



Gregorčičeva 20–25, SI-1001 Ljubljana

T: +386 1 478 1000

F: +386 1 478 1607

E: gp.gs@gov.si

<http://www.vlada.si/>

Številka: 61400-4/2015/4

Datum: 13. 11. 2015

Akcijski načrt za jezikovno opremljenost

Krovni dokument: Resolucija o nacionalnem programu za jezikovno politiko 2014–2018 (ReNPJP14-18)
(v nadaljnjem besedilu: Resolucija), točka 2.2 – Jezikovna opremljenost

KAZALO

0	Uvod	4
1	Pregled področja	5
1.1	<i>Splošno</i>	5
1.2	<i>Jezikovni opis</i>	5
1.3	<i>Standardizacija</i>	9
1.4	<i>Terminologija</i>	10
1.5	<i>Večjezičnost</i>	12
1.6	<i>Jezikovne tehnologije</i>	13
1.7	<i>Digitalizacija</i>	15
1.8	<i>Govorci s posebnimi potrebami</i>	16
2	Definiranje ciljev, aktivnosti in akcij	18
2.1	<i>Definiranje ciljev</i>	18
2.2	<i>Definiranje aktivnosti za uresničitev ciljev</i>	19
2.3	<i>Opredelitev akcij ter njihova finančna ocena in časovna umestitev</i>	19
3	Podroben opis akcij	27
	<i>Akcija S-1: Infrastrukturni center</i>	28
	<i>Akcija S-2: Zakonodaja za odprti dostop</i>	30
	<i>Akcija S-3: Vnašanje virov v Wikije</i>	31
	<i>Akcija P-1: Terminološki portal</i>	32
	<i>Akcija P-2: Aplikacija za nove terminološke zbirke</i>	34
	<i>Akcija P-3: Terminološko svetovanje</i>	35
	<i>Akcija P-4: Javna dostopnost terminoloških virov</i>	36
	<i>Akcija P-5: Samodejno luščenje terminologije</i>	37
	<i>Akcija P-6: Večjezični portal</i>	38
	<i>Akcija P-7: Slovnični portal za šolarje</i>	39
	<i>Akcija P-8: Spletna predstavitev normativnih podatkov</i>	40
	<i>Akcija P-9: Spodbujanje prožnih in inovativnih oblik učenja z razvojem jezikovnih virov in tehnologij</i>	41
	<i>Akcija K-1: Pisni korpusi</i>	42
	<i>Akcija K-2: Specializirani korpusi in konkordančniki</i>	44
	<i>Akcija K-3: Govorni korpus in baza</i>	47
	<i>Akcija K-4: Dvojezični korpusi</i>	49
	<i>Akcija K-5: Diahroni korpus</i>	51
	<i>Akcija L-1: Splošni slovar in slovarska baza</i>	52
	<i>Akcija L-2: Slovenski lingvistični atlas</i>	54
	<i>Akcija L-3: Slovar naselbinskih imen</i>	55
	<i>Akcija L-4: Terminološki slovarji različnih strok</i>	56
	<i>Akcija L-5: Zgodovinski slovar</i>	58
	<i>Akcija L-6: Imenske entitete v slovenščini</i>	59
	<i>Akcija L-8: Strojna slovnična analiza</i>	62
	<i>Akcija L-9: FrameNet za slovenščino</i>	64
	<i>Akcija L-10: sloWNet</i>	66
	<i>Akcija L-11: Anafora in koreference</i>	67

Akcija L-12: Zbirka normativnih zadreg	68
Akcija G-1: Zasnova slovnice	69
Akcija G-2: Šolska slovnica	70
Akcija N-1: Pravopisna pravila in slovar	71
Akcija D-1: Digitalizacija čistopisov pisne kulturne dediščine	73
Akcija D-2: Digitalizacija jezikovnih priročnikov	74
Akcija A-1: Razpoznavalnik tekočega govora	75
Akcija A-2: Strojni prevajalnik	77
Akcija A-3: Sistem dialoga	78
Akcija A-4: Bralnik sporočil gospodinjskih aparatov	80
Akcija A-5: Referenčni črkovalnik in slovnični pregledovalnik	81
Akcija Z-1: Korpus in slovar znakovnega jezika	83
Priloga 1: Pregled ciljev in aktivnosti po sklopih.....	85

0 Uvod

V skladu z Resolucijo, točka 2.2, ter na podlagi podrobne preučitve področja in potreb so bili definirani naslednji

tematski sklopi:

1. Splošni del
2. Jezikovni opis
3. Standardizacija
4. Terminologija
5. Večjezičnost
6. Jezikovne tehnologije
7. Digitalizacija
8. Govorci s posebnimi potrebami

Vsebina akcijskega načrta je organizirana v 3 poglavja. V prvem poglavju pregledamo področje za vsakega od tematskih sklopov in pričakovani razvoj. V drugem poglavju specificiramo cilje za vsakega od tematskih sklopov in za vsakega od ciljev opredelimo vse potrebne aktivnosti za njegovo uresničitev. Ker se aktivnosti pri mnogih ciljeh medsebojno prepletajo, naredimo zatem pregled po aktivnostih, opredelimo, za katere cilje so posamezne aktivnosti potrebne, jih finančno ovrednotimo ter časovno umestimo. V tretjem poglavju vsako od potrebnih aktivnosti, opredeljenih v poglavju dve, podrobneje opišemo. Akcijski načrt predlaga aktivnosti, ki bi jih bilo za uresničitev resolucije in v njej določenih ciljev izvesti, pri tem pa nikakor ne prejudicira morebitnih izvajalcev posameznih aktivnosti, temveč samo določa nosilce posameznih ciljev in ukrepov glede na pristojnosti posameznih resornih organov. Izvajalci posameznih aktivnosti bodo, tako kot to določajo pravni akti, izbrani na javnem razpisu.

1 Pregled področja

1.1 Splošno

V zadnjih nekaj desetletjih, najbolj izrazito pa prav v zadnjem desetletju, je digitalizacija prinesla izredne spremembe na področje jezikovne opremljenosti, ki ne zadevajo samo zunanje forme izdelkov za uporabnike, ampak pretresajo jezikoslovje vse do njegovih teoretskih temeljev, tehnologije pa razvijajo nove gospodarske panoge (zadnji čas to nišo najbolj prepoznavno zaseda Google) in prinašajo velike upe za ljudi s posebnimi potrebami.

Čeprav je področje jezikovne opremljenosti v resoluciji in tem akcijskem načrtu formalno razdeljeno v osem sklopov, se prav vsi sklopi med seboj močno prepletajo. Osrednja povezovalna nit med njimi so zaledni jezikovni viri (korpusi, slovarji in leksikalne baze, govorne baze ...) in osnovna JT-orodja (tokenizatorji, oblikoslovni označevalniki, skladijski razčlenjevalniki ...), ki so nujni tako za tehnološki razvoj kot za sodoben opis jezika in izdelavo sodobnih priročnikov, kakršne uporabniki danes pričakujejo.

Prehod v digitalno področje in razvoj številnih novih, v preteklosti nepoznanih načinov (p)opisa jezika, je sprožil dvoje stvari, ki ju lahko prepoznamo danes kot najbolj pereči skupni točki celotnega področja jezikovne opremljenosti: prva je potreba po centru, ki bo (digitalne) vire hranil, distribuiral in tudi vzdrževal (kar je zlasti zahtevno npr. pri raznih spletnih servisih); druga je potreba po čim bolj odprti in prosti dostopnosti izdelanih virov, orodij oz. aplikacij – vsaj tistih, ki so financirana iz javnih sredstev, s tem povezana pa je tudi pereča težava varovanja avtorskih pravic, zaradi katerega nekateri nadvse pomembni viri in baze ostajajo (vsaj delno) zaprti. Prvo potrebo naj bi, ob primernem financiranju, zadovoljila novo ustanovljena slovenska raziskovalna infrastruktura CLARIN.SI, ki bo nudila tudi trajen repozitorij jezikovnih virov. K zadovoljitvi druge potrebe naj bi pripomoglo dejstvo, da se je Slovenija kot članica EU in OECD zavezala k upoštevanju sporazumov o spodbujanju prostega dostopa do raziskovalnih podatkov, kar jezikovni viri v večini primerov nedvomno so, ter, da smo v Sloveniji pred kratkim sprejeli tudi Smernice za zajem, dolgotrajno ohranjanje in dostop do kulturne dediščine v digitalni obliki, ki tudi priporočajo odprt dostop do (pisne) kulturne dediščine. Ne nazadnje pa tudi pričujoči Akcijski načrt predvideva ukrepe za zagotavljanje prostega in odprtega dostopa do izdelanih jezikovnih virov.

1.2 Jezikovni opis

Področje jezikovnega opisa v skladu z Resolucijo obsega izdelavo slovarjev, slovnice, različnih korpusov za slovenski jezik, lingvističnih atlasov in opis zgodovinskega razvoja slovenskega jezika. Resolucija v zvezi z akcijskim načrtom eksplicitno navaja naslednje poudarke:

- o slovarskem opisu: »Akcijski načrt za jezikovno opremljenost mora predvideti tak slovarski opis sodobne slovenščine, ki bo prilagojen uporabi v spletnem in drugih digitalnih okoljih, pri njem pa bodo bodisi v eni ali v različnih slovarskih bazah na voljo različni podatki o slovenščini, od pomenskih, vezljivostnih, kolokacijskih, frazeoloških, idiomatskih, etimoloških, normativnih in drugih.«
- o drugih slovarjih: »Za dodatno jezikovno opremljenost bo treba dokončati, sestaviti oziroma prilagoditi digitalnemu okolju še slovarje, pri katerih smo

Slovenci v zamudi, zlasti sinonimni slovar, slovar besednih oblik vseh standardnih nelastnoimenskih besed, ki bo vseboval tudi podatke o dinamičnem in tonemskem naglasu, slovar zemljepisnih imen slovenskega narodnega prostora in eksonimov, slovar slovenskih osebnih imen (rojstnih in priimkov); za ohranitev slovenščine kot strokovnega jezika bo treba kontinuirano sestavljati terminološke slovarje različnih strok.«

- o korpusih: »Prihodnji akcijski načrt za jezikovno opremljenost mora upoštevati stalen in kontroliran razvoj temeljnih korpusov, ki zagotavljajo možnost kontinuirane analize slovenskega jezika za potrebe jezikovnega opisa. Med njimi so referenčni korpus slovenščine, širitev in nadgradnja govornega korpusa, korpus besedil s spleta, multimodalni korpusi (besedilo + slika + zvok), korpusi pisne produkcije učencev in dijakov ter drugi korpusi, ki jih bo predvidel prihodnji akcijski načrt.«
- o slovnici: »V obdobju, ki ga pokriva resolucija, je treba začeti načrtovanje izdelave znanstvene slovnice slovenskega jezika, ki bo prikazoval današnji slovnični ustroj slovenskega knjižnega/standardnega jezika kot povezovalnega jezika vseh Slovencev.«
- o uporabniških priročnikih: »Na osnovi teh temeljnih del je treba izdelati priročnike za druge ciljne uporabnike. To so predvsem slovarji in slovnice za različne stopnje izobraževanja rojenih govorcev, za tiste, katerih materni jezik ni slovenščina, in za osebe s posebnimi potrebami.«
- o dialektoloških in jezikovnozgodovinskih opisih: »|Treba je| posvetiti pozornost tudi dialektološkim raziskavam, zlasti pisanju lingvističnih atlasov, narečnih slovarjev in monografij o posameznih (zamirajočih) narečjih, ter zgodovinsko- in primerjalnojezikoslovnim raziskavam, zlasti sestavi zgodovinskega slovarja slovenskega jezika, zgodovinske slovnice, prenovi etimološkega slovarja itd.«

V zvezi s temi področji Resolucija povzame kot temeljni ukrep:

- »sprejetje temeljnih usmeritev za izdelavo temeljnih in specializiranih priročnikov sodobnega slovenskega knjižnega/standardnega jezika in temeljnih ter specializiranih priročnikov ter drugih slovenističnih del s področja dialektologije, zgodovinskega in primerjalnega jezikoslovja (5–10 let) na nacionalni ravni ter njihovo izdelovanje – kot del Akcijskega načrta za jezikovno opremljenost«.

Dodana pa sta še dva predvidena temeljna ukrepa:

- »sprejetje takega bibliometričnega vrednotenja slovarskih in drugih leksikografskih del, ki bo raziskovalce spodbujal k sodelovanju pri sestavljanju temeljnih in specializiranih priročnikov za slovenski jezik;«
- »oblikovanje prosto dostopnega spletnega portala s čim več dosegljivimi jezikoslovnimi podatki o slovenščini, namenjenega splošnim uporabnikom in strokovni javnosti«.

V nadaljevanju pregledamo obstoječe stanje na področju temeljnih in specializiranih del in priročnikov za slovenski jezik.

Temeljni informativno-normativni slovar slovenskega jezika

Temeljni informativno-normativni slovar slovenskega jezika je Slovar slovenskega knjižnega jezika, ki je bil v končni različici (in prvotisku) izdan po delih v letih 1970, 1975, 1979, 1985 in 1991, dopolnjen s Slovarjem novejšega besedja slovenskega jezika, ki je bil izdan 2012. SSKJ je na voljo v knjižni in prosto dostopni elektronski obliki, SNB pa poleg tega še pod pogoji Creative Commons 2,5. Inštitut za slovenski jezik Frana

Ramovša je oktobra 2014 predstavil drugo izdajo SSKJ, v kateri je gradivo obeh slovarjev združeno, pregledano, dodatno dopolnjeno in kolikor mogoče aktualizirano glede na današnjo podobo slovenščine. Ob že dlje časa prepoznani potrebi po izdelavi splošnega slovarja na osnovi aktualnega gradiva je bil v l. 2008 v okviru ZRC SAZU in SAZU organiziran Strokovni posvet o novem slovarju slovenskega jezika (zbornik prispevkov dostopen na <http://zalozba.zrc-sazu.si/sites/default/files/9789612541149.pdf>), l. 2013 pa je bila sprožena konkretna iniciativa treh avtorjev (Simon Krek, Iztok Kosem, Polona Gantar) s Predlogom za izdelavo Slovarja sodobnega slovenskega jezika (<http://www.sssj.si/>), tej pa je v začetku leta 2014 sledil še Posvet o novem slovarju slovenskega jezika, organiziran v okviru Ministrstva za kulturo (zbornik prispevkov dostopen na http://www.mk.gov.si/si/delovna_podrocja/sluzba_za_slovenski_jezik/predstavitev_podrocja/dogodki/posvet_o_novem_slovarju_slovenskega_jezika/), sprejeti pa so bili tudi sklepi o novem slovarju, kot so citirani v nadaljevanju:

1. »Posvet je izpostavil potrebo po nadaljevanju pogovorov o zasnovi novega slovarja slovenskega jezika, pri čemer je treba – prisotni pa se strinjajo, da je tudi mogoče – doseči interdisciplinarno sodelovanje strokovnjakov z različnih institucij – inštitutov, univerz in drugih.
2. Končna oblika slovarja v smislu digitalno : natisnjeno ni medsebojno izključujoča. Slovarska baza se zasnove digitalno, slovar bo prosto spletno dostopen, baza odprta za nekomercialno rabo, tisk slovarja oz. posameznih za to formo primernih oz. prirejenih delov pa je naknaden, vendar predviden že v zasnovi.
3. Slovar bo jezikovnotehnološko zasnovan. Ob tem je treba zagotoviti konceptualno enotno, predvidljivo in standardizirano metodologijo leksikografske obdelave podatkov ter dograjevati in dopolnjevati razvidne ter dokumentirane korpusne gradivne vire in orodja.
4. Slovar bo informativno-normativen.
5. Slovar bo poleg leksikografov združil moči strokovnjakov za zgodovino jezika, etimologijo, govor, normo, stilistiko, terminologijo, podatkovne baze, korpusne in druge, ki lahko skupaj izdelajo izdelek, primerljiv s sodobnimi evropskimi slovarji, in bo primeren ter ustrezno oblikovan za različne uporabnike.«

Po posvetu o novem slovarju sta se oblikovali dve iniciativi, ki želita sklepe slovarja uresničiti v novem slovarskem projektu. Prva je nastala na podlagi vabila Inštituta za slovenski jezik ZRC SAZU na spletnem forumu SlovLit (<http://mailman.ijs.si/pipermail/slovlit/2014/004721.html>), druga pa se je v okviru Centra za jezikovne vire in tehnologije organizirala v Konzorcij za jezikovne vire in tehnologije (<http://cjvt.si/projekti/>).

Drugi slovarji

Poleg osrednjega, splošnega slovarja so pomembnejši še naslednji slovarji, ki so najaktualnejši vsak za svoje področje:

- Slovenski pravopis (v prvotisku izdan l. 2001, l. 2003 izšel z vsebinskimi popravki v knjižni in elektronski obliki na CD-ROM-u; na spletu od l. 2010),
- Etimološki slovar slovenskega jezika (izdan 1976, 1982, 1995, 2005, 2007, dostopen v knjižni obliki),
- Slovenski etimološki slovar, v prvotisku izdan l. 2003 (dostopen v knjižni obliki),

- Slovar slovenskih frazemov, izdan 2011 (dostopen v knjižni obliki),
- Vezljivostni slovar slovenskih glagolov (dostopen v knjižni obliki in elektronsko, prek portala Termania),
- sinonimni slovar slovenskega jezika je v izdelavi in bo po navedbah izvajalca, Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, dokončan v času trajanja veljavnosti Resolucije,
- kot slovar besednih oblik 100.000 občno- in lastnoimenskih (10 %) besed trenutno služi leksikon Sloleks (dostopen v elektronski obliki), ki ni standardiziran in ni naglasno opremljen,
- po navedbah izvajalca, Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, je v izdelavi baza dinamično in tonemsko onaglašanih oblik slovenskih knjižnih/standardnih nelastnoimenskih besed, ki bo dokončana v času trajanja veljavnosti Resolucije.

Jezikovni opis lastnih imen

Naselbinska zemljepisna imena krajev znotraj RS so skupaj s pripadajočimi tvorjenkami dostopna v leksikonu Slovenska krajevna imena¹, imena zamejskih krajev v Italiji in Avstriji pa v Merkujevem² oz. Zdovčevem priročniku³, ki navajata tudi nekaj nenaselbinskih. Naselbinska imena iz Porabja niso nikjer sistematično opisana. Rojstna lastna imena in priimki so dostopni v bazi Statističnega urada RS, kjer seveda niso opremljeni z jezikovnimi podatki.

Slovnica

Za slovenščino imamo samo eno znanstveno slovnico, to je Toporišičeva Slovenska slovnica (aktualna je njena četrta izdaja iz leta 2004, temelji pa so v prvi izdaji iz leta 1976). Slovnica je po teoretskem izhodišču strukturalistična.

Dialektološki opis

Na področju dialektologije je eden večjih projektov Slovenski lingvistični atlas, ki temelji na zasnovi Frana Ramovša iz leta 1934, dejanske priprave nanj pa so se začele po drugi svetovni vojni na Inštitutu za slovenski jezik (ZRC) SAZU. Prvi zvezek SLA je izšel leta 2011, pokriva pa besedišče za področje človeka (telo, bolezni, družina). Ostala področja, ki jih predvideva popis SLA, so še obleka, hiša, vas, orodje, živina, rastline, planina, čas, pokrajina, štetje, razno. Projekt SLA deloma poteka znotraj raziskovalnega programa skupine na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU. Dialektološke raziskave in objave sicer tečejo tudi na več oddelkih za slavistiko ali slovenistiko v RS, delno pa tudi zunaj nje (npr. v Gradcu), kot zanimivo pobudo za prehod v izkoriščanje korpusnih orodij in metodologij pa omenimo poskusni osnutek dialektološkega korpusa (GOKO), predstavljen na primorski univerzi (l. 2012).

Jezikovnozdgovinski opis

Omeniti velja predvsem leta 2011 izdano Besedje slovenskega knjižnega jezika 16. stoletja, ki je nastalo na osnovi izpisov iz knjižnih izdaj iz obdobja 1550–1603 in je bilo deloma predhodno izdan tudi v tujini (Biblia Slavica, 2006). V ta sklop sodijo tudi Slovar jezika Janeza Svetokriškega, obrnjeni Kastelec-Vorenčev Dictionarium Latino-Carniolicum (1680–1710), drugi večji obrnjeni slovarji (Megiser 1592, 1603, Pohlin 1781, Gutsman 1789) ter gesla, ki jih skupaj s konceptom za zgodovinski slovar 16. stoletja prinaša Poskusni snopič Slovarja jezika slovenskih protestantskih piscev 16. stoletja (2001).

¹ Franc Jakopin, Tomo Korošec, Tine Logar, Jakob Rigler, Roman Savnik, Stane Suhadolnik, Miran Hladnik, Slovenska krajevna imena, Ljubljana, Cankarjeva založba, 1985.

² Pavle Merku, Slovenska krajevna imena v Italiji, Trst, Mladika, 1999.

³ Pavel Zdovc, Slovenska krajevna imena na avstrijskem Koroškem, Celovec, Slovenski znanstveni inštitut, 1993.

Navedeni slovarji so prosto dostopni tudi preko novega spletnega portala Fran, ki ga je pripravil Inštitut za slovenski jezik Frana Ramovša ZRC SAZU.

Korpusi in baze

Večji projekti na področju korpusov in baz (z izjemo zgodovinskih korpusov) so se zaključili z letom 2012, od takrat korpusno gradivo večinoma ni bilo posodobljeno.

Kot najnovejši pisni korpus sodobnih slovenskih besedil šteje Gigafida (več kot 1 milijarda besed), iz nje pa je vzorčen uravnoteženi korpus KRES (100 mio. besed). Drugi večji besedilni korpus je Nova beseda (318 mio besed). Kot referenčni korpus govornega jezika šteje korpus GOS (1 mio. besed). Specializirani korpusi so ŠOLAR (besedila učencev in dijakov) ter terminološki korpusi: Evroterm (besedila predpisov EU), korpus vojaških besedil Grizold, korpus besedil odnosov z javnostmi, korpus turističnih besedil, korpus Dnevi slovenske informatike ...

Med baze, pomembne za jezikovni opis v slovarjih in slovnici, lahko štejemo predvsem Leksikalno bazo za slovenščino, ki predstavlja pomenski in skladenjski opis izbranega nabora slovenskih besed, že omenjeni oblikoslovni leksikon Sloleks in semantično mrežo sloWNet, ki vsebuje opis pomenskih razmerij med besedami po predlogi Wordneta. K tem je treba dodati še nastajajoči bazi sinonimov slovenskega jezika ter dinamično in tonemsko onaglašeni oblik slovenskih knjižnih/standardnih nelastnoimenskih besed, ki sta po navedbah izvajalca, Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, v izdelavi in bosta dokončana v času trajanja veljavnosti Resolucije.

1.3 Standardizacija

Standardizacijska dejavnost za slovenščino tradicionalno poteka kot kolektivni projekt skupine strokovnjakov v okviru akademije znanosti (SAZU), kot v družbi splošno znane in nesporne avtoritete, neodvisno od politike, kar je praksa, uveljavljena v veliko evropskih državah s primerljivo vlogo akademije znanosti.

Slovenski okvir: Pravila aktualnega normativnega priročnika, Slovenskega pravopisa 2001 (SP 2001), so nastala v 80. letih 20. stoletja, ob njihovem usklajevanju pa je bila prvič vzpostavljena komunikacija s širšo strokovno in laično javnostjo. Konsenzualno usklajena pravila so bila l. 1989 predložena v potrditev SAZU, v 90. letih so bila nato večkrat dopolnjena in usklajena v skladu z družbenopolitičnimi spremembami. Pravopisni slovar pa, izhajajoč iz pravil, je izšel l. 2001 kot rezultat sodelavcev Inštituta za slovenski jezik Frana Ramovša Znanstvenoraziskovalnega centra SAZU (ISJFR ZRC SAZU) v okviru programa Naravna in kulturna dediščina slovenskega naroda, ki ga izvaja od svoje ustanoviteljice ZRC SAZU neodvisen samostojni javni raziskovalni zavod.

V obdobju po izidu SP 2001 se je sodelovanje med SAZU in ZRC SAZU na pravopisnem področju nadaljevalo v okviru aplikativnega raziskovalnega projekta »Sodobni pravopisni priročnik v knjižni, elektronski in spletni različici«. V okviru tega projekta je bilo l. 2011 ugotovljeno naslednje: kodifikacijska ambicija naj bo utemeljena na novih smernicah, ki naj jih predstavi vseslovenski znanstveni posvet glede sprememb v standardizaciji, dokončno pa oblikuje delovno telo, ki bi pripravilo izdajo posodobljenih pravopisnih pravil. Po zaključku projekta je pravopisna skupina na ZRC SAZU nadaljevala s prizadevanji za argumentiranje sprememb v pravilih, h kateremu je bila povabljen zainteresirana strokovna javnost (*Pravopisna stikanja*, 2012).

Sočasno je bila v okviru projekta »Sporazumevanje v slovenskem jeziku« pri aktivnosti »Slogovni priročnik« pripravljena zasnova zbirke normativnih zadreg na osnovi vprašanj

v izbranih jezikovnih svetovalnicah ter specializiranega luščenja podatkov iz korpusa (2011), pripravljen je bil strojno generiran in ročno pregledan *Leksikon besednih oblik* (Sloleks) s pogostnostnimi podatki iz strojno označenega korpusa Gigafida: Sloleks še ni standardiziran.

Od l. 2012 je na spletnem mestu ISJFR odprta jezikovna svetovalnica (<http://isjfr.zrc-sazu.si/sl/svetovalnica>), ki uporabnikom svetuje pri normativnih in slovničnih vprašanjih in ki ponuja osnovno zbirko normativnih zadreg v slovenščini.

V l. 2013 je bila ustanovljena Pravopisna komisija pri SAZU (<http://pravopisna-komisija.sazu.si>). Člane komisije je izbral in predlagal v sprejetje SAZU Znanstveni svet ISJFR.

Formalni okvir: Raba slovenščine je na ozemlju RS določena z Ustavo RS (*»Uradni jezik v Sloveniji je slovenščina.«*), z Zakonom o javni rabi slovenščine ter področnimi zakoni. Temeljni strateški dokument, ki postavlja programske smernice jezikovne politike in v tem okviru tudi standardizacije, je Resolucija o nacionalnem programu za jezikovno politiko 2014–2018.

Resolucijska obravnava standardizacije se osredinja na štiri dejavnosti, ki potekajo vzporedno skozi celotno obdobje. Te dejavnosti so:

1. **Jezikovnoteoretična** (PRAVILA: ustanovitev strokovnega telesa, prek tega pa ugotovitev stanja v kodifikaciji, preučitev nakazanih smernic, določitev hierarhije standardizacijskih meril)
2. **Jezikovnotehnološka** (TEHNOLOGIJE: izgradnja jezikovnotehnološke infrastrukture, ki omogoča pridobitev verodostojne podatkovne zbirke, pomembne za preučitev realnega stanja v jezikovni normi)
3. **Jezikovnoaplikativna** (PRIROČNIKI: koncipiranje temeljnih normativnih jezikovnih priročnikov in aplikativnih aplikativno-utemeljevalnih del z normativnim značajem)
4. **Jezikovnosvetovalna** (SVETOVANJE: vzpostavitev svetovalnega telesa, ki bi delovalo prek prosto dostopnega spletnega portala)

Utemeljitev kronološke vzporednosti: Dejavnosti so medsebojno prepletene; stranski produkt jezikovnega svetovanja je sprotno detektiranje uporabniških potreb, ki vplivajo na posodobitev kodifikacije. Z vpogledom v podatkovno zbirko realiziranih izraznih možnosti v jeziku skupina strokovnjakov, ki posodablja pravopisna pravila, pridobi podatke o aktualni jezikovni situaciji, ki jo ob koncipiranju pravil sooči z jezikovnosistemskimi kriteriji in kriteriji izročila ter gospodarnosti v jeziku. Koncipiranje novih normativnih del aplikativno-utemeljevalnega značaja izhaja iz analize uporabniških potreb in zadreg ob interpretiranju obstoječega pravopisa.

1.4 Terminologija

V Sloveniji se s terminologijo načrtno ukvarja več ustanov, v prvi vrsti Sekcija za terminološke slovarje ISJFR ZRC SAZU, v sklopu Univerze v Ljubljani delujeta Center za družboslovnoterminološko in publicistično raziskovanje na Fakulteti za družbene vede ter Katedra za leksikologijo, terminologijo in jezikovne tehnologije Oddelka za prevajalstvo na Filozofski fakulteti, na Univerzi na Primorskem se s terminološkim raziskovanjem ukvarjajo na Fakulteti za humanistične študije.

Med spletnimi servisi, ki ponujajo tudi terminološke vire, je največkrat uporabljana Termania (www.termania.net), ki jo je zasnovalo podjetje Amebis in kjer so slovarji različne veljave in strokovne potrditve (tako uradni priročniki strok kot individualni, dvojezični geslovniki ipd.), večji projekt je tudi spletni slovar informatike Islovar (www.islovar.org), ki ga ureja Slovensko društvo Informatika, na portalu Terminologišče (isjfr.zrc-sazu.si/en/terminologisce) pa že omenjena terminološka sekcija ISJFR ZRC SAZU ponuja terminološko svetovanje in dostop do razlagalnih terminoloških slovarjev, ki so v okviru posameznih strok vsi standardizacijske narave.

S terminologijo EU se ukvarjajo prevajalci in terminologi v okviru nacionalnih in evropskih institucij, v Sloveniji za to področje delujeta spletna servisa Evroterm in Evrokorpus, na ravni EU pa se terminologija, tudi slovenska, zbira v bazi IATE. V okviru prejšnjega nacionalnega programa za jezikovno politiko je bil poskusno vzpostavljen nacionalni mehanizem za potrjevanje terminologije EU, ki pa se v tedanji obliki ni najbolje obnesel.

S terminologijo se poleg naštetih v različnih oblikah in obsegi ukvarjajo tudi strokovnjaki različnih strok, prevajalci in jezikoslovci; tako različni glosarji nastajajo v okviru študijskih dejavnosti kot priloge diplomskih, magistrskih in doktorskih nalog, pa tudi v okviru raziskovalnih in gospodarskih dejavnosti.

Ker je terminologija ključna za učinkovito upravljanje z znanjem, lahko opredelimo nekaj temeljnih sporočanjских situacij, kjer je urejenost terminologije nujna za neokrnjeno funkcionalnost slovenščine:

1. Kjer znanje nastaja, tj. v znanstvenoraziskovalnem okolju; tu poteka specializirana komunikacija v slovenščini in tujih jezikih, za razvoj slovenskega strokovnega izrazja pa lahko veliko naredijo strokovnjaki sami, če imajo na voljo ustrezno infrastrukturo za ustvarjanje terminoloških virov in pomoč terminologov;
2. Kjer se znanje posreduje naprej, tj. v izobraževalnem okolju, še posebej v visokošolskem izobraževanju; ta del je sicer neločljivo povezan z jezikovno opremljenostjo, vendar se z jezikovno politiko v izobraževanju ukvarja drug akcijski načrt in so zato ti ukrepi iz pričujočega dokumenta izvzeti;
3. Kjer se specializirana besedila pisno ali govorno prevajajo iz enega v drug jezik, tj. v prevajalskih okoljih; tu sta za učinkovito prevajanje ključna hitro pridobivanje terminoloških informacij iz poenotenih, javno dostopnih virov, in dostop do ažurnega terminološkega svetovanja;
4. Kjer se o znanju piše in govori v vseh ostalih situacijah, tj. v medijih, gospodarskih družbah, javnih ustanovah itd.; vse te jezikovne situacije ustvarjajo potrebo po poenotenih, javno dostopnih terminoloških virih za različna strokovna področja.

Glede na zgornji opis na področju terminologije zaznavamo predvsem naslednje potrebe:

- skupno in poenoteno spletno mesto za vse (doslej razpršene) terminološke vire,
- usklajena in enotna načela za gradnjo terminoloških zbirk, ki omogočajo enostavno distribucijo in izmenjavo,
- zmogljivo in brezplačno spletno aplikacijo za izdelavo terminoloških baz,
- orodje za luščenje terminologije iz besedil, ki pospeši izdelavo terminoloških baz za še neobdelana področja,
- boljši dostop do terminološkega svetovanja za prevajalce evropskih institucij in za splošno javnost.

1.5 Večjezičnost

Slovenščina v okviru EU in širše sodi med manj vplivne jezike, saj je med petimi evropskimi jeziki z najmanj govorci in jo govori manj kot odstotek Evropejcev. Prav zaradi tega je sporazumevanje v tujih jezikih za Slovence že od nekdaj ključnega pomena, po podatkih študije Evrobarometra 386 iz leta 2012 pa se je kar 92 odstotkov Slovencev sposobnih sporazumevati v vsaj enem tujem jeziku, kar je krepko nad evropskim povprečjem (54 %) in s čimer se Slovenija uvršča na peto mesto držav EU.

Če se morda podatki o znanju tujih jezikov za Slovenijo zdijo optimistični, tega ne moremo trditi za jezikovno opremljenost, ki bi podpirala večjezičnost. Za uspešen gospodarski razvoj in nemoteno prehajanje znanja, blaga, storitev, idej in kulturnih dobrin prek jezikovnih meja znanje tujih jezikov ni edini pogoj. Sporazumevanje in dostop do informacij je treba učinkovito in hitro zagotavljati v vseh situacijah, tudi kadar tujejezične kompetence govorcev ne zadoščajo, prav tako pa je treba vsem državljanom Slovenije omogočiti, da imajo do čim več informacij in kulturnih dobrin dostop v maternem jeziku. Če govorimo o večjezičnosti, je osrednja dejavnost, ki omogoča učinkovito delovanje v večjezičnem okolju, prevajanje, za kakovostno (človeško in strojno) prevajanje pa so neobhodno pomembni jezikovni viri in tehnologije.

Jezikovna industrija, ki zajema prevajanje, tolmačenje, podnaslavljanje in sinhronizacijo multimedijskih vsebin, globalizacijo spletišč, lokalizacijo programja, poučevanje jezikov in jezikovne tehnologije, je po zadnji oceni EU leta 2008 obsegala približno 8,5 milijard evrov, v svetovnem merilu pa je v poročilu *Language Services Market 2012* ocenjena že na prek 24 milijard evrov, pri tem pa se je nedavna kriza tako rekoč ni dotaknila. Tudi v Sloveniji je prevajalstvo pomembna dejavnost, saj je v registru AJPes za prevajanje in tolmačenje registriranih 1.111 poslovnih subjektov, poleg teh pa se velik zalogaj prevajalskega dela opravlja še v prevajalskih službah EU in nacionalnih vladnih organih. Prevajalci in tolmači pri delu uporabljajo slovarje, korpuse in pomnilnike prevodov, ki morajo biti kakovostni, sodobni in čim bolj obsežni. Kljub velikemu številu človeških prevajalcev pa je potreba po kakovostnem strojnem prevajanju vse bolj izrazita, v razvojni strategiji EU za naslednje šestletno obdobje je prioriteta Automated Translation uvrščena v sam vrh ključnih informacijsko-komunikacijskih tehnologij.

Študija Slovenski jezik v digitalni dobi (<http://www.meta-net.eu/whitepapers/e-book/slovene.pdf>) slovenščino glede stopnje razvoja na področju strojnega prevajanja uvršča v najnižjo skupino, predvsem pa je stanje nezavidljivo pri omogočitvenih virih in tehnologijah, kamor sodijo dvojezični splošni in specializirani leksikoni in slovarji, vzporedni in primerljivi korpusi, oblikoslovni, skladenjski in semantični analizatorji itd.

Iz opisane situacije sledi, da so potrebe za boljšo opremljenost za večjezičnost usmerjene predvsem v zagotavljanje virov in tehnologij za prevajalce, tolmače in vse ostale, ki bi s pomočjo strojnega prevajalnika za slovenščino lažje dostopali do informacij v tujih jezikih in se lažje večjezično sporazumevali, vse to pa velja tudi za vse tiste govorce, ki jim slovenščina ni prvi jezik. Pri tem ugotavljamo, da je smiselno najprej poskrbeti za prioritete jezike, ti pa so: angleščina kot svetovno najbolj razširjeni jezik mednarodne komunikacije; sosednji jeziki: nemščina, italijanščina, hrvaščina, madžarščina; drugi veliki evropski jeziki: francoščina, španščina, ruščina; sledijo pa drugi veliki svetovni jeziki in jeziki, s katerimi obstajajo posebne kulturne, gospodarske ali zgodovinske povezave. Ta prioritetni seznam ni izključujoč, saj so vsakršni dvo- in večjezični viri s slovenščino dobrodošli in koristni.

V akcijskem načrtu smo tako predvideli naslednje ukrepe:

- Vzpostavitev večjezičnega portala, ki omogoča enotno vstopno točko do vseh virov in orodij za večjezično sporazumevanje,
- zagotavljanje dvo- in večjezičnih slovarjev in baz, kjer je prvi korak zagotavljanje avtorskih pravic za obstoječe dvo- in večjezične vire ter omogočanje dostopa do njih prek večjezičnega portala, drugi korak pa je gradnja novih in posodabljanje starih slovarjev,
- gradnjo dvojezičnih korpusov za prioritetne jezikovne pare, saj ti skupaj s slovarji in jezikovnimi tehnologijami predstavljajo nujen vir za gradnjo strojnih prevajalnikov,
- prosto dostopen strojni prevajalnik za angleščino in druge prioritetne jezikovne pare.

1.6 Jezikovne tehnologije

Dve od osrednjih evropskih organizacij s področja jezikovnih tehnologij, ELSNET, Evropska mreža odličnosti na področju jezikovnih tehnologij, in ELDA, evropska Agencija za evalvacijo in distribucijo jezikovnih virov, sta na pobudo nizozemske iniciative v letu 1999 sprožili iniciativo BLARK, katere cilj je bil definirati minimalen nabor jezikovnih virov in JT orodij za potrebe jezikovnih tehnologij, ki naj bi jih zagotovili za čim večje število jezikov. Nabor temelji na nizozemskih izkušnjah, ko so na prelomu tisočletja definirali t. i. zemljevid jezikovnih tehnologij za Nizozemsko. Na podlagi tega velja danes nizozemski koncept za vzor vsem srednje velikim in manjšim jezikom (nizozemščina je materni jezik ca. 23 milijonov ljudi) na področju jezikovnih tehnologij⁴. Koncept BLARK se sicer ni razširil tako, kot so morda upali, je pa bil pred par leti podrobno izdelan tudi za arabščino, ki zadnji čas doživlja zelo aktiven razvoj na področju jezikovnih tehnologij. Čeprav je osnovni nabor tehnologij in virov star že več kot desetletje, je za slovenščino še vedno delno aktualen, saj mnoge od tehnologij in virov s seznama še manjkajo ali pa so nezadostno razvite. Tehnologije in viri so razdeljeni na jezikovne tehnologije in govorne tehnologije. Jezikovne tehnologije zajemajo predprocesiranje besedila (tokenizacija, prepoznavanje imenskih entitet), oblikoslovno analizo in sintezo, oblikoslovno označevanje, skladiščno analizo in sintezo, semantično analizo (vključno z analizo koreferenc, razdvajanjem pomenov ipd.), pragmatično analizo in analizo konteksta, enojezični leksikon (oblikoslovni in fonetični), skladiščno drevesnico in evalvacijska merila. Govorne tehnologije zajemajo razpoznavanje govora (vključno z razpoznavanjem za nematerne govorce in prilagoditvijo na posamezne govorce), sintezo govora, izračunavanje intervalov in ravni zaupanja, identifikacijska orodja (prepoznavanje govorca, sledenje govorca ipd.), orodja za polavtomatsko označevanje korpusov, govorne korpusne za posamezne aplikacije, multimodalne govorne korpusne

⁴ Prim.: Erhard Hinrichs et al., Description of the BLARK, the Situation of Individual Languages D5C-4, 2010-02-12 Version: 2, <http://www-sk.let.uu.nl/u/d5c-4.pdf>;

Lars Borin, Martha D. Brandt, Jens Edlund, Jonas Lindh, Mikael Parkvall, The Swedish language in the digital age, META-NET White paper, <http://www.meta-net.eu/whitepapers/e-book/swedish.pdf>;

Kiril Simov et al., A Language Resources Infrastructure for Bulgarian, <http://www.mt-archive.info/LREC-2006-Prys.pdf>;

Delyth Prys, The BLARK Matrix and its relation to the language resources situation for the Celtic languages, <http://www.mt-archive.info/LREC-2006-Prys.pdf>;

Mojgan Seraji, Beáta Megyesi, Joakim Nivre, A Basic Language Resource Kit for Persian, http://www.lrec-conf.org/proceedings/lrec2012/pdf/338_Paper.pdf;

Mansour Alghamdi, Chafic Mokbel and Mohamed Mrayati, Arabic Language Resources and Tools for Speech and Natural Language, <http://www.mghamdi.com/KACST&UOB.pdf>;

Steven Krauwer, The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap, <http://www.elsnet.org/dox/krauwer-specm2003.pdf>.

(avdio + video), večmedijske govorne korpuse (govor + internet ipd.) in večjezične (vzporedne) govorne korpuse ter evalvacijska merila.

Naslednji pomemben evropski dokument s področja stanja jezikovnih tehnologij je izšel v zbirki Bela knjiga v okviru projekta META-NET. Namen zbirke Bela knjiga je preverjanje jedrnih tehnologij za vse evropske jezike. Za slovenščino (v kontekstu 31 evropskih jezikov) je pregled in oceno stanja na področju jezikovnih tehnologij pripravil Simon Krek (Slovenski jezik v digitalni dobi, 2012).

Za slovenščino so ugotovitve povzete takole:

- »Osnovna orodja za jezikovnotehnološko procesiranje besedil obstajajo, manjkajo pa vsi viri in orodja za naprednejše procesiranje.
- Na področju govornih tehnologij je sinteza govora trenutno najbolj razvita tehnologija. Razpoznavanje govora je omejeno na povsem osnovne aplikacije in orodja. Splošna dostopnost orodij in virov pri govornih tehnologijah je relativno nizka.
- Obsežnost vseh virov je resna težava. Celo v primerih, ko so viri visoko kvalitetni, niso dovolj obsežni.
- Ne obstaja skupna infrastruktura za hranjenje, vzdrževanje in distribucijo izdelanih virov in orodij ter skupna organizacijska platforma za sodelovanje akterjev na tem področju, kar je ključna težava.«

Bela knjiga META-NET-a napoveduje tudi strategijo razvoja jezikovnih tehnologij do leta 2020. Pričakuje velik razrast podobnih aplikacij, kot so iPhonov Siri, Microsoft Cortana ali upravljanje navigacije v avtomobilih, do leta 2020. V zvezi s tem izpostavlja kot posebej obetavni dve področji: robotiko in informacijske storitve. Pričakujemo lahko različne mobilne robote za osebne storitve, reševalne akcije, domačo oskrbo, varovanje in nadzor, ki bodo znali z nami komunicirati govorno – govoreče robote. Za razvoj tovrstne tehnologije je trenutno ena večjih ovir uspešna avtomatska razpoznavanje govora, ki bi morala med drugim premagovati tudi težave šumov v okolju in gibanja uporabnika in robota po prostoru.

Drugo, verjetno še bolj pomembno področje implementacije jezikovnih tehnologij so spletni iskalniki, ki že kažejo tendenco po izkoriščanju jezikovnotehnoloških virov in aplikacij s področja oblikoslovne, skladišne in semantične analize besedila za izboljšanje uspešnosti. Splet postaja nasploh vseprežemajoča sled naših aktivnosti in potenciali se iščejo tudi v smeri družabnega spleta. Pričakuje se nadaljnji razmah t. i. spleta 2.0, ki se nanaša na družabna omrežja, bloga ipd. Vendar družabni splet ne more izkoristiti vseh svojih potencialov, ker ob poplavi vsebin ne moremo več ohraniti pregleda. Do leta 2020 se pričakuje, da bodo razvita orodja, ki bodo med drugim sledila, analizirala, povzemala, strukturirala, dokumentirala in vizualno prikazovala dinamiko družabnega spleta.

Največji potencial JT se kaže na področju strojnega prevajanja, ki ima – če bi doseglo stopnjo ustrezne uspešnosti – skorajda nepredstavljivo velik pomen, ki je v večjezični evropski skupnosti in globalno povezanem svetu vsako leto bolj izrazit. Ob ustrezni pokritosti s potrebnimi viri in JT orodji se danes uspešnost strojnega prevajanja že bliža osnovni uporabnosti, tj. kot pomoč prevajalcem oz. za osnovno hitro razumevanje vsebine, kadar ni na voljo druge poti (zlasti splet), tako v dokumentih META-NET-a kot v mednarodni JT skupnosti nasploh pa prevladuje prepričanje, da obstaja še precejšen potencial za izboljšanje. Razvoj strojnega prevajanja se pričakuje tudi na področju govora, npr. za sestanke na daljavo. Konkretno aplikacijo na tem področju je prav pred kratkim (maj 2014) predstavil Microsoft, ki je izdelal dodatek za strojno prevajanje pogovorov prek Skypa za jezikovni par angleščina-nemščina.

Med ključnimi aplikacijami, v katere so integrirane JT, seveda ne smemo pozabiti na črkovalnike in slovnične pregledovalnike, ki danes še najhitreje od vseh JT-aplikacij najdejo pot do široke množice uporabnikov. V prihodnosti lahko pričakujemo razvoj tudi v smeri analize in popravljanja slogovne ustreznosti besedil glede na žanr. Pričakujemo lahko aplikacije, ki bodo pripravljale in predlagale osnutke e-sporočil, tako da jih bomo samo še uredili. Prejeta e-sporočila bodo semantično analizirana in povzeta v poročilih, prepoznani pa bodo tudi emocionalni elementi prejetih sporočil.

Kot vse bolj pomemben potencial uporabe se kaže tudi področje izobraževanja. Za širše uporabnike manj opazen, za prizadete pa tako rekoč življenjsko pomemben je tudi potencial JT za osebe s posebnimi potrebami.

1.7 Digitalizacija

Prehod v digitalno okolje predpostavlja, da sčasoma vsa besedilna ali druga gradiva (avdio, video itd.), ki predstavljajo pisno kulturno in znanstveno dediščino, postanejo dostopna v digitalni obliki. Prosta in odprta dostopnost digitaliziranih jezikovnih virov spodbuja njihovo uporabo na digitalnih medijih, vključno z medsebojno povezljivostjo, nudi pa tudi empirično osnovo za razvoj jezikovnih tehnologij slovenskega jezika. Odprti dostop omogočajo različne licence, od katerih so najbolj uveljavljene licence Creative Commons. Tu se avtorji odpovedo monopolu nad nadaljnjo uporabo digitalnega izvirnika, s čimer se brez vsakokratnega pridobivanja soglasja nosilca materialnih avtorskih pravic omogoči ne samo branje, temveč tudi prevzem/prenos (download) in nadaljnje razširjanje ter, odvisno od licence, predelava in komercialno izkoriščanje jezikovnih virov. Ker ni posebnih omejitev (še posebej za besedila državne uprave, znanstvena besedila, besedila, ki so jim potekle avtorske pravice, in jezikovne vire narejene za namene jezikovnih tehnologij), je najbolj primerna licenca za javno financirane jezikovne vire CC-BY 4.0 (priznanje avtorstva).

Digitalizacija obstoječih jezikovnih opisov, ki jih tradicionalno dojemamo v knjižni obliki, in zagotavljanje njihovega prostega in odprtega dostopa na spletu sta nujna. V času veljavnosti resolucije 2007–2011 je bil prek spletnega vmesnika zagotovljen prosti dostop do Slovarja slovenskega knjižnega jezika, Slovenskega pravopisa 2001 in nekaterih drugih priročnikov, v bodoče bo treba spletno dostopnost v okvirih, ki jih določajo slovenski pravni red in mednarodno primerljive uzance, urediti tudi za čim več preostalih ključnih, obstoječih in nastajajočih temeljnih in specialnih jezikovnih opisov in njihovih podatkovnih baz z namenom vgradnje v sekundarne aplikacije kot tudi za nadaljnji razvoj jezikovnih tehnologij za slovenski jezik. Odprtost je tudi tehnološko pogojena, zato naj bi bili podatkovne baze jezikovnih priročnikov in virov zapisane z upoštevanjem ustreznimi mednarodnih standardov in priporočil, kot so XML in standardi ISO, predvsem tistimi, izdelanimi v Tehničnem odboru TC 37 »Terminologija in drugi jezikovni viri«, pa tudi drugimi relevantnimi priporočili, kot na primer Priporočili za kodiranje in izmenjavo besedil Konzorcija za zapis besedil (Text Encoding Initiative, TEI).

V Sloveniji obstaja že vrsta digitalnih knjižnic, ki omogočajo digitalni dostop do polnega besedila. Osrednja zbirka je dLib, digitalna knjižnica Slovenije, ki pa za večino starejšega slovenskega slovstva ponuja skenograme oziroma iz njih avtomatsko zajeto polna besedila, ki imajo zaradi starih izvirnikov precej napačno optično prepoznanih delov. Korekcije so zamuden, vendar potreben del izdelave kvalitetnih čistopisov, ki jih je nato možno ponuditi v branje na raznovrstnih digitalnih platformah kot tudi jezikovnotehnološko digitalno umestiti. Izdelovanje čistopisov starejše jezikovne

produkcije je potrebno spodbujati tudi s pomočjo izkoriščanja moči množice pri vnašanju in popravljanju jezikovnih virov – primer dobre prakse predstavlja Slovenska leposlovna klasika na Wikiviru.

Za razvoj slovenskega znanstvenega jezika mora država podpirati znanstveno produkcijo v slovenskem jeziku, ki pa je tudi dragocen vir slovenskega strokovnega izrazja in strokovnega izražanja. Možnosti za digitalni zajem, procesiranje in distribucijo takih besedil doslej niso bile polno izkoriščene. Z omogočanjem dostopa do teh del, kot so na primer objave v znanstvenih in strokovnih zbornikih in revijah ter magistrske in doktorske disertacije, se bistveno poveča njihova dostopnost, ob tem pa je mogoče ta besedila nato tudi jezikovnotehnološko obdelati in dati na voljo posameznim znanstvenim skupnostim za namene upravljanja s terminologijami v okviru splošnega terminološkega portala kot dela prostodostopnega spletnega portala s čim več dosegljivimi jezikoslovnimi podatki o slovenščini.

Ekonomija obsega zahteva, da zniževanje stroškov razvoja dosežemo s čim več uporabniki, zaradi česar so manjše jezikovne skupine v deprivilegiranem položaju. S tem izzivom se sooča tudi Slovenija. Ministrstvo za javno upravo v Strategiji razvoja informatike v javni upravi predvideva centralizacijo informatike v državni upravi in izgradnjo treh računalniških oblakov (raziskovalni, testni in produkcijski). Razvoj novih jezikovnih tehnologij, storitev in podatkovnih baz je smiselno integrirati v tovrstne računalniške oblake. S tem bi prihranili precej sredstev, poleg tega pa bi podjetjem in raziskovalnim organizacijam omogočili testiranje in nadaljnji razvoj izdelanih storitev ali integracijo v nove storitve in aplikacije. Objava jezikovnih tehnologij bi bila smiselna tudi na portalu NIO (Nacionalni interoperabilnostni okvir, <https://nio.gov.si/nio/>), ki združuje vse odprte podatke javnega sektorja.

1.8 Govorci⁵ s posebnimi potrebami

Osebe s posebnimi potrebami, ki lahko pričakujejo največjo korist od jezikovnih tehnologij, so slepi in slabovidni ter gluhi in naglušni. V Sloveniji se ocenjuje število slepih in slabovidnih na 8000 do 10.000 (sem niso vštete osebe, ki zaradi starostne opešanosti vida prav tako čutijo določene omejitve pri branju in pisanju), število gluhih na prek 2.000, naglušnih pa je še nekajkrat več oseb. Slepí in slabovidni so organizirani pod krovno Zvezo društev slepih in slabovidnih Slovenije, gluhi in naglušni pa pod Zvezo društev gluhih in naglušnih Slovenije, pomembno vlogo za njihovo vključevanje v društvo pa imajo tolmači, organizirani v okviru Zavoda Združenje tolmačev za slovenski znakovni jezik.

Osrednje tehnologije, ki lahko pripomorejo k lažjemu vključevanju teh oseb v družbo, so sintetizator govora, razpoznavnik govora in tolmač znakovnega jezika. Sintetizator govora, prilagojen za slepe in slabovidne, je v svetu zelo razširjen v okviru aplikacij, kot so bralniki in povečevalniki zaslona (Jaws, Zoomtext, SuperNova, Window-eyes (komercialni) ali NVDA (odprtokodni)). Za bralnike in povečevalnike zaslona, ki znajo izkoriščati standard SAP15 v okolju operacijskega sistema Windows, za slovensko govorno področje obstajata dva komercialna sistema sinteze govora (Govorec in proteus na standardu MS SAPI 5), vendar njuna kakovost še ni zadovoljiva. Za aplikaciji Jaws in Zoomtext je podjetje Amebis že izdelalo prilagojeno različico starejše verzije slovenskega sintetizatorja (Govorec), ki pa še ne dosega optimalne razumljivosti. Posebne izzive

⁵ Beseda »govorci« se v akcijskem načrtu v zvezi z govorcí s posebnimi potrebami uporablja v pomenu »uporabnik jezika«, saj se enako kot na potrebe pri govorjenju in poslušanju nanaša tudi na potrebe pri branju in pisanju.

predstavlja družabni splet, npr. Facebook, kjer zaradi visoke grafičnosti obstoječi bralniki ne delujejo najbolje. Govorne vmesnike za družabna omrežja je tako treba razvijati posebej. Pomembno področje uporabe sintetizatorjev govora je torej računalniško okolje. Težave v vsakdanjem življenju, povezane z jezikom, za slepe in slabovidne predstavlja tudi domače okolje, ko je treba upravljati z gospodinjskimi aparati. Sodobni aparati izpisujejo informacije digitalno, v alfanumerični obliki, za slepe in slabovidne prilagojeni aparati pa imajo bistveno višjo ceno. Tudi v tej smeri sodobne tehnologije omogočajo cenejše rešitve.

Na področju slovenskega znakovnega jezika (SZJ) so dejavnosti trenutno razpršene in slabo usklajene. S SZJ se trenutno sistematično ukvarjajo naslednje ustanove:

- Zveza društev gluhih in naglušnih Slovenije (ZDGNS), kjer se že nekaj let gradi multimedijski spletni slovar SZJ,
- Zavod za gluhe in naglušne Ljubljana (ZGL), ki predstavlja nacionalno središče za poučevanje SZJ,
- Združenje tolmačev SZJ, kjer poteka izobraževanje tolmačev SZJ,
- Filozofska fakulteta UL, kjer teče projekt ARRS "Korpus in pilotna slovnica SZJ"; projekt se je iztekel junija 2014,
- Pedagoška fakulteta UL, kjer je že od leta 1980 prisoten predmet Osnove neverbalne komunikacije in predmet Slovenski znakovni jezik in kjer je nastalo več diplomskih del v zvezi s slovenskim znakovnim jezikom ter bila prvič uvedena transkripcija znakovnega jezika,
- Fakulteta za računalništvo in informatiko UL, kjer je v sodelovanju s podjetjem Zoom Promotion nastal še en spletni slovar SZJ ter številna spletna gradiva za poučevanje SZJ,
- v preteklosti se je z raziskavami SZJ ukvarjal tudi ISJFR ZRC SAZU.

Jezikovna opremljenost SZJ je nezadovoljiva. V okviru raziskovalnega projekta ARRS nastaja prvi reprezentativni korpus SZJ (SIGNOR), ki pa še ni dokončan; v njem so zbrani vzorci 80 gluhih informantov v obliki transkribiranih videoposnetkov. Za izdelavo sodobnih in kakovostnih jezikovnih priročnikov, predvsem boljšega slovarja in učbenikov SZJ, je najprej potrebno dopolniti korpus SIGNOR, nato pa izvesti temeljne raziskave SZJ na leksikalni, skladenjski in pragmatiki ravni.

Gluhi in naglušni imajo težave, povezane z jezikom, predvsem pri spremljanju televizijskih vsebin, pri šolanju ter v vsakovrstnih komunikacijskih situacijah. Težave premagujejo s pomočjo tolmačev za znakovni jezik. Zanje je pomembna tehnologija zlasti razpoznavalnik govora, ki je tudi potrebna vmesna točka do bolj kompleksne aplikacije, avtomatskega tolmača za znakovni jezik. V svetu obstaja tovrstna tehnologija le za redke jezike z največ govorci.

2 Definiranje ciljev, aktivnosti in akcij

2.1 Definiranje ciljev

Po posameznih sklopih so bili definirani cilji, kot prikazuje tabela 1.

Tabela 1: Cilji po sklopih

Splošno:	1 Infrastrukturni center
	2 Odprta dostopnost jezikovnih virov
	3 Spodbujanje razvoja slovenske Wikimedije
Jezikovni opis:	1 Novi informativno-normativni razlagalni slovar s slovarsko bazo
	2 Diahroni opis jezika
	3 Slovnični opis
	4 Slovar naselbinskih imen
	5 Slovenski lingvistični atlas
	6 Terminološki slovarji različnih strok
	7 Zgodovinski slovar
	8 Spodbujanje prožnih in inovativnih oblik učenja z razvojem jezikovnih virov in tehnologij
Standardizacija:	1 Pravopisni portal
	2 Zbirka normativnih zadreg
	3 Črkovalnik besed
	4 Korpus za normativna vprašanja
	5 Jezikovnosvetovalni portal
Terminologija:	1 Terminološki portal
Večjezičnost:	1 Večjezični portal
	2 Strojni prevajalnik
Jezikovne tehnologije:	1 Razpoznavalnik tekočega govora
	2 Avtomatsko procesiranje besedila v jezikovnih tehnologijah in spletnih iskalnih orodjih
	3 Slovnični pregledovalnik
	4 Sistemi dialoga
Digitalizacija:	1 Digitalni čistopisi slovenske pisne kulturne dediščine
	2 Digitalni jezikovni priročniki za slovenski jezik
Govorci s posebnimi potrebami:	1 Bralnik zaslona za slepe in slabovidne
	2 Branje sporočil gospodinjskih aparatov za slepe in slabovidne
	3 Razpoznavanje govora v pedagoškem procesu
	4 Slovar slovenskega znakovnega jezika (SZJ)

2.2 Definiranje aktivnosti za uresničitev ciljev

V drugem koraku so bile za vsakega od ciljev definirane vse potrebne aktivnosti, ki bi vodile k uresnitvi cilja. Popisane so v prilogi 1 tega dokumenta.

2.3 Opredelitev akcij ter njihova finančna ocena in časovna umestitev

Generalna pripomba: navedene akcije so finančno ovrednotene, njihova izvedba v predvidenih časovnih okvirih pa bo odvisna od uspešnosti prijav na javnih razpisih in razpoložljivosti proračunskih sredstev v ta namen. Ministrstvo za izobraževanje, znanost in šport ter Ministrstvo za kulturo v proračunih za leti 2016 in 2017 nimata načrtovanih posebnih dodatnih sredstev za izvedbo navedenih akcij.⁶

Pokazalo se je, da se nekatere aktivnosti ponavljajo pri več ciljih – predvsem aktivnosti, povezane z izdelavo temeljnih virov, kot so razni korpusi, leksikoni, JT-orodja za označevanje besedil ipd. V tretjem koraku so bile zato popisane vse različne predvidene aktivnosti in opredeljene glede na cilje, h katerim prispevajo. Smiselno so bile združene v krovne akcije. Nazadnje je bila vsaka akcija časovno umeščena in finančno ovrednotena, pri čemer so bile okvirno upoštevane smernice iz Resolucije za razdelitev po posameznih sklopih, končna vsota vseh akcij pa upošteva skupno vsoto za jezikovno opremljenost (in deloma za jezikovno izobraževanje, opredeljeno v Resoluciji).

Finančne vrednosti k akcijam pomenijo približno oceno stroškov. V skladu s pristopom matičnega dokumenta, Resolucije o nacionalnem programu za jezikovno politiko 2014–2018, bodo posamezni nosilci (tj. financerji) odgovorni za uresničevanje ukrepov s pomočjo evropskih sredstev v okviru finančne perspektive 2014–2020. Nekatere akcije se kljub temu že izvajajo, morebitno dodatno financiranje preko razpisov mora biti utemeljeno bodisi s prenehanjem financiranja ukrepa ali pa z izkazano potrebo po vzporedni nadgradnji akcije z razvidno in natančno definiranim specifičnim predmetom sofinanciranja, pri čemer morajo biti upoštevana določila 2. in 65. člena Zakona o javnih financah v zvezi z gospodarnostjo, racionalnostjo, zakonitostjo, namenskostjo in učinkovitostjo pri razpolaganju s proračunskimi sredstvi.

Poleg k posameznim akcijam navedenih nosilcev (ministrstva in agencija) je sonosilec akcij s področja jezikovnega opisa, standardizacije in terminologije, izhajajoč iz Resolucije o nacionalnem programu za jezikovno politiko 2014–2018, tudi SAZU, in sicer v okviru programa Naravna in kulturna dediščina slovenskega naroda – temeljne raziskave slovenskega jezika.

Tabela 2 podaja pregled akcij skupaj z aktivnostmi, ki so bile združene v posamezno akcijo, ter njihov finančni obseg in čas izvajanja.

⁶ Izjema je projekt Spodbujanje prožnih in inovativnih oblik učenja z razvojem jezikovnih virov in tehnologij, ki ga je Ministrstvo za kulturo pripravilo po usklajevanju z nosilnim Ministrstvom za izobraževanje, znanost in šport. Projekt je vlada potrdila z Odlokom o izvedbenem načrtu Operativnega programa za izvajanje evropske kohezijske politike za programsko obdobje 2014–2020.

Tabela 2: Opredelitev akcij

Akcije in aktivnosti	Sklop	Cilj	Ocena sredstev	Čas
SPLOŠNO				
Akcija S-1: Infrastrukturni center			854.500 €	2015-2018
delovanje infrastrukturnega centra	Splošno	Infrastrukturni center		
Akcija S-2: Zakonodaja za odprti dostop			0 €	2015-2018
prizadevanje za spremembo zakonodaje, da zagotavlja odprti dostop do vseh del, ki nastanejo z javnim financiranjem	Splošno	Odprta dostopnost jezikovnih virov		
Akcija S-3: Vnašanje virov v Wikije			49.000 €	2015-2018
prenos obstoječih virov v Wikimedijo	Splošno	Spodbujanje razvoja slovenske Wikimedije		
SPLETNI PORTALI				
Akcija P-1: Terminološki portal			98.500 €	2015-2018
terminološki portal	Terminologija	Terminološki portal		
Akcija P-2: Aplikacija za nove terminološke zbirke			49.000 €	2015-2016
spletna aplikacija za ustvarjanje novih terminoloških zbirk	Terminologija	Terminološki portal		
Akcija P-3: Terminološko svetovanje			118.500 €	2015-2018
terminološko svetovanje	Terminologija	Terminološki portal		
Akcija P-4: Javna dostopnost terminoloških virov			98.500 €	2015-2018
zagotavljanje javno dostopnih terminoloških virov	Terminologija	Terminološki portal		
Akcija P-5: Samodejno luščenje terminologije			98.500 €	2015-2016
samodejno luščenje terminologije	Terminologija	Terminološki portal		
Akcija P-6: Večjezični portal			49.000 €	2015-2018
večjezični portal	Večjezičnost	Večjezični portal		
Akcija P-7: Slovnčni portal za šolarje			197.500 €	2015-2018
nadgradnja slovnčnega portala za šolsko populacijo	Jezikovni opis	Slovnčni opis		

Akcija P-8: Spletna predstavitev normativnih podatkov			59.000 €	2015-2016
sistem za spletno predstavitev normativnih podatkov: pravila in slovar	Standardizacija	Pravopisni portal		
Akcija P-9: Spodbujanje prožnih in inovativnih oblik učenja z razvojem jezikovnih virov in tehnologij			2.000.000 €	2016-2018
posodobitev relevantnih vidikov jezikovnega opisa	Jezikovni opis			
posodobitev relevantnih vidikov kodifikacije	Standardizacija			
izdelava učnih e-orodij	Jezikovne tehnologije			
razvoj novih učnih pristopov in metod	Razvijanje jezikovne zmožnosti, glej ANJI			
izboljšanje učiteljskih kompetenc	Izobraževanje učiteljev, glej ANJI			
PISNI KORPUSI				
Akcija K-1: Pisni korpusi			296.500 €	2015-2018
nadgradnja pisnih korpusov sodobnih besedil	Jezikovne tehnologije	Razpoznavalnik tekočega govora		
nadgradnja pisnih korpusov sodobnih besedil	Jezikovne tehnologije	Avtomatsko procesiranje besedila v jezikovnih tehnologijah in spletnih iskalnih orodjih		
nadgradnja pisnih korpusov sodobnih besedil in njihovih konkordančnikov	Jezikovne tehnologije	Sistemi dialoga		
nadgradnja pisnih korpusov sodobnih besedil in njihovih konkordančnikov	Večjezičnost	Strojni prevajalnik		
nadgradnja pisnih korpusov sodobnih besedil in njihovih konkordančnikov	Jezikovni opis	Novi informativno-normativni razlagalni slovar s slovarsko bazo		
nadgradnja pisnih korpusov sodobnih besedil in njihovih konkordančnikov	Jezikovni opis	Slovnični opis		
nadgradnja korpusov strokovnih besedil	Terminologija	Terminološki portal		
Akcija K-2: Specializirani korpusi			227.000 €	2015-2018

nadgradnja in izdelava specializiranih korpusov	Jezikovni opis	Novi informativno-normativni razlagalni slovar s slovarsko bazo		
nadgradnja in izdelava specializiranih korpusov	Jezikovni opis	Slovnični opis		
izdelava korpusa slovenskega jezika, specializiranega za normativna vprašanja	Standardizacija	Korpus za normativna vprašanja		
izdelava korpusov znanstvenih besedil	Digitalizacija	Digitalna slovenska znanstvena besedila		
Akcija K-3: Govorni korpus in baza			395.500 €	2015-2018
nadgradnja referenčnega govornega korpusa	Jezikovni opis	Novi informativno-normativni razlagalni slovar s slovarsko bazo		
nadgradnja referenčnega govornega korpusa	Jezikovni opis	Slovnični opis		
večnamenski korpus in baza govorjenega jezika	Jezikovne tehnologije	Razpoznavalnik tekočega govora		
dialoški korpusi	Jezikovne tehnologije	Sistemi dialoga		
Akcija K-4: Dvojezični korpusi			395.500 €	2015-2018
dvojezični korpusi	Večjezičnost	Strojni prevajalnik		
Akcija K-5: Diahroni korpus			173.000 €	2015-2018
izdelava diahronnega korpusa slovenskega jezika	Jezikovni opis	Diahroni opis jezika		
SLOVARJI IN LEKSIKONSKE BAZE				
Akcija L-1: Splošni slovar in slovarska baza			4.452.500 €	2015-2018
Akcija L-2: Slovenski lingvistični atlas			118.500 €	2015-2018
izdelava spletne večpredstavnostne objave 1. in 2. zvezka SLA	Jezikovni opis	Slovenski lingvistični atlas		
Akcija L-3: Slovar naselbinskih imen			158.000 €	2015-2018

izdelava slovarja naselbinskih imen slovenskega narodnega prostora	Jezikovni opis	Slovar naselbinskih imen		
Akcija L-4: Terminološki slovarji različnih strok			300.000 €	2015-2018
izdelava terminoloških slovarjev različnih strok	Jezikovni opis	Terminološki slovarji različnih strok		
Akcija L-5: Zgodovinski slovar			300.000 €	2015-2018
izdelava zgodovinskega slovarja	Jezikovni opis	Zgodovinski slovar		
Akcija L-6: Imenske entitete v slovenščini			98.500 €	2015-2018
imenske entitete	Jezikovne tehnologije	Avtomatsko procesiranje		
imenske entitete	Jezikovne tehnologije	Sintetizator govora		
imenske entitete	Jezikovne tehnologije	Sistemi dialoga		
imenske entitete	Jezikovne tehnologije	Slovnični pregledovalnik		
imenske entitete	Večjezičnost	Strojni prevajalnik		
Akcija L-7: Večjezične baze in slovarji			247.000 €	2015-2018
dvo- in večjezične baze in slovarji	Večjezičnost	Strojni prevajalnik		
dvo- in večjezične baze in slovarji	Večjezičnost	Večjezični portal		
Akcija L-8: Strojna slovnična analiza			118.500 €	2015-2018
oblikoslovna in skladenjska analiza besedila	Jezikovne tehnologije	Avtomatsko procesiranje besedila v jezikovnih tehnologijah in spletnih iskalnih orodjih		
oblikoslovna in skladenjska analiza besedila	Jezikovne tehnologije	Razpoznavalnik tekočega govora		
oblikoslovna in skladenjska analiza besedila	Jezikovne tehnologije	Sintetizator govora		
oblikoslovna in skladenjska analiza besedila	Jezikovne tehnologije	Sistemi dialoga		

oblikoslovna in skladenjska analiza besedila	Jezikovne tehnologije	Slovnični pregledovalnik		
oblikoslovna in skladenjska analiza besedila	Večjezičnost	Strojni prevajalnik		
oblikoslovni sintetizator	Večjezičnost	Strojni prevajalnik		
oblikoslovni sintetizator	Jezikovne tehnologije	Sistemi dialoga		
Akcija L-9: FrameNet za slovenščino			148.000 €	2015-2018
semantična analiza vlog	Jezikovne tehnologije	Avtomatsko procesiranje besedila v jezikovnih tehnologijah in spletnih iskalnih orodjih		
semantična analiza vlog	Jezikovne tehnologije	Sintetizator govora		
semantična analiza vlog	Jezikovne tehnologije	Sistemi dialoga		
semantična analiza vlog	Jezikovni opis	Slovnični opis		
semantična analiza vlog	Večjezičnost	Strojni prevajalnik		
Akcija L-10: sloWNet			148.000 €	2015-2018
semantična mreža	Jezikovni opis	Slovnični opis		
semantična mreža	Večjezičnost	Strojni prevajalnik		
Akcija L-11: Anafora in koreference			148.000 €	2016-2018
razvozlanje anafore in analiza besedilnih koreferenc	Jezikovne tehnologije	Avtomatsko procesiranje besedila v jezikovnih tehnologijah in spletnih iskalnih orodjih		
razvozlanje anafore in analiza besedilnih koreferenc	Jezikovne tehnologije	Sistemi dialoga		
razvozlanje anafore in analiza besedilnih koreferenc	Jezikovne tehnologije	Slovnični pregledovalnik		
razvozlanje anafore in analiza besedilnih koreferenc	Jezikovni opis	Slovnični opis		
razvozlanje anafore in analiza besedilnih koreferenc	Večjezičnost	Strojni prevajalnik		
Akcija L-12: Zbirka normativnih zadreg			197.500 €	2015-2018

zasnova zbirke normativnih zadreg	Standardizacija	Zbirka normativnih zadreg		
gradnja zbirke (luščenje podatkov, kategorizacija podatkov)	Standardizacija	Zbirka normativnih zadreg		
preverjanje realizacije in ustreznosti pravopisnih, pravorečnih, oblikoslovnih podatkov (skupina strokovnjakov)	Standardizacija	Zbirka normativnih zadreg		
SLOVNICA				
Akcija G-1: Zasnova slovnice			197.500 €	2015-2018
koncept nove slovnice slovenskega jezika	Jezikovni opis	Slovnični opis		
slovnična podatkovna zbirka	Jezikovni opis	Slovnični opis		
Akcija G-2: Šolska slovnica			197.500 €	2015-2018
šolska slovnica / slovnični portal	Jezikovni opis	Slovnični opis		
NORMATIVNOST				
Akcija N-1: Pravopisna pravila in slovar			197.500 €	2015-2018
predstavitev normativnih podatkov: pravila in slovar	Standardizacija	Pravopisni portal		
jezikovno svetovanje	Standardizacija	Pravopisni portal		
predstavitev normativnih podatkov: spletni priročnik za splošne uporabnike	Standardizacija	Pravopisni portal		
DIGITALIZACIJA				
Akcija D-1: Digitalizacija čistopisov pisne kulturne dediščine			59.000 €	2015-2018
izdelava čistopisov z uporabo množičenja	Digitalizacija	Digitalni čistopisi slovenske pisne kulturne dediščine		
Akcija D-2: Digitalizacija jezikovnih priročnikov			118.500 €	2015-2018
digitalizacija ali prilagajanje digitalnemu okolju jezikovnih priročnikov za slovenščino	Digitalizacija	Digitalni jezikovni priročniki za slovenski jezik		

JT-APLIKACIJE				
Akcija A-1: Razpoznavalnik tekočega govora			316.500 €	2016-2018
razpoznavalnik tekočega govora z velikim slovarjem besed	Jezikovne tehnologije	Razpoznavalnik tekočega govora		
razpoznavalnik govora za šolsko okolje	Govorci s posebnimi potrebami	Razpoznavanje govora v šolskem okolju		
Akcija A-2: Strojni prevajalnik			148.000 €	2016-2018
razvoj strojnega prevajalnika	Večjezičnost	Strojni prevajalnik		
Akcija A-3: Sistem dialoga			69.000 €	2016-2018
sistem dialoga – aplikacija	Jezikovne tehnologije	Sistemi dialoga		
Akcija A-4: Bralnik sporočil gospodinjskih aparatov			59.000 €	2016-2016
bralnik sporočil gospodinjskih aparatov	Govorci s posebnimi potrebami	Branje sporočil gospodinjskih aparatov za slepe in slabovidne		
Akcija A-5: Referenčni črkovalnik in slovnični pregledovalnik			98.500 €	2016-2018
izdelava referenčnega črkovalnika za slovenščino	Standardizacija	Črkovalnik besed		
slovnični pregledovalnik - aplikacija	Jezikovne tehnologije	Slovnični pregledovalnik		
ZNAKOVNI JEZIK				
Akcija Z-1: Korpus in slovar znakovnega jezika			98.500 €	2015-2018
nadgradnja korpusa slovenskega znakovnega jezika (SZJ)	Govorci s posebnimi potrebami	Slovar slovenskega znakovnega jezika (SZJ)		
slovar slovenskega znakovnega jezika (SZJ)	Govorci s posebnimi potrebami	Slovar slovenskega znakovnega jezika (SZJ)		
		SKUPAJ:	13.153.500 €	

3 Podroben opis akcij

Na podlagi:

Deklaracije o dostopu do javno financiranih raziskovalnih podatkov (OECD) (<http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157&InstrumentPID=153>),

Priporočila Komisije z dne 17. julija 2012 o dostopu do znanstvenih informacij in njihovem arhiviranju

([http://eur-](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:SL:PDF)

[lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:SL:PDF](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:SL:PDF)),

Direktive 2013/37/EU Evropskega parlamenta in Sveta z dne 26. junija 2013 o spremembi Direktive 2003/98/ES o ponovni uporabi informacij javnega sektorja ([http://eur-lex.europa.eu/legal-](http://eur-lex.europa.eu/legal-content/SL/TXT/PDF/?uri=CELEX:32013L0037&qid=1401874406055&from=SL)

[content/SL/TXT/PDF/?uri=CELEX:32013L0037&qid=1401874406055&from=SL](http://eur-lex.europa.eu/legal-content/SL/TXT/PDF/?uri=CELEX:32013L0037&qid=1401874406055&from=SL)),

Resolucije o raziskovalni in inovacijski strategiji Slovenije 2011–2020 (<http://www.uradni-list.si/1/objava.jsp?urlid=201143&stevilka=2045>) in

Smernic za zajem, dolgotrajno hranjenje in dostop do kulturne dediščine v digitalni obliki (http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/Zakonodaja/2013/Smernice_za_zajem_dolgotrajno_ohranjanje_in_dostop_do_kulturne_dediscine_v_digitalni_obliki.pdf)

veljajo za vse akcije, navedene v nadaljevanju, naslednje skupne smernice, pri čemer pod »izdelke« razumemo vse jezikovne vire, priročnike, orodja, aplikacije itd., ki so izdelani v okviru akcij:

1. Vsi izdelki naj bodo opremljeni z natančnimi specifikacijami/tehničnim poročilom. Vlogo specifikacij lahko opravlja tudi uvodno poglavje vira, znanstveni članek ipd., če je v njem pojasnjeno, na kakšnih teoretično-uporabnostnih temeljih in načelih je vir, orodje, aplikacija, priročnik ipd. zasnovan, kako se ga uporablja ipd.
2. Vsi izdelki naj sledijo splošno sprejetim standardom, če ti obstajajo za njihovo področje, oz. modelom dobre prakse. Poročanje o upoštevanju standardov oz. o teoretično-uporabnostnih modelih, ki so bili pri oblikovanju izdelka pregledani in po katerih so se izvajalci (v določenih segmentih) zgledovali, je nujen sestavni del specifikacij/tehničnega poročila.
3. Vsi izdelki naj sledijo cilju čim večje usklajenosti in povezanosti z drugimi, tako obstoječimi kot tistimi, ki so predvideni s tem akcijskim načrtom. Poročanje o tem, v kolikšni meri in na kakšen način je izdelek usklajen z drugimi obstoječimi izdelki, je nujen sestavni del specifikacij/tehničnega poročila.
4. Vsi izdelki morajo dokazovati ustrezno kvaliteto. Poročilo o doseganju kvalitete je nujen sestavni del specifikacij/tehničnega poročila.
5. Vsi izdelki naj bodo čim bolj odprti (tj. v obliki odprte kode/izvornega vira) in prosto dostopni (vsaj brezplačna uporaba brez možnosti spreminjanja). O pogojih, pod katerimi bo izdelek dostopen, odloča financer projekta, v okviru katerega je izdelek nastal, ter nosilci materialnih in moralnih pravic. Izjeme so mogoče še, kjer odprti oz. prosti dostop onemogočajo avtorske pravice ali druge utemeljene omejitve.
6. Vsi izdelki naj bi se po poteku projekta skupaj s pripadajočo dokumentacijo, lahko pa že prej, prenesli v arhiviranje, distribucijo in vzdrževanje nacionalnemu infrastrukturnemu centru. O prenosu izdelka v arhiviranje, distribucijo in vzdrževanje nacionalnemu infrastrukturnemu centru odločata v dogovoru financer in izvajalec projekta, v okviru katerega je izdelek nastal. Izdelani viri, kjer je le mogoče oz. smiselno, se dodatno vključijo v primeren projekt Fundacije Wikimedia, kot so Wikislovar, Wikivir, Wikipodatki itd. Izdelana orodja pa se dodatno vključijo v enega od repozitorijev odprte kode, kot je npr. Github.

Akcija S-1: Infrastrukturni center

Delovanje infrastrukturnega centra

Ena od prioritet pri doseganju ustrezne opremljenosti slovenščine je organizacija zbiranja, evalvacije, hranjenja in distribucije jezikovnih virov (jezikovne baze, priročniki itd.) in orodij za slovenski jezik, vključno s slovenskim znakovnim jezikom in slovenskim lahkim branjem. Ta predvideva vzpostavitev osrednjega infrastrukturnega centra, ki deluje kot repozitorij in distribucijsko telo za vire in orodja, ki nastajajo v okviru različnih projektov in po koncu financiranja projektov tipično niso več na voljo. V EU je bil za namen vzpostavljanja osrednje evropske infrastrukture financiran projekt CLARIN (*Common Language Resources and Technology Infrastructure*), ki je od l. 2012 organiziran kot CLARIN ERIC (*European Research Infrastructure Consortium*). Akcijski načrt predvideva, da bo v času trajanja resolucije (2014-2018) organiziran slovenski konzorcij CLARIN, ki bo del evropskega konzorcija. Maja 2015 je bil slovenska raziskovalna infrastruktura CLARIN.SI priključena Evropskemu raziskovalnemu infrastrukturnemu konzorciju CLARIN ERIC.

Opis:

Vzpostavi se infrastrukturni center za zbiranje, evalvacijo, hranjenje in distribucijo jezikovnih virov (jezikovne baze, priročniki itd.), orodij za slovenščino, vključno s slovenskim znakovnim jezikom. V delovanje centra v okviru evropske infrastrukture CLARIN se vključi čim več deležnikov, ki se v Sloveniji ukvarjajo s temi področji. Za delovanje centra je treba zagotoviti ustrezno strojno in programsko opremo, programsko opremo prilagoditi za potrebe centra in vzpostaviti spletišče z informacijami o centru. Treba je izdelati repozitorij jezikovnih virov in orodij, ki mora biti skladen z zahtevami CLARIN ERIC, ki predvidevajo dolgoročno arhiviranje, uporabo trajnih identifikatorjev, enotno overjanje identitete uporabnikov, integracijo virov in njihovih metapodatkov v evropsko infrastrukturo ter možnost analize in označevanja deponiranih virov z jezikoslovnimi orodji. Treba je vzpostaviti tudi sistem spletnih storitev, bodisi takih, ki so namenjeni vizualizaciji vsebine jezikovnih virov (konkordančniki za korpuse in pregledovalniki različnih vrst slovarjev oz. leksikalnih baz) in tiste, katerih namen je obdelati nek jezikovni vir, npr. označiti korpuse za nadaljnje analize. Poleg tehnične infrastrukture je potrebno zagotoviti, da bo repozitorij ponujal dovolj kvalitetnih in za raziskovalce zanimivih jezikovnih virov, kar vključuje zbiranje dobrih praks in njihova diseminacija, aktivno zbiranje dostopnih jezikovnih virov, tako obstoječih kot tistih, ki bodo nastali v času trajanja tega akcijskega načrta. Treba je nuditi pomoč pri vključevanju virov v repozitorij, vključno z nadzorom kvalitete in financiranjem stroškov ponudnikov, nastalih z vključevanjem virov v repozitorij. Aktivno je treba poskrbeti za izdelavo spletnih storitev za vizualizacijo in označevanje jezikovnih virov, predvsem za pretvorbo obstoječih orodij v spletne servise. Delovanje centra predvideva stalno podporo tem aktivnostim.

Predvideni učinki:

- izboljššan dolgoročni izkoristek vloženih sredstev v jezikovne vire in orodja
- izboljššan dostop do jezikovnih virov in orodij za vse uporabnike
- zagotovljena trajna uporabnost jezikovnih virov in orodij
- izboljšano spremljanje usklajenosti razvoja in kakovosti jezikovnih virov in orodij

Nosilec:

MK⁷, MIZŠ (ARRS)

⁷ MK se vključuje v smislu spremljanja.

Časovni plan:
2015–2018

Obseg financiranja: obseg financiranja s strani MIZŠ (ARRS) je omejen na sredstva tekočega infrastrukturnega programa

Podrobna opredelitev aktivnosti s kazalniki:

1 Delujoč infrastrukturni center	da/ne
----------------------------------	-------

Akcija S-2: Zakonodaja za odprti dostop

Uveljavitev spremembe zakonodaje, ki zagotavlja odprti dostop do vseh del, ki nastanejo z javnim financiranjem

V Sloveniji je na spletu prisotnih že mnogo digitaliziranih slovenskih besedil, jezikovnih priročnikov, besedilnih korpusov in drugih jezikovnih virov, vendar so večinoma dostopni samo preko določenega portala, ki mnogokrat navaja omejitve uporabe, ki presegajo omejitve, ki izhajajo iz avtorskopravne zaščite samih besedil in to kljub temu, da so vsebine nastale izključno z javnim financiranjem. V bodoče bo zato treba dostopnost v okvirih, ki jih določajo slovenski pravni red in mednarodno primerljive uzance, urediti za vse nastajajoče jezikovne vire.

Opis:

Odprti dostop omogočajo različne licence, od katerih so najbolj uveljavljene licence Creative Commons. Tu se avtorji odpovedo monopolu nad nadaljnjo uporabo digitalnega izvirnika, s čimer se brez vsakokratnega pridobivanja soglasja nosilca materialnih avtorskih pravic omogoči ne samo branje, temveč tudi prevzem/prenos (download) in nadaljnje razširjanje ter, odvisno od licence, predelava in komercialno izkoriščanje jezikovnih virov. Kjer ni posebnih omejitev (še posebej za besedila državne uprave, znanstvena besedila, besedila, ki so jim potekle avtorske pravice, in jezikovne vire narejene za namene jezikovnih tehnologij), je najbolj primerna licenca za javno financirane jezikovne vire CC-BY 4.0 (priznanje avtorstva).

Za dolgoročno ureditev pravnega okvira odprte dostopnosti bo potrebno slovensko zakonodajo spremeniti tako, da bo dovoljevala čim bolj liberalen dostop do vseh jezikovnih virov financiranih iz javnih sredstev, kot tudi predvidevajo npr. slovenske Smernice za zajem, dolgotrajno ohranjanje in dostop do e-vsebin kulturne dediščine.

Predvideni učinki:

- Bistveno povečanje izkoriščenosti digitalnih virov slovenskega jezika, tako neposredno za uporabnike kot za ponovno izkoriščanje.

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Podrobna opredelitev aktivnosti s kazalniki:

1 formulacija člena o odprtem dostopu jezikovnih virov, nastalih z javnim financiranjem	da/ne
2 sprejetje zakonodaje v DZ	da/ne

Akcija S-3: Vnašanje virov v Wikije

Prenos obstoječih virov v projekte Fundacije Wikimedia

Projekti Fundacije Wikimedia, predvsem Wikipedija postajajo izjemno pomemben vir informacij, od enciklopedičnih do slovarskih, zaradi svoje odprtosti pa so tudi koristni za jezikovne tehnologije, zato je treba poskrbeti za sistematičen razvoj slovenskih baz. Smiselno je zagotoviti predvsem prenos podatkov iz obstoječih enciklopedičnih, leksikalnih in drugih baz, ki vsebinsko ustrezajo, v slovensko Wikipedijo, Wikislovar, Wikivir in druge baze Wikimedia. Gre predvsem za ustrezno upravljanje z avtorskimi pravicami, pri čemer je prva naloga pregled možnih enciklopedičnih, biografskih in drugih baz, ki so možni kandidati za prenos.

Wikislovar je primarni večjezični vir jezikovnotehnološka podjetja in za raziskovalce, ki se ukvarjajo s tehnologijami, pri katerih potrebujejo večjezične baze podatkov. Smiselno je zagotoviti prenos dela podatkov, ki obstajajo v drugih, specifično organiziranih slovarskih bazah (npr. razlagalni slovarji), tudi v slovenski Wikislovar.

Opis:

Financira se odkup pravic za obstoječe enciklopedične in slovarske vire za potrebe prenosa informacij v Wikipedijo in Wikislovar. Odkupljene podatke se pretvori v ustrezno obliko in vnese v spletne baze.

Predvideni učinki:

- izboljššan dostop do odprto dostopnih virov znanja na spletu
- izboljššan možnosti za razvoj jezikovnih tehnologij za slovenščino

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Odkupljene pravice	da/ne
2 Vneseni podatki v Wikipedijo in Wikislovar	da/ne

Akcija P-1: Terminološki portal

Terminološki portal

Terminologija je ključen del specializirane komunikacije in omogoča, da se znanje ustvarja, posreduje, poučuje in uporablja. Na sodobno urejenih terminoloških bazah temeljijo številne storitve, kot so prevajanje, tolmačenje in gospodarsko delovanje v globalnem okolju, prav tako pa so terminološki viri potrebni za razvoj drugih jezikovnih tehnologij, kot je strojni prevajalnik. Trenutno so terminološki viri razpršeni, predvsem pa so podatki v njih zapisani v zelo raznolikih podatkovnih strukturah.

Terminološki portal je spletno mesto, ki združuje specializirane eno- in večjezične terminološke vire, daje navodila, smernice in vzorčne podatkovne zbirke za učinkovito upravljanje s terminologijo, ponuja prosto dostopno strežniško infrastrukturo za gradnjo novih slovarjev in orodja za luščenje terminologije iz besedilnih zbirk. Na portalu je poleg tega na voljo terminološka svetovalnica, ki ažurno svetuje ob terminoloških zagatah prevajalcev v ustanovah EU in drugih uporabnikov terminologije.

Portal se vzpostavi po temeljiti javni strokovni razpravi o njegovi vlogi pri usklajevanju terminologije.

Opis:

Akcija nadgrajuje rezultate pilotnega raziskovalnega projekta L6-9778 Slovenski terminološki portal (1. 1. 2007–31. 12. 2009). Vzpostavitev terminološkega portala, ki bi izpolnjeval vse naštetе funkcije, najprej zahteva preučitev potreb, obstoječih virov in njihovih formatov ter možnosti pretvorb, ki bodo potrebne za integracijo v enotno iskalno storitev. Nato je treba izdelati specifikacije podatkovnih struktur v skladu z mednarodnimi standardi (TBX), vhodnih in izhodnih formatov ter izdelati vzorčna gradiva za različne scenarije uporabe.

Zatem se vzpostavi spletišče z ustrezno strežniško infrastrukturo, iskalnikom po vseh integriranih virih ter spletnimi storitvami, kot so luščenje terminologije iz besedilnih korpusov, samodejno prepoznavanje terminov v specializiranih besedilih, spletni urejevalnik terminoloških slovarjev in terminološko svetovanje. Posebna pozornost se mora nameniti povezljivosti z drugimi portali v okviru tega akcijskega načrta, še posebej z večjezičnim portalom.

Predvideni učinki:

- boljši dostop do terminoloških virov
- lažje vključevanje obstoječih terminoloških virov v druga orodja in storitve
- vzpostavljena infrastruktura za povezane terminološke ukrepe (spletna aplikacija za izdelavo slovarjev, luščenje terminologije, terminološko svetovanje)

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Specifikacije podatkovnih struktur, vhodnih in izhodnih formatov ter oblikovanje priporočil in vzorčnih scenarijev	da/ne
2 Spletišče s strežniško infrastrukturo, iskalnikom in konkordančnikom	da/ne

Akcija P-2: Aplikacija za nove terminološke zbirke

Spletna aplikacija za ustvarjanje novih terminoloških zbirk

Terminološki viri nastajajo v različnih okoljih, ustvarjajo jih prevajalci, raziskovalci, strokovnjaki v javnih in zasebnih ustanovah, študenti in drugi. Poleg organiziranega terminološkega dela je zainteresirani javnosti smiselno ponuditi tudi okolje, kjer lahko v spletni aplikaciji strokovnjak ali skupina strokovnjakov ustvari terminološki slovar in prispeva specializirano izrazje, ob tem pa ima prek terminološkega portala na voljo tudi svetovanje v zvezi s strukturo in metodologijo dela. Razumljivo je, da bodo imeli takšni viri zelo različno uporabno vrednost, zato bi bile tovrstne terminološke baze javnosti dostopne z ustreznim pojasnilom o njihovem nastanku, po drugi strani pa so lahko takšni viri osnova za nadaljnjo obdelavo s strani usposobljenega terminografa.

Opis:

Spletna aplikacija za ustvarjanje novih terminoloških zbirk mora biti uporabniku prijazno okolje, ki ponuja možnost za definiranje različnih oblik terminoloških glosarjev, eno- in večjezičnih, pojmovno zasnovanih, z možnostjo uvoza podatkov iz običajnih datotečnih formatov in prav tako izvoza vanje (Excel, Multiterm itd.). Omogočati mora večuporabniško delo z različnimi vlogami uporabnikov, prav tako mora omogočati dodajanje večpredstavnih podatkov in povezav na zunanje vire. Aplikacija mora biti opremljena s podrobno dokumentacijo in ustrezati sodobnim terminološkim standardom.

Predvideni učinki:

- zbiranje terminoloških virov iz različnih strokovnih krogov,
- boljša terminološka opremljenost za področja, ki doslej še niso bila obdelana,
- boljša povezljivost terminoloških podatkov iz različnih virov.

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Razvoj spletne aplikacije za ustvarjanje terminoloških slovarjev	da/ne
--	-------

Akcija P-3: Terminološko svetovanje

Terminološko svetovanje

Številni uporabniki terminologije v obstoječih virih in slovarjih ne najdejo odgovorov na svoje dileme v zvezi s strokovnimi poimenovanji ali njihovo rabo v različnih kontekstih. Mnogo takšnih vprašanj se poraja tudi prevajalcem in terminologom, ki delajo v institucijah EU. Terminološko svetovanje je na voljo prek spletišča Terminološke sekcije ZRC SAZU z imenom Terminologišče, kjer strokovnjaki odgovarjajo tudi na vprašanja, povezana s prevajanjem terminologije EU. To storitev sodelavci omenjene ustanove zaenkrat opravljajo ob svojih ostalih zadolžitvah, hitro rastoče potrebe prevajalcev in dinamika pa narekujeta razširitev dejavnosti, okrepitev obstoječe ekipe sodelavcev, saj bi le tako lahko odgovorili na bistveno večji pritok vprašanj. Obstaja torej potreba po razširitvi in formalizaciji te storitve.

Opis:

Terminološko svetovanje je storitev, kjer uporabnik prek spletnega vmesnika postavi s terminologijo povezano vprašanje, vmesnik pa uporabnika spodbuja tudi k čim bolj natančni formulaciji vprašanja z navedbo vira, sobesedila, morebitnih tujejezičnih ustreznikov itd. Za učinkovito delovanje te storitve je treba zagotoviti predvsem infrastrukturo, se pravi spletni vmesnik z arhiviranjem že odgovorjenih vprašanj, in okrepiti ustrezno strokovno usposobljeno osebje, ki na vprašanja odgovarja čim bolj ažurno, po potrebi pa za mnenje vpraša še zunanje strokovnjake. Svetovalna dejavnost tako vključuje tudi vzdrževanje imenika zunanjih svetovalcev po posameznih področjih, ki jih je mogoče vključiti v reševanje posameznih dilem.

Predvideni učinki:

- dostopnost storitve terminološkega svetovanja, ki je brezplačna, zanesljiva in ažurna

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Svetovalna dejavnost	število odgovorjenih vprašanj
------------------------	-------------------------------

* del sredstev za prvo dejavnost se prekriva z jezikovnim svetovanjem

Akcija P-4: Javna dostopnost terminoloških virov

Zagotavljanje javno dostopnih terminoloških virov

Za mnoga strokovna področja obstajajo obsežni terminološki priročniki, slovarji ali baze. Žal je le malo teh virov javno dostopnih, nekateri so dostopni le v tiskani obliki, predvsem pa je skoraj vsak terminološki vir izdelan po drugačni metodologiji, zato je močno otežena uporaba več virov naenkrat, prav tako jih ni mogoče vključevati v druge aplikacije ali prevajalske programe. Doslej je bil izveden le en poskus integracije različnih virov v skupno iskalno okolje, in sicer na portalu Termania, vendar ta vsebuje le majhno število terminoloških slovarjev, ki so pri iskanju pomešani s splošnimi viri in zato nepregledni. Za celovito rešitev težav z integracijo obstoječih virov v skupno okolje je v okviru tega akcijskega načrta predvidena vzpostavitev terminološkega portala, ki bi ponujal enovit dostop do vseh doslej ponujenih terminoloških slovarjev, obenem pa bi bilo nujno nabor dostopnih slovarjev čim bolj razširiti.

Opis:

Cilj akcije je zagotovitev javnega dostopa do čim večjega števila terminoloških slovarjev prek enovitega in zmogljivega iskalnega vmesnika na terminološkem portalu, ter dostopa do baz teh slovarjev za vgrajevanje v jezikovne tehnologije in prevajalska orodja. Prvi korak je pridobivanje avtorskih pravic za že objavljene slovarje, pri čemer bo treba najprej izdelati prioritetni seznam teh slovarjev in se nato povezati z založbami in/ali avtorji, ki so lastniki avtorskih pravic. V drugem koraku sledi pretvorba pridobljenih slovarjev v standardizirani format, ki bo specificiran v okviru terminološkega portala, ter integracija v terminološki portal. Kot kažejo dosedanje izkušnje, ta pretvorba zaradi izjemno različnih slovarskih metodologij nikakor ne bo enostavna; še posebej pri tiskanih slovarjih starejšega datuma podatkovne strukture namreč niso dosledno uporabljane ali dokumentirane.

Predvideni učinki:

- brezplačni javni dostop do terminoloških virov za številna strokovna področja
- možnost vključevanja terminoloških virov v druge jezikovne aplikacije (strojni prevajalnik, razpoznavalnik govora itd.)

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015-2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Pridobivanje avtorskih pravic za obstoječe slovarje	število slovarjev z urejenimi/pridobljenimi avtorskimi pravicami
2 Pretvorba v standardizirani format in integracija v terminološki portal	število integriranih slovarjev

Akcija P-5: Samodejno luščenje terminologije

Samodejno luščenje terminologije

Sodobni terminološki viri morajo odsevati jezikovno dinamičnost strok. Mnoge stroke se namreč hitro razvijajo, zato s klasičnimi priročniki ni mogoče spremljati razvoja strokovne terminologije, poleg tega je ustvarjanje terminoloških slovarjev na tradicionalen način zelo dolgotrajno opravilo. Podobno kot se pri izdelavi splošnih slovarjev uporabljajo korpusi, ki kažejo sodobno jezikovno rabo in razmerja med pogostnostmi posameznih jezikovnih enot, se tudi pri izdelavi terminoloških slovarjev uporabljajo specializirani korpusi, iz katerih je mogoče s samodejnimi metodami izluščiti terminološke kandidate in jih predati v nadaljnjo obdelavo terminologu ali strokovnjaku. Namen orodja za samodejno luščenje terminologije je torej skrajšanje in poenostavitev postopka izdelave terminološkega slovarja, kar je še posebej relevantno za tiste stroke, ki doslej sploh še niso imele slovarskih priročnikov. Orodje je primerno tudi za uporabo pri večjih interdisciplinarnih prevajalskih projektih.

Opis:

Cilj akcije je zagotovitev orodja za samodejno luščenje terminologije iz slovenskih besedil, v perspektivi pa še razširitev tega orodja za luščenje terminologije iz dvojezičnih besedilnih zbirk. Programska oprema za delovanje potrebuje temelje jezikovne tehnologije za slovenščino, in sicer tokenizacijo, lematizacijo in oblikoskladenjsko označevanje, pri statističnem vrednotenju terminološkosti pa se za primerjavo uporablja lematizirani frekvenčni seznam referenčnega korpusa, kot je na primer Gigafida. Orodje je mogoče prilagajati posamezni domeni s pomočjo prilagajanja oblikoskladenjskih terminoloških vzorcev in izbire statistike za vrednotenje terminološkosti. Orodje podpira dva načina delovanja: pri prvem načinu se terminološki kandidati iz korpusa uporabniku izpišejo v obliki seznama, ki ga uporabnik lahko obdeluje naprej, pri drugem načinu pa se izluščeni termini označijo v samem besedilu.

Predvideni učinki:

- hitrejša izdelava terminoloških virov za nova in interdisciplinarna področja,
- možnost vključevanja luščilnika terminologije v druge napredne jezikovne aplikacije,
- izvirne znanstvene objave.

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015-2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Razvoj orodja za samodejno luščenje terminologije	da/ne
---	-------

Akcija P-6: Večjezični portal

Večjezični portal

Namen večjezičnega portala je vzpostavitev spletnega mesta, kjer so zbrani večjezični viri in tehnologije za prevajalce in druge uporabnike, obenem pa so večjezični viri primerni za vgradnjo v naprednejše večjezične aplikacije. Najpomembnejši viri so dvo- in večjezični slovarji, dvo- in večjezični korpusi ter pomnilniki prevodov, od tehnologij pa je zagotovo najpomembnejši prosto dostopni strojni prevajalnik. Mnogi dvo- in večjezični viri že obstajajo, a so zapisani v različnih formatih ali pa zanje niso rešena vprašanja avtorskih pravic.

Opis:

Cilj akcije je vzpostavitev večjezičnega portala. V prvi fazi je potrebno izdelati specifikacije za različne tipe jezikovnih virov ter definirati njihove zapise v skladu z mednarodnimi standardi (TEI, TMX), vzporedno s tem pa določiti tudi tehnične specifikacije portala vključno z iskalnimi orodji. Posebna pozornost se mora nameniti poveztivosti z drugimi portali v okviru tega akcijskega načrta, še posebej s terminološkim portalom. V drugi fazi sledi vzpostavitev samega portala z vso pripadajočo strežniško in programsko infrastrukturo, v tretji fazi pa je načrtovana integracija vseh obstoječih večjezičnih virov v portal. Akcija je neposredno povezana z akcijama K-4 Dvojezični korpusi in L-5 Dvo- in večjezične baze in slovarji, v katerih je cilj zagotavljanje in gradnja večjezičnih virov.

Predvideni učinki:

- enotna vstopna točka za večjezične vire in tehnologije,
- boljši dostop do virov in tehnologij za prevajanje,
- možnost uporabe večjezičnih virov v naprednih večjezičnih aplikacijah.

Nosilec:

MK

Časovni plan:

2015-2018

Podrobna opredelitev aktivnosti s kazalniki:

1 Specifikacije portala, vhodnih in izhodnih formatov	da/ne
2 Vzpostavitev spletišča, zagotovitev strežniških kapacitet, zagotovitev iskalnika in konkordančnika	da/ne
3 Integracija obstoječih večjezičnih virov v portal	število integriranih virov

Akcija P-7: Slovnični portal za šolarje

Nadgradnja slovničnega portala za šolsko populacijo

Slovnični portal je namenjen jezikovnemu opismenjevanju učencev v slovenskih šolah. Temelji na obstoječem slovničnem portalu, ki je bil že razvit v okviru že financiranih projektov v preteklem obdobju. Nadgradnja portala mora biti usklajena s konceptom novega informativno-normativnega razlagalnega slovarja s slovarsko bazo, novim konceptom slovničnega opisa sodobne slovenščine ter novim konceptom šolske slovnice. V okviru nadgradnje portala je predvidena korpusna analiza specializiranih korpusov z normativnimi popravki (Šolar, Lektor) ter variantnosti pri rabi v splošnih korpusih, kakršna je bila razvita v okviru omenjenih projektov.

Opis:

Nadgradi se slovnični portal, ki je primarno namenjen jezikovnemu opismenjevanju učencev v slovenskih šolah. Portal temelji na že obstoječih rešitvah, ki so bile razvite v preteklih projektih.

Predvideni učinki:

- izboljšanje pismenosti slovenske šolske populacije
- izboljšan slovnični opis sodobnega slovenskega jezika

Nosilec:

MK, MIZŠ

Časovni plan:

2015-2018

Podrobna opredelitev aktivnosti s kazalniki:

1 Nadgrajeni slovnični portal	da/ne
-------------------------------	-------

Akcija P-8: Spletna predstavitev normativnih podatkov

Sistem za spletno predstavitev normativnih podatkov: pravila in slovar (Cilj: Pravopisni portal)

Namen *Pravopisnega portala* je vzpostavitev spletnega mesta, kjer so zbrani in s pomočjo sistema za pregledovanje spletnih vsebin organizirano predstavljeni normativni podatki kot rezultat delovanja standardizacijskega telesa oz. pravopisne komisije. Pravopisna pravila so predstavljena po problemskih sklopih (pisavoslovje, pravopis – mala/velika začetnica, prevzemanje, ločila, pisanje skupaj in narazen, pripadajoča poglavja iz glasoslovja, oblikoslovja in besedotvorja) z vzajemno povezavo na slovarsko predstavitev normativne informacije.

Opis:

Cilj akcije je priprava sistema za pregledovanje spletnih vsebin, tj. osnovnih pravopisnih pravil s povezavami na tipične zglede, posebnosti, izjeme, torej na slovarsko pravopisno zbirko. Sistem je namenjen uporabnikom različnih profilov.

Akcija **P-8** je neposredno povezana z akcijo **N-1** (*Predstavitev normativnih podatkov: pravila in slovar*) se v sklopu cilja *Pravopisni portal* kronološko umešča za njo: prvi problemski sklopi prenovljenih pravopisnih pravil (kot rezultat dela pravopisne komisije) in pripadajočega slovarja naj bi bili konec leta 2016 s pomočjo sistema že dodani na spletno mesto *Pravopisni portal*. Do leta 2018, ko poteka prenova pravopisnih pravil in akcija **N-1**, pa naj bi vzpostavljeni sistem omogočal pregledovanje več sto problemskih sklopov.

Predvideni učinki:

- enotna vstopna točka za pravopisna pravila in slovar
- pregleden prikaz pravopisnih pravil in pripadajočih slovarskih sestavkov
- povezave med pravili in slovarjem v okviru enega problemskega sklopa

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Sistem za pregledovanje pravil	da/ne
2 Sistem za pregledovanje slovarja	da/ne
3 Modul za dodajanje problemskih sklopov	da/ne

Akcija P-9: Spodbujanje prožnih in inovativnih oblik učenja z razvojem jezikovnih virov in tehnologij

Spodbujanje prožnih in inovativnih oblik učenja z razvojem jezikovnih virov in tehnologij

Cilj akcije je s pomočjo novih učnih jezikovnih e-orodij razviti inovativne in prožne učne pristope in metode, ki bodo pripomogli k zvišanju jezikovne zmožnosti učencev.

Nova učna jezikovna e-orodja bodo razvita na podlagi opisa izbranih in relevantnih vidikov slovenščine, tako z vidika slovnice, slovarja kot pravopisnih informacij. Novi opisi bodo prosto in v dogovoru s financerjem odprto dostopni z namenom povezljivosti in integracije z drugimi obstoječimi aktualiziranimi opisi jezika. Nova učna orodja bodo utemeljena na veljavnih učnih načrtih in standardih. Ob gradnji orodij bodo izvajalci zasnovali tudi nove pristope in metode, s katerimi bodo seznanili in usposobili učitelje.

Akcija je naravnana celostno, upoštevajoč tako vidike jezikovnega opisa, standardizacije, jezikovnih tehnologij oz. aplikacij kot tudi jezikovnega izobraževanja.

Opis:

1. izbrani vidik(i) slovarsko-slovničnega opisa in jezikovnih normativnih informacij v digitalni obliki
2. učna e-orodja, temelječa na sodobnem jezikovnem opisu in aktualiziranih jezikovnonormativnih informacijah
3. razvoj inovativnih in prožnih učnih pristopov in metod
4. izobraževanje učiteljev

Predvideni učinki:

- posodobitev relevantnih vidikov jezikovnega opisa in kodifikacije
- izdelava učnih jezikovnih e-orodij
- razvoj novih učnih pristopov in metod
- izboljšanje jezikovne zmožnosti
- izboljšanje učiteljskih kompetenc v smislu prilagojenosti potrebam učencev v digitalni dobi

Nosilec:

MK

Časovni plan:

2016–2023

Obseg financiranja: 2.000.000,00 EUR

Podrobna opredelitev aktivnosti s kazalniki:

1 Izbrani vidiki jezikovnega opisa	da/ne
2 Izbrani vidiki prenovljene jezikovne kodifikacije	da/ne
2 Izdelana e-orodja (s področja slovnice, slovarja, pravopisa, drugo)	4 sklopi
3 Izdelava smernic za nove učne pristope in metode	da/ne
4 Delavnice usposabljanj za učitelje	Število

Akcija K-1: Pisni korpusi

Nadgradnja pisnih korpusov sodobnih besedil in njihovih konkordančnikov

Večji slovenski enojezični korpusi z vsaj delno žanrsko raznolikostjo so Gigafida, Kres, slWaC in Nova beseda. Za naslednje obdobje je treba predvideti oblikovanje in nadgradnjo osrednjega korpusa slovenskega jezika po zgledu drugih evropskih jezikov. To vprašanje vključuje predvsem časovno opredelitev sodobne slovenščine, odločitev o vključitvi besedil, objavljenih pred opredeljeno mejo, ki izkazujejo vpliv na sodobni jezik (branost), distribucijo žanrov, označenost in podobno. Odločitve o vprašanju sodobne slovenščine določajo mejo med korpusom starejše slovenščine in korpusom sodobnega jezika.

Referenčni korpus: pri referenčnem korpusu sodobnega jezika je treba zagotoviti sprotno obnavljanje korpusa zaradi možnosti spremljanja sprememb v besedišču. Ker se javno objavljane vse bolj premika v digitalne medije, predvsem na splet, je v prvi vrsti treba zagotoviti redno in nadzorovano pajkanje (*web crawling*) slovenskega spleta, razen pri žanrih, ki se na spletu pojavljajo v manjši meri, predvsem leposlovje, stvarna literatura oz. nasploh monografske publikacije. Pri teh je treba predvideti redno nadgrajevanje korpusa z zbiranjem besedil pri izdajateljih.

Korpusi besedil s spleta: za slovenščino trenutno obstajata internetni del korpusa Gigafida ter korpus sloWaC. Vzpostaviti je treba mehanizem za sprotni zajem, čiščenje, označevanje in uvoz v konkordančnik. Za procesiranje spletnega korpusa je treba v okviru jezikovnotehnoloških orodij predvideti tudi izdelavo orodja za avtomatsko prepoznavanje novih besed in pomenov. Korpusi besedil s spleta so neodvisna kategorija od osrednjega korpusa sodobnega jezika, ki za razliko od splošnega spletnega korpusa stremlje k uravnoteženosti, spletni korpus pa k čim večjemu zajemu slovenskega spleta.

Korpusi strokovnih besedil: primeri obstoječih enojezičnih terminoloških korpusov so DSI in SDJT (informatika), KoRP (odnosi z javnostmi), Grizold (vojaška terminologija), EduKorp (izobraževanje) in nekateri manjši (KONJI, FILMI). Med večjezičnimi terminološkimi korpusi je največji korpus Turk (turistika). Terminološki korpusi so pomembni za proces izdelave terminoloških slovarjev in baz, ker pomenijo izhodišče za avtomatsko pridobivanje terminoloških kandidatov ter drugih jezikovnih informacij. Akcijski načrt predvideva vzpostavitev platforme za procesiranje besedil s posameznih področjih vednosti v okviru terminološkega portala, ki vključuje strojno luščenje terminologije, skupaj s kolokacijami in zgledi pri prepoznanih terminoloških kandidatih. Akcijski načrt je glede prednostnih področij gradnje terminoloških korpusov neopredeljen, bistveno je, da je vzpostavljen uporabniško prijazen proces vključevanja besedil v terminološki korpus ter celoten proces do vključevanja terminoloških kandidatov v terminološki portal, skupaj z drugimi leksikalnimi informacijami. Končni cilj je korpus znanstvenih besedil v slovenskem jeziku za vsa znanstvena področja.

Konkordančniki: za raziskovalno relevanten in uporabniško prijazen dostop do korpusov je v prihodnji razvoj treba vključiti nadgradnjo spletnih konkordančnikov ter oblikovanje enotnih izhodišč pri ravnanju z avtorskimi pravicami pri besedilih v korpusih. Pri tem je treba zagotoviti nadzorovan prosti dostop do korpusov za raziskovalne namene ter čim bolj nerestriktiven javni dostop do korpusov na spletu, z upoštevanjem varovanja avtorskih pravic in zaščite osebnih podatkov. Za ta namen je treba predvideti izdelavo smernic pri zbiranju gradiva za korpuse, tipskih pogodb za različne tipe korpusov, stopnje varovanja avtorskih pravic in podobno, pri čemer zgled predstavljajo evropske prakse iz projektov ELRA, CLARIN, META-NET (META-SHARE) in drugih.

Opis:

Vzpostavi se mehanizem za redno vzdrževanje referenčnega korpusa sodobnega jezika in korpusa besedil s spleta. Nadgradi se obstoječe spletne konkordančnike.

Predvideni učinki:

- izboljššan dostop do osnovnih virov za analizo sodobnega slovenskega jezika
- izboljššan slovarski opis sodobnega slovenskega jezika
- izboljššan slovnični opis sodobnega slovenskega jezika
- izboljšane možnosti za razvoj jezikovnotehnoloških aplikacij za slovenščino

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015-2018

Obseg financiranja: možnost izvedbe preko CRP**Podrobna opredelitev aktivnosti s kazalniki:**

1 Nadgrajeni in vzdrževani referenčni korpus sodobnega jezika	da/ne
2 Nadgrajeni in vzdrževani korpus besedil s spleta	da/ne
3 Nadgrajeni in vzdrževani spletni konkordančnik za pisne korpuse	da/ne

Akcija K-2: Specializirani korpusi in konkordančniki

Nadgradnja pisnih korpusov sodobnih besedil, nadgradnja in izdelava specializiranih korpusov, izdelava korpusov znanstvenih besedil, izdelava korpusa slovenskega jezika, specializiranega za normativna vprašanja

Oblikoskladenjsko in skladenjsko označeni korpusi (odvisnostne drevesnice): Pri nivojih označevanja ročno preverjenih učnih korpusov manjka raven razreševanja koreferenčnosti, treba je povečati dele korpusa z označenimi imenskimi entitetami, v obstoječih korpusih manjka tudi označevanje semantičnih vlog ter stalnih besednih zvez (*multi-word expressions*), tako na ravni imenskih entitet kot tudi idiomatskih izrazov, frazeologije itd. Pri skladenjskem razčlenjevanju podatki kažejo, da je povedi s skladenjsko označenimi drevesnicami načeloma dovolj, manjkajo pa oblikoskladenjsko in skladenjsko označeni učni korpusi za različne (neformalne) žanre: govorni jezik, socialna omrežna ipd. Pri jezikovnotehnoloških učnih korpusih je tipično, da se označevanje novih ravni nalaga na že ročne pregledane podatke, zato so izhodišče za označevanje dodatnih ravni korpusi ssj500k, jos100k in jos1M, kar velja za sodobni pisni jezik, v primeru označevanja in razčlenjevanja starejših besedil je izhodišče korpus goo300k, pri govornem jeziku in socialnih omrežjih izhodiščnih učnih korpusov še ni. Izhodišče za te učne korpusa sta korpusa Gos in Tweet-sl.

Semantično in diskurzno označeni korpusi: Naslednjo raven označevanja učnih korpusov predstavljajo semantično in diskurzno označeni korpusi. Semantični bazi, ki sta večinoma rabljeni za označevanje učnih korpusov, sta WordNet in FrameNet. Slovenska različica sloWNet je bila že uporabljena za semantično označevanje učnega korpusa jos100k, ti podatki pa niso bili upoštevani pri sestavljanju učnega korpusa ssj500k. Ker je korpus jos100k integriran v korpus ssj500k, je pri nadgradnji semantičnega označevanja treba upoštevati podatke, ki so že na voljo. Slovenski FrameNet trenutno obstaja na ravni pilotnega korpusa z 225 stavki (4.000 besed), označenimi na vseh ravneh (oblikoslovna, skladenjska, semantična), ki jih predvideva sistem označevanja. Za nadaljnje označevanje je smiselno že označene stavke integrirati v osrednji učni korpus, treba pa je zagotoviti bolj učinkovit sistem ročnega označevanja, saj obstoječa programska oprema FrameNet Desktop ni najbolj primerna za označevanje.

Dialoški korpusi: predvidi se izdelava dialoškega korpusa ali več specifičnih korpusov za namen izdelave sistemov dialoga za slovenščino. Korpus vključuje posnetke in transkripcije, ki omogočajo raziskave medčloveške interakcije, usmerjene v izgradnjo sistemov dialoga. Ta naloga vključuje tudi pregled možnosti izdelave dialoškega podkorpusa iz korpusa Gos.

Multimodalni korpusi (besedilo + slika / zvok): akcijski načrt predvideva izdelavo multimodalnih korpusov za potrebe izdelave jezikovnotehnoloških aplikacij, kot so virtualni agenti, ter v tem kontekstu tudi za raziskave mimike in gestikulacij. Označevanje teh segmentov se dodaja korpusom, ki že vsebujejo druge ravni transkripcije, predvsem govora.

Korpusi z oznakami popravkov: tovrstni korpusi so namenjeni učenju in poučevanju jezika ter raziskovanju težav, ki jih imajo učenci bodisi pri usvajanju tujega jezika ali pri opismenjevanju v maternem jeziku. Tipično ti korpusi vsebujejo tudi oznake napak pri produkciji besedil v tujem ali maternem jeziku. Za slovenščino trenutno obstajajo korpusi Šolar (korpus pisne produkcije učencev in dijakov), Lektor (korpus lektorskih popravkov) in Kus (korpus usvajanja slovenščine). V prihodnje je treba načrtovati povečanje vseh treh korpusov, prvih dveh do velikost vsaj 10 milijonov besed, korpus usvajanja slovenščine do velikosti 5 milijonov besed.

Korpusi znanstvenih besedil: V zadnjem času postaja digitalno dostopnih vse več znanstvenih besedil v slovenskem jeziku, bodisi neposredno s strani organizatorjev znanstvenih srečanj ali skozi portale, kot so dLib.si, v novejšem času pa skozi digitalne knjižnice univerz preko Nacionalnega portala odprte znanosti. Ob naporih za uporabo slovenščine tudi kot jezika znanosti se je potrebno zavedati, da so takšna digitalizirana in dostopna znanstvena besedila tudi izjemno dragocen vir slovenskega strokovnega izrazja, ki je sicer nedostopno. To še posebej velja za terminologijo, ki jo zaradi hitrega razvoja znanosti klasični terminološki slovarji ne morejo dohajati, je pa mnogokrat vsebovana v slovenskih publikacijah posameznih področij.

Ostali korpusi: (a) enojezični in večjezični korpusi podnapisov – tovrstni korpusi postajajo pomemben del jezikovne infrastrukture, predvsem v kombinaciji z avtomatsko razpoznavo govora in strojnim prevajanjem. Predvidevamo lahko, da bo razvoj pri evropskih jezikih šel v smer zagotavljanja podnapisov v realnem času, vključno s strojnim prevajanjem podnapisov v druge jezike. Korpusi podnapisov zagotavljajo možnost razvoja tovrstnih tehnologij. (b) Korpusi, označeni za potrebe strojnega analize razpoloženja (sentiment analysis) in luščenja mnenj (opinion mining) – strojna analiza razpoloženja in luščenje mnenj vedno bolj nadomešča tradicionalno anketiranje, predvsem v okolju spletnih forumov, socialnih omrežij itd. (c) Korpusi za strojno določanje berljivosti besedil – strojno določanje berljivosti besedil je pomembno za učenje jezika in druge aplikacije, kjer je zahtevnost besedila dejavnik, denimo pri rangiranju spletnih strani glede na berljivost, ocenjevanju zahtevnosti besedil v šolskih besedilih, (spletnih) anketah itd. (d) Korpusi za prepoznavanje sovražnega govora – strojno prepoznavanje sovražnega govora zaradi preprostega dostopa lahkim dostopom vseh uporabnikov spleta do spletnih forumov postaja pomembna tehnologija, ki za statistične aplikacije za prepoznavanje sovražnega govora. (e) Drugi korpusi: primer specializiranih korpusov je denimo korpus za prepoznavanje avtorstva in podobni - prednostne teme so prepuščene iniciativi raziskovalcev.

Konkordančniki: za raziskovalno relevanten in uporabniško prijazen dostop do korpusov je v prihodnji razvoj treba vključiti nadgradnjo spletnih konkordančnikov ter oblikovanje enotnih izhodišč pri ravnanju z avtorskimi pravicami pri besedilih v korpusih. Pri tem je treba zagotoviti nadzorovan prosti dostop do korpusov za raziskovalne namene ter čim bolj nerestriktiven javni dostop do korpusov na spletu, z upoštevanjem varovanja avtorskih pravic in zaščite osebnih podatkov. Za ta namen je treba predvideti izdelavo smernic pri zbiranju gradiva za korpuse, tipskih pogodb za različne tipe korpusov, stopnje varovanja avtorskih pravic in podobno, pri čemer zgled predstavljajo evropske prakse iz projektov ELRA, CLARIN, META-NET (META-SHARE) in drugih.

Opis:

Vzpostavi se mehanizem za redno vzdrževanje oblikoskladenjsko in skladenjsko označenih korpusov, semantično in diskurzno označenih korpusov, dialoških, multimodalnih in ostalih korpusov. Nadgradi se obstoječe spletne konkordančnike.

Pri izdelavi korpusov znanstvenih besedil naj se iz obstoječih portalov, ob upoštevanju avtorskih pravic nad izvirnimi besedili, zajame znanstvena besedila in se jih pretvori v primerno jezikoslovno označeni korpus oz. korpuse ter jih naredi prosto dostopne preko spletnih konkordančnikov za namene preučevanja slovenskega znanstvenega in strokovnega izrazja. Iz korpusa se naknadno lahko avtomatsko izlušči terminologijo, kjer je to možno skupaj s tujejezičnimi ustrezniciami in termine ponudi v prosti in odprti dostop. Izluščene terminološke baze so nato uporabne za posamezne znanstvene skupnosti za namene upravljanja s terminologijami npr. v okviru splošnega terminološkega portala kot dela prostodostopnega spletnega portala s čim več dosegljivimi jezikoslovnimi podatki o slovenščini.

Predvideni učinki:

- izboljššan dostop do osnovnih virov za analizo sodobnega slovenskega jezika
- izboljššan slovarski opis sodobnega slovenskega jezika
- izboljššan slovnični opis sodobnega slovenskega jezika
- izboljššane možnosti za razvoj jezikovnotehnoloških aplikacij za slovenščino
- zagotovljena korpusna infrastruktura za preučevanje slovenskega znanstvenega izrazja
- možnost izdelave avtomatsko izluščene terminološke baze slovenskega znanstvenega izrazja kot neposredno uporabni priročnik za doslej pomanjkljivo pokrita področja in kot osnova za izdelavo slovenske terminologije posameznih znanstvenih področij

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015-2018

Obseg financiranja: možnost izvedbe preko CRP**Podrobna opredelitev aktivnosti s kazalniki:**

1 Nadgrajeni in vzdrževani spletni konkordančnik za specializirane korpuse	da/ne
2 Nadgrajeni in vzdrževani oblikoskladenjsko in skladenjsko označeni korpusi	da/ne
3 Izdelani in vzdrževani semantično in diskurzno označeni korpusi	da/ne
4 Izdelani in vzdrževani dialoški korpusi	da/ne
5 Izdelani in vzdrževani multimodalni korpusi	da/ne
6 Dostopen korpus znanstvenih del	da/ne
7 Avtomatsko izdelan nabor terminoloških kandidatov	da/ne
8 Vključitev terminoloških kandidatov v terminološki portal	da/ne
9 Izdelani in vzdrževani ostali korpusi	da/ne

Akcija K-3: Govorni korpus in baza

Nadgradnja referenčnega govornega korpusa, večnamenski korpus in baza govornenega jezika ter dialoški korpus

Govorni korpusi, ki se razvijajo vzporedno z govorno bazo, so nujno potrebni, da se lahko začnejo razvijati postopki za kakršno koli avtomatsko procesiranje govornenega besedila – bodisi kot del jezikovnih modelov v razpoznavalniku ali prevajalniku bodisi kot del bolj bazičnih orodij – za izdelavo oblikoslovnih označevalnikov in skladišnih razčlenjevalnikov za govorna besedila, za izdelavo leksikonov s podatki o govornem jeziku, seveda tudi za vse višje ravni analize besedila (semantična, pragmatična analiza) ipd. Govorni korpus z govorno bazo pa ni samo eden najbolj osnovnih virov za razvoj govornih tehnologij, ampak tudi temeljni vir za jezikoslovne raziskave govornenega jezika. Te imajo zdaj na voljo korpus v obsegu 1 mio. besed (GOS), ki pa je zaradi majhnosti le pogojno referenčni. Standard v svetu se na področju govornih korpusov bliža številki 10 mio. besed. Dialoški korpusi so tesno povezani z govornimi korpusi in so potrebni za razvoj sistemov dialoga ali diskurzne in jezikoslovne raziskave dialoga.

Opis:

Govorna baza in korpus v obsegu nekaj 100 ur, ki zajema različne tipe javnega ter nejavnega (tudi zasebnega) diskurza. Vsebina baze in korpusa se skrbno uravnateži glede na vsa osrednja področja uporabe: v jezikoslovju, analizi diskurza, razpoznavanju tekočega govora, sistemih dialoga ter za razvoj multimodalnih govornih agentov. Pri zasnovi specifikacij za izdelavo je tako treba upoštevati predvsem uporabnost za akustično in jezikovno modeliranje na eni strani ter jezikoslovne raziskave (vključno s fonetično ravno) na drugi strani (kar vključuje med drugim avtomatsko oblikoslovno, skladiščno, semantično, pragmatično in diskurzno označevanje in analizo besedil, jezikoslovne in diskurzne analize govora), določen del baze pa mora naslavljal tudi potrebe multimodalnih JT sistemov in sistemov dialoga. Baza in korpus se izdelata v obliki več podbaz/podkorpusov, ki pa vsi sledijo enotnemu standardu. Vsi posnetki v bazi se segmentirajo in transkribirajo v skladu z obstoječimi standardi. Izdelani vir se poleg izvirne oblike ponudi tudi prek spletnega servisa.

Predvideni učinki:

- izboljšana kvaliteta razpoznavalnika govora za tekoči govor
- spodbujanje gospodarskega razvoja na področju IT produktov z vključenim razpoznavanjem tekočega govora (npr. avtomatsko podnaslavljanje, narekovalniki ipd.)
- spodbujanje razvoja virtualnih multimodalnih agentov
- spodbujanje razvoja sistemov dialoga
- spodbujanje jezikoslovnih in diskurzni raziskav govornenega jezika
- podpora opisu govornenega jezika v jezikovnih priročnikih

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015-2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Govorna baza za akustične modele za splošno domeno	da/ne
2 Govorna baza za akustične modele za šolsko domeno	
3 Govorni korpus za jezikovne modele	da/ne
4 Razširjen referenčni govorni korpus slovenščine	da/ne
5 Multimodalna baza (označene geste)	da/ne
6 Dialoški korpus	da/ne
7 Nadgrajen konkordančnik za govorni korpus	da/ne
8 Orodje za avtomatsko pretvorbo pogovornega v standardizirani zapis (referenca: korpus GOS) kot odprtokodna aplikacija	da/ne
9 Orodje za segmentacijo govora na izjave (na akustični ravni)	da/ne

Akcija K-4: Dvojezični korpusi

Dvojezični korpusi

Dvojezični korpusi predstavljajo temeljni vir za izdelavo dvojezičnih slovarjev in baz, poleg tega pa so temeljni vir za gradnjo številnih večjezičnih tehnologij. V prvi vrsti so vzporedni korpusi neobhodno potrebni za gradnjo strojnega prevajalnika. Obstoječi strojni prevajalniki za slovenščino (Google, Bing, Presis) so namreč komercialni in jih zato ne moremo prištevati med splošno dostopna orodja, poleg tega gre pri tem za tako pomembno tehnologijo, da je vanjo vsekakor smiselno investirati javna sredstva. Pri zagotavljanju dvojezičnih korpusov upoštevamo trenutno stanje, ki zajema obstoj številnih dvojezičnih korpusov (Evrokorpus, JRC-ACQUIS, IJS-Elan, Trans itd.) večinoma za jezikovni par slovenščina-angleščina, namen akcije pa je bistveno povečati in uravnotežiti obstoječe korpusne za prioritete jezikovne pare z namenom zagotavljanja dovolj velike količine podatkov za razvoj strojnih prevajalnikov.

Opis:

Prva prioriteta je gradnja dvojezičnega, stavčno poravnane vzporednega korpusa za jezikovni par angleščina-slovenščina v čim večjem obsegu, predvidoma minimalno 100 milijonov besed za en jezik. Korpusu se priključijo tudi (deli) obstoječih korpusov, vendar je zgornja številka namenjena novim besedilom. Vsebinsko mora biti korpus zasnovan tako, da pokriva čim širši spekter različnih besedil. Poseben poudarek je treba nameniti kakovosti poravnave besedil. Za oba jezika morajo biti besedila oblikoskladenjsko označena in skladenjsko razčlenjena. Poleg vzporednega korpusa se mora izdelati tudi pripadajoči enojezični korpus vseh zbranih besedil (tudi tistih, ki iz poravnave izpadejo) za vsak jezik.

Drugi del akcije je vključuje več aktivnosti za gradnjo stavčno poravnanih vzporednih korpusov za ostale prednostne jezikovne pare, med temi obvezno za jezikovni par z nemščino. Obseg naj bi bil čim večji, predvidoma pa minimalno 20 milijonov besed za en jezik. Tudi ti korpusi morajo biti čim bolj reprezentativni in uravnoteženi ter čim bolj natančno poravnani. Za vse jezike, kjer je na voljo oblikoskladenjsko označevanje in skladenjsko razčlenjevanje, morajo biti dodane tudi te oznake. Poleg vzporednega korpusa se mora izdelati tudi pripadajoči enojezični korpus vseh zbranih besedil (tudi tistih, ki iz poravnave izpadejo) za vsak jezik.

Poseben podsklop akcije je zasnova in izdelava korpusa tolmačenj, obvezno za jezikovna para angleško-slovensko in nemško-slovensko, lahko pa tudi za dodatne jezikovne pare, v minimalnem obsegu 100.000 besed za en jezik v paru z angleščino in 100.000 besed v paru s slovenščino.

Vsi korpusi, ki bodo nastali v okviru te akcije, morajo biti zapisani v skladu s specifikacijami akcije P-7. Vsi korpusi morajo imeti urejene avtorske pravice, tako da bo do njih mogoče prosto dostopati prek večjezičnega portala ter da bodo hkrati odprti in prosto dostopni za razvoj tehnologij.

Predvideni učinki:

- zagotovitev dragocenih jezikovnih virov za prevajalce, strokovnjake, jezikoslovce in prevodoslovce
- zagotovitev temeljnih jezikovnih virov za strojni prevajalnik in druge večjezične aplikacije
- zagotovitev temeljnih jezikovnih virov za gradnjo novih dvojezičnih baz in slovarjev

Nosilec:
MK, MIZŠ (ARRS)

Časovni plan:
2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Veliki vzporedni, stavčno poravnani, oblikoskladenjsko označen in skladienjsko razčlenjen korpus za jezikovni par AN-SL	obseg korpusa v milijonih besed za en jezik
2 Vzpostavljenemu korpusu pripadajoči enojezični korpus vseh besedil, zbranih v angleškem jeziku, oblikoskladenjsko označen in skladienjsko razčlenjen	obseg korpusa v milijonih besed
3 Vzpostavljenemu korpusu pripadajoči enojezični korpus vseh besedil, zbranih v slovenskem jeziku, oblikoskladenjsko označen in skladienjsko razčlenjen	obseg korpusa v milijonih besed
4 Vzpostavljeni, stavčno poravnani, oblikoskladenjsko označen in skladienjsko razčlenjen* korpus za jezikovni par NE-SL	obseg korpusa v milijonih besed za en jezik
5 Vzpostavljenemu korpusu pripadajoči enojezični korpus vseh besedil, zbranih v nemškem jeziku, oblikoskladenjsko označen in skladienjsko razčlenjen*	obseg korpusa v milijonih besed
6 Vzpostavljenemu korpusu pripadajoči enojezični korpus vseh besedil, zbranih v slovenskem jeziku, oblikoskladenjsko označen in skladienjsko razčlenjen	obseg korpusa v milijonih besed
7 Vzpostavljeni, stavčno poravnani korpusi za druge prioritete jezikovne pare ter pripadajoči enojezični korpusi (IT-SL, ŠP-SL, HR-SL, HU-SL, FR-SL), oblikoskladenjsko označeni in skladienjsko razčlenjeni*	obseg korpusov v milijonih besed
8 Vzpostavljeni, stavčno poravnani korpus tolmačenj za jezikovni par AN-SL	obseg korpusa v št. besed
9 Vzpostavljeni, stavčno poravnani korpus tolmačenj za jezikovni par NE-SL	obseg korpusa v št. besed

* Oblikoskladenjsko označijo in skladienjsko razčlenijo se korpusi samo za tiste jezike, kjer so tovrstna orodja prosto na voljo za nekomercialno ali komercialno rabo.

Akcija K-5: Diahroni korpus

Izdelava diahronnega korpusa slovenskega jezika

V naslednjem obdobju je predvidena gradnja enotnega diahronnega korpusa slovenskega jezika, ki zajema digitalizirane vire od prvih pisnih virov v slovenskem jeziku do predvidene meje sodobnega jezika. Za diahroni korpus je treba predvideti gradnjo sekundarnih virov in tehnologij (leksikoni, označevalniki itd.), ki podpirajo strojno obdelavo starejših besedil – primer so rezultati projekta IMP ter IMPACT. V diahroni korpus je smiselno vključiti čim več že obstoječih virov, ki so nastali v okviru različnih projektov (dLib, Wikivir, eZISS, NRSS, eZMono, korpus besedil protestantskih piscev itd.), opredeliti je treba kriterije vključevanja besedil, format, označevanje itd. ter osnovno odločitev ali gre za združevanje samostojnih korpusov ali enoten korpus.

Konkordančniki: za raziskovalno relevanten in uporabniško prijazen dostop do korpusov je v prihodnji razvoj treba vključiti nadgradnjo spletnih konkordančnikov za prikaz podatkov v diahronih korpusih. To vključuje predvsem možnost iskanja po različnih nivojih podatkov, ki jih vsebujejo tovrstni korpusi, npr. posodobljene oblike besed in podobno.

Opis:

Vzpostavi se mehanizem za gradnjo enotnega diahronnega korpusa slovenskega jezika. Nadgradi se obstoječe spletne konkordančnike. Na začetku gradnje diahronih korpusov je treba izdelati splošni načrt gradnje enega ali več diahronih korpusov.

Predvideni učinki:

- izboljššan dostop do osnovnih virov za analizo starejšega slovenskega jezika
- izboljšane možnosti za slovarski opis starejšega slovenskega jezika
- izboljšane možnosti za slovnični opis starejšega slovenskega jezika
- izboljšane možnosti za razvoj jezikovnotehnoloških aplikacij za obdelavo starejše slovenščine

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015-2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Nadgrajeni in vzdrževani diahroni korpusi	da/ne
2 Nadgrajeni in vzdrževani spletni konkordančnik za diahronne korpuse	da/ne

Akcija L-1: Splošni slovar in slovarska baza

Stanje v zvezi s splošnim slovarjem slovenskega jezika je že nekoliko podrobneje opisano v pregledu področju v sekciji 1.2.

V stroki in širši javnosti (v medijih) je bila zadnja leta toliko jasno izražena potreba po izdelavi splošnega slovarja slovenskega jezika na novo, da lahko to akcijo označimo za eno od prioriteten. Resolucija v zvezi z obstoječim slovarjem navede dve pomanjkljivosti:

- (1) »predvsem zaradi jezikovnih sprememb zadnjih desetletij ne zadošča[...] več vsem potrebam, ki jih pogojuje današnja jezikovna stvarnost« (ta pomanjkljivost je bila po navedbah izvajalca, Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, deloma odpravljena z drugo izdajo SSKJ oktobra 2014, vendar slovar ni uporaben kot izhodišče za bodoči informativno-normativni slovar, saj tako aktualnost slovarskega opisa kot status avtorskih pravic preprečujeta rabo slovarja ali njegovih naslednikov za namene, ki zahtevajo opis sodobne slovenščine in odprti dostop do slovarske baze),
- (2) je temeljni slovar sicer prešel v digitalno okolje, vendar mu »zaradi prvotno knjižne zasnove ni prilagojen v polni meri«.

Pomemben korak h konceptu novega splošnega slovarja slovenskega jezika je bil narejen na Posvetu o novem slovarju slovenskega jezika, organiziran v okviru Ministrstva za kulturo februarja 2014. Prispevki s posveta so objavljeni na naslovu http://www.mk.gov.si/si/delovna_podrocja/sluzba_za_slovenski_jezik/predstavitev_podrocja/dogodki/posvet_o_novem_slovarju_slovenskega_jezika/.

Opis:

Izdela se nov splošni slovar slovenskega jezika z okvirno 100.000 gesli. Pri izdelavi je treba upoštevati določila iz Resolucije. Ta pravijo:

- (1) Da mora biti to opis sodobne slovenščine, zasnovan tako, da »bo prilagojen uporabi v spletnem in drugih digitalnih okoljih, pri njem pa bodo bodisi v eni ali v različnih slovarskih bazah na voljo različni podatki o slovenščini, od pomenskih, vezljivostnih, kolokacijskih, frazeoloških, idiomatskih, etimoloških in drugih. Osnovo za opis jezika predstavljajo zbrani empirični podatki o jezikovni rabi, ki zajemajo različne pojavnne oblike jezika.«
- (2) Da je treba na osnovi teh temeljnih del »izdelati priročnike za druge ciljne uporabnike. To so predvsem slovarji in slovnice za različne stopnje izobraževanja rojenih govorcev, za tiste, katerih materni jezik ni slovenščina, in za osebe s posebnimi potrebami.«

Poleg določil Resolucije, omenjenih zgoraj, mora novi slovar slediti tudi skupnim določilom tega akcijskega načrta, opredeljenim na začetku poglavja 3, ki veljajo za vse akcije.

Prav tako mora novi slovar slediti sklepom udeležencev Posveta o novem slovarju slovenskega jezika, Ministrstvo za kulturo, 12. 2. 2014, ki so jih soglasno sprejeli prisotni predstavniki vseh zainteresiranih institucij in se glasijo tako:

1. »Posvet je izpostavil potrebo po nadaljevanju pogovorov o zasnovi novega slovarja slovenskega jezika, pri čemer je treba – prisotni pa se strinjajo, da je tudi mogoče – doseči interdisciplinarno sodelovanje strokovnjakov z različnih institucij – inštitutov, univerz in drugih.
2. Končna oblika slovarja v smislu digitalno : natisnjeno ni medsebojno izključujoča. Slovarska baza se zasnove digitalno, slovar bo prosto spletno dostopen, baza odprta za nekomercialno rabo, tisk slovarja oz. posameznih za

to formo primernih oz. prirejenih delov pa je naknaden, vendar predviden že v zasnovi.

3. Slovar bo jezikovnotehnoško zasnovan. Ob tem je treba zagotoviti konceptualno enotno, predvidljivo in standardizirano metodologijo leksikografske obdelave podatkov ter dograjevati in dopolnjevati razvidne ter dokumentirane korpusne gradivne vire in orodja.
4. Slovar bo informativno-normativen.
5. Slovar bo poleg leksikografov združil moči strokovnjakov za zgodovino jezika, etimologijo, govor, normo, stilistiko, terminologijo, podatkovne baze, korpusne in druge, ki lahko skupaj izdelajo izdelek, primerljiv s sodobnimi evropskimi slovarji, in bo primeren ter ustrezno oblikovan za različne uporabnike.«

Predvideni učinki:

- sodoben slovarski opis slovenskega knjižnega/standardnega jezika in drugih zvrsti oz. pojavnih oblik sodobne slovenščine
- izhodišče za sodoben slovnični opis slovenskega knjižnega/standardnega in drugih zvrsti oz. pojavnih oblik jezika
- jezikovni podatki o sodobni slovenščini, zbrani na enem mestu
- izboljšana bralna pismenost
- izboljšane možnosti za razvoj jezikovnotehnoških aplikacij za sodobno slovenščino

Nosilec:

MK, MIZŠ (ARRS v okviru raziskovalnega programa Slovenski jezik v sinhronem in diahronem razvoju do leta 2020)

Časovni plan:

2015–2018

Podrobna opredelitev aktivnosti s kazalniki:

1 Splošni slovar in slovarska baza ⁸	da/ne
---	-------

⁸ Podrobnejša opredelitev brez natančne teoretske in metodološke zasnove slovarja ni mogoča. Temeljni okvir zasnove podaja Resolucija o nacionalnem programu za jezikovno politiko 2014–2018 in sklepi udeležencev posveta o novem slovarju (MK, 12. 2. 2014).

Akcija L-2: Slovenski lingvistični atlas

Izdelava spletne večpredstavnostne objave 1. in 2. zvezka SLA

Ob projektu *SLA* so se od njegove zasnove v letu 1934 do danes razvijale različne metode kartiranja – od prvih rokopisnih osnutkov lingvističnih kart, nastalih že v petdesetih letih 20. stoletja, do prvih objav takih simbolnih kot izoglosno-napisnih fonetičnih in leksičnih kart. Ob prelomu tisočletja so nastale prve karte z računalniškimi oblikovalskimi orodji, kot je CorelDraw, uporaba programskih orodij za urejanje ustreznih podatkovnih baz pa je ob povezavi z geografskim informacijskim sistemom (GIS) omogočila tudi avtomatizirano kartiranje in sodobnejšo prostorsko vizualizacijo narečnih pojavov. Sodobna računalniška orodja (podatkovna baza SlovarRed, vnašalni sistem ZRCola, program ArcGIS) so olajšala/omogočila tudi pripravo 1. zvezka *SLA* (2011).

Opis:

Akcija predvideva aktivnosti, ki niso opredeljene v okviru osnovnega zagotavljanja jezikovnega opisa za potrebe *SLA*. V ta namen bo podatkovna baza dopolnjena s podatki o obsegu posameznih krajevnih govorov, torej z zamejitvijo t. i. poligonov, določenih z inovativnimi spremenljivkami, ki omogočajo razmejitve govorov ob upoštevanju zgodovinskih in geografskih kriterijev ter podatkov GURS o naseljih v Sloveniji. Ta izpopolnitev bo med drugim omogočila pripravo natančnejših izoglosnih in ploskovnih jezikovnih kart, s tem pa tudi osnovo za različne dialektometrične raziskave. Dopolnitev podatkovne baze in izpopolnitev možnosti za povezavo njenih atributov z GIS pa lahko omogočijo tudi večpredstavnostni prikaz jezikovnega gradiva v prostoru – z dostopom do digitaliziranega arhivskega gradiva, zvočnih in video posnetkov ter drugih spletnih povezav na podatke o raziskavah krajevnega govora in o kraju. Uporaba jezikovnih tehnologij naj bi pripomogla k širitvi znanstvenih spoznanj, aplikaciji jezikoslovnih vsebin tako v šoli kot v nestrokovni javnosti, hkrati pa pomagala raziskovalcem samim pri njihovem znanstvenem delu.

Predvideni učinki:

- ponatančenje dialektološkega opisa celotnega narodnega prostora
- razvoj sodobnega dialektološkega vnašalnega sistema
- dostopnost atlasa v interaktivni spletni obliki
- uzaveščanje javnosti o bogastvu slovenskih narečij
- izboljšane možnosti za nadaljnje, interpretativno jezikoslovno delo

Nosilec:

MK

Časovni plan:

2015–2018

Podrobna opredelitev aktivnosti s kazalniki:

Izboljšanje atlasa z uvedbo poligonov, določenih z inovativnimi spremenljivkami	da/ne
Večpredstavnostna spletna objava <i>SLA</i> 1 in 2	da/ne
Razvijanje vnašalnega sistema za dialektološke zapise in njegovo prilagajanje novim programskim okoljem	da/ne

Akcija L-3: Slovar naselbinskih imen

Izdelava slovarja naselbinskih imen slovenskega narodnega prostora

V Sloveniji in zamejstvih je prek 10.000 naseljenih krajev z uradnim statusom naselja, katerih imena je treba zbrati in opremiti z imenotvornimi in normativnimi podatki. Osnova za delo so monografije *Slovenska krajevna imena* F. Jakopina, T. Korošca, T. Logarja, J. Riglerja in R. Savnika, Merkujev priročnik *Slovenska krajevna imena v Italiji* in Zdovčev *Slovenska krajevna imena na avstrijskem Koroškem*. K tem bo treba dodati še imenja na avstrijskem Štajerskem in iz Porabja, pri vseh pa bo treba preveriti aktualnost in pravilnost podatkov.

Opis:

Priročnik bo na enem mestu zbral vsa slovenska naselbinska imena in jim pripisal najnujnejše jezikoslovne podatke: naglasno mesto in samoglasniško kvaliteto ter kvantiteto, sklanjatev, vezljivost, pripadajoči pridevnik, prebivalsko ime (moško, žensko in množinsko obliko), vse z morebitnimi dvojnicami, ki bodo normativno ovrednotene. Naselbinska imena z dvojezičnih področij bodo opremljena tudi z osnovno tujejezično obliko (oz. dvema). Identifikacija naselbinskega imena bo izpeljana z geokoordinatami. Slovar bo odprto dostopen na spletu, izšel pa bo tudi v knjižni obliki.

Predvideni učinki:

- vzpostavljena normativno-informativna baza slovenskih naselbinskih imen z zanesljivimi jezikovnimi in jezikoslovnimi podatki
- izid slovarja v spletni in knjižni obliki
- povezljivost baze z drugimi jezikovnimi viri in priročniki

Nosilec:

MK

Časovni plan:

2015–2018

Podrobna opredelitev aktivnosti s kazalniki:

Digitalizacija in integracija podatkov iz obstoječih virov	da/ne
Dopolnitev, aktualizacija in preverba podatkov	da/ne
Normativno vrednotenje podatkov	da/ne
Oprema baze z geokoordinatami	da/ne
Spletna in knjižna objava slovarja	da/ne

Akcija L-4: Terminološki slovarji različnih strok

Izdelava terminoloških slovarjev različnih strok

Za učinkovito sporazumevanje med strokovnjaki posameznih strok je zelo pomembna usklajena terminologija, sicer lahko pride do nesporazumov, ki so vsaj neprijetni, v nekaterih primerih pa tudi nevarni (promet, medicina, vojska itd.). Urejanje in usklajevanje terminologije je torej dejavnost, ki ima neposredne učinke na kakovost sporazumevanja v stroki. Kadar posamezni termini iz različnih razlogov prestopijo meje strokovnega jezika, to vpliva tudi na jezik v najširšem smislu. Urejena (v idealnem primeru tudi usklajena) terminologija je lahko prikazana v terminoloških slovarjih, terminoloških bazah, terminoloških podatkovnih zbirkah itd. Med terminološkimi viri je največ terminoloških slovarjev, ki s svojim načinom prikazovanja terminov, odnosov med termini, definicijami in pogosto tudi tujejezičnimi ustrezniki strokovnjakom nudijo najpomembnejše informacije o pojmi in izrazih, ki jih označujejo, torej o terminih.

Opis:

Delo za terminološki slovar se vedno začne s premislekom o zasnovi slovarja. Ko je ta določena, je prva faza vedno pregled rabe terminov na posameznem strokovnem področju. Pred razvojem jezikovnih tehnologij so termine zbirali z ročnim izpisovanjem iz relevantne strokovne literature, danes pa je prva faza terminološkega dela izdelava korpusa strokovnih besedil področja, ki se uslovarja. Na podlagi korpusa se s pomočjo jezikovnotehnoloških orodij izdela osnutek geslovnika, ki ga mora pregledati in dopolniti še strokovnjak oz. skupina strokovnjakov. Delo se nadaljuje z obravnavo posameznih terminov po pojmovnih skupinah. S tem se poveča konsistentnost slovarskih definicij, pa tudi konsistentnost slovarja v celoti. Glede na zasnovo terminološkega slovarja so uslovarjeni termini lahko tudi normativno ovrednoteni. Terminološki slovar po navadi vključuje tudi tujejezične ustreznike. Delo se zaključi s temeljitim usklajevalnim pregledom, sledi pa izdaja v elektronski in/ali tiskani obliki. Po izidu je zaželeno dopolnjevanje in posodabljanje slovarja.

Predvideva se izdelava slovarjev tistih strok, za katere področni strokovnjaki izkažejo velik interes in katerih financiranje ni predvideno iz drugih virov.

Predvideni učinki:

- opis terminologije na posameznih strokovnih področjih
- poenotenje terminologije na posameznih strokovnih področjih
- večja dostopnost do strokovno relevantnih informacij o terminologiji

Nosilec:

MK⁹, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

Zasnova slovarjev	Število zasnov
-------------------	----------------

⁹ MK se vključuje v smislu spremljanja.

Izdelava specializiranih korpusov	Število izdelanih korpusov
Izdelava geslovníkov na podlagi korpusa	Število predlaganih terminov po posameznih strokah
Pisanje in urejanje slovarskih sestavkov	Število urejenih slovarskih sestavkov po posameznih strokah
Izdaja slovarjev v tiskani in elektronski obliki	Število objavljenih slovarjev

Akcija L-5: Zgodovinski slovar

Izpeljava aktivnosti za hitrejši in vsebinsko popolnejši začetek izhajanja Slovarja slovenskega knjižnega jezika 16. stoletja in vsebinska priprava korpusa besedil slovenskega knjižnega jezika 16. stoletja

Slovar slovenskega knjižnega jezika 16. stoletja je trenutno največji slovenski zgodovinskoslovaropisni projekt. Glede na relativno maloštevilno gradivo iz 17. ter 18. stoletja ter relativno dobro popisano gradivo 19. stoletja gre za slovar, ki bo temeljni zgodovinski slovar slovenskega jezika. Šele njegovo dokončanje bo omogočalo izdelavo priročnikov, kot je npr. Oxford English Dictionary, tudi za slovenski jezik.

Opis:

Dodatno financiranje dela na tem slovarju bi omogočilo temeljitejši premislek o slovarskem konceptu, izdelavo dodatnih raziskav, pomembnih za pripravo dokončnih gesel za prvi zvezek slovarja ter izdelavo priročnika za bodoče leksikografe. Ker bo slovar narejen z najnaprednejšo računalniško tehnologijo, bi dodatno financiranje ponudilo tudi rešitve za številna jezikovnotehnološka vprašanja, ki jih odpira izdelava tako velikega slovarja v današnji dobi. Obenem s prvim zvezkom slovarja bi bila tako lahko pripravljena tudi sodobna objava slovarja na spletu, ki bi bil v celoti odprto dostopen. Spletna verzija slovarja ne bi strogo sledila abecednemu redu priprave, kot ga zahteva knjižna izdaja, ampak bi na podlagi podatkov o iskanju vključevala tudi druga gesla in se bo v prihodnosti vsako leto dopolnjevala z novimi gesli. V okviru dodatnega financiranja bi se pripravila tudi vsebinska in jezikovnotehnološka izhodišča za pripravo korpusa slovenskega jezika 16. stoletja, za katerega je večina prepisov že pripravljenih. Korpus bi bil odprto dostopen.

Predvideni učinki:

- izdelane raziskave za razrešitev problematike, ki jo odpira začetek izhajanja slovarja
- dodatna vzorčna gesla
- priročnik za bodoče leksikografe
- spletna verzija slovarja
- korpus besedil slovenskega knjižnega jezika 16. stoletja z ustreznimi iskalniki ter z obliko, primerno za vgradnjo v druge korpus

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP ali raziskovalnega projekta

Podrobna opredelitev aktivnosti s kazalniki:

Izboljšava slovarskega koncepta na osnovi dodatnih raziskav	Da/ne
Izdelava priročnika za bodoče leksikografe	Da/ne
Spletna progresivna objava slovarja	Da/ne
Vsebinska in jezikovnotehnološka zasnova korpusa slovenskega knjižnega jezika 16. stoletja	Da/ne

Akcija L-6: Imenske entitete v slovenščini

Imenske entitete

Razpoznavanje imenskih entitet je ena temeljnih tehnologij predprocesiranja, uvrstimo jo lahko na sorodno raven kot postopek tokenizacije, in je potrebno tako rekoč na vseh ravneh procesiranja besedila, od oblikoslovnega in skladenjskega označevanja do semantične analize besedila in strojnega prevajanja. Razpoznavanje imenskih entitet je bilo za slovenščino šele pred kratkim osnovano v okviru projekta SSJ na podlagi korpusa ssj500k (del korpusa z ročno označenimi entitetami). Obstoječe orodje za avtomatsko razpoznavanje entitet obsega trenutno splošno uspešnost okrog 80%, od tega najslabšo za imena organizacij, manj kot 60%. Nadaljnja nadgradnja orodja je zato nujna. Poleg orodja je pomemben vir tudi leksikon imenskih entitet, ki med drugim omogoča nadaljnje izboljšave procesiranja imenskih entitet, razvoj orodij, specializiranih za analizo tujejezičnih imenskih entitet v besedilih, skupaj s fonetičnim zapisom je nujen vir za kvalitetno sintezo govora itd.

Opis:

Izdela se orodje za razpoznavanje imenskih entitet za splošno domeno. Čeprav so zaželeni tudi preskusi drugih pristopov (s pravili oz. hibridni pristopi) je priporočljiva uporaba predvsem pristopov osnovanih na statističnih metodah oz. metodah strojnega učenja, saj so ti dokazano področno manj omejeni in bolj primerni za splošno domeno. Vendar pa potrebujejo za indukcijo modela označeni korpus, zato sklop predvideva tudi izdelavo večjega in raznovrstnega korpusa z ročno označenimi imenskimi entitetami, ob tem pa razmislek o najboljšem naboru tipov imenskih entitet. Na podlagi korpusnih in drugih podatkov se izdelava seznam imenskih entitet, ki je osnova za leksikon imenskih entitet s podatki o vrsti, z oblikoslovnimi podatki in s fonetičnim zapisom. Za izvajanje akcije je nujno, da se povezuje z akcijami, v katerih se izdelujejo drugi korpusni in leksikalni viri za slovenščino (akcije K in L), tako da se skrbi za skladnost v formatu, v vsebini pa za to, da se podatki, kolikor je treba, lahko avtomatsko pretakajo med viri, po drugi strani pa ne podvajajo, kjer ni potrebno.

Predvideni učinki:

- izboljšana možnost luščenja informacij iz besedila
- izboljšano avtomatsko označevanje
- izboljšano procesiranje besedila v JT aplikacijah, kot so sintetizator govora, razpoznavnik tekočega govora, strojni prevajalnik, sistem dialoga itd.
- spodbujanje jezikoslovnih raziskav lastnih poimenovanj v slovenščini
- podpora standardizaciji jezika

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP ali raziskovalnega projekta

Podrobna opredelitev aktivnosti s kazalniki:

1 Učni korpus z označenimi imenskimi entitetami	da/ne
2 Leksikon imenskih entitet	da/ne
3 Orodje za označevanje imenskih entitet	da/ne

Akcija L-7: Večjezične baze in slovarji

Dvo- in večjezične baze in slovarji

Akcija naslavlja vrzel v jezikovni opremljenosti za večjezičnost, kajti nekoč je bila izdelava dvojezičnih slovarjev zanimiva in ekonomsko smiselna predvsem za založbe in knjigotržce, danes pa je ta dejavnost v celoti ukinjena in ni videti, da bi jo v slovenskem prostoru zapolnil kak drug akter.

V spodnji tabeli so podatki o obstoječih (pretežno tiskanih) virih, za slabše opremljene jezikovne pare pa si številni prevajalci raje pomagajo s sodobnejšimi in obsežnejšimi slovarji v kombinaciji z angleščino. Poleg naštetih se na spletu za nekatere jezikovne pare ponujajo tudi slovarji tujih (PONS) in neznanih založnikov, za katere podatki o obsegu in načinu izdelave niso na voljo. Zato je treba vlagati predvsem v zagotavljanje dostopnosti obstoječih dvojezičnih slovarjev prek odkupa avtorskih pravic, digitalizacije in distribucije, ter v nadgradnjo in posodabljanje slovarskih baz za slabše opremljene jezike, tudi z uporabo sodobnih jezikovnih tehnologij in izrabo vzporednih korpusov.

Jezik	Najsodobnejši obstoječi slovar	Leto izida/izdelave	Obseg
angleščina	Veliki angleško-slovenski slovar OXFORD®-DZS	2005/2006	117.500
	Slovensko-angleški slovar, DZS (Grad/Leeming)	1990	42.000
nemščina	Veliki nemško-slovenski slovar, DZS	2001	126.000
	Veliki slovensko-nemški slovar, DZS	2003	86.000
francoščina	Francosko-slovenski slovar, DZS	1975	40.000
	Slovensko-francoski slovar, DZS	1990	34.000
italijanščina	Veliki italijansko-slovenski slovar, DZS	1997	60.500
	Veliki slovensko-italijanski slovar, DZS	2006	90.500
španščina	Špansko-slovenski slovar, DZS	1969	46.500
	Slovensko-španski slovar, DZS	1979	34.000
hrvaščina	Veliki sodobni slovensko-hrvaški in hrvaško-slovenski slovar, Mozaik	2014	60.000
srbsščina	Mali srbsko-slovenski in slovensko-srbski slovar, DZS	2005	30.000
madžarščina	Slovensko-madžarski slovar, DZS	1996	40.000
	Madžarsko-slovenski slovar, Pomurska založba	1998	51.000

Opis:

Ker razpoložljiva sredstva ne zadoščajo, da bi s kakovostnimi dvojezičnimi slovarji premostili vse vrzeli in pokrili vse jezikovne pare, akcija predvideva financiranje enega dvojezičnega slovarja, ki je bil ugotovljen kot prioriteta, in sicer slovensko-angleškega slovarja. Preostala sredstva, namenjena zagotavljanju dvo- in večjezičnih splošnih slovarskih virov, se bodo deloma porabila za pridobivanje avtorskih pravic in digitalizacijo obstoječih zastarelih slovarjev, deloma pa za podporo projektom, kjer se obstoječi dvojezični slovar posodobi, nadgradi in dopolni z uporabo jezikovnotehnoloških metod ter množičenja.

Predvideni učinki:

- zagotovitev sodobnega slovensko-angleškega slovarja,
- zagotovitev dvo- in večjezičnih slovarjev za prevajalce, strokovnjake, jezikoslovce in druge uporabnike,

- zagotovitev temeljnih jezikovnih virov za strojni prevajalnik in druge večjezične aplikacije.

Nosilec:

MK

Časovni plan:

2015–2018

Podrobna opredelitev aktivnosti s kazalniki:

1 Gradnja slovensko-angleškega dvojezičnega slovarja	da/ne (število vnosov min. 40.000)
2 Zagotavljanje avtorskih pravic za obstoječe dvo- in večjezične slovarje	število slovarjev, za katere so bile pridobljene pravice
3 Nadgradnja ali gradnja dvo- ali večjezičnega slovarja za enega od prioriternih jezikovnih parov (AN-SL, NE-SL, FR-SL, IT-SL, ŠP-SL, HR-SL, HU-SL)	število novih vnosov v dvo- ali večjezičnem slovarju; min. 2.000

Akcija L-8: Strojna slovnična analiza

Oblikoslovna in skladenjska analiza besedila in oblikoslovni sintetizator

Oblikoslovni označevalnik in lematizator sodita med najbolj temeljna jezikovnotehnološka orodja ter omogočata in podpirata tako rekoč vse jezikovnotehnološke aplikacije. Oblikoslovni označevalniki in lematizatorji za slovenščino že obstajajo, povprečno dosegajo uspešnost nad 90 %. S tem se razvoj na področju oblikoslovja ne sme ustaviti: treba je iskati možnosti za nadaljnje izboljšanje, tako s stališča uspešnosti kot hitrosti, predvsem pa s prilagajanjem na različne registre (zlasti za širjenje na neformalni jezik – družabni splet – in govorjeni jezik). Kot posebna aplikacija se mora ponuditi tudi tokenizator, ki je sicer pogosto vključen kot del lematizatorja, v sklopu tokenizatorja oz. kot posebna aplikacija se izdelava tudi orodje, ki označuje meje med povedmi. Temeljni manko na področju oblikoslovja pa je oblikoslovni sintetizator, ki bi znal na podlagi leme in oblikoskladenjske oznake izpisati pravilno besedno obliko. Tako orodje je pomembno za tvorjenje besedila, zlasti v strojnem prevajanju in v sistemih dialoga.

Med podobno osnovna orodja danes sodi tudi skladenjski razčlenjevalnik, ki podpira postopke, kot so semantična analiza besedila, luščenje informacij, avtomatsko povzemanje, strojno prevajanje ipd. Skladenjski razčlenjevalnik za slovenščino sicer obstaja in je prosto dostopen, enako kot pri označevalniku pa velja, da je treba iskati možnosti za nadaljnje izboljšanje, tako s stališča uspešnosti kot hitrosti, predvsem pa s prilagajanjem na različne registre. Za govorjeni jezik se izdelava posebno orodje, ki segmentira besedilo govora na izjave.

Opis:

Izvajajo se nadaljnje temeljne in aplikativne raziskave na področju avtomatskega oblikoslovnega označevanja (predvsem v smeri dodatno ročno označenih korpusov, izboljšanja ekspertnih pravil, povečanja leksikona in izboljšanja časovne komponente). Označevalnik se širi na nove registre, predvsem na neformalna spletna besedila ter govorjeni jezik. S pomočjo oblikoslovnega slovarja in drugih dostopnih virov se izdelava oblikoslovni sintetizator. Nadaljujejo se temeljne in aplikativne raziskave na področju avtomatskega skladenjskega razčlenjevanja (npr. s povečanjem ročno označenega korpusa ali z aktivnim učenjem posameznih tipov povedi, besednih zvez ali povezav, s preskušanjem hibridnih pristopov k razčlenjevanju, uporabo podatkov iz drugih razpoložljivih virov, optimiranje časovne komponente razčlenjevalnika), razčlenjevalnik se širi na nove registre, predvsem na neformalna spletna besedila ter govorjeni jezik.

Predvideni učinki:

- izboljšana možnost luščenja informacij iz besedila
- izboljšano avtomatsko označevanje
- izboljšano procesiranje besedila v vseh naprednejših JT aplikacijah
- podpora zahtevnejšim korpusnojezičnim analizam za potrebe slovarja in slovnice
- spodbujanje jezikoslovnih raziskav
- podpora standardizaciji jezika

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP**Podrobna opredelitev aktivnosti s kazalniki:**

1 Tokenizator in označevalnik mej med povedmi v pisnem besedilu	da/ne
2 Razširitev korpusa z ročno označenimi oblikoslovnimi oznakami	da/ne
3 Prenovljena in izboljšana verzija oblikoslovnega označevalnika	da/ne
4 Prilagojena/e verzija/e oblikoslovnega označevalnika za posebne registre	da/ne
5 Oblikoslovni sintetizator	da/ne
6 Razširitev korpusa z ročno označenimi skladejskimi povezavami	da/ne
7 Prenovljena in izboljšana verzija skladejskega razčlenjevalnika	da/ne
8 Prilagojena/e verzija/e skladejskega razčlenjevalnika za posebne registre	da/ne

Akcija L-9: FrameNet za slovenščino

Semantična analiza vloga

Semantična analiza vlog je poleg semantičnih mrež temeljni postopek semantične analize besedila. Trenutno prepoznamo vsaj tri popularnejše načine za tovrstno analizo. PropBank pristopa k analizi na podlagi skladnje in je osredotočen na glagole. Izhaja iz tradicije Penn Treebanka in frazne gramatike. Označevanje samostalnikov v skladu s PropBankom se razvija kot poseben projekt – NomBank. Avtomatični postopki za označevanje za angleški jezik dosegajo uspešnost okrog 80 %. Semantična analiza v okviru FrameNeta temelji na kontekstnih vlogah. FrameNet omogoča označevanje tudi drugih besednih vrst, ne samo glagolov, čeprav se pogosteje označujejo predvsem glagoli. Angleški FrameNet obsega 6.100 označenih leksikalnih enot v 135.000 stavkih. Uspešnost orodij za avtomatsko označevanje po načelu FrameNeta za angleščino dosega povprečno vrednosti okrog 80%. Močno uveljavljena je tudi praška odvisnostna drevesnica, ki vključuje t. i. tektogramatično (tj. pomenskoskladenjsko) analizo, vendar je v drugih jezikih redkeje aplicirana kot PropBank ali FrameNet. Za slovenščino obstaja prvi osnutek FrameNeta, ki vsebuje 200 ročno semantično označenih povedi oz. 908 besed. Gre predvsem za individualni raziskovalni projekt. Drugi načini semantične analize vlog še niso bili aplicirani na slovenščino. Ker je FrameNet hkrati eden najbolj razširjenih načinov za semantično analizo in obstajajo orodja za avtomatsko označevanje za angleščino in še nekatere jezike, je najbolj smiselno nadaljevanje semantične analize vlog na podlagi FrameNeta.

Opis:

FrameNet ločuje med dvema različnima metodologijama, prvi je leksikografsko označevanje, prek katerega pridobivamo popolne opise vezljivostnih vzorcev, ki jih izbrane leksikalne enote izkazujejo, drugi je označevanje celih besedil, kjer vsaki potencialni leksikalni enoti (tj. glagolu, samostalniku, pridevniku, prislovu v točno določenem pomenu) določimo vse semantične (in skladenjske) argumente. V okviru aktivnosti se nadaljuje nadgradnja obstoječega FrameNeta po obeh načinih. Prav tako se izdelava orodje za semantično analizo vlog v besedilu.

Predvideni učinki:

- spodbujen razvoj naprednejših JT aplikacij za slovenščino
- izboljšano procesiranje besedil
- spodbujene jezikoslovne raziskave pomenske vloge besed in njihovih vezljivostnih vzorcev
- sestavni del slovnične podatkovne baze
- omogočitev izdelave jezikovnih priročnikov s podatki o besedah in njihovih vezljivostnih vzorcih

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Korpus z oznakami po principu FrameNet	da/ne
2 Leksikon vezljivostnih vzorcev po sistemu FrameNet	da/ne
3 Orodje za semantično analizo na podlagi FrameNeta	da/ne

Akcija L-10: sloWNet

Semantična mreža

Napredne aplikacije jezikovnih tehnologij, kot so luščenje informacij, odgovarjanje na vprašanja, avtomatsko povzemanje, analiza koreferenc, strojno prevajanje, sistem dialoga, potrebujejo za podporo tudi semantično analizo besedila. V tej smeri so bile v svetu razvite številne ontologije, semantične mreže, mreže vedenja ... ter korpusi z različnimi semantičnimi oznakami, prav tako pa tudi številne tehnologije za avtomatsko označevanje teh informacij. Semantične mreže so prva pomembnejša skupina tovrstnih jezikovnih virov in orodij. Vsebujejo inventar besed in relacij med njimi v smislu sopomenskosti znotraj posameznih sinsetov, med sinseti pa v smislu protipomenskosti, nad- in podpomenskosti ipd. So eden od ključnih elementov semantičnega opisa v jezikovnih tehnologijah, pomembne pa so seveda tudi za jezikoslovno raziskovanje jezika in se tesno povezujejo z raziskovanjem sinonimije. Najbolj razširjena semantična mreža v svetu je princetonski WordNet (vsebuje 117.000 sinsetov), ki je v zasnutku že apliciran tudi na slovenski jezik – sloWNet (20.000 sinsetov, le delno ročno pregledan). Na njegovi podlagi sta semantično označena tudi korpusa ssj500k in SPOOK. Za širšo uporabnost pa je nujna njegova ustrezna nadgradnja.

Opis:

Izvedejo se pregledovanje in popraviljanje obstoječega sloWNeta in njegova nadgradnja. Večji poudarek se nameni samostalnikom, tako da se v tem segmentu pokrije večina besedišča, ostale besedne vrste se delajo predvsem poskusno, da se razvije sistem označevanja (pri besednih vrstah, kot so glagoli, je namreč zaradi razlik z izhodiščnim angleškim jezikom treba iskati tudi jezikovno specifične rešitve). Prav tako se na podlagi WordNetovih shem razvije orodje za razdvoumljanje besednih pomenov (angl. word sense disambiguation).

Predvideni učinki:

- spodbujen razvoj naprednejših JT aplikacij za slovenščino
- izboljšano procesiranje besedil
- spodbujene jezikoslovne raziskave pomenskih razmerij med besedami
- sestavni del slovarske baze
- omogočitev izdelave jezikovnih priročnikov s podatki o pomenskimi razmerji med besedami

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Popravljen in razširjen sloWNet	da/ne
2 Orodje za ročno pregledovanje korpusa s semantičnimi oznakami	da/ne
3 Ročno pregledan učni korpus s semantičnimi oznakami	da/ne
4 Orodje za razdvoumljanje besednih pomenov	da/ne

Akcija L-11: Anafora in koreference

Razvozlavanje anafore in analiza besedilnih koreferenc

Analiza besedilnih koreferenc pomeni analizo enot besedila, ki se nanašajo na istega referenta (npr. isto osebo). Obstajajo trije temeljni načini izražanja besedilnih koreferenc: prek anafore, tj. z zaobljiko – npr. zaimkom – se nanašamo na entiteto, ki je bila predhodno omenjena v besedilu; prek katafore, tj. z zaobljiko, se nanašamo na entiteto, ki jo bomo v besedilu imenovali šele v nadaljevanju; prek eksofore, tj. z zaobljiko, se nanašamo na entiteto, ki ni omenjena v besedilu, je pa prisotna v kontekstu komunikacijske situacije. Med temi je bistvena anaforična analiza besedila, saj je to najpogostejši način koreferiranja. Prvi korak k uspešni analizi besedilnih koreferenc je tako razvozlavanje anaforičnih navezovanj. Razvozlavanje anafore in analiza besedilnih koreferenc je eden od pomembnih zalednih postopkov za kompleksnejšo semantično analizo besedila. Anaforična in koreferenčna analiza besedila v slovenskem jezikovnotehnološkem prostoru še ni širše podprta, poteka pa v povezavi z oblikoslovnim in sintaktičnim označevanjem besedila, ki je za slovenščino že aktivno. Tesno povezano je tudi s semantično in pragmatično ravno besedila, tako da se pričakujejo temeljnoraziskovalni nastavki za višje ravni procesiranja besedila na ravni semantike, konteksta in pragmatike. Tudi v mednarodnem prostoru je koreferenčna in anaforična analiza besedila še vedno zelo živo in aktualno temeljnoraziskovalno področje.

Opis:

Znanstvena poglobitev v problematiko analize besedil na ravni anafore in koreference s sodelovanjem jezikoslovcev in inženirjev ter definiranje specifikacij za označevanja anaforičnih povezav in koreference v besedilu. V praktičnem delu je potrebno ročno označevanje besedila, temu sledi učenje strojnih postopkov za avtomatsko prepoznavanje anaforičnih povezav in označevanje koreference. Rezultati so izdelane specifikacije, učni jezikovni viri ter orodje za anaforično analizo besedila in označevanje koreference, prav tako pa tudi izvirni znanstveni članki s tega področja.

Predvideni učinki:

- spodbujen razvoj naprednejših JT aplikacij za slovenščino
- izboljšano procesiranje besedil
- spodbujene jezikoslovne in diskurzne raziskave znotrajbesedilnega navezovanja
- sestavni del slovnične podatkovne baze

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2016-2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Specifikacije za označevanje anafore in koreferenc	da/ne
2 Učni korpus z označenimi anaforičnimi povezavami in koreferencami	da/ne
3 Orodje za označevanje anafore in koreferenc	da/ne

Akcija L-12: Zbirka normativnih zadreg

Zbirka normativnih zadreg

Z zbirko normativnih zadreg se vzpostavlja gradivska osnova za pripravo normativno specializiranega slovarja, ki ga na Slovenskem tradicionalno poznamo pod imenom pravopisni slovar, in izhajajočih poljudnih normativnih priročnikov za strokovno manj radovedne/zahtevne jezikovne uporabnike. Zbirka normativnih zadreg ponuja verodostojno gradivsko osnovo, z različnimi empirično preverljivimi postopki omogočajo upoštevanje uporabnikovih jezikovnih zadreg in posledično njihovih potreb. Na osnovi tako pridobljenih podatkov je mogoče zasnovati glede obvestilnosti uporabniku prilagojene normativne priročnike, knjižne in spletne.

Opis:

Akcija **L-12** se neposredno navezuje na akcijo **N-1**, saj se z zbirko normativnih zadreg vzpostavlja gradivska osnova za pripravo pravopisnega slovarja.

Geslovnik zbirke je zasnovan problemsko, podatke zagotavlja iz različnih virov, in sicer: **1.** iz že ugotovljenih normativnih zadreg jezikovnih uporabnikov (vprašanja, zastavljena po spletnih forumih in jezikovnih svetovalnicah), ki jih je treba dopolniti z novimi jezikovnimi vprašanji, **2.** iz dvojnic, ki jih je mogoče izluščiti iz oblikoslovno označenih referenčnih korpusov (pisne, oblikoslovne, skladišne dvojnice), **3.** iz nadgrajenih specializiranih korpusov za normativna vprašanja (Šolar, Lektor) (povezava z akcijo **K-2**), **4.** iz poizvedb po že spletno objavljenih slovarjih, **5.** iz izrecno normativno ovrednotenih podatkov v aktualnih in starejših priročnikih (za obdobje do izida SSKJ) z normativnim značajem. Navedeni podatki morajo biti v zbirki metajezikovno dokumentirani in povezani s podatki v sodobnem referenčnem korpusu sinhrono slovenščine.

Predvideni učinki:

- vzpostavitev pogojev za prenovo pravopisnih pravil s pomočjo induktivnih in empiričnih metod
- pridobitev osrednje zbirke normativnih zadreg jezikovnih uporabnikov, iz katere izhajajo teoretična dela in aplikacije
- standardizacija zbirke normativnih zadreg z vidika pravopisa, pravorečja, oblikoslovja

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Konceptualna zasnova zbirke	da/ne
2 Zbirka normativnih zadreg	število enot zbirke po problemskih sklopih
3 Povezava zbirke z referenčnim korpusom	da/ne

Akcija G-1: Zasnova slovnice

Koncept nove slovnice slovenskega jezika in slovnična podatkovna zbirka

Za temeljni slovnični opis velja Toporišičeva Slovenska slovnica (aktualna je njena četrta izdaja iz leta 2004, prva izdaja je iz leta 1976). V zadnjih dveh ali treh desetletjih je v jezikoslovju opazen premik znanstvene paradigme iz raziskovanja jezikovnega sistema, kakršen je bil značilen predvsem za strukturalizem, v celostno in empirično naravnano obravnavo jezika, ki skuša v enoten sistem zajeti delovanje jezika v realnih okoliščinah, v povezavi s področji, kot so psihologija, nevrobiologija, umetna inteligenca itd. Treba je oblikovati interdisciplinarno skupino strokovnjakov, ki bo na osnovi dosedanjih spoznanj, ob upoštevanju novejših podatkov in uporabi sodobnih orodij sestavila koncept za novo slovnico slovenskega jezika.

Dostopnost podrobnih statističnih jezikovnih podatkov je nujni predpogoj za začetek dela na novem slovničnem opisu slovenskega jezika, ki bo v bistveno večji meri kot do sedaj izhajal iz jezikovne realnosti. Pri tem se je smiselno zgledovati po projektih, ki so pri drugih jezikih omogočili t. i. komunikativni slovnični opis, kot je bil denimo *The Survey of English Usage* v 60. letih prejšnjega stoletja ali *COBUILD* v 80. letih, pri angleščini.

Opis:

Izdela se koncept novega slovničnega opisa sodobne knjižne/standardne slovenščine na podlagi podatkov o sodobni slovenščini. Izdela se zbirko podatkov o rabi sodobne slovenščine, primerno za izdelavo nove slovenske slovnice.

Predvideni učinki:

- izboljššan slovnični opis sodobnega slovenskega jezika
- izboljššan dostop do osnovnih virov za analizo sodobnega slovenskega jezika
- izboljššan slovarski opis sodobnega slovenskega jezika
- izboljšane možnosti za razvoj jezikovnotehnoloških aplikacij sodobno slovenščino

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP ali raziskovalnega projekta

Podrobna opredelitev aktivnosti s kazalniki:

1 Koncept novega slovničnega opisa sodobne slovenščine	da/ne
2 Zbirka podatkov o rabi sodobne slovenščine za novi slovnični opis	da/ne

Akcija G-2: Šolska slovnica

Šolska slovnica

Šolska slovnica je namenjena jezikovnemu opismenjevanju učencev v slovenskih šolah. Zasnovana mora biti tako, da jo je mogoče uporabljati tako individualno kot tudi v sklopu pouka slovenščine. Upoštevati mora različne zmožnosti abstraktnega mišljenja v razvoju učeče se populacije ter nove možnosti tako pri predstavitvi slovničnih informacij v digitalnem mediju kot tudi pri zbiranju slovničnih podatkov v slovnični podatkovni zbirki. Šolska slovnica ne nadomešča učbenikov za slovenski jezik, zasnovana mora biti kot pedagoška aplikacija novega slovničnega opisa slovenščine.

Opis:

Izdela se koncept šolske slovnice, ki je namenjen jezik(osl)ovnemu opismenjevanju učencev v slovenskih šolah. Slovnica je izvirno zasnovana za rabo v digitalnem mediju, z možnostjo tiskanja posameznih elementov. Na podlagi slovnice se nadgradi obstoječi slovnični portal, ki je prosto dostopen na spletu.

Predvideni učinki:

- izboljššan šolski slovnični opis sodobnega slovenskega jezika
- izboljšana pismenost šolske populacije

Nosilec:

MK

Časovni plan:

2015–2018

Podrobna opredelitev aktivnosti s kazalniki:

1 Izdelan koncept šolske slovnice	da/ne
2 Izdelana šolska slovnica	da/ne

Akcija N-1: Pravopisna pravila in slovar

Predstavitev normativnih podatkov: pravila in slovar (Cilj: Pravopisni portal)

Namen *Pravopisnega portala* je vzpostavitev spletnega mesta, kjer so zbrani in s pomočjo sistema za pregledovanje spletnih vsebin organizirano predstavljeni normativni podatki kot rezultat delovanja standardizacijskega telesa, tj. predstavljena so pravopisna pravila po problemskih sklopih (pisavoslovje, pravopis – mala/velika začetnica, prevzemanje, ločila, pisanje skupaj in narazen, pripadajoča poglavja iz glasoslovja, oblikoslovja in besedotvorja) s povezavo na slovarsko predstavitev normativne informacije (kot rezultat pričujoče akcije) in obratno.

Problemski sklopi izhajajo iz analize normativnih zadreg jezikovnih uporabnikov (dvojnice v jezikovni rabi, poizvedbe po korpusih, specializiranih za normativna vprašanja). Slovarsko reševanje pravopisne problematike poteka vzporedno s prenovo pravopisnih pravil. Za manj zahtevne uporabnike je predvidena zbirka preprosteje opisno pojasnjenih pravopisnih zadreg (*Spletni priročnik za splošne uporabnike*), ki izhaja tudi iz svetovalnih stikov z jezikovnimi uporabniki (*Jezikovno svetovanje*). Ti so povezani tudi s težavami pri razumevanju slovarskega metajezika, pogosto razkrijejo, da je tolmačenje jeziko(slo)vnih podatkov zaradi nepoznavanja načina posredovanja informacij lahko povsem zgrešeno.

Opis:

Cilj akcije je predstavitev povezanih normativnih podatkov, tj. problemskih sklopov pravopisnih pravil in pripadajočega nabora slovarskih sestavkov na *Pravopisnem portalu*. Slovarska makrostruktura je oblikovana za vsak problemski sklop posebej, mikrostruktura vključuje tako tipične zglede kot izjeme in posebnosti, slednje s komentarjem; slovarska ideja sledi normativni obvestilnosti tudi na opisni ravni (normativno orientirana kvalifikatorska pojasnila). Izhodišče za aktivnost je široko zasnovana zbirka normativnih zadreg, vzpostavljena z akcijo **L-12**.

Akcija **N-1** je torej neposredno povezana z delovanjem standardizacijskega telesa ter prenovo kodifikacije) in tehnično odvisna od akcije **P-8 (spletna predstavitev normativnih podatkov)**, ki vzpostavlja Pravopisni portal z možnostjo zajemanja obeh sklopov. V sklopu cilja *Pravopisni portal* se kronološko umešča za akcijo **N-1**: posamezni problemski sklopi prenovljenih pravopisnih pravil in pripadajočega slovarja naj bi bili konec leta 2016 s pomočjo sistema že dodani na spletno mesto *Pravopisni portal*. Do leta 2018, ko poteka prenova pravopisnih pravil in akcija **N-1**, pa naj bi vzpostavljeni sistem (akcija **P-8**) omogočal pregledovanje več sto problemskih sklopov.

Predvideni učinki:

- uporabniku prijazen portal kot enotna vstopna točka ter izhodišče za povezave pravopisnih pravil in pripadajoče slovarske zbirke
- omogočen vpogled v aplikativno-utemeljevalno zasnovo pravopisa: pravila -- slovar
- omogočena usmerjena problemska razprava in takojšnja povratna informacija
- boljši dostop do normativnih informacij

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Koncept spletne predstavitve pravopisnih pravil po problemskih sklopih	da/ne
2 Koncept spletne predstavitve slovarja po problemskih sklopih	da/ne
3 Pravopisni slovar po problemskih sklopih	število slovarskih sestavkov
4 Povezave pravil in slovarja	da/ne

Akcija D-1: Digitalizacija čistopisov pisne kulturne dediščine

Izdelava čistopisov starejših besedil z uporabo množičenja

V Sloveniji obstaja že vrsta digitalnih knjižnic, ki omogočajo dostop do besedil naše pisne kulturne dediščine. Osrednja zbirka je digitalna knjižnica Slovenije dLib.si, ki pa za starejše slovensko slovstvo ponuja le skenograme oziroma iz njih avtomatsko zajeta besedila, vendar imajo ta, zaradi starih izvirnikov veliko napačno optično prepoznanih delov. Korekcije takih prepisov so zamuden, vendar potreben del izdelave kakovostnih čistopisov starejših slovenskih besedil, ki jih je nato možno digitalno umestiti.

V svetu kot tudi pri nas se je kot model izdelave poceni, a kvalitetnih jezikovnih virov uveljavilo množičenje (ang. crowdsourcing), kjer pri delu preko interneta sodeluje veliko število posameznikov, z Wikipedijo kot najbolj znanim primerom. V Sloveniji imamo tudi že primer dobre prakse na področju takšne izdelave jezikovnih virov, in sicer projekt »Slovenska leposlovna klasika«, kjer na Wikiviru študentje že vrsto let opravljajo redakcijo avtomatsko izdelanih transkripcij. Ta bogata zbirka je popolnoma prosto dostopna, kdorkoli in kadarkoli pa lahko tudi naknadno popravi napake v čistopisih.

Opis:

Ukrep predvideva redne letne razpise, ki imajo za cilj zagotovitev oz. izdelavo digitalnih faksimilov del slovenske pisne kulturne dediščine in predvsem izdelavo čistopisov z uporabo množičenja. Faksimili in čistopisi naj bodo prosto in odprto dostopni preko spleta. Za pomembnejša dela je poleg enostavnega prepisa dodatno zaželeno vključiti elemente tekstnokritičnih izdaj, predvsem izdelavo tako diplomatičnega kot kritičnega prepisa.

Projektne prijave naj bi vsebovale naslednje elemente:

- izbira del glede na pomembnost za zgodovino Slovenije, slovenskega slovstva in jezika z upoštevanjem avtorskih pravic na izvirnih besedilih
- zagotovitev oz. izdelava digitalnih faksimilov in avtomatskih (OCR) prepisov
- izbira platforme za množičenje, kjer imajo prednost že izdelane odprtokodne rešitve
- organizacija in izvedba razpisa in usposabljanja popraviljalcev
- vsebinsko in finančno izvajanje in spremljanje dela
- zagotovitev dostopa in prevzema v standardnem formatu na trajnem repozitoriju.

Predvideni učinki:

- boljša dostopnost del slovenske pisne kulturne dediščine

Nosilec:

MK

Časovni plan:

2015–2018 (letni razpisi za izvajanje dejavnosti)

Podrobna opredelitev aktivnosti s kazalniki:

1 Izdelava čistopisov z uporabo množičenja	da/ne
--	-------

Akcija D-2: Digitalizacija jezikovnih priročnikov

Digitalizacija in prilagajanje digitalnemu okolju jezikovnih priročnikov za slovenščino

Mnogo ključnih jezikovnih priročnikov za slovenski jezik, predvsem slovarjev, je dostopnih samo za branje, bodisi na spletu v obliki PDF ali pa celo samo v knjižni obliki. Digitalizacija oz. prilagajanje digitalnemu okolju obstoječih jezikovnih opisov, ki jih tradicionalno dojemamo v knjižni obliki, in zagotavljanje njihovega prostega in odprtega dostopa na spletu sta nujna, saj se bodo javna sredstva vložena v njihovo izdelavo šele tako polno osmislila v kontekstu informacijske družbe. S prehodom v digitalno okolje postanejo priročniki bistveno bolj dostopni in jih je lažje uporabljati, služili pa bodo lahko ne samo kot referenčni viri človeškim uporabnikom, dostopni na različnih digitalnih platformah, temveč tudi kot jezikovni (leksikalni) viri za jezikovnotehnološke raziskave in aplikacije.

Cilj ukrepa je torej spodbujanje digitalizacije, kvalitetnega zapisa in omogočanje prostega in odprtega dostopa za obstoječe jezikovne priročnike slovenskega jezika.

Opis:

V okviru akcije je predvidena digitalizacija obstoječih jezikovnih priročnikov za slovenščino, pri čemer je poudarek na slovarskih publikacijah. Akcija ne izključuje digitalizacije priročnikov, ki trenutno obstajajo samo v knjižni obliki, vendar imajo prednost dela, za katere že obstaja kakovosten digitalni vir, saj to bistveno poceni izdelavo referenčnega priročnika. Izdelani digitalni referenčni priročniki morajo biti čim bolj kvalitetni (brez napak in da čim bolj verno odslikavajo informacije iz originala) in pomensko označeni z upoštevanjem mednarodnih standardov za zapis besedilnih virov. Ključno je, da so izdelani priročniki prosto dostopni na spletu, poleg tega pa tudi odprto dostopni za prevzem. Sredstva namenjena za to akcijo lahko, poleg stroškov dela, zajemajo tudi odkup avtorskih pravic.

Prevzem virov naj ima čim manj omejitev (kot npr. prepoved predelav ali komercialne uporabe), pri čemer je zaželena licenca CC-BY 4.0 (priznanje avtorstva). Za zagotovitev trajnosti dostopa naj bodo izdelani digitalni priročniki v izvornem XML deponirani v repozitoriju slovenske raziskovalne infrastrukture CLARIN.

Predvideni učinki:

- Povečanje števila odprto dostopnih digitalnih priročnikov za slovenščino
- Olajšana zmožnost pisne komunikacije v slovenščini celotni populaciji
- Zagotovitev odprtih leksikalnih jezikovnih virov za uporabo v jezikovnotehnoloških raziskavah in aplikacijah

Nosilec:

MK, MIZŠ

Časovni plan:

2015–2018 (razpisi za digitalizacijo posameznih del ali zbirk del: 4 x 30.000 EUR)

Podrobna opredelitev aktivnosti s kazalniki:

1 Število novih prosto dostopnih jezikovnih priročnikov za slovenščino	da/ne
2 Število novih odprto dostopnih in standardno označenih jezikovnih priročnikov za slovenščino	da/ne

Akcija A-1: Razpoznavalec tekočega govora

Razpoznavalec tekočega govora z velikim slovarjem besed in razpoznavalec govora za šolsko okolje

Razpoznavanje govora je po eni strani nujen sestavni del kompleksnejših aplikacij, kot sta strojno prevajanje govora ali sistemi dialoga, hkrati pa je z večanjem obsega medijske produkcije tudi široko uporabno kot samostojni modul (npr. za avtomatsko indeksiranje). Potencialno je tudi zelo pomemben pripomoček za gluhe in naglušne osebe in njihovo lažje vključevanje v družbo. Ti imajo največje težave v procesu šolanja, saj lahko do učnih vsebin dostopajo samo prek pisnega kanala ali znakovnega tolmačenja, vendar je slednje zaradi stroškov tako rekoč neizvedljivo, tako da lahko vsebini pouka ali predavanj praviloma sledijo le prek zapiskov. Idealna rešitev bi bil avtomatski znakovni tolmač, predstopnja take aplikacije pa je razpoznavalec tekočega govora z velikim slovarjem besed, prilagojen za šolsko domeno. Obstoječi razpoznavalniki govora za slovenščino niso prosto dostopni in so večinoma omejeni na razpoznavanje izoliranih besed oziroma pri tekočem govoru ne dosegajo dovolj dobre natančnosti razpoznavanja. Določen preboj na področju tehnologije za tekoči govor z velikim slovarjem besed bo omogočila večnamenska govorna baza, predvidena v akciji K-3, s katero se ta akcija tesno povezuje. Nepodprto pa bo še vedno ostalo področje namenskih algoritmov za razpoznavanje govora v visoko pregibnih jezikih z relativno prostim besednim redom, kjer bodo še vedno potrebne nadaljnje temeljne raziskave, na podlagi katerih lahko pričakujemo dodatno izboljšanje rezultatov.

Opis:

Nadaljujejo se temeljne raziskave na področju razpoznavanja tekočega slovenskega govora z velikim slovarjem besed. Izdelava razpoznavalec tekočega govora z velikim slovarjem besed, ki deluje kot prosto dostopna internetna storitev za nekomercialno rabo. Razpoznavalec govora v obliki internetne storitve omogoča njegovo uporabo v največjem možnem številu scenarijev, ne glede na uporabniško strojno opremo (osebni računalnik, pametni telefon, tablica, STB). Uporabniki lahko preko posebej definiranega protokola oz. domače strani pošljejo na razpoznavalec govora svoje zvočne posnetke v primernem zvočnem formatu, po opravljenem procesiranju govora pa dobijo kot rezultat razpoznano besedilo. Za delovanje internetne storitve oz. domače strani razpoznavalnika tekočega govora skrbi nacionalni center za jezikovne tehnologije v sodelovanju z razvijalci sistema. Za potrebe gluhih in naglušnih, zlasti na področju šolanja, se posebna različica razpoznavalnika prilagodi na domeno šolanja (robustno procesiranje govora, adaptacija, jezik) ter prav tako izvede kot internetna storitev. Pri tem se pričakuje, da bodo v okviru prilagoditve izvedene tudi potrebne temeljne raziskave razpoznavanja govora za specifično domeno izobraževanja. Dolgoročno se zasleduje cilj izdelave razpoznavalnika, ki bo za izbrano domeno razpoznaval govor v realnem času.

V okviru akcije se mora tudi prizadevati za vzpostavitev dogovora med zakonodajnimi organi, pisarno informacijske pooblaščenke, zvezo gluhih in naglušnih, nacionalnim centrom za jezikovne tehnologije ter osnovnimi in srednjimi šolami in visokošolskimi ustanovami, ki bo gluhim in naglušnim dovoljeval snemanje govora učiteljev, predavateljev in drugih izvajalcev šolskega procesa med šolsko uro, seminarjem, tečajem, vajami, predavanjem ipd., v namene kasnejše tehnične obdelave posnetka, s končnim ciljem transkribiranja posnetka, na tak način, da se po eni strani ustrezno varujejo osebni podatki na posnetku in avtorske pravice, po drugi strani pa se gluhim in naglušnim omogoči pretvorba posnetka v transkripcijo, jezikovnotehnoški skupnosti pa uporaba posnetka za nadaljnji razvoj tehnologije in opis jezika. Slednje je ključnega pomena za nadaljnji uspešen razvoj sistema.

Predvideni učinki:

- spodbujanje gospodarskega razvoja na področju IT produktov z vključenim razpoznavanjem tekočega slovenskega govora (npr. avtomatsko podnaslavljanje, narekovalniki ipd.)
- olajšano sledenje učnemu procesu v šolah in na fakultetah za gluhe in naglušne osebe

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2016–2018

Obseg financiranja: možnost izvedbe preko CRP**Podrobna opredelitev aktivnosti s kazalniki:**

1 Jezikovni model za slovenski jezik	da/ne
2 Razpoznavalnik tekočega govora z velikim slovarjem besed, dostopen kot internetna storitev in preko domače strani, povprečen pričakovani delež napak manjši od 30 % na medijskem govoru	da/ne
3 Razpoznavalnik tekočega govora, prilagojen za potrebe gluhih in naglušnih pri šolanju, dostopen kot internetna storitev in preko domače strani	da/ne
4 Pravilnik o snemanju govora v izobraževalnem procesu	da/ne
5 Analiza problema in dolgoročni načrt izdelave razpoznavalnika govora z velikim slovarjem besed v realnem času za domeno izobraževanja	da/ne

Akcija A-2: Strojni prevajalnik

Razvoj strojnega prevajalnika

Področje jezikovnih tehnologij, ki ima potencialno največjo uporabno vrednost in največji potencialni vpliv na družbo, je strojno prevajanje. Gre tudi za eno prvih področij jezikovnih tehnologij, njegovi začetki segajo takoj v obdobje po koncu druge svetovne vojne. Z nekaj periodičnimi upadi zanimanja se vse od takrat aktivno razvija in tehnologija je do danes dosegla stopnjo, ko so strojni prevodi že uporabni ali za temeljno razumevanje pomena besedila ali kot podlaga prevajalcu za hitrejše prevajanje besedila. Strojno prevajanje je za zdaj tudi eno redkih področij jezikovnih tehnologij, ki jih (poleg črkovalnika in delilnika besed) za slovenski jezik pokrivajo tudi mednarodne korporacije – Googlov prevajalnik in Microsoftov Bing namreč vključujeta tudi slovenski jezik. Čeprav se zdi to na videz razbremenitev nacionalnih naporov za razvoj na tem področju, ne smemo spregledati nevarnosti, ki lahko sledi iz posedovanja tovrstne tehnologije izključno v rokah zasebnih mednarodnih korporacij, kar je zaznala tudi Evropska unija in reagirala med drugim s financiranjem projekta Moses, s katerim je ponudila močno, tudi za komercialne potrebe prosto dostopno orodje za razvoj strojnega prevajanja.

Opis:

Temeljne raziskave na področju strojnega prevajanja in izdelava novega strojnega prevajalnika (predvidoma na platformi Moses) za slovenščino v povezavi z angleščino ali dodatno tudi za druge aktualne jezikovne pare. Prevajalnik se ponudi v brezplačno neomejeno uporabo prek internetnega strežnika. Ob tem se upoštevajo možnosti, ki jih na tem področju že ponuja Evropska komisija: na Generalnem direktoratu za prevajanje preko IKT tehnologij omogočajo strojno prevajanje besedil in dokumentov. Smiselno bi se bilo povezati za namene skupnega razvoja strojnega prevajanja.

http://ec.europa.eu/dgs/translation/translationresources/machine_translation/index_en.htm

Predvideni učinki:

- hitrejši in cenejši dostop do tujejezičnih vsebin
- učinkovitejša medkulturna komunikacija
- neodvisnost tehnološkega razvoja, povezanega s strojnim prevajanjem, od zasebnih mednarodnih korporacij
- spodbujanje gospodarskega razvoja na področju IT-produktov s strojnim prevajanjem

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan:

2016–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Razvoj AN-SL in SL-AN strojnega prevajalnika in evalvacija	uspešnost prevajalnika glede na obstoječe sisteme
2 Razvoj strojnih prevajalnikov za druge jezikovne pare s slovenščino	da/ne

Akcija A-3: Sistem dialoga

Sistem dialoga

Sistem dialoga pomeni tehnologijo, ki je v ozadju katerekoli aplikacije, v kateri lahko uporabnik komunicira z računalnikom v naravnem jeziku. Razvrstimo jih lahko glede na različne kriterije, in sicer:

- glede na vrsto naloge: sistemi, ki samo podajajo informacije (npr. o gledališčih in njihovem programu, restavracijah, letalskem in železniškem prometu, vremenski napovedi itd.), in sisteme, ki poleg tega tudi rešujejo probleme (npr. iščejo najprimernejšo množico povezav, recimo v železniškem potniškem prometu najboljše povezave, odpravljajo napake, recimo v električnem omrežju, itd.),
- glede na modalnost: omogoča govorjeni ali pisni dialog oz. multimodalni dialog (vključuje npr. govor in sliko),
- glede na napravo, v katero je vgrajen: mobilni telefon, avtomobil, robot, prenosni računalnik, brskalnik, virtualno okolje, gospodinjski aparat ...,
- glede na vrsto aplikacijo, v katero je vključen: informacijske storitve, upravljanje, zabava, izobraževanje, opomnik, zdravstveno varstvo, varstvo starejših ...

Ideja o tem, da bi naredili računalnik, ki bi se pogovarjal s človekom, sega v leto 1950, ko je Alan Turing objavil znameniti članek *Computing Machinery and Intelligence*, v katerem je zapisal, da bo v naslednjih 50 letih računalnik opravil t. i. Turingov test in s tem dokazal, da je inteligenčen.

V slovenščini je verjetno najbolj kompleksen sistem dialoga SecondEGO, ki vključuje tehnologijo Piflar in odgovarja na vprašanja (t. i. question-answering system), deluje pa preko pisnega kanala. Obstaja (ali je obstajalo) pa še nekaj drugih, podobnih sistemov, npr. sistem Sara na FERI UM, ki ima tudi multimodalno komponento, sistem Vida pri DURS, razvitih pa je bilo tudi nekaj predvsem eksperimentalnih akademskih sistemov.

Osnovni gradniki sistema dialoga, ki omogoča komunikacijo z računalnikom v naravnem jeziku, so naslednji:

- Razpoznavanje govora: Če poteka komunikacija govorno, računalnik v prvem koraku zapiše govor v besedilo.
- Semantična analiza: V zahtevnejših komunikacijah moramo nato semantično analizirati razpoznano besedilo, da ugotovimo, kaj uporabnik želi od računalnika.
- Vodenje dialoga: Ta modul je srce sistema dialoga in upravlja celoten potek interakcije z uporabnikom in reakcije sistema. Vzdržuje zgodovino dialoga, določa strategijo vodenja dialoga, ima povezavo s podatkovnimi bazami v ozadju, iz katerih črpa informacije in vsebine ...
- Tvorjenje odgovora in sinteza govora: Ko ima računalnik vse informacije, ki jih potrebuje, tvori odgovor na uporabnikovo zahtevo v obliki besedila. Če je interakcija govorna, bomo odgovor računalnika še sintetizirali v govor.

Sistem dialoga je kompleksna tehnologija, ki se že dolgo razvija in v prihodnosti lahko pričakujemo njeno integracijo vsaj v nekaterih aplikacijah, najprej verjetno na mobilnih telefonih (tak aktualen primer iz tujine je Siri).

Opis:

Izvedejo se raziskave, ki ustrezno celostno predstavijo problematiko področja, izpostavijo najbolj problematične točke in pregledajo tehnične rešitve. Izdela se prototip sistema dialoga, apliciran v ustrezno uporabniško aplikacijo. Zaželeno so tudi temeljne raziskave, ki predvidijo tehnične rešitve, zlasti take, ki so vezane na posebnosti slovenskega jezika.

Predvideni učinki:

- spodbujen razvoj tehnologije sistema dialoga za slovenski jezik
- spodbujanje nadgradnje tehnologij z aplikacijami, ki temeljijo na sistemu dialoga

Nosilec:

MK, MIZŠ (ARRS), MDDSZ

Časovni plan:

2016–2018

Obseg financiranja: možnost izvedbe preko CRP**Podrobna opredelitev aktivnosti s kazalniki:**

1 Javno projektno poročilo s pregledom področja	da/ne
2 Pregledne in izvirne znanstvene objave	min. 2
3 Izdelan prototip sistema dialoga	da/ne

Akcija A-4: Bralnik sporočil gospodinjskih aparatov

Bralnik sporočil gospodinjskih aparatov

Možnost upravljanja gospodinjskih aparatov je ključna za večjo samostojnost slepih in slabovidnih. Sodobni aparati (tehtnice, pralni in pomivalni stroji, štedilniki itd.) pogosto prikazujejo informacije v obliki števil (ali tudi besedila) na prikazovalnik, ki ga slepi in slabovidni ne morejo ali ga težko preberejo. Mobilna aplikacija, ki informacijo (vsaj številčno) prepozna in prebere, prinaša veliko dodano vrednost za uporabnike. Pričakuje se, da bo dodatno potrebno izvesti temeljne raziskave, kako s postopki digitalnega procesiranja signalov robustno prepoznati informacije na prikazovalnikih gospodinjskih aparatov.

Opis:

Izdela se aplikacija za mobilni telefon, ki prek fotoaparata prepozna informacijo na prikazovalnikih gospodinjskih aparatov, nato pa informacijo posreduje v obliki govornega sporočila.

Predvideni učinki:

- olajšano upravljanje sodobnih gospodinjskih aparatov za slepe in slabovidne

Nosilec:

MK, MIZŠ (ARRS), MDDSZ

Časovni plan:

2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Aplikacija za mobilni telefon za branje informacij z gospodinjskih aparatov	da/ne
---	-------

Akcija A-5: Referenčni črkovalnik in slovnični pregledovalnik

Izdelava referenčnega črkovalnika in slovničnega pregledovalnika za slovenščino

Črkovalniki, lokalizirani za posamezne jezike, so sedaj postali že del standardne opreme predvsem, vendar ne samo odprtokodnih programov, kot so urejevalniki besedil in spletni brskalniki. Za preverjanje pravilnega zapisa slovenskih besed pa morajo ti črkovalniki vsebovati referenčen seznam slovenskih besed in ponavadi tudi vzorce za njihovo pregibanje. Bolj znani črkovalniki, kot npr. ispell, aspell, myspell in hunspell sicer podpirajo slovenščino, vendar trenutni sezname slovenski besed niso dovolj veliki, pomanjkljivi pa so tudi pregibni vzorci. V zadnjem času pa se pojavljajo tudi odprtokodni programi za preverjanje skladnje povedi, kjer slovenski jezik izrazito zaostaja, saj zaenkrat obstaja preverjanje skladnje (vključno s pravilno postavitvijo vejic) samo v okviru komercialnega produkta Besana.

Raziskave potreb jezikovnih uporabnikov so pokazale, da je slednjim prikaz pravopisnih dejstev, integriranih v orodja, ki ga najpogosteje spremljajo pri njegovih pisnih dejavnostih, vse bolj dobrodošel pripomoček. V perspektivi je treba pričakovati, da se bo uporabnik še redkeje zazrl v slovar, saj se bo zanesel na tisti »standardizacijski priročnik«, ki jim ga ponuja Microsoft v črkovalniku ali pa iskalnik Google, ki vam s sintagmo "Ali ste morda mislili ..." ponudi statistično pogostejšo, a ne nujno s pravopisnim standardom poenoteno možnost. Tudi črkovalniki v urejevalnikih besedil ne ponujajo vedno rešitev, usklajenih s standardno različico.

Slovnični pregledovalnik za slovenščino prav tako še ni prosto dostopen. Obstoja komercialni slovnični pregledovalnik, ki pa tudi nudi še veliko prostora za izboljšanje, zlasti v smeri integracije statističnih metod, implementirana pravila pa bi potrebovala ponovno strokovno oceno – tj. v primeru, da bi se pokazala za bolj racionalno možnost nadgradnja tega kot izgradnja novega orodja. Slovnični pregledovalnik je eno od jezikovnotehnoloških orodij, ki lahko prek uvoza v besedilne urejevalnike podobno kot črkovalnik najde pot do najširšega kroga uporabnikov.

Opis:

V okviru akcije predvidevamo izdelavo novega seznama slovenskih besed, ki bo nastal na osnovi referenčnega oblikoslovnega leksikona, predvidenega v akciji **L-1**, in bo primeren za vgradnjo v odprtokodne črkovalnike, dodatno pa bi bilo tudi koristno poboljšanje pregibnih vzorcev teh črkovalnikov. Ti viri morajo biti v sklopu akcije tudi vgrajeni v črkovalnike in razne odprtokodne aplikacije, ki jih uporabljajo, kot sta npr. LibreOffice in Firefox, tako da bodo postali del standardne distribucije teh programov. Namen te akcije je torej pridobiti referenčni črkovalnik, ki ga potrdi standardizacijsko telo in bo prosto dostopen na voljo jezikovnim uporabnikom. Dodatno bi bilo koristno dopolniti bazo pravil za odprtokodno orodje za preverjanje in lektoriranje besedil LanguageTool, ki je vgrajeno v odprtokodne programe LibreOffice, OpenOffice in Firefox.

Drugi del akcije je namenjen izdelavi odprtokodnega slovničnega pregledovalnika za slovenščino, ki bo na voljo za vgradnjo v programe, kot so MS Word, LibreOffice ipd.

Predvideni učinki:

- izboljšano procesiranje besedil
- usklajenost normativnih podatkov
- spodbujen razvoj naprednejših JT aplikacij za slovenščino

Nosilec:
MK, MIZŠ (ARRS)

Časovni plan:
2015–2018

Obseg financiranja: možnost izvedbe preko CRP

Podrobna opredelitev aktivnosti s kazalniki:

1 Nabor relevantnega besedja in pripadajočih oblik	da/ne
2 Potrditev normativne ustreznosti orodja	da/ne
3 Tehnična priprava za nameščanje orodja	da/ne
4 Slovnčni pregledovalnik za slovenščino	da/ne

Akcija Z-1: Korpus in slovar znakovnega jezika

Nadgradnja korpusa slovenskega znakovnega jezika (SZJ) in slovar

Gluhi za sporazumevanje uporabljajo slovenski znakovni jezik (SZJ), ki je z zakonom opredeljen kot eden od uradnih jezikov Republike Slovenije, gluhi pa imajo pravico do njegove uporabe v vseh javnih situacijah. Kljub temu je opremljenost s temeljnimi jezikovnimi viri, ki bi bili tudi osnova za temeljne raziskave, vse prej kot zadovoljiva.

Korpus v sodobnem jezikoslovju predstavlja temeljni vir za jezikovni opis, prav za SZJ pa sta opis in izvedba temeljnih raziskav nujno potrebna. Še posebej nujen je korpus tudi za izdelavo sodobnih izobraževalnih gradiv, ki bi temeljil na specifičnih lastnostih SZJ. Na področju razvoja SZJ, izdaj učbenikov, slovarja in raziskovanja kretenj je bila doslej najbolj dejavna Zveza društev gluhih in naglušnih Slovenije (ZDGNS), pri kateri je ustanovljen tudi Center za razvoj SZJ. V okviru raziskovalnega projekta, ki ga je financirala ARRS, z naslovom "Korpus in pilotna slovnica SZJ" je bil zgrajen korpus SIGNOR, ki predstavlja reprezentativen vir dejanske rabe SZJ. Ker pa je SIGNOR označen zgolj z osnovnimi nivoji analize, še niso mogoče kvantitativno podprte raziskave skladišne strukture SZJ in slovničnih značilnosti.

Slovar predstavlja temeljni jezikovni vir, ki zagotavlja osnovno raven standardizacije jezika, podpira njegovo uporabo v čim več življenjskih situacijah in je neobhodno potreben pri poučevanju SZJ. Doslej sta se z gradnjo slovarja ukvarjali dve ekipi:

- Zveza društev gluhih in naglušnih Slovenije (ZDGNS), kjer se že nekaj let gradi multimedijski spletni slovar SZJ, ki je tudi najobsežnejši slovarski vir SZJ,
- Fakulteta za računalništvo in informatiko UL, kjer je v sodelovanju s podjetjem Zoom Promotion nastal še en spletni slovar SZJ ter številna spletna gradiva za poučevanje SZJ.

Poleg omenjenih obstaja še nekaj manjših slovarskih zbirk SZJ. Spletni slovar ZDGNS je bil pred kratkim na podlagi ponudbe v postopku enostavnega naročanja Ministrstva za kulturo (MK) prenovljen, kljub temu pa bi bilo za potrebe gluhe skupnosti slovar nadaljnje posodobiti in metodološko prenoviti, ga dopolniti z novimi kretnjami in ga uskladiti s podatki o dejanski rabi.

Opis:

Kot prvo aktivnost načrtujemo (nad)gradnjo korpusa SZJ s segmentacijo na izjave ter označevanjem učnega korpusa za raziskovanje skladnje. Nato načrtujemo izgradnjo leksikalne baze SZJ. Prav tako je potrebno zagotoviti, da bo korpus SZJ javen in s tem dostopen širši raziskovalni in pedagoški sferi.

Druga aktivnost predvideva (nad)gradnjo slovarja SZJ. Pri tem mora biti slovar metodološko osnovan na dejanskih podatkih o rabi iz korpusa SZJ, tehnična izvedba slovarja bi morala zagotavljati tudi večjo odprtost in preglednost podatkov, s poudarkom na čim boljšem sodelovanju različnih akterjev pri gradnji slovarja. Sodobni slovarji znakovnih jezikov omogočajo tudi iskanje po kretnji, ne zgolj po njeni pomenski oznaki; tako lahko uporabnik ugotovi, kaj neznana kretnja pomeni. Slovar mora zajemati čim več besedišča, tudi regionalne različice kretenj in strokovno terminologijo, biti mora prilagojen za uporabo na mobilnih napravah, omogočati mora napredne načine iskanja, v perspektivi pa bi bila zaželena tudi povezava med slovarjem in korpusom SZJ.

Predvideni učinki:

- omogočitev temeljnih raziskav fonoloških, oblikoslovnih in skladišnih vidikov SZJ;
- možnost izdelave boljših izobraževalnih gradiv,
- sodoben in poenoten slovar SZJ,

- izvirne znanstvene objave.

Nosilec:

MK, MIZŠ (ARRS)

Časovni plan: možnost izvedbe preko raziskovalnega projekta
2015–2018

Podrobna opredelitev aktivnosti s kazalniki:

1.1 Označevanje celotne baze posnetkov korpusa SZJ	označenih min. 30.000 kretenj
1.2 Označevanje skladišne strukture v korpusu SZJ	označenih min. 3.000 izjav
1.3 Izdelava leksikalne baze SZJ	min. 2.500 leksikalnih vnosov
1.4 Pregled in integracija baze obstoječega slovarja ZDGNS z leksikalno bazo SZJ	da/ne
1.5 Izdelava multimedijske spletne slovarske aplikacije z iskalnikom	da/ne

Priloga 1: Pregled ciljev in aktivnosti po sklopih

Razpredelnica prikazuje, katere aktivnosti so bile definirane kot potrebne za uresničitev posameznih ciljev, opredeljenih v poglavju 2.

Sklop	Cilj	Aktivnosti
Splošno	Infrastrukturni center	Delovanje infrastrukturnega centra
	Odprta dostopnost jezikovnih virov	Prizadevanje za spremembo zakonodaje, da zagotavlja odprti dostop do vseh del, ki nastanejo z javnim financiranjem
	Spodbujanje razvoja slovenske Wikimedije	Prenos obstoječih virov v Wikimedijo
Jezikovni opis	Splošni slovar in slovarska baza	Aktivnosti bodo definirane s konceptom slovarja.
	Diahroni opis jezika	Izdelava diahronnega korpusa slovenskega jezika
	Slovnični opis	Nadgradnja pisnih korpusov sodobnih besedil
		Nadgradnja in izdelava specializiranih korpusov
		Nadgradnja referenčnega govornega korpusa
		Oblikoslovni in glasoslovni leksikon
		Semantična mreža
		Semantična analiza vlog
		Razvozavljanje anafore in analiza besedilnih koreferenc
		Slovnična podatkovna zbirka
		Koncept nove slovnice slovenskega jezika
		Šolska slovnica/slovnični portal
	Slovar naselbinskih imen	Izdelava slovarja naselbinskih imen slovenskega narodnega prostora
	Slovenski lingvistični atlas	Izboljšanje atlasa z uvedbo poligonov, določenih z inovativnimi spremenljivkami
		Večpredstavna spletna objava 1. in 2. zvezka SLA
		Razvijanje vnašalnega sistema za dialektološke zapise in njegovo prilagajanje novim programskim okoljem
	Terminološki slovarji različnih strok	Zasnova slovarjev
		Izdelava specializiranih korpusov
		Izdelava geslovníkov na podlagi korpusov
		Pisanje in urejanje slovarskih sestavkov
		Izdaja slovarjev v tiskani in elektronski obliki
	Zgodovinski slovar	Izboljšava slovarskega koncepta na

		osnovi dodatnih raziskav
		Izdelava priročnika za bodoče leksikografe
		Spletna progresivna objava slovarja
		Vsebinska in jezikovnotehnološka zasnova korpusa slovenskega knjižnega jezika 16. stoletja
	Spodbujanje prožnih in inovativnih oblik učenja z razvojem jezikovnih virov in tehnologij	Izbrani vidik(i) slovarsko-slovničnega opisa in jezikovnih normativnih informacij v digitalni obliki
		Učna e-orodja, temelječa na sodobnem jezikovnem opisu in aktualiziranih jezikovnonormativnih informacijah
		Razvoj inovativnih in prožnih učnih pristopov in metod
Standardizacija	Pravopisni portal	Izobraževanje učiteljev
		Predstavitev normativnih podatkov: pravila in slovar
		Predstavitev normativnih podatkov: spletni priročnik za splošne uporabnike
	Jezikovnosvetovalni portal	Jezikovno svetovanje
	Zbirka normativnih zadreg	Zasnova zbirke normativnih zadreg
		Gradnja zbirke (luščenje podatkov, kategorizacija podatkov)
		Preverjanje realizacije in ustreznosti pravopisnih, pravorečnih, oblikoslovnih podatkov (skupina strokovnjakov)
	Črkovalnik besed	Izdelava referenčnega črkovalnika za slovenščino
	Korpus za normativna vprašanja	Izdelava korpusa slovenskega jezika, specializiranega za normativna vprašanja
		Specialni moduli za komplicirane poizvedbe pri korpusih (mala/velika začetnica, ločila ...)
Terminologija	Terminološki portal	Terminološki portal
		Spletna aplikacija za ustvarjanje novih terminoloških zbirk
		Samodejno luščenje terminologije
		Zagotavljanje javno dostopnih terminoloških virov
		Terminološko svetovanje
		Nadgradnja korpusov strokovnih besedil
Večjezičnost	Večjezični portal	Večjezični portal
		Dvo- in večjezične baze in slovarji
	Strojni prevajalnik	Dvojezični korpusi

		Dvo- in večjezične baze in slovarji
		Imenske entitete in stalne besedne zveze
		Oblikoslovna in skladenjska analiza besedila
		Oblikoslovni leksikon
		Referenčni korpusi slovenščine
		Razvozlavanje anafore in analiza besedilnih koreferenc
		Semantična mreža
		Semantična analiza vlog
		Oblikoslovni sintetizator
		Razvoj strojnega prevajalnika
Jezikovne tehnologije	Razpoznavalnik tekočega govora	Oblikoslovna in skladenjska analiza besedila
		Referenčni korpus slovenščine za učenje jezikovnih modelov
		Večnamenski korpus in baza govorjenega jezika
		Razpoznavalnik tekočega govora z velikim slovarjem besed
	Avtomatsko procesiranje besedila v jezikovnih tehnologijah in spletnih iskalnih orodjih	Imenske entitete in stalne besedne zveze
		Oblikoslovna in skladenjska analiza besedila
		Oblikoslovni in glasoslovni leksikon
		Referenčni korpusi slovenščine
		Razvozlavanje anafore in analiza besedilnih koreferenc
		Semantična mreža
		Semantična analiza vlog
	Slovnični pregledovalnik	Imenske entitete in stalne besedne zveze
		Oblikoslovna in skladenjska analiza besedila
		Oblikoslovni leksikon
		Razvozlavanje anafore in analiza besedilnih koreferenc
		Imenske entitete in stalne besedne zveze
		Slovnični pregledovalnik – aplikacija
	Sistemi dialoga	Imenske entitete in stalne besedne zveze
		Oblikoslovna in skladenjska analiza besedila
		Oblikoslovni sintetizator
		Oblikoslovni in glasoslovni leksikon
		Referenčni korpusi slovenščine
		Razvozlavanje anafore in analiza besedilnih koreferenc
		Semantična mreža
		Semantična analiza vlog

		Dialoški korpusi
		Sistem dialoga - aplikacija
Digitalizacija	Digitalni čistopisi slovenske pisne kulturne dediščine	Izdelava čistopisov z uporabo množičenja
	Digitalni jezikovni priročniki za slovenski jezik	Digitalizacija ali prilagajanje digitalnemu okolju jezikovnih priročnikov za slovenščino
	Digitalna slovenska znanstvena besedila	Vključitev novih znanstvenih del v obstoječe portale
		Izdelava korpusov znanstvenih besedil
Govorci s posebnimi potrebami	Bralnik zaslona za slepe in slabovidne	Sintetizator govora za bralnik zaslona v slovenskem jeziku
	Branje sporočil gospodinjskih aparatov za slepe in slabovidne	Bralnik sporočil gospodinjskih aparatov
	Razpoznavanje govora v pedagoškem procesu	Razpoznavalnik govora za pouk in predavanja
	Slovar slovenskega znakovnega jezika (SZJ)	Nadgradnja korpusa SIGNOR
		Spletni slovar SZJ