Oznaka poročila: ARRS-RPROJ-ZP-2013/266

ZAKLJUČNO POROČILO RAZISKOVALNEGA PROJEKTA

A. PODATKI O RAZISKOVALNEM PROJEKTU

1.Osnovni podatki o raziskovalnem projektu

Šifra projekta	Z6-3668
Naslov projekta	Jezikovno-neodvisne metode za avtomatsko gradnjo semantičnih leksikonov s pomočjo primerljivih korpusov
Vodja projekta	26294 Darja Fišer
Tip projekta	Zt Podoktorski projekt - temeljni
Obseg raziskovalnih ur	3400
Cenovni razred	A
Trajanje projekta	05.2010 - 04.2012
Nosilna raziskovalna organizacija	581 Univerza v Ljubljani, Filozofska fakulteta
Raziskovalne organizacije - soizvajalke	
Raziskovalno področje po šifrantu ARRS	6 HUMANISTIKA 6.05 Jezikoslovje
Družbeno-ekonomski cilj	13.06 Humanistične vede - RiR financiran iz drugih virov (ne iz SUF)

2.Raziskovalno področje po šifrantu FOS[1]

Šifra	6.02
- Veda	6 Humanistične vede
- Področje	6.02 Jeziki in književnost

B. REZULTATI IN DOSEŽKI RAZISKOVALNEGA PROJEKTA

3.Povzetek raziskovalnega projekta[2]

SLO

V okviru podoktorske raziskave smo preizkusili pristope avtomatskega luščenja prevodnih ustreznic iz besedil, ki niso vzporedna (izvirnik in prevod), temveč zgolj primerljiva (podobna tema, čas objave, komunikacijski namen). Primerljivi korpusi postajajo vse bolj priljubljen vir jezikovnega znanja, saj jih je z razvojem vsebin na svetovnem spletu vse lažje zbirati, predvsem

pa je primerljivih vsebin na voljo veliko več kot vzporednih, še posebej za jezikovne kombinacije s slovenščino. Raziskava je temeljila na predpostavki, da se besede v izvornem jeziku pojavljajo v podobnih kontekstih kot njihovi prevodi v ciljnem jeziku. Iskanje besed z najbolj podobnimi konteksti smo izvedli tako, da smo za vse besede v izvornem in ciljnem delu korpusa izdelali kontekstne vektorje na podlagi sobesedila, v katerem se pojavljajo. Nato smo s pomočjo dvojezičnega slovarja izvorne kontekstne vektorje prevedli v ciljni jezik in jih primerjali s ciljnimi kontekstnimi vektorji. Vektor tiste besede, ki je v tovrstni primerjavi dobil najboljši rezultat, smo obravnavali kot najverjetnejši prevod besede v izvirniku, saj so si njuni kontekstni vektorji najbolj podobni.
Kolikor nam je znano, raziskav na primerljivih korpusih na področju slovenske leksikalne semantiče še ni opravljenih, zato zaključen projekt nedvomno predstavlja pomemben mejnik v slovenskih korpusnih, pa tudi jezikovnotehnoloških raziskavah. Ne samo, da je rezultat projekta utemeljena, preizkušena in jezikovno neodvisna metodologija luščenja prevodnih ustreznic iz primerljivih korpusov, je zaključen projekt prinesel tudi oprijemljive rezultate v obliki semantičnega leksikona sloWNet, ki je poravnan z wordneti za številne druge jezike in tako uporaben za eno- in večjezične računalniške aplikacije. V okviru projekta smo razvili tudi orodje sloWTool za pregledovanje, popravljanje in vizualizacijo wordnetov in orodje sloWCrowd za reševanje leksikalno-semantičnih nalog s pomočjo izkoriščanja množic. Vsi jezikovni viri in orodja, ki so bili razviti v okviru projekta, so objavljeni in javno dostopni v raziskovalne namene pod licenco Creative Commons: http://nl.ijs.si/slownet/. Izdelani viri in orodja s tem zapolnjujejo vrzel v jezikovnih virih za slovenščino in postavljajo temelje za širšo, semantično obogateno izrabo slovenskih korpusnih virov (Fišer 2012).

ANG

The main goal of the postdoctoral project was to test and develop approaches to automatic extraction of translation equivalents from texts which are not parallel (original and its translation) but comparable (similar topic, time of publication, communicative purpose). Comparable corpora have become an increasingly popular source of linguistic knowledge because they are much easier to compile directly from the web than parallel corpora, and because many existing monolingual corpora can be used as (weakly) comparable data. The research was based on the premise that words in the source language appear in similar contexts than its translations in the target language. Words with the most similar contexts across languages were identified by constructing context vectors for all the words in the source corpus and all the words in the target corpus. Next, all context vector features in the source language were translated into the the target languages with a seed dictionary, after which the translated source context vectors were compared to the target context vectors. The target word with the most similar context vector for a given source word was considered as its most likely translation equivalent.
To my knowledge, there has been no previous research into comparable corpora in the field of Slovene lexical semantics. This is why the postdoctoral project presents an important milestone in Slovene corpus linguistics as well as human language technologies. Not only is the result of the project be an established, tested and language-independent methodology for the extraction of translation equivalents from comparable corpora; the project has also brought a highly palpable result in the form of a Slovene semantic lexicon that is aligned to wordnets in many other languages and therefore useful for mono- as well as multilingual applications. In this way, the developed wordnet has significantly narrowed the gap in the field of Slovene language resources and has provided the foundations for a broader, semantically-enriched exploitation of Slovene corpus resources.

4.Poročilo o realizaciji predloženega programa dela na raziskovalnem projektu[3]

Podoktorska raziskava je bila sestavljena iz več faz:
(1) Izdelava primerljivega korpusa
V okviru projekta smo uporabili dva primerljiva korpusa, enega strokovnega, drugega pa splošnejšega v štirih jezikih. Prvi korpus je bil izdelan iz spletne enciklopedije Wikipedija za slovenščino, angleščino, francoščino in hrvaščino. Kot drugi primerljivi korpus pa smo uporabili zdravstveni podkorpus, ki je bil izdelan iz spletnega korpusa za iste 4 jezike. slWac in hrWac sta bila pridobljena s svetovnega spleta s spletne domene .si oziroma .hr, lematizirana in oblikoskladenjsko označena (Ljubešić in Erjavec 2011). Angleški ukWac in francoski frWac sta bila razvita v okviru iniciative Wacky (Baroni idr. 2009) in sta prosto dostopna za raziskovalne

namene. V zdravstveni podkorpus so bila zajeta vsa besedila, ki so vključena v te spletne korpuse in so bila objavljena v revijah o zdravju in zdravem načinu življenja oz. so jim glede na jezikovni model, ki je bil na podlagi zdravstvenih člankov zgrajen, dovolj podobna (Fišer idr. 2011).
(2) Izdelava izhodiščnega leksikona
Za izhodiščne dvojezične leksikone, ki so pri raziskavi potrebni za prevajanje kontekstnih vektorjev, smo preizkusili več variant. Uporabili smo leksikone, zgrajene na podlagi prosto dostopnega spletnega slovarja Wiktionary, pa tudi avtomatsko izdelana leksikona za angleško-slovenski in angleško-francoski jezikovni par, ki smo ju izluščili iz besedno poravnanih vzporednih korpusov iz prejšnjih raziskav (Apidianaki idr. 2012). Ker vzporedni korpus za hrvaško-slovenski jezikovni par ni bil na voljo, smo leksikon zanj avtomatsko izluščili iz identičnih besed in sorodnic med jezikoma, ki smo ju našli v primerljivih korpusih (Fišer in Ljubešić 2011).
(3) Gradnja in primerjava kontekstnih vektorjev
Za gradnjo vektorjev smo preizkusili več statističnih mer, kot najboljši pa se je izkazal logaritem verjetnosti. Za primerjavo vektorjev smo prav tako preizkusili različne statistične mere, pri čemer je v eksperimentih najboljše rezultate dosegla divergenca Jensen-Shannon (Ljubešić idr. 2011). Sistem za vse jezikovne pare deluje za enobesedne izraze brez upoštevanja večpomenskisti, na paru angleščina-slovenščina pa smo razvili tudi pristop za večpomenske besede (Fišer idr. 2012, Apidianaki idr. 2012) ter pokazala, da luščenje po enakem principu deluje tudi za večbesedne terimne (Ljubešić et al. 2012). Posebnost uporabljenega pristopa za luščenje prevodnih ustreznic za jezikovni par slovenščina-hrvaščina je v tem, da za razliko od ostalih jezikovnih parov v tem primeru za prevajanje kontekstnih vektorjev nismo uporabili nobenega izhodiščnega slovarja, temveč smo izkoristili podobnosti med jezikoma in se naslonili zgolj na ortografsko identične besede in sorodnice, ki se pojavljajo v korpusih za oba jezika (Fišer in Ljubešić 2011). S tem je pristop postal povsem neodvisen od katerega koli zunanjega vira znanja, kar je za številne jezikovne pare zelo atraktivno, saj obsežni dvojezični slovarski viri za večino jezikov še vedno niso na voljo za raziskovalne namene.
(4) Luščenje definicij in semantično povezanih besed
Za potrebe slovenskega wordneta smo preizkusili tudi luščenje definicij in semantično povezanih besed iz korpusov, v katerih smo s pomočjo leksiko-semantičnih besedilnih vzorcev v tekočih besedilih prepoznali potencialne definicijske stavke, te pa nato s pomočjo orodja za strojno učenje Weka klasificirali kot prave definicije oz. kot nedefinicije (Fišer idr. 2011).
(5) Identifikacija lažnih prijateljev
Z upoštevanjem osnovne predpostavke distribucijske semantike, ki se glasi, da se besede s podobnim pomenom v različnih jezikih pojavljajo v podobnih kontekstih, ni mogoče avtomatsko identificirati zgolj prevodnih ustreznic, temveč tudi lažne prijatelje, ki so si na videz sicer zelo podobni, vendar so njihove korpusne frekvence in kontekstni vektorji zelo različni. S tem smo pravzaprav razvili učinkovito metodo za avtomatsko prepoznavanje lažnih prijateljev, ki je zelo koristna za jezikovnotehnološke aplikacije, pa tudi v leksikografiji in pri poučevanju tujih jezikov (Fišer in Ljubešić, v tisku).
(6) Vrednotenje rezultatov
Avtomatsko vrednotenje rezultatov je bilo opravljeno s pomočjo zlatih standardov, ki smo jih pripravili za vsak jezikovni par posebej, in sicer tako, da smo iz tradicionalnih slovarskih virov izbrali določeno število naključnih slovarskih vnosov ter nato primerjali, ali se prevodne ustreznice, pridobljene s pristopi, ki smo jih razvili v raziskavi, ujemajo s slovarskimi. Merili smo natančnost, priklic in f-mero. Za vse jezikovne pare, tipe korpusov in eksperimentalne nastavitve smo dosegli najboljše rezultate za samostalnike, ki z najboljšimi parametri celo presega 80 %. Najvišjo stopnjo izboljšanja smo dosegli z razširitvami avtomatsko izluščenega slovensko-hrvaškega leksikona, na splošno pa najboljše rezultate daje luščenje angleško-slovenskih prevodnih ustreznic iz obsežnih spletnih primerljivih korpusov z uporabo izhodiščnega slovarja, minimalno frekvenco 50 za prevajane besede, logaritmom verjetnosti za gradnjo ter divergenco Jensen-Shannon za primerjavo kontekstnih vektorjev.
Ročno vrednotenje rezultatov smo opravili na naključno izbranem vzorcu prevodnih ustreznic, kjer smo poleg vrednotenja ustreznosti izluščene prevodne ustreznice opravili še klasifikacijo in kvalitativno analizo napak ter identificirali največje pomanjkljivosti uporabljenega pristopa ter možnosti za izboljšave. Rezultati ročnega vrednotenja kažejo, da se najbolje odrežejo konkretni samostalniki, za katere je mogoče pridobiti zadostno količino sobesedila. Slabše se odrežejo abstraktni samostalniki in pridevniki, ki v besedilih pogosto nimajo stabilnega konteksta. Zanimiva ugotovitev ročnega vrednotenja rezultatov je ta, da se med napačnimi prevodnimi

ustreznicami nikoli ne ponavljajo povsem nepovezani kandidati, temveč so s pravilnim vsi v tesni semantični povezavi (nadpomenke, podpomenke, kohiponimi, protipomenke), kar pomeni, da primerjava kontekstnih vektorjev pravzaprav zelo dobro deluje in je uporabna tudi za iskanje elementov določenega pomenskega polja.
Vrednotenje rezultatov s pomočjo aplikacije pa smo izvedli s pomočjo avtomatskega razreševanja večpomenskosti s pomočjo izdelanega wordneta, rezultate pa nato uporabili za izboljšanje strojnega prevajanja večpomenskih besed v prevajalnikih GoogleTranslate in Presis (Vintar idr. 2012). Rezultati vrednotenja z aplikacijo UKB (Agirre in Soroa 2009) so pokazali, da je s pomočjo pravilno razdvoumljenih večpomenskih izvirnih besed v obeh jezikovnih smereh mogoče dobiti boljše prevode, kot jih predlaga Presis, v nekoliko manjši meri pa tudi GoogleTranslate.
Izdelan wordnet smo skušali ovrednotiti s pomočjo ročnega semantičnega označevanja vzporednega korpusa SPOOK (Fišer in Bizjak 2012). Z raziskavo smo skušali ugotoviti, v kolikšni meri slovenski wordnet, ki je bil izdelan na podlagi angleškega, pokriva besedišče iz korpusa, ter ali je uporaben za označevanje pomenov besed različnih zvrsteh literarnih besedil. Ugotovili smo, da je v sloWNetu pogosto besedišče odlično zastopano, redkejše pa je v veliki meri še potrebno vanj dodati. Prav tako smo opazili, da se pri semantičnem označevanju enega dela vzporednega korpusa pojmi ohranijo tudi v prevodu in da s semantičnim označevanjem ni večjih težav, s čimer potrjujemo utemeljenost metodologije, ki je temeljila na prevajanju tujejezičnega leksikalnega vira v slovenščino. Pri tem pa je potrebno poudariti, da moramo slovenske kulturno-specifične pojme, ki v tako nastali semantični mreži zaenkrat manjkajo, čim prej dodati neodvisno od angleškega vira, ki nam je služil kot iztočnica za izdelavo sloWNeta.
(7) Primerjava slovenskega wordneta z drugimi leksikalnimi viri
Izdelan wordnet smo s sodelavci projekta Sporazumevanje v slovenskem jeziku na vzorčnem naboru besedišča primerjali in povezali s Slovensko leksikalno bazo, s čimer smo wordnet obogatili s kolokacijami in skladenjskimi informacijami, v leksikalno bazo pa nadgradili s semantičnimi relacijami in angleškimi prevodnimi ustreznicami (Fišer idr. 2012). Izdelan wordnet smo prav tako primerjali s semantično mrežo, ki bi jo dobili, če bi za izdelavo uporabili izključno jezikovne podatke iz slovenskega referenčnega korpusa, ne pa tudi angleškega wordneta in večjezičnih virov, na podlagi katerih smo iskali slovenske ustreznice (Fišer idr. 2012). Pri tem smo sodelovali s sodelavci Politehniko v Wroclawu, ki so na ta način razvili poljski wordnet. Ugotavljamo, da je pristop zelo koristen predvsem za razširitev obstoječega wordneta, saj je z uporabljeno metodo na podlagi korpusnih podatkov mogoče najti najverjetnejše mesto v semantični mreži, kamor se uvršča nova beseda, ki je doslej v wordnetu ni bilo.
(8) Vizualizacija rezultatov
Za lažjo uporabo razvitega wordneta za slovenščino smo s sodelavci z Univerze v Mariboru razvili specializiran brskalnik po semantičnem leksikonu sloWTool (Fišer in Novak 2011), ki omogoča eno- in večjezično iskanje, iskanje po posameznih poljih, megleno iskanje in sledenje semantičnim relacijam. sloWTool je kot brskalnik, pa tudi celotna programska koda zanj, prosto dostopen na spletu: http://nl.ijs.si/slowtool/. Poleg brskalnika sloWTool omogoča tudi popravljanje in dodajanje sinsetov ter vizualizacijo, ki s pomočjo hiperboličnih grafov besede v wordnetu prikazuje kot vozlišča, semantične relacije med njimi pa kot povezave med njimi. V sloWTool so integrirane tudi številne podatkovne zbirke, ki so jih razvili na drugih inštitucijah, kot so: področne oznake za sinsete WordNet Domains (Bentivogli idr. 2004), splošna vrhnja ontologija pojmov SUMO/MILO (Niles in Pease 2001), gruče pomenov (Navigli 2006) in povezava na galerijo ImageNet (Deng idr. 2009). Poleg slovenskega in angleškega wordneta so v istem brskalniku na našem strežniku na voljo tudi francoski, poljski in japonski wordnet, v prihodnje nameravamo dodati tudi wordnete za druge jezike, ki so prosto dostopni za raziskovalne namene.
(9) Validacija rezultatov
Za vrednotenje rezultatov, dobljenih s preizkušenimi pristopi za luščenje prevodnih ustreznic iz heterogenih večjezičnih virov, ki niso nujno vzporedni, smo s sodelavci z Inštituta Jožef Stefan razvili orodje sloWCrowd (Tavčar idr. 2012), ki je namenjen validaciji pravilno izluščenih prevodnih ustreznic oz. izločanju napak v avtomatsko generiranih semantičnih leksikonih. Orodje je je zasnovano tako, da odgovore za problematične literale omogoča zbirati iz široke množice uporabnikov. Orodje je prosto dostopno in temelji na razširjenih tehnologijah, kot sta PHP in MySQL. Sestavljata ga administratorski in uporabniški vmesnik. V administratorskem vmesniku izdelamo projekt, sledimo poteku projekta in izvažamo rezultate, v uporabniškem vmesniku pa reševalci glasujejo o (ne)pravilnosti naključno izbranih literalov. Rezultati prvega eksperimenta, ki smo ga izvedli z orodjem sloWCrowd, so spodbudni, saj so bili uporabniki orodja z njim zadovoljni, odločitev o dokončnem izbrisu nekega literala na podlagi ujemanja njihovih

odgovorov pa enostavna, hitra in zanesljiva. Dodatna prednost razvitega orodja je, da ga je mogoče prilagoditi za najrazličnejše naloge, pri katerih je koristno sodelovanje večjega števila reševalcev. Orodje je prosto dostopno na spletu: http://nl.ijs.si/slowcrowd/.
(Zaradi prostorskih omejitev je seznam bibliografije v priponki.)

5. Ocena stopnje realizacije programa dela na raziskovalnem projektu in zastavljenih raziskovalnih ciljev[4]

Ocenjujem, da smo zastavljen program v podoktorskem projektu v celoti uresničili. Utemeljenost in uspešnost uporabljenih metod smo potrdili s temeljitim avtomatskim, ročnim in aplikativnim vrednotenjem rezultatov. V zadnji različici izdelanega wordneta za slovenščino je tako 42.722 množic sinonimov, od katerih smo jih skoraj četrtino (9.626) že tudi ročno pregledali in popravili. Program je bil realiziran tudi z objavo znanstvenih prispevkov v številnih monografijah, revijah in konferenčnih zbornikih (glej seznam bibliografije v prejšnjem razdelku). Že sprejet, a še v tisku, je tudi prispevek za revijo Slovenščina 2.0 "Best friends or just faking it? Corpus-based extraction of Slovene-Croatian translation equivalents and false friends", v recenzentskem postopku pa je tudi prispevek za revijo s faktorjem vpliva Landuage Resources and Evaluation "Building a Poor Man's Wordnet".

6. Utemeljitev morebitnih sprememb programa raziskovalnega projekta oziroma sprememb, povečanja ali zmanjšanja sestave projektne skupine[5]

Sprememb v programu raziskovalnega projekta ali sestavi projektne skupine ni bilo.

7. Najpomembnejši znanstveni rezultati projektne skupine[6]

	Znanstveni dosežek			
1.	COBISS ID		46703714	Vir: COBISS.SI
	Naslov	SLO	Luščenje dvojezičnih leksikonov iz primerljivih korpusov za sorodne jezike	
		ANG	Bootstrapping bilingual lexicons from comparable corpora for closely related languages	
	Opis	SLO	Predstavitev pristopa za luščenje dvojezičnih leksikonov iz primerljivih korpusov za sorodne jezike	
		ANG	Presentation of an approach for bootstrapping bilingual lexicons from comparable corpora for closely related languages	
	Objavljeno v		Springer; Text, speech and dialogue; Lecture notes in computer science; 2011; Str. 91-98; Avtorji / Authors: Fišer Darja	
	Tipologija		1.01 Izvirni znanstveni članek	
2.	COBISS ID		50261858	Vir: COBISS.SI
	Naslov	SLO	Jezikovni viri in orodja za semantično obogateno procesiranje slovenščine	
		ANG	Language resources and tools for semantically enhanced processing of Slovene	
	Opis	SLO	Predstavitev jezikovnih virov in orodij za semantično obdelavo slovenščine	
		ANG	Presentation of language resources and tools for semantic processing of Slovene	
	Objavljeno v		Cambridge Scholars; Multilingual processing in eastern and southern EU languages; 2012; Str. 92-118; Avtorji / Authors: Fišer Darja	
	Tipologija		1.16 Samostojni znanstveni sestavek ali poglavje v monografski publikaciji	
3.	COBISS ID		47786850	Vir: COBISS.SI
	Naslov	SLO	sloWNet 3.0: Gradnja, razširitev in čiščenje	
		ANG	sloWNet 3.0: Development, extension and cleaning	

	Opis	SLO	Predstavitev pristopov za širitev in čiščenje zadnje različice slovenskega wordneta
		ANG	Presentation of the approaches for the extension and cleaning of Slovene wordnet
	Objavljeno v		The Global WordNet Association; 6th International Global Wordnet Conference; 2012; Str. 113-117; Avtorji / Authors: Fišer Darja, Novak Jernej, Erjavec Tomaž
	Tipologija		1.08 Objavljeni znanstveni prispevek na konferenci
4.	COBISS ID		47260258 Vir: COBISS.SI
	Naslov	SLO	Avtomatsko luščenje hrvaško-slovenskega leksikona iz primerljivih korpusov
		ANG	Automatic extraction of Croatian-Slovene lexicon from comparable corpora
	Opis	SLO	V prispevku predstavljamo metodo za avtomatsko luščenje hrvaško-slovenskega leksikona iz primerljivega časopisnega korpusa s predpostavko, da se besede in njihove prevodne ustreznice pojavljajo v podobnih sobesedilih. Izhodiščni leksikon za primerjavo kontekstnih vektorjev z izkoriščanjem podobnosti med jezikoma zgradimo kar iz korpusa, nato pa opravimo še razvrščanje rezultatov glede na stopnjo sorodnosti med izvorno besedo in njenimi prevodnimi kandidati. Rezultati so zelo spodbudni in odpirajo številne možnosti uporabe za druge sorodne jezike.
		ANG	In this paper we present a method for extracting a bilingual lexicon for closely related languages from comparable corpora. We take advantage of the similarities between languages to build a seed lexicon to compare context vectors in both languages and use cognates for reranking translation candidates. The results are very encouraging, suggesting that other similar languages could benefit from the same approach.
	Objavljeno v		Znanstvena založba Filozofske fakultete; Meddisciplinarnost v slovenistiki; 2011; Str. 137-144; Avtorji / Authors: Fišer Darja, Ljubešić Nikola
	Tipologija		1.16 Samostojni znanstveni sestavek ali poglavje v monografski publikaciji
5.	COBISS ID		50058338 Vir: COBISS.SI
	Naslov	SLO	Polisemija in luščenje dvojezičnih leksikonov iz primerljivih korpusov
		ANG	Addressing polysemy in bilingual lexicon extraction from comparable corpora
	Opis	SLO	Prispevek predstavlja pristop za iskanje prevodnih ustreznic polisemnih samostalnikov.
		ANG	This paper presents an approach to extract translation equivalents from comparable corpora for polysemous nouns. As opposed to the standard approachesthat build a single context vector for all occurrences of a given headword, we first disambiguate the headword with third-party sense taggers and then build a separate context vector for each sense of the headword. Sincestate-of-the-art word sense disambiguation tools are still far from perfect, we also tried to improve the results by combining the sense assignments provided by two different sense taggers. Evaluation of the resultsshows that we outperform the baseline (0.473) in all the settings we experimented with, even when using only one sense tagger, and that the best-performing results are indeed obtained by taking into account the intersection of both sense taggers (0.720).
	Objavljeno v		ELRA; LREC 2012; 2012; Str. 3031-3035; Avtorji / Authors: Fišer Darja, Ljubešić Nikola, Kubelka Ozren
	Tipologija		1.08 Objavljeni znanstveni prispevek na konferenci

8.Najpomembnejši družbeno-ekonomski rezultati projektne skupine[7]

	Družbeno-ekonomski dosežek			
1.	COBISS ID		47343714	Vir: COBISS.SI
	Naslov	*SLO*	Poročilo o poletni šoli ESSLLI 2011	
		ANG	Report on ESSLLI 2011	
	Opis	*SLO*	Poročilo predsednice organizacijskega odbora 23. poletne šole ESSLLI, ki je avgusta leta 2011 potekala v Ljubljani, udeležilo pa se je je 400 udeležencev in 80 predavateljev.	
		ANG	Report of the president of the organizing committee of the 23rd ESSLLI summer school which took place in August 20011 in Ljubljana and was attended by 400 students and 80 lecturers.	
	Šifra		B.01 Organizator znanstvenega srečanja	
	Objavljeno v		EACL; EACL Newsletter; 2011; Issue 14; str. 1; Avtorji / Authors: Fišer Darja	
	Tipologija		1.25 Drugi članki ali sestavki	
2.	COBISS ID		25363239	Vir: COBISS.SI
	Naslov	*SLO*	sloWNet, wordnet za slovenščino	
		ANG	sloWNet, a wordnet for Slovene	
	Opis	*SLO*	Baza slovenskega semantičnega leksikona sloWNet	
		ANG	The database of the Slovene semantic lexicon called sloWNet	
	Šifra		F.15 Razvoj novega informacijskega sistema/podatkovnih baz	
	Objavljeno v		Filozofska fakulteta; 2011; Avtorji / Authors: Fišer Darja, Erjavec Tomaž	
	Tipologija		2.20 Zaključena znanstvena zbirka podatkov ali korpus	
3.	COBISS ID		25364007	Vir: COBISS.SI
	Naslov	*SLO*	sloWTool, orodje za iskanje, popravljanje in vizualizacijo wordnetov	
		ANG	sloWTool, a tool for browsing, editing and visualization of wordnets	
	Opis	*SLO*	Programska koda in namestitvena navodila za sloWTool, orodje za iskanje, popravljanje in vizualizacijo wordnetov	
		ANG	The code and installation instructions of sloWTool, a tool for browsing, editing and visualization of wordnets	
	Šifra		F.15 Razvoj novega informacijskega sistema/podatkovnih baz	
	Objavljeno v		Filozofska fakulteta; 2011; Avtorji / Authors: Novak Jernej, Fišer Darja, Erjavec Tomaž	
	Tipologija		2.21 Programska oprema	
4.	COBISS ID			Vir: vpis v poročilo
	Naslov	*SLO*	sloWCrowd, orodje za validacijo leksikalno-semantičnih nalog z množicanjem	
		ANG	sloWCrowd, a crowdsourcing tool for validation of lexico-semantic tasks	
	Opis	*SLO*	orodje za validacijo leksikalno-semantičnih podatkov z izkoriščanjem množic	
		ANG	a crowdsourcing tool for tasks in lexical semantics	
	Šifra		F.15 Razvoj novega informacijskega sistema/podatkovnih baz	
	Objavljeno v		http://nl.ijs.si/slowcrowd/	

	Tipologija	2.21 Programska oprema	
5.	COBISS ID		Vir: vpis v poročilo
	Naslov	*SLO*	Članica uredniškega odbora mednarodne znanstvene revije Journal of Language Modelling
		ANG	Member of the Editorial Board of the Journal of Language Modelling
	Opis	*SLO*	Revija Journal of Language Modelling je prosto dostopna, odprta, recenzirana mednarodna revija za interdisciplinarne raziskave iz teoretičnega jezikoslovja in računalniške obdelave naravnega jezika.
		ANG	Journal of Language Modelling is a free (no publication fees) open-access peer-reviewed journal aiming to bridge the gap between theoretical linguistics and natural language processing.
	Šifra		C.06 Članstvo v uredniškem odboru
	Objavljeno v		http://nlp.ipipan.waw.pl/ojs/index.php/JLM/about/displayMembership/4
	Tipologija		3.25 Druga izvedena dela

9.Drugi pomembni rezultati projetne skupine[8]

V času trajanja projekta sem bila poleg znanstvenega dela aktivna tudi na uredniškem, recenzentskem in organizacijskem področju. Sem članica Stalnega odbora ESSLLI FoLLI ter tajnica Slovenskega društva za jezikovne tehnologije. Poleg članstva v uredniškem odboru in recenziranja mednarodne revije Journal of Language Modelling sem tudi članica uredniškega odbora in recenzentka elektronske jezikoslovne revije Slovenščina 2.0. Redno recenziram tudi za revije Journal of Language Resources and Evaluation, International Journal of Corpus Linguistics in Informatica. Vsa ta leta recenziram tudi za številne mednarodne conference: ACL, EACL, RANLP, TSD, *SEM, GWA, LTC, IS-LTC, eLEX in ITI. Bila sem predsednica organizacijskega obora poletne šole ESSLLI leta 2011, ki se je je udeležilo preko 400 udeležencev ter 80 predavateljev. Sodelovala sem pri organizaciji mednarodne conference eLEC 2011 ter poletne sole TransTech 2013.

10.Pomen raziskovalnih rezultatov projektne skupine[9]

10.1.Pomen za razvoj znanosti[10]

SLO

Glede na to, da na primerljivih korpusih na področju slovenske leksikalne semantike še ni bila opravljena nobena raziskava, pričujoči projekt nedvomno predstavlja pomemben mejnik v slovenskih korpusnih, pa tudi jezikovnotehnoloških raziskavah. Pomen projekta za razvoj znanosti oziroma stroke je dvojen:
(1) projekt je prinesel utemeljeno, preizkušeno in jezikovno-neodvisno metodologijo za luščenje prevodnih ustreznic iz primerljivih korpusov;
(2) zelo oprijemljiv rezultat projekta pa je tudi težko pričakovan semantični leksikon za slovenščino. Ker je izdelan wordnet poravnan z wordneti za številne druge jezike, je tako uporaben tako za eno- kot tudi za večjezične računalniške aplikacije. Izdelan wordnet je s tem zapolnil vrzel v jezikovnih virih za slovenščino in postavil temelje za širšo, semantično obogateno izrabo slovenskih korpusnih virov.
Najpomembnejše ugotovitve opravljene raziskave so:
- za uspešno luščenje prevodnih ustreznic ne potrebujemo primerljivih korpusov, ki so posebej za to nalogo izdelani po strogih kriterijih primerljivosti, temveč zadoščajo obsežni spletni korpusi, ki so za številne jezike že zgrajeni oz. je njihova gradnja precej enostavnejša;
- za uspešno luščenje prevodnih ustreznic ne potrebujemo enake količine podatkov za oba jezika, kar je v praksi razen za nekaj svetovnih jezikov težko doseči, temveč je s pazljivo izbiro statističnih mer za primerjavo kontekstnih vektorjev izvedljiv tudi scenarij, ko imamo za en jezik (npr. angleščino) na voljo bistveno večjo količino jezikovnih podatkov kot za drugega (npr. slovenščino).

- za uspešno luščenje prevodnih ustreznic med sorodnimi jeziki ne potrebujemo izhodiščnega slovarja, nedostopnost katerega je pogost problem, temveč ga lahko z upoštevanjem leksikalnega prekrivanja in drugih podobnosti med jezikoma izluščimo neposredno iz korpusa;
- z upoštevanjem osnovne predpostavke distribucijske semantike, ki se glasi, da se besede s podobnim pomenom v različnih jezikih pojavljajo v podobnih kontekstih, ni mogoče avtomatsko identificirati zgolj prevodnih ustreznic, temveč tudi lažne prijatelje, ki so si na videz sicer zelo podobni, vendar so njihove korpusne frekvence in kontekstni vektorji zelo različni. V skladu s to predpostavko smo razvili učinkovito metodo za avtomatsko prepoznavanje lažnih prijateljev, ki je zelo koristna za jezikovnotehnološke aplikacije, pa tudi v leksikografiji in pri poučevanju tujih jezikov.
Vsi leksikalni viri in orodja, razviti v okviru podoktorske raziskave, so pod licenco Creative Commons prosto dostopni v raziskovalne namene in bodo nedvomno zanimivi tudi za raziskovalce zunaj domače ustanove.

ANG

Since there has been no research into comparable corpora in the field of Slovene lexical semantics, the completed project presents a major milestone in Slovene corpus linguistics as well natural language processing. The importance of the results of the proposed project is two-fold:
(1) the project resulted in an established, tested and language-independent methodology for the extraction of translation equivalents from comparable corpora; and
(2) the semantic lexicon for Slovene has been released. It is aligned to wordnets in many other languages and is therefore useful for mono- as well as multilingual applications. In this way, the developed wordnet is narrowing the gap in the field of Slovene language resources and provide the foundations for a broader, semantically-enriched exploitation of Slovene corpus resources.
The key findings based on the completed research are:
- For successful extraction of translation equivalents, custom-made comparable corpora are not required. Instead, web corpora, which already exist for a number of languages or can be build relatively easily, suffice.
- For successful extraction of translation equivalents, the same amount of data for both languages, which is extremely difficult to achieve for most languages, is not required. Instead, given a careful selection of statistical similarity measures, substantially more data may be available for one language (e.g. English) than for the other (e.g. Slovene) with equal success.
- For successful extraction of translation equivalents from closely related languages a seed dictionary, which is unavailable for many language pairs, is not required. Instead, it can be induced directly from the comparable corpus by taking into account lexical overlap and other similarities between the two languages.
- According to the main premise of distributional semantics, words with a similar meaning appear in similar contexts. This allows us not only to identify translation equivalents, but also false friends, which are orthographically very similar in both languages but normally have very different frequencies in corpora and very different context vectors. Based on this principle we have also developed an efficient methodology for automatic identification of false friends, which are useful in human language technologies as well as in lexicography and language pedagogy.
The impact of the first semantic lexicon for Slovene, developed within the proposed post-doctoral project, extends beyond the host institution, since the all the created resources are freely available for all other researchers under the Creative Commons licence which has so far not been common practice in Slovenia.

10.2.Pomen za razvoj Slovenije[11]

SLO

Semantične leksikone tipa wordnet so doslej uporabili v številne namene in se z njihovo pomočjo lotili reševanja zelo različnih nalog tako raziskovalci kot tudi industrijski uporabniki. Med industrijskimi uporabniki, ki wordnet s pridom izkoriščajo za spletno iskanje in ciljno oglaševanje, je najvidnejši predstavnik Google. Njihovemu zgledu bi lahko sledilo tudi katero slovensko podjetje, ki se ukvarja s ponujanjem spletnih storitev.
Neposredna možnost uporabe wordneta, razvitega v okviru predlaganega projekta, bi bila ponudba aplikacije za avtomatsko razreševanje večpomenskosti in pojmovno zasnovanih spletnih storitev, kot so (medjezično) iskanje informacij, odgovarjanje na vprašanja in strojno prevajanje. Tovrstne storitve bi v družbi, temelječi na znanju, ki ima visoko razvito

informacijsko tehnologijo, bile zelo dobrodošle, saj slovenščina v primerjavi z drugimi evropskimi jeziki, na tem področju še precej zaostaja.

ANG

Semantic lexicons such as wordnet have been used in several different tasks and for several different purposes by researchers in the academia as well as by the industry. Among the companies to have gained the most advantage with the use of wordnet is Google. They have used wordnet for refining web searching and targeted advertising, and example that could well be followed by any Slovene company that offers web services.
A direct way to use the developed wordnet is its integration into an application for automatic word-sense disambiguation and semantically-aware web services, such as (cross-language) information retrieval, question-answering and machine translation. Such applications would be highly welcome in a society, based on knowledge and with highly evolved information technology because Slovene language is still lagging behind its other European counterparts in this field.

11. Samo za aplikativne projekte in podoktorske projekte iz gospodarstva! Označite, katerega od navedenih ciljev ste si zastavili pri projektu, katere konkretne rezultate ste dosegli in v kakšni meri so doseženi rezultati uporabljeni

Cilj		
F.01	**Pridobitev novih praktičnih znanj, informacij in veščin**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.02	**Pridobitev novih znanstvenih spoznanj**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.03	**Večja usposobljenost raziskovalno-razvojnega osebja**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.04	**Dvig tehnološke ravni**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.05	**Sposobnost za začetek novega tehnološkega razvoja**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.06	**Razvoj novega izdelka**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	

	Uporaba rezultatov	
F.07	**Izboljšanje obstoječega izdelka**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.08	**Razvoj in izdelava prototipa**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.09	**Razvoj novega tehnološkega procesa oz. tehnologije**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.10	**Izboljšanje obstoječega tehnološkega procesa oz. tehnologije**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.11	**Razvoj nove storitve**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.12	**Izboljšanje obstoječe storitve**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.13	**Razvoj novih proizvodnih metod in instrumentov oz. proizvodnih procesov**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.14	**Izboljšanje obstoječih proizvodnih metod in instrumentov oz. proizvodnih procesov**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.15	**Razvoj novega informacijskega sistema/podatkovnih baz**	
	Zastavljen cilj	◯ DA ◯ NE

	Rezultat	
	Uporaba rezultatov	
F.16	**Izboljšanje obstoječega informacijskega sistema/podatkovnih baz**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.17	**Prenos obstoječih tehnologij, znanj, metod in postopkov v prakso**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.18	**Posredovanje novih znanj neposrednim uporabnikom (seminarji, forumi, konference)**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.19	**Znanje, ki vodi k ustanovitvi novega podjetja ("spin off")**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.20	**Ustanovitev novega podjetja ("spin off")**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.21	**Razvoj novih zdravstvenih/diagnostičnih metod/postopkov**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.22	**Izboljšanje obstoječih zdravstvenih/diagnostičnih metod/postopkov**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.23	**Razvoj novih sistemskih, normativnih, programskih in metodoloških rešitev**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	

F.24	**Izboljšanje obstoječih sistemskih, normativnih, programskih in metodoloških rešitev**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.25	**Razvoj novih organizacijskih in upravljavskih rešitev**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.26	**Izboljšanje obstoječih organizacijskih in upravljavskih rešitev**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.27	**Prispevek k ohranjanju/varovanje naravne in kulturne dediščine**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.28	**Priprava/organizacija razstave**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.29	**Prispevek k razvoju nacionalne kulturne identitete**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.30	**Strokovna ocena stanja**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.31	**Razvoj standardov**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.32	**Mednarodni patent**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	

	Uporaba rezultatov	
F.33	**Patent v Sloveniji**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.34	**Svetovalna dejavnost**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	
F.35	**Drugo**	
	Zastavljen cilj	◯ DA ◯ NE
	Rezultat	
	Uporaba rezultatov	

Komentar

12. Samo za aplikativne projekte in podoktorske projekte iz gospodarstva!
Označite potencialne vplive oziroma učinke vaših rezultatov na navedena področja

	Vpliv	**Ni vpliva**	**Majhen vpliv**	**Srednji vpliv**	**Velik vpliv**	
G.01	**Razvoj visokošolskega izobraževanja**					
G.01.01.	Razvoj dodiplomskega izobraževanja	◯	◯	◯	◯	
G.01.02.	Razvoj podiplomskega izobraževanja	◯	◯	◯	◯	
G.01.03.	Drugo:	◯	◯	◯	◯	
G.02	**Gospodarski razvoj**					
G.02.01	Razširitev ponudbe novih izdelkov/storitev na trgu	◯	◯	◯	◯	
G.02.02.	Širitev obstoječih trgov	◯	◯	◯	◯	
G.02.03.	Znižanje stroškov proizvodnje	◯	◯	◯	◯	
G.02.04.	Zmanjšanje porabe materialov in energije	◯	◯	◯	◯	
G.02.05.	Razširitev področja dejavnosti	◯	◯	◯	◯	
G.02.06.	Večja konkurenčna sposobnost	◯	◯	◯	◯	
G.02.07.	Večji delež izvoza	◯	◯	◯	◯	
G.02.08.	Povečanje dobička	◯	◯	◯	◯	
G.02.09.	Nova delovna mesta	◯	◯	◯	◯	
G.02.10.	Dvig izobrazbene strukture zaposlenih	◯	◯	◯	◯	
G.02.11.	Nov investicijski zagon	◯	◯	◯	◯	

G.02.12.	Drugo:					
G.03	**Tehnološki razvoj**					
G.03.01.	Tehnološka razširitev/posodobitev dejavnosti					
G.03.02.	Tehnološko prestrukturiranje dejavnosti					
G.03.03.	Uvajanje novih tehnologij					
G.03.04.	Drugo:					
G.04	**Družbeni razvoj**					
G.04.01	Dvig kvalitete življenja					
G.04.02.	Izboljšanje vodenja in upravljanja					
G.04.03.	Izboljšanje delovanja administracije in javne uprave					
G.04.04.	Razvoj socialnih dejavnosti					
G.04.05.	Razvoj civilne družbe					
G.04.06.	Drugo:					
G.05.	**Ohranjanje in razvoj nacionalne naravne in kulturne dediščine in identitete**					
G.06.	**Varovanje okolja in trajnostni razvoj**					
G.07	**Razvoj družbene infrastrukture**					
G.07.01.	Informacijsko-komunikacijska infrastruktura					
G.07.02.	Prometna infrastruktura					
G.07.03.	Energetska infrastruktura					
G.07.04.	Drugo:					
G.08.	**Varovanje zdravja in razvoj zdravstvenega varstva**					
G.09.	**Drugo:**					

Komentar

13.Pomen raziskovanja za sofinancerje[12]

	Sofinancer			
1.	Naziv			
	Naslov			
	Vrednost sofinanciranja za celotno obdobje trajanja projekta je znašala:			EUR
	Odstotek od utemeljenih stroškov projekta:			%
	Najpomembnejši rezultati raziskovanja za sofinancerja			Šifra
		1.		
		2.		

		3.		
		4.		
		5.		
	Komentar			
	Ocena			

14.Izjemni dosežek v letu 2012[13]

14.1. Izjemni znanstveni dosežek

FIŠER, Darja. Language resources and tools for semantically enhanced processing of Slovene. V: VERTAN, Cristina (ur.), HAHN, Walther von (ur.). Multilingual processing in eastern and southern EU languages : low-resourced technologies in translation. [S.l.]: Cambridge Scholars, 2012, str. 92-118.
V poglavju predstavimo metode, jezikovne vire in orodja, ki smo jih razvili za semantično obogateno procesiranje slovenščine. Za izdelavo 1. semantičnega leksikona za slovenščino smo uporabili heterogene večjezične jezikovne vire. S pomočjo izdelanega wordneta smo 1. za slovenščino semantično označili korpus. Razvili smo tudi orodje za pregledovanje in vizualizacijo semantične mreže, s katerim je mogoče proučevati leksikalno-semantične lastnosti slovenščine v primerjavi z drugimi jeziki, wordnet pa smo prav tako uporabili za izboljšanje strojnega prevajanja večpomenskih besed. Zadnja različica slovenskega wordneta vsebuje 42722 sinsetov, od katerih smo jih 25 % že ročno pregledali in popravili.

14.2. Izjemni družbeno-ekonomski dosežek

TAVČAR, Aleš, FIŠER, Darja, ERJAVEC, Tomaž. sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. V: ERJAVEC, Tomaž (ur.), ŽGANEC GROS, Jerneja (ur.). Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012, [Ljubljana, Slovenia] : zbornik 15. mednarodne multikonference Informacijska družba - IS 2012, zvezek C : proceedings of the 15th International Multiconference Information Society - IS 2012, volume C, (Informacijska družba). Ljubljana: Institut Jožef Stefan, 2012, str. 197-202.
Za hitrejši razvoj leksikalnih virov smo izdelali orodje sloWCrowd, ki je zasnovano tako, da nam odgovore omogoča zbirati iz široke množice uporabnikov. Zanesljivost zbranih rezultatov dosežemo z večkratnim ponavljanjem istega vprašanja pri različnih uporabnikih, odgovore uporabnikov z nizko stopnjo natančnosti glede na zlati standard pa izločimo. Trenutno št. registriranih uporabnikov orodja je 280, doslej pa smo zbrali 30000 odgovorov za 6500 besed.

C. IZJAVE

Podpisani izjavljam/o, da:

- so vsi podatki, ki jih navajamo v poročilu, resnični in točni
- se strinjamo z obdelavo podatkov v skladu z zakonodajo o varstvu osebnih podatkov za potrebe ocenjevanja ter obdelavo teh podatkov za evidence ARRS
- so vsi podatki v obrazcu v elektronski obliki identični podatkom v obrazcu v pisni obliki
- so z vsebino zaključnega poročila seznanjeni in se strinjajo vsi soizvajalci projekta

Podpisi:

zastopnik oz. pooblaščena oseba raziskovalne organizacije:	in	*vodja raziskovalnega projekta:*
Univerza v Ljubljani, Filozofska fakulteta		Darja Fišer
_____		_____

ŽIG

Kraj in datum:	Ljubljana	14.3.2013

Oznaka prijave: ARRS-RPROJ-ZP-2013/266

---

[1] Opredelite raziskovalno področje po klasifikaciji FOS 2007 (Fields of Science). Prevajalna tabela med raziskovalnimi področji po klasifikaciji ARRS ter po klasifikaciji FOS 2007 (Fields of Science) s kategorijami WOS (Web of Science) kot podpodročji je dostopna na spletni strani agencije (http://www.arrs.gov.si/sl/gradivo/sifranti/preslik-vpp-fos-wos.asp). Nazaj

[2] Napišite povzetek raziskovalnega projekta (največ 3.000 znakov v slovenskem in angleškem jeziku) Nazaj

[3] Napišite kratko vsebinsko poročilo, kjer boste predstavili raziskovalno hipotezo in opis raziskovanja. Navedite ključne ugotovitve, znanstvena spoznanja, rezultate in učinke raziskovalnega projekta in njihovo uporabo ter sodelovanje s tujimi partnerji. Največ 12.000 znakov vključno s presledki (približno dve strani, velikost pisave 11). Nazaj

[4] Realizacija raziskovalne hipoteze. Največ 3.000 znakov vključno s presledki (približno pol strani, velikost pisave 11) Nazaj

[5] V primeru bistvenih odstopanj in sprememb od predvidenega programa raziskovalnega projekta, kot je bil zapisan v predlogu raziskovalnega projekta oziroma v primeru sprememb, povečanja ali zmanjšanja sestave projektne skupine v zadnjem letu izvajanja projekta, napišite obrazložitev. V primeru, da sprememb ni bilo, to navedite. Največ 6.000 znakov vključno s presledki (približno ena stran, velikost pisave 11). Nazaj

[6] Navedite znanstvene dosežke, ki so nastali v okviru tega projekta. Raziskovalni dosežek iz obdobja izvajanja projekta (do oddaje zaključnega poročila) vpišete tako, da izpolnite COBISS kodo dosežka – sistem nato sam izpolni naslov objave, naziv, IF in srednjo vrednost revije, naziv FOS področja ter podatek, ali je dosežek uvrščen v A'' ali A'. Nazaj

[7] Navedite družbeno-ekonomske dosežke, ki so nastali v okviru tega projekta. Družbeno-ekonomski rezultat iz obdobja izvajanja projekta (do oddaje zaključnega poročila) vpišete tako, da izpolnite COBISS kodo dosežka – sistem nato sam izpolni naslov objave, naziv, IF in srednjo vrednost revije, naziv FOS področja ter podatek, ali je dosežek uvrščen v A'' ali A'.

Družbeno-ekonomski dosežek je po svoji strukturi drugačen kot znanstveni dosežek. Povzetek znanstvenega dosežka je praviloma povzetek bibliografske enote (članka, knjige), v kateri je dosežek objavljen.

Povzetek družbeno-ekonomskega dosežka praviloma ni povzetek bibliografske enote, ki ta dosežek dokumentira, ker je dosežek sklop več rezultatov raziskovanja, ki je lahko dokumentiran v različnih bibliografskih enotah. COBISS ID zato ni enoznačen, izjemoma pa ga lahko tudi ni (npr. prehod mlajših sodelavcev v gospodarstvo na pomembnih raziskovalnih nalogah, ali ustanovitev podjetja kot rezultat projekta … - v obeh primerih ni COBISS ID). Nazaj

[8] Navedite rezultate raziskovalnega projekta iz obdobja izvajanja projekta (do oddaje zaključnega poročila) v primeru, da katerega od rezultatov ni mogoče navesti v točkah 7 in 8 (npr. ker se ga v sistemu COBISS ne vodi). Največ 2.000 znakov, vključno s presledki. Nazaj

[9] Pomen raziskovalnih rezultatov za razvoj znanosti in za razvoj Slovenije bo objavljen na spletni strani: http://sicris.izum.si/ za posamezen projekt, ki je predmet poročanja Nazaj

[10] Največ 4.000 znakov, vključno s presledki Nazaj

[11] Največ 4.000 znakov, vključno s presledki Nazaj

[12] Rubrike izpolnite / prepišite skladno z obrazcem "izjava sofinancerja" http://www.arrs.gov.si/sl/progproj/rproj/gradivo/, ki ga mora izpolniti sofinancer. Podpisan obrazec "Izjava sofinancerja" pridobi in hrani nosilna raziskovalna organizacija – izvajalka projekta. Nazaj

[13] Navedite en izjemni znanstveni dosežek in/ali en izjemni družbeno-ekonomski dosežek raziskovalnega projekta v letu 2012 (največ 1000 znakov, vključno s presledki). Za dosežek pripravite diapozitiv, ki vsebuje sliko ali drugo slikovno gradivo v zvezi z izjemnim dosežkom (velikost pisave najmanj 16, približno pol strani) in opis izjemnega dosežka (velikost pisave 12, približno pol strani). Diapozitiv/-a priložite kot priponko/-i k temu poročilu. Vzorec diapozitiva je objavljen na spletni strani ARRS http://www.arrs.gov.si/sl/gradivo/, predstavitve dosežkov za pretekla leta pa so objavljena na spletni strani http://www.arrs.gov.si/sl/analize/dosez/. Nazaj

Obrazec: ARRS-RPROJ-ZP/2013 v1.00
71-E3-4D-5B-86-02-77-06-DB-B0-C3-B8-9C-A7-D6-29-CD-AD-F6-1A

APIDIANAKI, Marianna, LJUBEŠIĆ, Nikola, FIŠER, Darja. Disambiguating vectors for bilingual lexicon extraction from comparable corpora. V: ERJAVEC, Tomaž (ur.), ŽGANEC GROS, Jerneja (ur.). *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012.*

FIŠER, Darja, BRODA, Bartosz, PIASECKI, Maciej. Weaving sloWNet using window-based co-occurence features. V: ERJAVEC, Tomaž (ur.), ŽGANEC GROS, Jerneja (ur.). Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012.

FIŠER, Darja, GANTAR, Polona, KREK, Simon. Using explicitly and implicitly encoded semantic relations to map Slovene wordnet and Slovene lexical database. V: 8th International Conference on Language Resources and Evaluation, 21-27 May 2012, Istanbul, Turkey.

FIŠER, Darja, LJUBEŠIĆ, Nikola, KUBELKA, Ozren. Addressing polysemy in bilingual lexicon extraction from comparable corpora. V: 8th International Conference on Language Resources and Evaluation, 21-27 May 2012, Istanbul, Turkey.

FIŠER, Darja, LJUBEŠIĆ, Nikola, VINTAR, Špela, POLLAK, Senja. Building and using comparable corpora for domain-specific bilingual lexicon extraction. V: 4th Workshop on BUCC, 24 June, 2011, Portland, Oregon. *Comparable Corpora and the Web : proceedings of the workshop*. Portland: Association for Computational Linguistics, 2011.

FIŠER, Darja, LJUBEŠIĆ, Nikola. Avtomatsko luščenje hrvaško-slovenskega leksikona iz primerljivih korpusov. V: KRANJC, Simona (ur.). *Meddisciplinarnost v slovenistiki*, (Obdobja, Simpozij, = Symposium, 30). 1. natis. Ljubljana: Znanstvena založba Filozofske fakultete, 2011.

FIŠER, Darja, LJUBEŠIĆ, Nikola. Best friends or just faking it? Corpus-based extraction of Slovene-Croatian translation equivalents and false friends (v tisku).

FIŠER, Darja, NOVAK, Jernej. Visualising sloWNet. V: KOSEM, Iztok (ur.), KOSEM, Karmen (ur.). *Electronic lexicography in the 21st century : new applications for new users : proceedings of eLex 2011, 10-12 November 2011, Bled, Slovenia*. Ljubljana: Trojina, Institute for Applied Slovene Studies, 2011.

FIŠER, Darja, POLLAK, Senja, VINTAR, Špela. Samodejno luščenje definicij iz specializiranih besedil. V: KRANJC, Simona (ur.). *Meddisciplinarnost v slovenistiki*, (Obdobja, Simpozij, = Symposium, 30). 1. natis. Ljubljana: Znanstvena založba Filozofske fakultete, 2011, str. 145-150.

FIŠER, Darja. Language resources and tools for semantically enhanced processing of Slovene. V: VERTAN, Cristina (ur.), HAHN, Walther von (ur.). *Multilingual processing in eastern and southern EU languages : low-resourced technologies in translation*. [S.l.]: Cambridge Scholars, 2012, str. 92-118.

FIŠER, Darja. Semantično označevanje korpusov. V: VINTAR, Špela (ur.). *Slovenske korpusne raziskave*, (Zbirka Prevodoslovje in uporabno jezikoslovje). 1. natis. Ljubljana: Znanstvena založba Filozofske fakultete, 2010, str. 110-130.

LJUBEŠIĆ, Nikola, FIŠER, Darja, VINTAR, Špela, POLLAK, Senja. Bilingual lexicon extraction from comparable corpora : a comparative study. V: First International Workshop on Lexical Resources, An ESSLLI 2011 Workshop, Ljubljana, Slovenia - August 1-5, 2011.

LJUBEŠIĆ, Nikola, VINTAR, Špela, FIŠER, Darja. Multi-word term extraction from comparable corpora by combining contextual and constituent clues. V: 8th International Conference on Language Resources and Evaluation, 21-27 May 2012, Istanbul, Turkey.

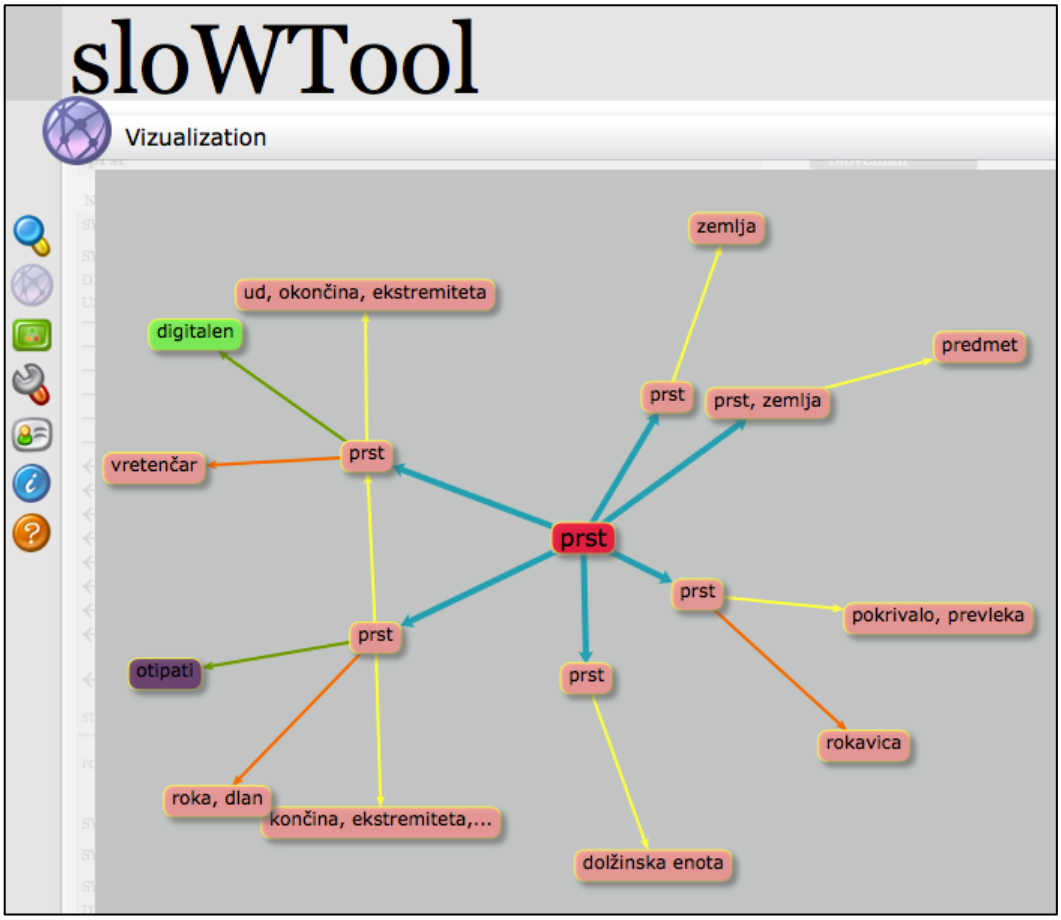
TAVČAR, Aleš, FIŠER, Darja, ERJAVEC, Tomaž. sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. V: ERJAVEC, Tomaž (ur.), ŽGANEC GROS, Jerneja (ur.). *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012.*

VINTAR, Špela, FIŠER, Darja, VRŠČAJ, Aljoša. Were the clocks striking or surprising? Using WSD to improve MT performance. V: 13th Conference on the European Chapter of the Association for Computational Linguistics, Avignon, April 23-17 2012. *Proccedings of the workshop*. Avignon: ACL, 2012, str. 87-92.

VEDA: Humanistične vede
Področje: 6.05.00 - Jezikoslovje

Izjemni znanstveni dosežek: **Razvoj virov in orodij za semantično obogateno procesiranje slovenščine**

Vir: FIŠER, Darja. Language resources and tools for semantically enhanced processing of Slovene. V: VERTAN, Cristina (ur.), HAHN, Walther von (ur.). Multilingual processing in eastern and southern EU languages : low-resourced technologies in translation. [S.l.]: Cambridge Scholars, 2012, str. 92-118.



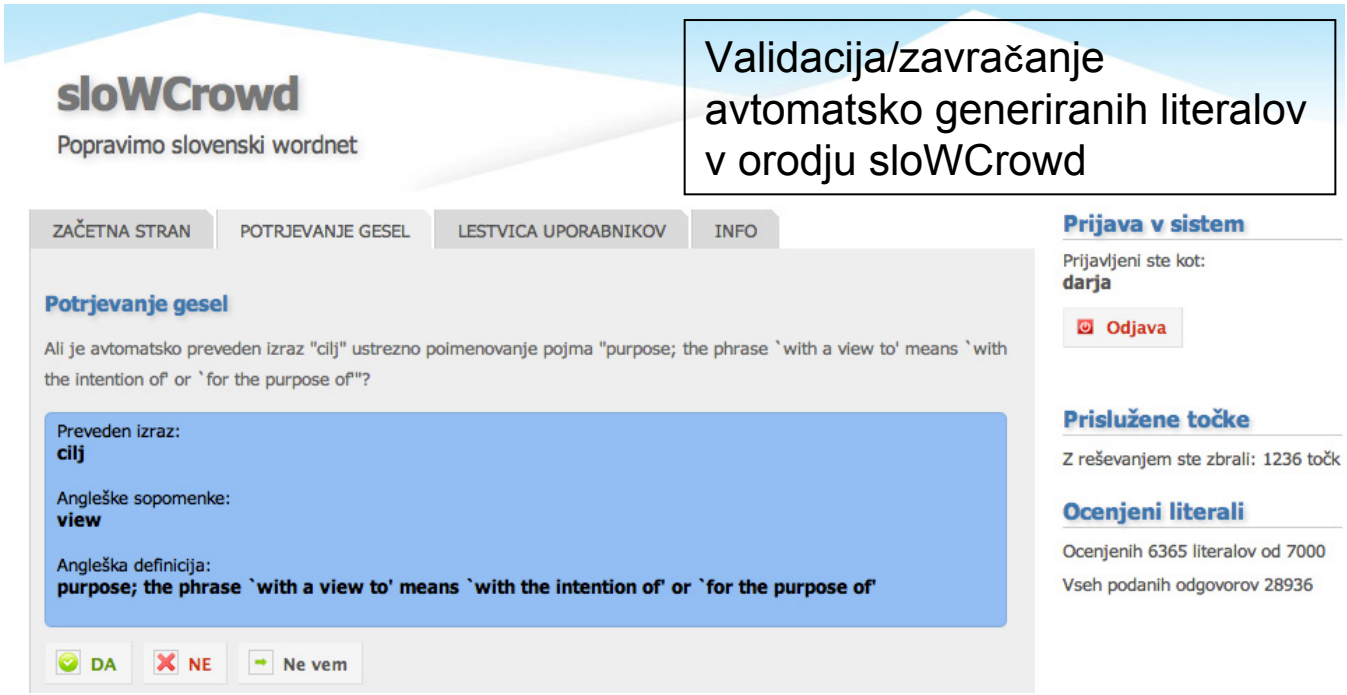
Vizualizacija semantične mreže za različne pomene besede “prst” v slovenskem wordnetu

V poglavju predstavimo metode, jezikovne vire in orodja, ki smo jih razvili za semantično obogateno procesiranje slovenščine. Za izdelavo prvega semantičnega leksikona za slovenščino smo uporabili heterogene večjezične jezikovne vire, kot so slovarji, enciklopedije, vzporedni in primerljivi korpusi. S pomočjo izdelanega wordneta smo semantično označili dele korpusov JOS in SPOOK, ki sta tako prva korpusa za slovenščino, označena na pomenski ravni. Poleg virov smo razvili tudi orodje za iskanje po wordnetu in vizualizacijo semantične mreže, s katerim je mogoče proučevati leksikalno-semantične lastnosti slovenščine v primerjavi z drugimi jeziki, izdelan wordnet pa smo prav tako uporabili za izboljšanje strojnega prevajanja večpomenskih besed. Zadnja različica slovenskega wordneta vsebuje 42722 množic sinonimov, od katerih smo jih skoraj četrtino (9626) že tudi ročno pregledali in popravili.

VEDA: Humanistične vede
Področje: 6.05.00 - Jezikoslovje

Izjemni družbeno-ekonomski dosežek: **Čiščenje leksikalnih virov z množicanjem**

Vir: TAVČAR, Aleš, FIŠER, Darja, ERJAVEC, Tomaž. sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. V: ERJAVEC, Tomaž (ur.), ŽGANEC GROS, Jerneja (ur.). Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012, [Ljubljana, Slovenia] : zbornik 15. mednarodne multikonference Informacijska družba - IS 2012, zvezek C : proceedings of the 15th International Multiconference Information Society - IS 2012, volume C, (Informacijska družba). Ljubljana: Institut Jožef Stefan, 2012, str. 197-202.



Za hitrejše in cenejše čiščenje leksikalnih virov z množicanjem smo razvili orodje sloWCrowd, ki je zasnovano tako, da nam odgovore na leksikalno-semantična vpračanja omogoča zbirati iz široke množice uporabnikov. Orodje je prosto dostopno in temelji na razširjenih tehnologijah, kot sta PHP in MySQL. Zanesljivost zbranih rezultatov dosežemo z večkratnim ponavljanjem istega vprašanja pri različnih uporabnikih, odgovore uporabnikov z nizko stopnjo natančnosti glede na zlati standard pa izločimo. Trenutno število registriranih uporabnikov orodja je 280, doslej pa smo zbrali skoraj 30000 odgovorov za okoli 6500 besed. Dodatna prednost razvitega orodja je, da ga je mogoče prilagoditi za najrazličnejše naloge, kar že preizkušamo v sodelovanju z inštitutom INRIA iz Francije za popravljanje francoskega wordneta WOLF in s sodelavci projekta Sporazumevanje v slovenskem jeziku za Slovensko leksikalno bazo.