

SLOVENSKI TERMINOLOŠKI PORTAL
V OKVIRU PROJEKTA
*RAZVOJ SLOVENŠČINE V DIGITALNEM
OKOLJU**Mateja JEMEC TOMAZIN*

ORCID 0000-0003-4053-1570

Tanja FAJFAR

ORCID 0000-0001-9410-7076

Mojca ŽAGAR KARER

ORCID 0000-0001-5302-1663

DOI 10.51587/2522-980X-2020-004-49-54

V prispevku predstavljamo zasnovo Slovenskega terminološkega portala, ki ga pripravlja konzorcij partnerjev projekta Razvoj slovenščine v digitalnem okolju. Portal bo omogočal iskanje po povezanih in vključenih terminoloških virih, izdelavo specializiranih korpusov ter lastnih terminoloških virov in zastavljanje terminoloških vprašanj. Terminološki portal bo na voljo javnosti v letu 2022.

Ključne besede: terminologija, terminološki portal, terminološki vir, terminološko svetovanje.

SLOVENE TERMINOLOGY PORTAL IN THE CONTEXT OF THE PROJECT *DEVELOPMENT OF SLOVENE IN A DIGITAL ENVIRONMENT*

Mateja JEMEC TOMAZIN

Tanja FAJFAR

Mojca ŽAGAR KARER

In the article, we present the concept of Slovene terminology portal which is prepared by the consortium of partners in the project Development of Slovene in a Digital Environment is considered in the article. The portal will enable searching for terminology in linked and incorporated terminology resources, creating specialised corpora, compiling terminology resources and terminology counselling. The terminology portal will be available to the users in 2022.

Keywords: terminology, terminology portal, terminology resource, terminology counselling .

Uvod

Projekt *Razvoj slovenščine v digitalnem okolju* (maj 2020–december 2022), ki ga financirata Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj, želi zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik, ki bodo omogočile nadaljnje raziskave, pomagale pa bodo tudi gospodarstvu pri razvoju novih izdelkov in storitev ter širši javnosti v vsakdanjem življenju. Končni izdelki bodo na uporabniku prijazen način pomagali pri sporazumevanju, sodelovanju, poslovanju, izmenjavi znanja in udeleževanju v družabnih in političnih razpravah. Zlasti pa bodo prispevali k premagovanju jezikovnih meja in pomagali pri sporazumevanju s pametnimi napravami v jezikih, ki jih pametne naprave obvladajo.¹ Jeziki z manjšim številom govorcev, kakor je slovenščina, se tako lahko v digitalnem svetu hitro izgubijo, če se programska oprema ne razvija tudi zanje. V konzorciju, ki izvaja projekt, sodeluje 12 ustanov iz Republike Slovenije, in sicer tri univerze, tri raziskovalne ustanove ter šest podjetij.

¹ Ta čas je komunikacija s pametnimi napravami mogoča le v 8 jezikih v okviru programa Amazon Alexa, v 13 jezikih za Google Assistant in v 21 jezikih za Apple Siri.

1. Slovenski terminološki portal

Sodelavci Terminološke sekcije Inštituta za slovenski jezik Frana Ramovša ZRC SAZU vodimo 5. delovni sklop omenjenega projekta, ki se imenuje Terminološki portal. V okviru delovnega sklopa je predvidena izdelava samostojnega terminološkega portala z *iskalnikom po terminoloških virih*. Prav tako bodo del portala spletni *konkordančnik za analizo specializiranih korpusov*, ki jih bo uporabnik lahko sestavil sam ali uporabil besedila z javno dostopnih spletnih mest, kakršno je Nacionalni portal odprte znanosti (openscience.si); *orodje za luščenje terminoloških kandidatov* iz prej omenjenih korpusov ter spletni *urejevalnik terminoloških virov*. Na portalu bodo objavljeni tudi smernice in navodila za sestavljanje terminoloških virov z vzorčnimi podatkovnimi zbirkami, ves čas pa bo delovala tudi svetovalnica za reševanje terminoloških vprašanj, na katera bomo odgovarjali sodelavci Terminološke sekcije. Sestavni del projekta je tudi priprava načrta razvoja in nadgradnje terminološkega portala po zaključku projekta decembra 2022.

2. Izhodišča

Osnovno vprašanje, ki smo si ga na Terminološki sekciji zastavili pred začetkom projekta, je, komu je namenjen terminološki portal in kakšni viri naj bodo na njem predstavljeni, saj je eno od izhodišč projekta tudi odprta znanost in odprto dostopni rezultati projekta, ki bodo na voljo za nadaljnjo uporabo v znanosti in gospodarstvu. Primarni uporabniki terminologije, to so področni strokovnjaki, želijo praviloma najti odgovore na svoja vprašanja, povezana z izbiro najprimernejšega poimenovanja, ustreznika v tujem jeziku ali preverjanja vsebine pojma na določenem strokovnem področju. Drugi uporabniki terminologije, npr. prevajalci, pa želijo vse podatke z različnih področij imeti zbrane na enem mestu, hkrati pa potrebujejo več dodatnih informacij o vsebini pojma, ker ne poznajo pojmovnega sistema stroke.

Od sistematičnega prikaza vseh podatkov, ki so v vključenih terminoloških virih, je odvisna celotna zasnova portala, funkcije iskalnika, prikaz podatkov, razvrščanje zadetkov, hkrati pa morajo takšno zasnovo dopolnjevati tudi navodila za uporabo iskalnika in smernice za samostojno izdelavo terminološkega vira. Praktični del smernic bodo predstavljale tudi vzorčne zbirke, vključno s terminološko svetovalnico.

3. Funkcionalnosti terminološkega portala

Storitve terminološkega portala bodo dostopne na dva načina – z registracijo uporabnikov in brez nje. Vsakdo bo lahko brez registracije iskal po slovarskih sestavkih, dostopnih na terminološkem portalu, za uporabo luščilnika in urejevalnika terminoloških virov ter pošiljanje vprašanj v terminološko svetovalnico pa bo zahtevana registracija uporabnika z veljavnim elektronskim naslovom.

3.1 Iskalnik

Privzeta oblika iskanja bo *osnovno iskanje*, kjer bo uporabnik lahko iskal po poljubnem nizu znakov, pri čemer bo iskalnik preverjal zadetke najprej med slovenskimi termini, nato med tujimi termini in nato tudi med drugimi elementi (npr. v definiciji, med sinonimi). Ker bodo vključeni tudi dvojezični terminološki slovarji, bo možno na začetku izbrati jezik, vendar bo iskanje tudi brez izbire jezika potekalo po vseh elementih.

Najprej bodo prikazani popolni zadetki, nato zadetki, ki vsebujejo iskani niz, in sicer najprej v elementu *iztočnica*, nato *tuji termin* in na koncu še v vseh drugih elementih, zlasti definiciji.

Iskanje ne bo mogoče po slovarjih, ki jih avtorji (še) ne bodo odprli za javnost.

3.2. Uporabniški korpusi in luščenje terminoloških kandidatov

Posebna spletna storitev bo omogočala oblikovanje lastnih specializiranih korpusov iz besedil, ki jih ima uporabnik na svojem računalniku, ali besedil, ki so že objavljena na javnih repozitorijih pod ustrezno licenco.² Zaželeno oblika besedila je txt-datoteka, ki omogoča najboljše rezultate, vendar bodo možni tudi drugi splošni besedilni formati (npr. .doc, .docx, .pdf). Modul luščenja bo vseboval dve osnovni orodji. Prvo orodje bo iz besedil izluščilo terminološke kandidate v obliki seznama, ki ga bo uporabnik lahko obdeloval naprej, drugo pa bo v besedilih označilo terminološke kandidate. Pri procesu nalaganja besedil v spletni vmesnik, namenjen luščenju, bo izvedena tudi segmentacija, tokenizacija, lematizacija in skladiščno označevanje. Za omenjene procese se v času projekta uporablja cevovod CLASSLA (<https://pypi.org/project/classla/0.0.4/>), ki bo po zaključku projekta zamenjan z novim, izboljšanim cevovodom, ki se

² Pogoji uporabe portala sledijo odprti znanosti, vendar bo moral vsak uporabnik sam označiti, da ima potrebne materialne avtorske pravice za uporabo besedil. Takšna rešitev je v skladu z načeli odprte znanosti in mednarodnih infrastruktur, kot sta CLARIN ERIC ter DARIAH.

razvija v času projekta. Vsak uporabnik bo lahko imel shranjenih do 5 svojih korpusov, nato pa se bodo najstarejše verzije izbrisale. Rezultate luščenja bo moč izvoziti na lokalni računalnik.

3.3. Urejevalnik terminoloških virov

S spletnim urejevalnikom za izdelavo terminoloških virov bodo lahko posamezni uporabniki ali skupina avtorjev urejala terminologijo svojega področja. Poleg vzorčnih zbirk, ki bodo prikazovale način sestavljanja slovarskih sestavkov, in smernic bodo lahko uporabniki sestavili svojo shemo z elementi, ki bodo na voljo. Minimalna zahteva je, da ima vsak terminološki vir na portalu najmanj dva elementa – poleg slovenskega termina še definicijo ali termin v tujem jeziku. Možno pa je izbrati še dodatne elemente, npr. pojasnila, sinonime in druge povezane termine, dodati pa bo mogoče tudi aktivne povezave do zvočnih ali video posnetkov ter slik. Slovarski sestavki ne bodo objavljeni sproti, ampak jih bo moral avtor aktivno objaviti, objavo pa potrditi administrator terminološkega portala.

Pomemben del nastajanja terminološkega vira so tudi metapodatki, ki bodo beleženi avtomatsko (npr. datum vnosa sestavka, datum spremembe sestavka, avtorji sestavka), podatke o fazi urejanja pa bo treba vnašati ročno. Pravice sodelovanja pri nekem slovarju bo vedno dodeljeval tisti, ki bo slovar začel oblikovati, saj bo s tem avtomatsko dobil vlogo *glavnega urednika*, drugi vključeni uporabniki pa bodo dobili vlogo *urednika*.

Za lažjo komunikacijo med avtorji bo omogočeno oddajanje komentarjev.

Vsak vir bo glavni urednik uvrstil na določeno področje po klasifikaciji CERIF (<https://www.arrs.si/sl/gradivo/sifranti/sif-cerif-cercs.asp>).

Vsak terminološki vir, ki bo nastal na portalu, bo možno izvoziti tudi lokalno.

3.4. Terminološka svetovalnica

V Terminološki sekciji že vrsto leto opravljamo terminološko svetovanje (<https://isjfr.zrc-sazu.si/sl/terminologisce/svetovanje#v>). Povezava do te storitve bo vključena tudi na novi terminološki portal, odgovor pa bomo pripravili sodelavci Terminološke sekcije. Del vzorčnih zbirk bo tudi celotna zbirka terminoloških odgovorov, ki je objavljena na zgornji povezavi.

4. Odprtost rezultatov projekta

Vsa programska koda in zbirke podatkov, ki bodo nastale med projektom, bodo dostopne javno pod odprtokodno licenco CC BY-SA 4.0 na GitHubu (<https://github.com/clarinsi>). Vse aplikacije (prepoznavna govora, transkripcija, strojno prevajanje, terminološki portal in luščenje terminologije) pa bodo (po vzoru latvijskega portala hugo.lv) na voljo tudi na javnem portalu RSDO, kjer jih bo vsak lahko preizkusil in uporabljal, viri pa bodo naloženi na repozitorij Clarin.si (<https://www.clarin.si/repository/>). To zagotavlja tudi dolgoročno dostopnost in vzdrževanje rezultatov po koncu projekta.

Sklep

Slovenski terminološki portal bo omogočil sodobne tehnološke rešitve za samostojno zbiranje in obdelavo terminologije različnim skupinam uporabnikov. Omogočil bo nastajanje manjših področnih slovarjev, za katere se oblikuje potreba med skupino strokovnjakov, ki pa nimajo orodij, s katerimi bi lahko te vire izdelali. Prav tako bo na enem mestu ponudil zadetke iz doslej na več mestih objavljenih terminoloških virov, kar bo olajšalo iskanje, seveda pa se sestavljalcem zavedamo, da obstajajo tudi omejitve. S partnerji, ki poleg Inštituta za slovenski jezik Frana Ramovša ZRC SAZU sodelujejo v projektu terminološkega portala (podjetje Amebis, Institut "Jožef Stefan", Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru ter Fakulteta za upravo, Fakulteta za računalništvo in informatiko ter Filozofska fakulteta Univerze v Ljubljani), skušamo zmanjšati tehnične ovire ter narediti uporabniku prijazne spletne vmesnike ter oblikovati čim bolj preprosta in jasna navodila za samostojno delo – pri tem pa se zavedamo, da bo kakovost terminologije še vedno najbolj odvisna od ljudi.