu konkordance.

5.2 Pozicijski atributi

Za vse pozicijske atribute Gigafide velja, da imajo lahko več vrednosti, ločilo med njimi pa je presledek. Več vrednosti je potrebnih zaradi tega, ker imajo nekateri korpusi normalizirane besede (starejša slovenščina, uporabniško generirane vsebine), kjer se lahko ena izvorna beseda preslika v več normaliziranih ali obratno.

metaFida ima naslednje pozicijske atribute:

1. **word**: standardni atribut za pojavnice, kot se pojavijo v besedilu.
2. **norm**: normalizirani (posodobljeni, standardizirani) zapis besede. V korpusih, ki nimajo tega atributa, je vrednost enaka pojavnici.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

3. **lemma**: osnovna oblika besede.
4. **tag_en**: oblikoslovna oznaka MULTEXT-East v angleškem jeziku.
5. **tag**: oblikoslovna oznaka MULTEXT-East v slovenskem jeziku.
6. **lc**: pojavnica v malih črkah. Dinamičen atribut (ni prisoten v vertikalni datoteki).
7. **norm_lc**: normalizirana pojavnica v malih črkah. Dinamičen atribut.
8. **lemma_lc**: lema v malih črkah. Dinamičen atribut.

V konfiguracijski datoteki je zapisana poizvedba za enostavno iskanje, da se to preslika v [lc="%s" | norm_lc="%s" | lemma_lc="%s"], torej bodisi pojavnica, normalizirana oblika pojavnice bodisi lema, ne glede na velike/male črke.

6 Kvantitativna analiza

V Tabeli 3 podajamo glavne kvantitativne značilnosti korpusa metaFida 1.0.

Tabela 3: Velikost korpusa metaFida 1.0

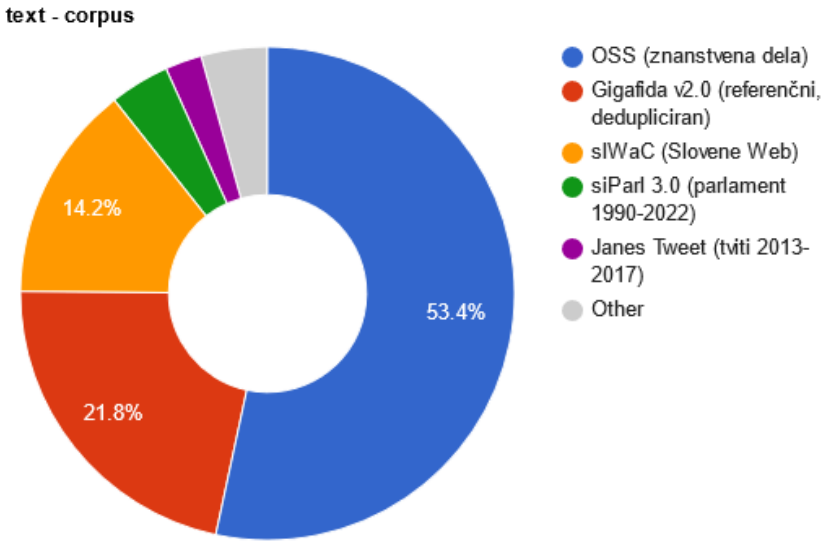
Pojavnic	6.094.189.351
Besed	4.727.457.629
Povedi	281.218.815
Odstavkov	51.822.749
Besedil	15.454.886

V procesu deduplikacije je bilo identificiranih in odstranjenih 6.438.044 podvojenih odstavkov (11 % glede na velikost nededupliciranega korpusa), s čimer je 1.160.996 besedil ostalo brez odstavkov, torej vsebine, in so bili posledično ravno tako odstranjeni (7 %). S tem je korpus izgubil 14.085.270 besed (0,3 %) oz. 83.174.557 pojavnic (1,3 %). Kot vidimo, je, procentualno gledano, izgubljenih besed zelo malo, saj so tipično odstranjeni kratki odstavki, kjer je možnost podvojenosti največja. Zanimivo je tudi, da je število izbrisanih besedil primerljivo s številom izbrisanih odstavkov, kar pomeni, da so tipično podvojena celotna besedila. Vključno z izvornimi vrzelmi (element <gap>) vsebuje korpus 12.633.604 vrzeli.

Na Sliki 1 podajamo število pojavnic, ki jih posamezen korpus prispeva v korpus metaFida. Več kot polovico pojavnic je iz korpusa OSS, pri čemer opozarjamo, da korpus vsebuje dosti šuma, saj so bila besedila pretvorjena neposredno iz PDF-jev in je zato število pojavnic še večje, kot bi bilo sicer.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

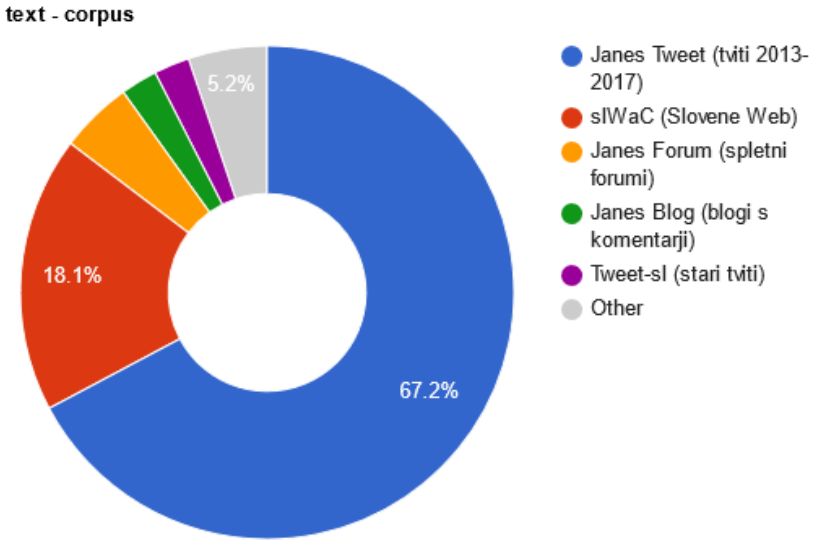
Naslednji, ki prispeva okoli petino pojavnic, je korpus Gigafida, temu pa, z nekaj manj kot šestino pojavnic, sledi korpus slovenskega spleta slWaC. Naslednji, s 3,9 % pojavnic, je parlamentarni korpus siParl, zadnji od vrhnjih petih pa je, morda presenetljivo, korpus tvitov Janes-Tweet z 2,4 % pojavnic. Kljub temu da vseh ostalih 29 korpusov prispeva samo 4,3 % pojavnic, je bila njihova vključitev vseeno smiselna, saj je (a) tudi to velik korpus z 264 milijoni pojavnic in (b) ti korpusi prispevajo k raznovrstnosti metaFide, saj vsebujejo tudi zelo drugačna besedila (in besede) kot prvih pet, npr. starejšo slovenščino (IMP), izdelke šolarjev (Šolar) itd.



Slika 1: Prispevek pojavnic posameznih korpusov v korpus metaFida 1.0

Slika 2 podaja število besedil, ki jih posamezen korpus prispeva v korpus metaFida. Tu je stanje zelo drugačno od pokritja s pojavnicami, saj je na prvem mestu korpus tvitov, ki prispeva več kot dve tretjini besedil. Razlog je, da je en tvit označen kot eno besedilo in so zato ta besedila izredno kratka, medtem ko npr. korpus OSS vsebuje mdr. celotne monografije, kot so npr. doktorske disertacije. Drugi je korpus spletni besedil slWaC, ki tudi vsebuje krajša besedila (spletne strani) in je tudi sicer velik, saj je tretji po številu pojavnic. Tretje, četrto in peto mesto ravno tako zasedajo besedila s spleta, in sicer korpusi Janes-Forum (4,8 %), Janes-Blog (2,4 %) in starejši (in razmeroma majhen) korpus starejših tvitov Tweet-sl (2,3 %). Vsi ostali korpusi nato prispevajo samo okoli dvajsetino besedil.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.



Slika 2: Prispevek besedil posameznih korpusov v korpus metaFida 1.0

7 Ocena uspešnosti in prioritete za nadaljnji razvoj

Izdelava korpusa metaFida je bila v celoti uspešna, in to po rezultatu kot tudi glede na načrt iz projektne prijave, z eno samo izjemo. V načrtu je bilo tudi odstranjevanje bližnjih duplikatov odstavkov, ki pa smo ga nadomestili z odstranjevanjem polnih duplikatov. Razlog je bil, poleg izredne računske zahtevnosti tega procesa glede na velikost korpusa, tudi to, da za dva največja vključena korpusa, OSS (oz. v različici 0.1, KAS) in Gigafida, odstranjevanje bližnjih duplikatov odstavkov ni potrebno, saj OSS vsebuje integralna besedila in bi takšno odstranjevanje škodilo celovitosti besedil, Gigafida pa je bližnje duplikate odstavkov že tako ali tako imela odstranjene. S tem bi ta dodatni korak prinesel le majhne koristi oz. bi pri uporabnosti korpusa celo škodil.

metaFida omogoča uporabnikom dostop do vseh pomembnih jezikovnih korpusov na enem mestu. S tem jezikovni uporabniki hitro in enostavno dostopajo do koristnih jezikovnih informacij znotraj največjega korpusa slovenskega jezika, ki jih lahko uporabijo za razumevanje jezika, prevajanje, razvijanje aplikacij za obdelavo jezika, šolsko delo itd.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Korpus metaFida 0.1 se je že v času projekta uspešno uporabljal, npr. v okviru dela Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, pa tudi s strani posameznih raziskovalcev na Filozofski fakulteti Univerze v Ljubljani. Predstavljen je bil tudi na več dogodkih, z odprto objavljenimi materiali:

- Tomaž Erjavec, Katja Meden, Jakob Lendardič: Predstavitev storitev CLARIN.SI. Predavanje na izobraževanju knjižničarjev, NUK, 22. 2. 2022: 1. CLARIN.SI [PDF], 2. Repozitorij [PDF], 3. Konkordančniki [PDF]
- Tomaž Erjavec: **The CLARIN.SI research infrastructure**. Predavanje za študente DigiLing joint masters program in digital linguistics , Univerza v Ljubljani, 25. 10. 2022. [PDF]
- Tomaž Erjavec: **Korpus metaFida**. Predavanje v okviru dogodka Mednarodni dan slovarjev 2022: Slovarji včeraj, danes, jutri, 12. 10. 2022 [PDF]
- Tomaž Erjavec: **Konkordančniki CLARIN.SI**. Predavanje v okviru Podiplomske šole ZRC SAZU, doktorski modul *Leksikologija, leksikografija, slovničarstvo*, 23. 5. 2022 [PDF]

Za nadaljnji razvoj korpusa metaFida je potrebno naslednje:

- **Sprotna vključitev novonastalih korpusov v metakorpus**
  Vsi slovensko pomembni korpusi, naj gre za nove različice obstoječih korpusov ali pa povsem nove, bi se morali sproti vključevati v korpus metaFida.
- **Posodabljanje postopka gradnje metakorpusa**
  Programsko opremo za izdelavo metakorpusa je treba vzdrževati in dopolnjevati.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.