SISTEM ZA RAZREŠEVANJE KOREFERENC PRI ANALIZI SLOVENSKIH BESEDIL IN MOŽNOSTI NJEGOVE UPORABE

Peter HOLOZAN
Amebis, d. o. o., Kamnik

Holozan, P. (2015): Sistem za razreševanje koreferenc pri analizi slovenskih besedil in možnosti njegove uporabe. Slovenščina 2.0, 3 (1): 60–89.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2015/1/Slo2.0_2015_1_04.pdf.

Razreševanje koreferenc je pomemben del jezikovnih tehnologij, vendar za slovenščino ta tehnologija še ni bila razvita. Obstajajo različne vrste koreferenc, članek se osredotoča predvsem na anafore pri osebnih zaimkih. Uporabljenih je bilo sedem metod razreševanja, ki se med seboj dopolnjujejo, najpomembnejša temelji na metodah na osnovi aktivacije. Prvi rezultati so obetavni, za podrobnejšo analizo delovanja pa bo potreben korpus z označenimi primeri.

Razreševanje koreferenc je bilo uporabljeno tudi v sistemu za odgovarjanje na vprašanja Piflar, ki zna s tem odgovoriti na več vprašanj, ker mu uspe nadomestiti osebne zaimke, hkrati pa je bil Piflar dopolnjen še z drugimi dodatki, npr. z odgovarjanjem na posamične stavčne člene in na trdilne povedi, izboljšano pa je bilo tudi tvorjenje dolgih odgovorov pri odločevalnih vprašanjih.

Razreševanje koreferenc je izboljšalo tudi delovanje strojnega prevajalnika Presis, in sicer pri določanju spola osebnih zaimkov in pri razdvoumljanju prilastkovih odvisnikov.

Ključne besede: razreševanje koreferenc, odkrivanje koreferenčnosti, anafora, odgovarjanje na vprašanja

1 UVOD[1]

Razreševanje koreferenc[2] (oz. odkrivanje koreferenčnosti) je pomemben del jezikovnih tehnologij. Koreference lahko razumemo kot elemente kohezije besedila, ki je eden od sedmih kriterijev besedilnosti in ki pove, kako so sestavine površinskega besedila med seboj povezane (de Beaugrande, Dressler 1992; prim. tudi Zuljan Kumar 2007; Zuljan Kumar 2010; Bucik 2001). Koreference lahko razdelimo na anafore, ki jih lahko interpretiramo z nečim, kar je v besedilu že bilo navedeno, in katafore, pri katerih je interpretacija pogojena z nečim, kar je v besedilo uvedeno kasneje (Gorjanc 1999; Korošec 1981). Tipičen primer koreferenc so zaimki, niso pa koreference omejene le na zaimke, čeprav se največkrat omejimo le nanje. Širši pregled slovenskega besediloslovja, katerega del so tudi koreference, je v Korošec (2006), novejša raziskava, ki poleg slovenščine obravnava še hrvaščino in uporablja korpusni pristop, pa je v Balažic Bulc, Gorjanc (2015).

Razreševanje koreferenc lahko pomaga tudi pri razdvoumljanju besedil, saj v precej primerih ne moremo razdvoumljati brez razrešenih koreferenc, po drugi strani pa iz tega sledi, da pri razreševanju koreferenc ne moremo izhajati iz že razdvoumljenega vhodnega besedila, ampak se morata razdvoumljanje in razreševanje koreferenc dopolnjevati (McShane, Beale, Nirenburg 2010). Tak primer sta npr. povedi »Miha je videl matico, ki jo je privil Janez.« in »Miha je videl matico, ki jo je vzgojil Janez.«, kjer je pomen večpomenske besede »matica«, v drugem delu povedi pozaimljene v »jo«, razdvoumljen s povedkom v prilastkovem odvisniku.

Ker je anaforična raba zaimkov pogostejša kot kataforična, je več raziskav posvečenih razreševanju prve (npr. Mitkov 1999; Němčík 2006).

[1] Članek je dopolnjena različica prispevka Holozan (2014b).
[2] V Holozan (2014b) je bil uporabljen izraz sklic, vendar se je pokazalo, da je v besediloslovju že uveljavljen termin koreferenca, zato je v tem članku uporabljen ta izraz. Razmislek je bil še o terminu referenca (tudi v angleščini se uporabljata termina reference resolution in co-reference resolution), vendar je tudi to presplošen izraz in je ustreznejša koreferenca v pomenu besedilne reference.

Za slovenščino programski sistem za avtomatsko razreševanje koreferenc še ni bil narejen, zato je smiselno preizkusiti, kako uspešno se da tak sistem vgraditi v analizator, ki prevaja naravni jezik v Amebisov vmesni jezik, katerega podrobni opis je v točki 6.2 v Holozan (2011). Ta vmesni jezik uporabljajo mnogi izdelki podjetja Amebis, npr. strojni prevajalnik Presis in sistem za odgovarjanje Piflar, kar pomeni, da bo ta izboljšava vplivala tudi nanje, če dodamo možnost, da uporabijo to dodatno informacijo.

2 KOREFERENČNOST: OSEBNI ZAIMKI

V raziskavi smo se osredotočili le na osebne zaimke. V Gigafidi predstavljajo osebni zaimki 1,4 % vseh pojavnic, vendar je treba upoštevati še to, da je v slovenščini osebek pogosto le nakazan z glagolsko obliko (Toporišič 2004: 607) in torej pozaimljanje posledično ni tako pogosto kot npr. v angleščini (test je pokazal, da je v povedih iz korpusa jos100k (Erjavec, Krek 2008), ki jih je analizatorju uspelo analizirati, izpuščenih osebnih zaimkov več kot dvakrat toliko kot neizpuščenih). Pri avtomatski analizi besedila je treba vseeno rekonstruirati tudi osebke oz. osebne zaimke, ki jih izraža končnica povedka, če želimo dobiti prave besedilne vezi in posledično pravi smisel stavkov, povedi ter celotnega besedila.

Slika 1 prikazuje primer besedila s koreferencami. Med poševnicami so dodani sicer izpuščeni osebni zaimki (da se pokaže delež elips, pri katerih je treba razrešiti koreference), odebeljeno pa so označene besedilne reference, ki jih je treba povezati. V istem primeru lahko v drugo gručo povežemo besedilne reference »avtodomu«, »tovorni avtomobil«, »ki«, »avtomobila«, »vozilo« in »njem«.

V avtodomu sta /**onadva/** prevažala opij

Policisti in kriminalisti so v sodelovanju s cariniki na avtocestnem počivališču v bližini Murske Sobote zaustavili tovorni avtomobil, ki je bil predelan v avtodom. Po pregledu avtomobila so /oni/ v njem našli 390 gramov opija, tri vrečke obrezanih makovih glavic in tri grame halucinogenih gobic. Vozilo je bilo registrirano v Franciji. V njem pa sta se vozila **Francoza**, stara 32 in 33 let, so sporočili /oni/ s policijske uprave v Murski Soboti.

Zaradi utemeljenega suma neupravičene proizvodnje in prometa s prepovedanimi drogami sta bila **tujca** s kazensko ovadbo privedena pred preiskovalnega sodnika okrožnega sodišča v Murski Soboti, ki je zoper **oba** odredil /***on***/ pripor.

Slika 1: Besedilo s primeri koreferenčnih osebnih zaimkov, ki jih mora razrešiti avtomatski prevajalnik.

Programi v slovenščini težje razrešujejo koreferenčnost, ki nima izraženega zaimka, kot to velja za angleščino, ker je v angleščini mogoče uporabiti podatek, da je pred samostalniško frazo določni člen, kar potem zoži koreferenčne možnosti in olajša iskanje besedilnih referenc. To pomeni, da za slovenščino ne moremo neposredno uporabiti metod za angleščino in da se moramo v slovenščini veliko bolj zanesti na pomene, delno pa tudi na besedni red.

Težava so tudi koreference, ki so v angleščini zapisani z besedo »one«, npr.: »If you cannot attend a tutorial in the morning, you can go for an afternoon one.« (Mitkov 1999). Slovenski prevod bi bil: »Če se ne moreš udeležiti vaj dopoldne, greš lahko na popoldanske.« V slovenščini tukaj ni posebne besede, na katero bi lahko vezali koreferenco, ampak bi tukaj lahko rekli, da gre za elipso besede »vaje« v drugem stavku. Podrobnejša slovensko-angleška kontrastivna analiza angleškega »one« je v Kocijančič Pokorn (1997).

Koreference lahko povezujejo več predhodnih besed v eno besedo ali obratno. V primeru »Srečal sem Johna in Mary. Bila sta zelo vesela, saj smo dobri prijatelji.« se John in Mary najprej povežeta v izpuščeni zaimek »onadva«, nato

pa še skupaj s pripovedovalcem (1. osebo) v izpuščeni zaimek »mi«. Obratno pa je v primeru »Starejši par je hodil po parku in moški se je nenadoma spotaknil.«, kjer je »moški« najverjetneje del »para« iz predhodnega stavka.

Razreševanje koreferenc je zelo odvisno od pomenov. Če uporabim primer iz Němčík (2006): »John je skril Billove ključe. Bil je pijan.«, se ljudem zdi najverjetnejša interpretacija, da se drugi stavek nanaša na Billa, ker pač sklepamo, da je vožnja pod vplivom alkohola nevarna in je Johna skrbelo za Billa, zato mu je skril ključe, da ne bi mogel odpeljati. Ni pa to edina možna interpretacija, morda je bil pijan John in je hotel nagajati Billu ter mu je zato skril ključe hiše. Taki primeri kažejo na to, da je razreševanje koreferenc res zahteven problem za računalnike.

3 NEKATERE METODE RAZREŠEVANJA KOREFERENC

Za razreševanje koreferenc je bila razvita množica metod in nekatere bodo na kratko predstavljene v nadaljevanju tega razdelka.

3.1 Hobbsovo sintaktično iskanje

Hobbsovo sintaktično iskanje (Hobbs 1978) je bila prva metoda, ki je uporabila jezikovno znanje in je kljub starosti in relativni preprostosti (že sam Hobbs je menil, da je to le naivna metoda) še vedno primerljivo uspešna v primerjavi z modernejšimi metodami (Němčík 2006).

Osnova za postopek je drevo izpeljav za poved. Hobbsovo iskanje določi vrstni red, v katerem samostalniške fraze postanejo kandidati za razreševanje koreferenc. V drevesu začnemo iskati levo od zaimka, za katerega želimo razrešiti koreferenco, potem pa se dvigamo in vsakič iščemo v širino od leve proti desni, pri kandidatih pa moramo preveriti, ali se ujemajo v številu, spolu, osebi in živosti.

Metodo lahko dopolnimo s pomenskimi omejitvami pri kandidatih.

Težava pri metodi je, da lahko vedno najdemo primere, v katerih ne deluje,

dodatno pa je izdelava drevesa izpeljav sama po sebi zapleten problem.

3.2 Algoritem BFP

Algoritem BFP (Brennan, Friedman, Pollard 1987) temelji na teoriji fokusa (ang. *centering theory*), ki je bila prvič opisana v (Joshi, Kuhn 1979). Ta opisuje, kako se spreminja fokus diskurza, ena od metod fokusiranja pa je tudi uporaba zaimkov, ki nas usmerjajo na fokus. Ta se lahko spreminja z različnimi vrstami prehodov.

Pokazalo pa je se, da razvoj v smeri vedno bolj kompleksnih pravil slepa ulica, ker ni bilo mogoče dovolj podrobno zajeti splošnega znanja in opisati jezika, zato so se metode usmerile v smeri, ki zahteva manj znanja (Němčík 2006).

3.3 Faktorji poudarka

Postopek s faktorji poudarka (ang. *salience factors*) je bil predlagan v (Lappin, Leass, 1994). Ti faktorji so uteži, ki so prirejene posamičnim možnostim koreferenc in potem kombinirane, da se določi najpomembnejši element diskurza. Dodatno postopek ugotavlja, kateri zaimki so del fraz in nimajo koreferenc (npr. »it« v »It's raining.«) in določa povratne zaimke (Němčík 2006).

Uteži je treba določiti z eksperimentiranjem, kar pomeni, da potrebujemo korpus primerov, da lahko avtomatsko preverjamo različne uteži.

3.4 Robustni sistemi z malo potrebnega znanja

Primer za tak sistem je MARS (Mitkov, Evans, Orasan 2002). Sistem temelji na množici predhodnostnih kazalnikov (ang. *set of antecedent indicators*). Vsak od njih opisuje določen pogoj, ki se nanaša na danega kandidata za koreferenco, in vpliv, ki ga ima na verjetnost, da je to verjetni izvor koreference (Němčík 2006).

Prednost te metode je, da ne potrebuje zunanjega skladenjskega razčlenjevalnika, za večino kazalnikov pa se zdi, da je jezikovno neodvisna, zato

je bila ta metoda uporabljena tudi za druge jezike, kot so francoščina, poljščina, arabščina in bolgarščina (Němčík 2006).

3.5 Statistične metode

Po letu 1990 so se za razreševanje koreferenc začele uporabljati tudi statistične metode (in tudi druge metode strojnega učenja). Primer je (Ge, Hale, Charniak 1998).

Zanimiv primer je tudi sistem SkipCor (Žitnik 2014), ki je sicer namenjen odkrivanju koreferenčnosti, kar je gručenje omenitev oz. združevanje omenitev, ki se sklicujejo na isto entiteto. Algoritem v sistemu SkipCor najprej identificira omenitve v vhodnem besedilu in jih pretvori v zaporedja izpuščenih omenitev. Nato za vsak tip zaporedja uporabi ustrezen linearno-verižni model CRF in vrne označena zaporedja, ki se uporabijo v postopku gručenja. Kot končni rezultat vrne seznam entitet, ki so predstavljene kot množice omenitev (Žitnik 2014).

Vendar vse te metode zahteva korpus učnih primerov, ki ga za slovenščino še nimamo, zato se za zdaj nismo usmerili v to smer, ampak raje v metode, ki ne potrebujejo korpusa učnih primerov.

4 UPORABLJENE METODE RAZREŠEVANJA

Ideja razreševanja koreferenc, ki jo opisujemo v nadaljevanju, je uporaba množice metod, od katerih vsaka razrešuje določene koreference, metode pa se uporabljajo od bolj proti manj zanesljivim (v tem vrstnem redu so tudi opisane, poudariti pa je treba, da je zanesljivost v tem trenutku le ocena, ki potrebuje še bolj temeljito preverjanje na večjem številu primerov).

Izbrane so bile metode, ki ne potrebujejo učnega korpusa, ker tega za slovenščino še ni. Obstaja pa po drugi strani možnost, da bi si lahko s temi za zdaj uporabljenimi metodami pomagali, da se naredi osnutek korpusa primerov koreferenc, ki se potem še ročno dopolni, da ni treba celotnega izdelati ročno.

Koreference so v vmesnem jeziku opisane z novim elementom ORI, ki je dodan

k obstoječemu elementu (največkrat je to osebni zaimek (OSZ) oz. navidezni osebni zaimek (tj. osebni zaimek, ki se skriva v osebni glagolski obliki; NOZ), lahko pa tudi drug samostalniški zaimek (SAZ) ali samostalnik (SAM)) v element JED (jedro dela samostalnike fraze). Element ORI vsebuje element SFR (samostalniško frazo). Slika 2 prikazuje primer, ko je koreferenca dodana osebnemu zaimku v vlogi osebka (element OSB).

(1OSB:(-SFR:(-DSF:(-JED:(-OSZnemt:[10])**,(-ORI:(-SFR:(-DSF:(-JED:(-SAMe:{7d62a7;4207ac9}[/]<dc>)))))**))))

Slika 2: Primer zapisa koreference v vmesnem jeziku.

Za preizkušanje so bili uporabljeni umetno skonstruirani primeri, pravljica Rdeča kapica, Cankarjev Na klancu, testno besedilo iz priročnika Pravipis Aleksandre Kocmut, Wikipedija, šala neznanega izvora in prispevek iz črne kronike.

4.1 Izpusti (elipse) osebka

To je vrsta koreferenc, ki jih je mogoče zelo zanesljivo razrešiti. Gre za zaporedna stavka, pri čemer je v drugem izpuščen osebek, tako da se uporabi kar osebek iz prvega stavka: »Miha je prišel do vrat in pozvonil.« V teh primerih se običajno izpusti še pomožni glagol, lahko pa tudi veznik: »Metka je rekla, da rada pleše in poje.«

4.2 Prilastkovi odvisniki

Tudi pri prilastkovih odvisnikih vemo, da se zaimek (»ki«, »kateri« ali pa naslonska oblika osebnega zaimka ob »ki«) v odvisniku nanaša na besedo, ki je jedro ob tem odvisniku. V primeru »Miha je videl sliko, ki jo je naslikal Janez.« tako vemo, da se »jo« nanaša na besedo »sliko«.

Težava lahko nastopi le v primerih, ko ni jasno, kaj je jedro: »Bila sta privedena pred preiskovalnega sodnika okrožnega sodišča v Kamniku, ki je zoper oba

odredil pripor.« V takih primerih se lahko zgodi, da analizator označi kot jedro »Kamnik«, kar morajo potem razrešiti pomenske omejitve.

4.3 Prva in druga oseba dobesednega navedka premega govora

Iz spremnega stavka premega govora je mogoče določiti, na koga se nanašata prva in druga oseba v dobesednem navedku, in sicer se prva oseba tipično nanaša na osebek spremnega stavka, druga oseba pa na predmet v dajalniku, primer je npr.: »Miha je rekel Janezu: "Pismo ti bom poslal jutri."« Če je spremni stavek spredaj, gre za katafori: »"Pismo ti bom poslal jutri," je Miha rekel Janezu.«

Podatek o tem, kateri element predloge glagola spremnega stavka se nanaša na katero osebo, je bil dodan kot atribut elementa predloge, kar je splošnejša rešitev kot zanašanje na osebek oz. predmet v dajalniku.

Določanje za zdaj deluje le pri primerih, ko imamo spremni stavek, ni pa še analize besedilne zgradbe, ki bi to ugotavljala za dobesedne navedke brez spremnih stavkov.

4.4 Delna osebna imena

Še posebej v časopisnem poročanju je običajno, da se oseba prvič navede s polnim imenom, v nadaljevanju pa le s priimkom (v bolj neformalnih besedilih pa tudi le z imenom), npr.: »Darko Krašovec je bil ponoči, na prvi seji pravkar oblikovane vlade Mira Cerarja, potrjen za generalnega sekretarja. Čeprav do zdaj sodnik, pa Krašovec v politiki ni novo ime.« Pri časopisnih naslovih so pogoste tudi katafore take vrste, saj je oseba v naslovu omenjena le s priimkom, v samem članku pa je potem navedena s polnim imenom.

Postopek za to vrsto koreferenc pravzaprav ni posebej zapleten, če imamo podatek, kaj so osebna imena, vsa imena oseb je treba shraniti v seznam in potem pogledati po seznamu, kadar naletimo le na posamičen priimek oz. ime.

Zapis, ki ga uporablja Amebisov vmesni jezik, ki prvi del imena osebe (običajno

torej osebno ime) zapiše v elementu JED (jedro dela samostalniške fraze), priimek pa v elementu IMP (imenski prilastek), po drugi strani pa sam priimek postane JED (če pa je pred imenom še kakšna druga beseda, npr. »matematik Josip Plemelj«, pa celo tako osebno ime kot priimek postaneta IMP), sicer pomeni, da se postopek malo zaplete in je treba pri izvedbi paziti na vse te pretvorbe. Dodatna težava so primeri, kjer bi morali koreferenco vezati na element IMP, kar za zdaj še ni podprto (če je torej posamičen priimek uporabljen kot prilastek za drugo besedo, npr. »matematik Plemelj« kot koreferenco za »matematik Josip Plemelj«).

4.5 Anafore pri osebnih zaimkih

Postopek za razreševanje anafor je bil zasnovan na podlagi metod na osnovi aktivacije (activation-based methods), kakor so opisane v (Němčík 2006) in ki izhajajo iz dela Eve Hajičove in sodelavcev (Hajičová 1987), vendar v tem trenutku še v precej poenostavljeni in predelani obliki.

Postopek je tak, da se gradi kontekst analize, ki vsebuje seznam kandidatov za razreševanje anafor, pri čemer ima vsak kandidat shranjeno analizo ustrezne samostalniške fraze, mesto zadnje uporabe (npr. osebek, predmet v tožilniku, prislovno določilo), podatke o spolu, številu, osebi in živosti ter oceno. Ko se pride do osebnega zaimka, ki še nima razrešene koreference, se poišče, ali obstaja kakšen kandidat, ki ustreza glede spola, števila, osebe in živosti, če jih je več, se izbere tisti, ki ima višjo oceno oz. se je pojavil zadnji, dodatno pa oceno zviša še ujemanje mesta uporabe (če npr. razrešujemo koreferenco pri osebku, ima prednost kandidat, ki je bil že prej osebek).

Uporaba kandidata mu poviša oceno, z začetno oceno se na seznam kandidatov dodajo tudi vse samostalniške fraze, ki nastopajo v analizi. Na koncu vsakega stavka, povedi in odstavka se znižajo (prepolovijo) ocene vseh obstoječih kandidatov, kandidati, katerih ocena pade na 0, se izbrišejo iz konteksta analize.

Ta osnovni postopek je bil dopolnjen z dodatnimi pravili, ki so opisana v

nadaljevanju.

4.5.1 PREMI GOVOR

Premi govor prekine tok pripovedovanja z drugim tokom, zato konteksta iz spremnega besedila ne smemo uporabiti pri analizi premega govora in obratno. Rešitev je, da ima analizator dva konteksta – enega za osnovno besedilo in drugega za premi govor, pri čemer se kontekst za premi govor vsakič ponastavi (dokler ne bo izdelana boljša analiza besedilne zgradbe, ki bi določila, kdo se s kom pogovarja).

Dopolnitev za prihodnost je še, da se iz spremnega stavka v kontekst premega govora preneseta prva in druga oseba (iz »Janez je rekel Micki: 'Jutri ti bom prinesel to knjigo.'« bi tako lahko ugotovili, da bo Janez prinesel knjigo Micki).

4.5.2 POMENSKE OMEJITVE

Samo informacije o skladnji in osnovne omejitve (oseba, spol, število) ne zadoščajo vedno za razreševanje koreferenc.

Metka je prebrala knjigo, ki jo je napisala Karmen, in jo povabila na kavo.
Metka je prebrala knjigo, ki jo je napisala Karmen, in jo vrgla stran.

Slika 3: Pomenske omejitve pri koreferencah.

Čeprav sta si povedi na sliki 3 enaki do drugega »jo«, je razrešitev te koreference vseeno drugačna. V prvi povedi se drugi »jo« nanaša na »Karmen«, v drugi pa na »knjiga«.

Podobno je v realnem primeru »Zadremala je že skoro, ali zgodilo se ji je, kakor da bi polagoma drsala navzdol, kakor da bi se skrinja nagibala, nagibala … in prestrašila se je in se je prebudila.«, kjer se ni prestrašila skrinja, ampak oseba, ki sicer ni navedena v tej povedi.

V precejšnjem delu primerov si bo dalo pomagati že s tem, da imajo glagolske

predloge lahko pri parametrih omejitve, ali so ti parametri obvezno osebe (oz. organizacije) oz. niso osebe. Vendar pa to vedno ne zadošča, v primeru »Hm, lahko bi kar takoj pojedel to deklico, ampak je premajhna, da bi mi potešila lakoto. Če odigram pravilno, bom lahko pojedel njo, pa tudi njeno babico!« je tako postopek najprej menil, da se »njo« nanaša na »lakoto« kar pomeni, da je treba v predlogi omejiti, da se ne da pojesti lakota. V takih primerih bi si lahko pomagali tudi s korpusom, vendar oseb ne jemo prav pogosto, če ne gre za pravljico.

4.5.3 STAVKI BREZ ANALIZE

Pojavi se vprašanje, kaj narediti v primeru, ko analizatorju ne uspe analizirati katerega od vmesnih stavkov. Tak primer je bil »Sončni žarki so se že igrali na strehi županove hiše. Francka je bila vsa nemirna, srce ji je utripalo od sreče in obenem od straha, da bi zamudila voz.«, kjer analizator ni prepoznal stavka »Francka je bila vsa nemirna« (ker še ni podpiral kombinacije zaimka »ves« s pridevnikom na mestu povedkovega določila), zato je potem postopek priredil zaimku »ji« vrednost »županova hiša«.

Idealna rešitev je, da se dopolni analizator, vendar ni mogoče pričakovati, da mu bo v dogledni prihodnosti uspelo analizirati vse (še posebej pri izpustih) zato je varianta, ki je vredna razmisleka, ta, da se v takih primerih ponastavi (pobriše) stanje konteksta. Na ta način sicer lahko izgubimo nekatere razrešitve koreferenc, ki izhajajo še iz prejšnjih stavkov, vendar se izognemo napakam, kar je v večini primerov bolj pomembno (torej povečamo natančnost na račun priklica).

Vsekakor pa je dolgoročna rešitev izboljševanje analizatorja.

4.5.4 PONAVLJANJE V STAVKU

V primeru »Ko sva zapuščala hišo, se je mačka nekako med nogami izmuznila

nazaj v hišo. Nisva jo[3] želela pustiti v hiši, ker se neprestano trudi požreti papigo.« je analizator poskušal razrešiti »jo« s »hiša« namesto »mačka«. Pomenske omejitve ni (morala bi biti precej podrobna, kjer lahko nekje pustiš tako osebo kot predmet), možno pa je postaviti dodatno zahtevo, da ne smemo znotraj posameznega stavka razrešiti zaimka z besedo, ki v tem stavku še nastopa, s čimer se potem zaimek »jo« razreši v »mačka«.

4.5.5. Pomenske razlike med izpuščenimi in navedenimi osebnimi zaimki

Lastnost, ki nam lahko pomaga pri razreševanju koreferenc, je, da se na neživo vedno nanašamo le z naslonsko obliko osebnega zaimka. Tako ne moremo reči »Knjiga je bila zelo zanimiva in njo sem prebral v dveh urah.« ampak le »Knjiga je bila zelo zanimiva in prebral sem jo v dveh urah.« Ker v imenovalniku ni naslonskih oblik, to pomeni, da neživo v osebku ne more biti nadomeščeno z osebnim zaimkom, ampak le z izpustom osebnega zaimka ali pa s kazalnim zaimkom.

S pomočjo tega pravila lahko v besedilu »Njegova jeza je v Naomi zbujala občutek krivde. Ona je bila tista, ki je vztrajala na prenovi strehe.« ugotovimo, da »Ona« ne smemo razrešiti z »njegova jeza«, ampak z »Naomi«.

4.6 Prislovni zaimki

Tukaj se razrešujejo koreference, ki so vezane ne prislovne zaimke, in sicer na »tam«, »tja« in »takrat«. Prva dva se sklicujeta na kraj, drugi pa na čas.

Slika 4 vsebuje primere prislovnih zaimkov s koreferencami. V prvih treh primerih zaimek nadomešča prislovno določilo iste vrste (z isto vprašalnico), četrti primer pa kaže, da je pri krajevnih prislovnih določilih treba imeti možnost, da se pretvarja med prislovnimi določili kraja za »kje« in »kam« (torej je treba »v katedrali sv. Sofije« pretvoriti v »v katedralo sv. Sofije«), za kar je bila dopolnjena podatkovna baza Ases s povezavami med ustreznimi pari

[3] »jo« namesto »je« je bil v originalnem besedilu, vsekakor pa mora biti analizator čim bolj neobčutljiv na pogoste nestandardne oblike, ker pač se pojavljajo v besedilih.

pomenov predlogov oz. prislovov.

1) Oblaki nastajajo poleti *nad večjimi ognjeniki*. **Tam** nastanejo zato, ker se topli zrak dviga in ohlaja.
2) Princ Borjatinski, guberner Jakutska je leta 1670 zaupal Dežnjovu odpravo *v Moskvo*. **Tja** je moral odnesti »sobolji zaklad« in uradne dokumente.
3) Biologija se je začela hitro razvijati in rasti*, ko je Anton van Leeuwenhoek izboljšal mikroskop*. **Takrat** so učenjaki odkrili semenčice, bakterije, infuzorije in raznovrstnost mikroskopskega življenja.
4) Večina dragocenosti, ki jo jih Slovani naropali v Hersonu, se je znašla v Novgorodu, kjer so jih vse do 20. stoletja hranili *v katedrali sv. Sofije*. **Tja** so prišle morda po zaslugi prvega novgorodskega škofa Joahima Hersonskega, katerega ime kaže na njegovo povezanost s tem mestom.
5) Prosti čas je mojster izkoristil za obisk *Londona*. **Tja** je prišel kot izrazit skladatelj italijanske opere.

Slika 4: Primeri s prislovnimi zaimki.

Zadnji, peti primer pa kaže, da ni nujno, da se prislovni zaimki sklicujejo na prislovna določila, ampak se lahko sklicujejo tudi na samostalnike, npr. na zemljepisna imena.

Razmislek pri prislovnih zaimkih je, da so relativno redki v besedilih, uporabljamo jih le, če želimo povezavo posebej poudariti. Veliko pogostejše je implicitno navezovanje, da se naslednji stavek dogaja v istem času in prostoru, zato bo treba razmišljati tudi v smeri, kako najti te implicitne koreference.

4.7 Katafora v osebku odvisnika

Pri katafori je zaimek pred samostalniško frazo, ki jo nadomešča. Primer za to je poved »Ker jo je zeblo, je Mojca oblekla jopico.«

Vidimo lahko, da je pri tem tipu zaimek, ki ga želimo razrešiti, v odvisniku, razrešitev koreference pa moramo poiskati v osebku glavnega stavka, ki sledi, pri čemer pa moramo paziti še na to, da je ta osebek na začetku stavka, drugače

imamo težave pri primeru »Ker jo je zeblo, ji je Mojca oblekla jopico.« ali pa »Ker jo je zeblo, ju je Mojca poslala domov.«

Ta metoda ne deluje vedno, če imamo primer »Ker jo je zeblo, je Mojca Ano poslala domov.« bo program sklepal, da je zeblo Mojco, čeprav je bolj logično, da je zeblo Ano. Za rešitev takih primerov bi moral program veliko bolje razumeti pomen povedi. Po drugi strani pa se da tudi reči, da ta primer krši načelo členjenja po aktualnosti, ker ni spredaj že prej omenjena Ana, torej: »Ker jo je zeblo, je Ano Mojca poslala domov.« (v tem primeru program »jo« poveže na »Ano«).

5 REZULTATI

Pojavil se je problem, kako preizkušati delovanje razreševanje koreferenc. Idealna možnost bi bila primerjava rezultatov z referenčnim korpusom z označenimi koreferencami, vendar ta za slovenščino še ne obstaja, zato je bilo treba preizkušati posamezne primere.

5.1 Korpus z označenimi koreferencami

Za slovenščino korpusa z označenimi koreferencami, s katerim bi lahko preizkušali razreševanje koreferenc (in ga tudi uporabili za učenje pri statističnih metodah) še nimamo, zato je smiselno premisliti, kakšen korpus bi pravzaprav želeli.

Na prvi pogled se zdi smiselno, da bi uporabili npr. kar ssj500k, ki je že najbolj označen ročno pregledan korpus za slovenščino (in vsebuje tudi skladenjsko razčlembo in označene imenske entitete), in dodatno označili še koreference. Na ta način bi lahko hitro preizkusili tudi metode, ki potrebujejo skladenjsko razčlembo. Vendar je pri ssj500k težava to, da je sestavljen iz odstavkov iz različnih besedil, nima pa večjih kosov iz istega besedila, kar je velika omejitev, ker je treba pri razreševanju koreferenc iskati tudi prek meja odstavkov.

S problemom, kako izbrati primere za označevanje koreferenc, se je ukvarjala

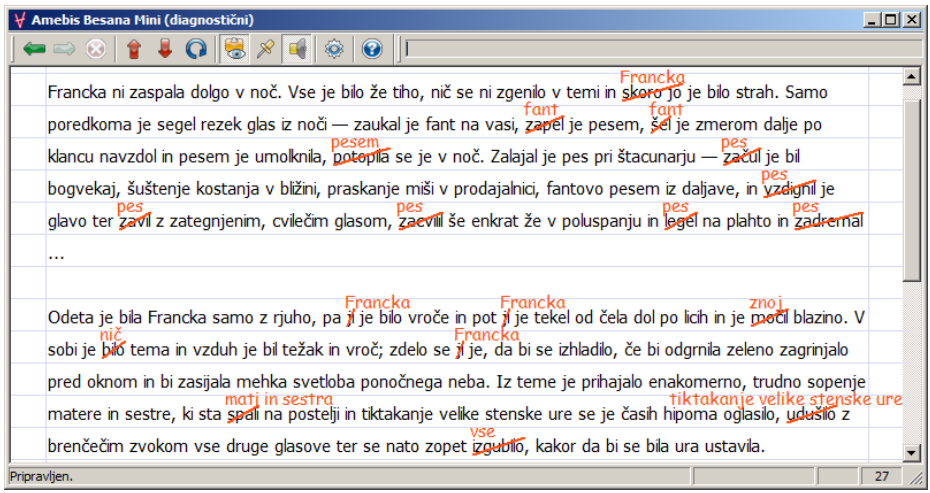
(Bucik, 2001), tam je tudi nekaj že ročno označenih besedil, manjkajo pa koreference zaimkov, pa tudi sicer je besedil relativno malo, da bi lahko preizkušali statistične metode.

To kaže, da bi bilo treba pripraviti namenski korpus z označenimi koreferencami ali pa pri kakšnem naslednjem ročno označenem korpusu izbrati besedila tako, da bodo v korpusu tudi večji kosi besedil oz. celotna kratka besedila. In šele s takim korpusom bo mogoče res preizkusiti različne metode razreševanja koreferenc in videti, kako uspešne so v slovenščini.

5.2 Pregled rezultatov v Besani Mini

Za lažje preizkušanje delovanja razreševanja koreferenc (neposredno branje Amebisovega vmesnega jezika ni posebej preprosto, saj je bolj prilagojen temu, da ga berejo računalnik) je bila zgrajena posebna verzija slovničnega pregledovalnika Besana, in sicer z vmesnikom Besana Mini. Koreference se izpisujejo kot ena od napak, ki jih program išče (razrešitve (izpisane vedno v imenovalniku) nadomestijo osebni zaimek, če pa gre za izpust osebnega zaimka, pa nadomestijo glagol), za boljšo preglednost se izključi izpis vseh drugih napak. Na ta način se besedilo, s katerim želimo preveriti delovanje razreševanje koreferenc, le skopira v odložišče in Besana takoj izpiše najdene razrešitve, kot je primer na sliki 5.

Prvi rezultati (kot na primer na sliki 5) so videti obetavno, na rezultatih se vidi tudi uporaba sopomenk (ker se »pot« nadomesti z »znoj«, ki je prva beseda, na katero je vezan ta pomenski koncept v bazi Ases). Ni pa tudi v tem primeru razreševanje brez napak, v zadnji vrstici se na »izgubilo« poveže »vse« namesto »tiktakanje velike stenske ure«.

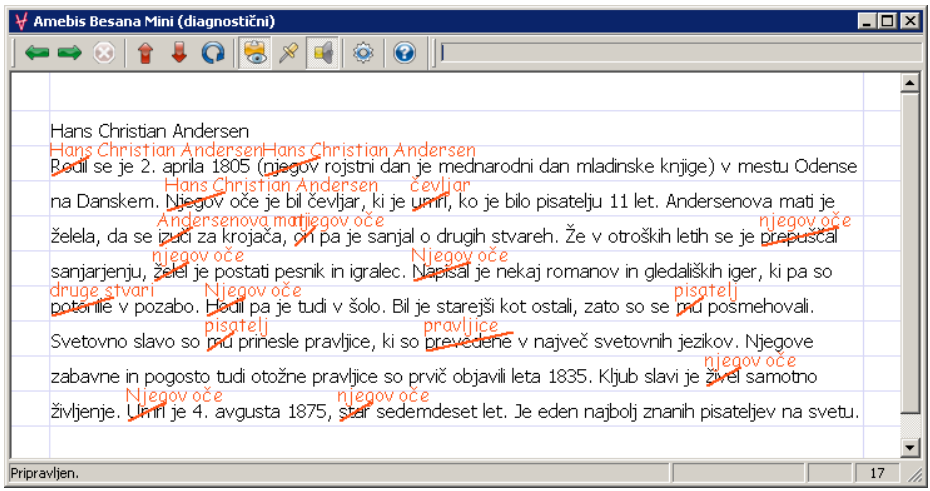


Slika 5: Primer razreševanja koreferenc na prvih dveh odstavkih romana Na klancu.

Vsekakor se bo treba bolj posvetiti izboljšavam analizatorja, saj napake pri skladenjski razčlembi zelo vplivajo na razreševanje koreferenc. Pri primeru »Jetra so za vretenčarje značilen organ. Imajo osrednjo vlogo v presnovi in številne druge naloge« tako v drugi povedi ni našlo izpusta osebnega zaimka, ker je analizator napačno določil, da je osebek »naloge«, ta primer je zdaj že popravljen, seveda pa zanesljivo obstaja še množica problematičnih primerov

Težave nastanejo tudi zato, ker sistem še ne vsebuje dovolj pomenskih omejitev, ki pomagajo pri izbiranju prave koreference. Te omejitve bi tudi v splošnem pomagale pri razdvoumljanju, zato bo dopolnjevanje baze Ases v tej smeri zelo koristno.

Občasno se pokažejo tudi težave, ker starejši kontekst preveč vpliva na nove stavke, kar kaže, da bo smiselno preizkusiti hitrejše pozabljanje konteksta. Točne nastavitve teh parametrov pa bodo zahtevale več preizkušanja in predvsem pripravo korpusa primerov razrešenih koreferenc, kar bo omogočilo hitrejše preizkušanje različic in tudi primerjavo z drugimi metodami razreševanja koreferenc.



Slika 6:. Primer iz Wikipedije s težavami pri razreševanju koreferenc.

Primer na sliki 6 kaže, kako lahko ena napaka vpliva naprej. Ker program zgreši in pripiše, da je o drugih stvareh sanjal njegov oče in ne Hans Christian Andersen, se potem ta napaka prenese še na vse naslednje koreference. Še pred tem je tudi nejasno, za koga je njegova mati želela, da se izuči za krojača.

6 UPORABA V SISTEMU PIFLAR

Piflar je sistem za odgovarjanje na vprašanja v naravnem jeziku, ki se med drugim uporablja na Amebisovem portalu za virtualne asistente SecondEgo[4], in sicer za več namenov: odgovarjanje s pomočjo splošnega znanja (ki vsebuje povedi tipa »France Prešeren se je rodil v Vrbi 3. decembra 1800.«), dopolnitev iskanja po spletnih straneh, odgovarjanje iz Wikipedije in splošni odgovori iz Gigafide, načrtovana je še možnost, da lahko avtor agenta vnese povedi, ki se jih agent nauči (možno pa bi bilo tudi, da bi si agent s pomočjo Piflarja zapomnil, kaj mu je povedal sogovornik med pogovorom, vendar bi bilo za to treba vzpostavljati začasne baze za trajanje pogovora).

Sistem kot osnovo uporablja Amebisov vmesni jezik in pomene iz baze Ases,

[4] www.secondego.com

zato je v načelu jezikovno neodvisen (taka uporaba sicer zahteva natančnejše razdvoumljanje, zato je lahko manj zanesljiva od enojezične) in omogoča, da se nauči v enem jeziku in odgovarja v drugem (trenutno so podprte slovenščina, angleščina in nemščina).

6.1 Uporaba razreševanja koreferenc v Piflarju

Na začetku so bili velika omejitev zaimki v vhodnem besedilu, ker se je sistem naučil znanje z zaimki namesto z njihovimi pravimi pomeni (Holozan 2014a)

Že v (Vicedo, Ferrández 2000) je bilo pokazano, da je razreševanje koreferenc pomembno za odgovarjanje na vprašanja. Zato je bil Piflar dopolnjen s podporo za element ORI, ki je bil dodan v vmesni jezik za zapisovanje razrešenih koreferenc, tako da zdaj uporablja ta element namesto originalnega jedra (JED). S to dopolnitvijo zdaj pravilno odgovarja tudi v primerih, ko je treba upoštevati koreference.

Primer, ki pokaže, kako pomembne so lahko koreference pri odgovarjanju, prikazuje tabela 1, kjer so odebeljeno označena vprašanja, na katera je mogoče odgovoriti zaradi razrešenih koreferenc, prej pa bi bili uporabljeni osebni zaimki oz. Piflar ni imel odgovora na vprašanje. V primerjavi s podobno tabelo v (Holozan 2014b) zdaj Piflar tudi pri odločevalnih vprašanjih odgovarja v polnih stavkih, in sicer zato, ker prej kdaj ni bilo jasno, kaj pravzaprav pomeni »Da.« oz. »Ne.«

Pri nekaterih vprašanjih je besedni red zaznamovan (npr. »Kdaj kosilo je pojedel Miha?«), ker so tudi vprašanja generirana avtomatsko, kar je dodatna možnost Piflarja, ki omogoča, da lahko hitro vidimo, na katera vprašanja lahko Piflar odgovori glede na dano vhodno besedilo. Niso pa to vsa vprašanja, na katera lahko odgovori v tem primeru, možna so namreč še skrajšana vprašanja (npr. »Ali je Miha naletel na Janeza?« brez »v Ljubljani«), ki pa jih Piflar še ne zna zgenerirati.

Vprašanje	**Odgovor**
Ali je Miha prebral knjigo?	Da, Miha je prebral knjigo.
Kdo je prebral knjigo?	Knjigo je prebral Miha.
Kaj je Miha prebral?	Miha je prebral knjigo.
Ali je Miha potem pojedel kosilo?	Da, Miha je potem pojedel kosilo.
Kdaj kosilo je pojedel Miha?	Kosilo je pojedel Miha potem.
Kaj je Miha pojedel potem?	Miha je pojedel kosilo potem.
Kdo je pojedel kosilo potem?	Kosilo je pojedel Miha potem.
Ali je Miha šel v Ljubljano?	Da, Miha je šel v Ljubljano.
Kdo je šel v Ljubljano?	V Ljubljano je šel Miha.
Kam je šel Miha?	Miha je šel v Ljubljano.
Ali je Miha naletel na Janeza v Ljubljani?	Da, Miha je naletel na Janeza v Ljubljani.
Kje na Janeza je naletel Miha?	Na Janeza je naletel Miha v Ljubljani.
Na koga je Miha naletel v Ljubljani?	Miha je naletel na Janeza v Ljubljani.
Kdo na Janeza je naletel v Ljubljani?	Na Janeza je naletel Miha v Ljubljani.
Ali je Miha pozdravil Janeza?	Da, Miha je pozdravil Janeza.
Kdo je pozdravil Janeza?	Janeza je pozdravil Miha.
Koga je Miha pozdravil?	Miha je pozdravil Janeza.
Ali je Miha naletel na Janeza?	Da, Miha je naletel na Janeza v Ljubljani.
Na koga je Miha naletel?	Miha je naletel na Janeza.
Ali je bilo kosilo pojedeno?	Da, kosilo je bilo pojedeno potem.

Tabela 1: Vprašanja in odgovori, ki jih najde Piflar za primer »Miha je prebral knjigo. Potem je pojedel kosilo in šel v Ljubljano. Tam je srečal Janeza in ga pozdravil.«.

Zanimiv primer, ki vsebuje še prislovni zaimek, je, če imamo vhodno besedilo »Matematik Josip Plemelj se je rodil 11. decembra 1873 na Bledu. Tam je obiskoval osnovno šolo.« Na vprašanje »Kje je Josip Plemelj obiskoval osnovno šolo?« tako dobimo odgovor »Josip Plemelj je hodil v osnovno šolo na Bledu.« Težava pri teh krajevnih (in podobno časovnih) določitvah pa je, da so največkrat implicitne (se prenašajo iz prejšnjih stavkov brez izrecne uporabe prislovnih zaimkov), česar Piflar še ne zna uporabiti.

Razreševanje koreferenc pri premem govoru pa je Piflarju omogočilo, da pri vhodnem besedilu »"Poklicala te bom," je rekla Mojca Janezu.« na vprašanje »Kdo bo poklical Janeza?« odgovori z »Janeza bo poklicala Mojca.«.

6.2 Druge dopolnitve Piflarja

Sistem Piflar je bil dodatno dopolnjen s tem, da zna odgovarjati tudi na vprašanja, ki niso zastavljena v obliki stavka, ampak le kot posamičen stavčni člen. Tako npr. kot odziv na vprašanje »Miha« poišče dejstvo, ki vsebuje samostalniško frazo »Miha«, npr. »Miha je šel v Ljubljano.« ali pa »Miha je pozdravil Janeza.«, če ima v bazi primer iz Tabele 1. Na ta način se Piflar bolj uspešno odziva na način iskanja, na katerega so uporabniki navajeni iz običajnih iskalnikov.

Pri tovrstnih vprašanjih je možno kombinirati tudi npr. osebek in prislovno določilo. Če imamo npr. učni stavek »Isaac Newton je umrl 20. marca 1727 v Kensingtonu.«, zdaj Piflar na vprašanje »Isaac Newton 1727« odgovori: »Isaac Newton je umrl 20. marca 1727.«

Zadnja verzija Piflarja zna narediti tudi ustrezne dolge odgovore pri odločitvenih vprašanjih, kjer je prej odgovarjal preprosto z »Da.« (ali pa ni odgovoril), zdaj pa odgovor dopolni s polnim odgovorom, pri čemer navede vse znane podatke iz dejstva. Tako npr. zdaj na vprašanje »Ali se je matematik Josip Plemelj rodil na Bledu?« vrne dolgi odgovor »Da, matematik Josip Plemelj se je rodil na Bledu 11. decembra 1873.«

Dodana je bila tudi možnost, da se Piflar odziva tudi na trdilne povedi. Tako na poved »Matematik Josip Plemelj se je rodil na Bledu.« odgovori: »Da, matematik Josip Plemelj se je rodil na Bledu 11. decembra 1873.«, na poved »Matematik Josip Plemelj se ni rodil na Bledu.« pa odgovori »Ne, matematik Josip Plemelj se je rodil na Bledu 11. decembra 1873.« Potencialna uporaba tega je, da se virtualni agent s seznamom povedi nauči določenih podatkov o sebi in tako potem s pomočjo Piflarja odgovarja na vprašanja oz. trditve v zvezi s temi podatki (dodatna funkcionalnost, ki je bila v ta namen vgrajena v Piflarja, je zmožnost obračanja prve in druge osebe v vprašanju, vprašanje »Ali si neumen?« se tako prevede v »Ali sem neumen?«). Če tako Piflarja naučimo »Nisem neumen.«, bo lahko na podlagi tega odgovoril na vsa vprašanja, ki so navedena v tabeli 2.

Vprašanje	**Odgovor**
Ali si ti neumen?	Ne, nisem neumen.
Ali ti si neumen?	Ne, nisem neumen.
Ali si neumen?	Ne, nisem neumen.
Ali nisi ti neumen?	Pač, nisem neumen.
Ali ti nisi neumen?	Pač, nisem neumen.
Ali nisi neumen?	Pač, nisem neumen.
Kdo ni neumen?	Nisem neumen.
Neumen si.	Ne, nisem neumen.
Nisi neumen.	Pač, nisem neumen.
Torej nisi neumen.	Pač, torej nisem neumen.

Tabela 2: Seznam vprašanj in odgovorov, ki jih Piflar pozna, če se nauči »Nisem neumen.«.

Težava se je pokazala pri zanikanih vprašanjih, pri čemer ni bilo jasno, ali

odgovoriti z »da« ali »ne«. Zato je bila narejena spletna anketa (za izpolnjevanje so bili naprošeni člani dveh prevajalskih e-poštnih seznamov). Rezultati so pokazali[5], da bi približno polovica izbrala »da«, polovica pa »ne«, zato je bila na koncu uporabljena ideja Janeza Perka, da se v teh primerih raje uporabi »pač« po zgledu francoskega »si«. Anketa je pokazala tudi to, da Piflar v odgovorih izpušča osebni zaimek »jaz« tudi v primerih, ko bi ga moral poudariti (značilen primer je odgovor na vprašanje »Kdo ni neumen?«), za rešitev tega bo treba dopolniti vmesni jezik, da se bo dalo označiti poudarke, kar bo potem omogočilo, da Piflar označi poudarjene zaimke, ki jih generator ne bo izpuščal.

6.3 Možnosti nadaljnjega razvoja Piflarja

Ena možnost za dopolnitev Piflarja, povezana z razreševanjem koreferenc, bi bila to, da bi razreševanje koreferenc našlo tudi dodatne povezave, npr. nadpomenke/podpomenke oz. celo parafraze. S tem bi Piflar lahko dobil možnost, da poveže različne izraze, ki označujejo isti koncept, kar pomeni, da bi lahko tudi pri spraševanju upošteval vse te načine, kako imenovati neki koncept).

Zanimiva možnost je, da se Piflar dopolni še z uporabo protipomenk, vendar bi se bilo treba pri tem omejiti le na določene glagolske predloge. Tako bi se Piflar naučil še zanikanega stavka s protipomenko, iz »Sem pošten.« bi se ta hkrati naučil še »Nisem nepošten.« Tega bi se bilo morda smiselno lotiti tako, da se v bazo dodata tako osnovni stavek kot njegova negacija s protipomenko, s čimer bi dosegli to, da samega iskanja ne bi bilo treba nič spreminjati.

Naslednja potencialna dopolnitev je odgovarjanje na vprašanja oz. trditve tipa: »Matematik Josip Plemelj se je rodil v Celju.« Piflar ve, da se je rodil na Bledu in nima podatka, da bi se rodil v Celju. Vendar bi za odgovor potreboval dodatno informacijo, in sicer v tem primeru to, da se vsaka oseba lahko rodi le v enem

[5] https://goo.gl/3mNXg4

kraju in da dejstvo, da se je rodil v nekem kraju, hkrati pomeni, da se ni rodil v nobenem drugem kraju (to seveda v splošnem ne velja pri vseh glagolih, dejstvo, da je npr. nekdo obiskal Bled, ne pove nič o tem, ali je ta oseba kdaj obiskala Celje).

Pri uporabi v sistemu Piflar se z bolj kompleksnimi odgovori kaže tudi to, da bo treba dopolniti tudi generator, ki prevaja vmesni jezik v naravni jezik, in sicer v smeri, da bo poskrbel za naravnejše odgovore s tem, da bo dodajal izpuste in po potrebi tudi koreference z osebnimi zaimki, da bodo odgovori zveneli bolj naravno. Zdaj npr. pri učnem besedilu »Oblaki nastajajo poleti nad večjimi ognjeniki. Tam nastanejo zato, ker se topli zrak dviga in ohlaja.« na vprašanje »Zakaj nastanejo oblaki nad večjimi ognjeniki?« odgovori »Do oblakov pride nad večjimi ognjeniki, ker se dviguje topli zrak in ker se ohlaja.«, namesto »Do oblakov nad večjimi ognjeniki pride, ker se topli zrak dviguje in ohlaja.«

7 UPORABA V PREVAJALNIKU PRESIS

Pokazalo se je, da razreševanje koreferenc pomaga tudi strojnemu prevajalniku Presis[6]. Ker Presis deluje tako, da prevede vhodno besedilo v vmesni jezik in potem vmesni jezik v izhodno besedilo, vsaka izboljšava analizatorja ali generatorja takoj izboljša še delovanje Presisa.

Prva izboljšava je pravilno prevajanje osebnih zaimkov, ki se nanašajo na predmete in so tako lahko različnega spola v različnih jezikih. Tako se zdaj poved »Pobral sem knjigo in jo začel brati.« prevede v »I picked up a book and started to read it.« namesto v »I picked up a book and started to read her.« kot do zdaj.

Koreference lahko pomagajo tudi pri določanju spola osebka v dobesednem navedku premega govora, če je ta v prvi oz. drugi osebi (ker npr. v angleščini v prihodnjiku in pretekliku v splošnem ni mogoče ugotoviti spola osebka, v slovenščini pa je spol določen in tako vpliva na prevod). Tak primer je »"I will

[6] presis.amebis.si

go there," she said.«, ki se zdaj prevede v »"Šla bom tja," je rekla.«, prej pa se je prevedlo v »"Šel bom tja," je rekla.«.

Druga izboljšava pa je razdvoumljanje prilastkovih odvisnikov glede na jedro, kar je prikazano v tabeli 3.

Poved	**Prevod**
To je bilo dekle, ki sem ga osvojil.	This was a girl, that I got to go out with.
To je bilo mesto, ki sem ga osvojil.	This was a town, that I occupied.
To je bilo prvo mesto, ki sem ga osvojil.	This was first place, that I won.
To je bilo občinstvo, ki sem ga osvojil.	This was audience, that I won.
To je bila gora, ki sem jo osvojil.	This was a mountain, that I scaled.
To je bila snov, ki sem jo osvojil.	This was substance, that I learned.

Tabela 3: Različno prevedeni prilastkovi odvisniki glede na jedro.

Zadnji primer pa kaže, da bo treba razdvoumljanje dopolniti še s tem, da bodo prilastkovi odvisniki lahko vplivali tudi na razdvoumljanje jedra, kar bo zamenjalo besedo »substance« s »subject«.

8 SKLEP

Poleg Piflarja in Presisa je potencialna možnost uporabe našega sistema še pri iskanju po korpusih, npr. pri iskanju kolokacij, kjer bi z razširitvijo iskanja na osebne zaimke z razrešenimi koreferencami lahko povečali število zadetkov pri isti velikosti korpusa.

Naslednja zanimiva možnost je uporaba pri napredni funkciji »išči in zamenjaj« (ang. *search and replace*). Ta je v pregibnih jezikih precej zahtevna, saj je treba v primeru zamenjave spola (če npr. zamenjujemo »sklic« s »koreferenca«) zamenjati ne le samo besedo, ampak tudi ustrezne pridevnike in glagole, ki so vezani nanjo, kar zahteva ne le oblikoskladenjsko analizo in

lematizacijo, ampak tudi skladenjsko razčlembo (Nevěřilová, Suchomel 2014). Dodatno pa je treba zamenjati še morebitne zaimke, ki se sklicujejo na menjano besedo (»Program poišče *sklic* in *ga* prikaže.« v »Program poišče *koreferenco* in *jo* prikaže.«), ter popraviti glagole pri elipsah menjane besede v osebku, za kar pa potrebujemo razreševanje koreferenc.

Sistemu še ne uspe razrešiti vseh koreferenc (delno je to tudi zato, ker analizatorju ne uspe analizirati vse povedi, vendar je to ločen problem), zato je še veliko možnosti za izboljšave, še posebej to velja za koreference, ki niso zaimki, niti dotaknil pa se še ni tudi bolj zapletenih povezanih koreferenc (par – moški). Manjka tudi še razreševanje kazalnih zaimkov.

Razreševanje koreferenc je možno izboljšati tudi z analizo besedilne zgradbe, predvsem dialogov, s čimer bi se lahko bolje povezale informacije v različnih odstavkih in v premem govoru, npr. pri dialogih. Tak primer je na Sliki 7.

Rdeča kapica je vprašala volka: »Zakaj imaš tako velike oči?«

»Da te bolje vidim.«

Slika 7: Primer dialoga Rdeče kapice z volkom.

Na podlagi tega učnega besedila bi moral biti Piflar sposoben na vprašanje »Zakaj ima volk tako velike oči?« odgovoriti z »Volk ima tako velike oči, da bolje vidi Rdečo kapico.«. Taka osnovna analiza besedilne zgradbe, pri kateri bi se med zaporednimi dobesednimi navedki dveh oseb izmenjavale koreference za prvo in drugo osebo, niti ne bi bila tako zahtevna, težava je bolj v tem, kako ugotoviti, da gre res samo za dve osebi in ne za več, kar lahko analizo veliko bolj zaplete.

Pogoj za nadaljnji razvoj sistema avtomatskega razreševanja koreferenc v slovenščini pa bo verjetno tudi priprava korpusa primerov razrešenih koreferenc, ki bi omogočil hitro primerjavo delovanja različnih postopkov. Pri pripravi takega korpusa bi bilo pomembno, da se ne omejimo le na osebne

zaimke, ampak v njem označimo tudi druge vrste koreferenc, s čimer bi postal uporaben tudi za preizkušanje razreševanja bolj zapletenih kohezivno-koherenčnih besediloslovnih nalog.

LITERATURA

Balažic Bulc, T., in Gorjanc, V. (2015): The position of connectors in Slovene and Croatian student academic writing : a corpus-base approach. V S. Starc, C. Jones, in A. Maiorani (ur.): Meaning making in text: multimodal and multilingual functional perspectives: 51–71. New York: Palgrave Macmillan.

Brennan, S. E., Friedman, M. W., in Pollard, C. J. (1987): A centering approach to pronouns. Proceedings of the 25th Annual Meeting of the AC: 155–162. Stanford.

Bucik, K. (2001): Strukture kohezije: strukture koreferenc in tematske progresije: Diplomsko delo. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

De Beaugrande, R. A., in Dressle, W. U. (1992): Uvod v besediloslovje. Prev. Derganc A., Miklič, T. Ljubljana: Park.

Erjavec, T., in Krek, S. (2008): Oblikoskladenjske specifikacije in označeni korpusi JOS. V T. Erjavec in J. Žganec Gros (ur.): Zbornik 6. konference Jezikovne tehnologije: 49–53. Ljubljana: IJS.

Ge, N., Hale, J., in Charniak E. (1998): A statistical approach to anaphora resolution. V Proceedings of the Sixth Workshop on Very Large Corpora: 161–170. Montreal.

Gorjanc, V. (1999): Kohezivni vzorec matematičnih besedil. Slavistična revija, 2 (47): 139–159.

Hajičová, E. (1987): Focusing – a meeting point of linguistics and artificial intelligence. V P. Jorrand in V. Sgurev (ur.): Artificial Intelligence, II: Methodology, Systems, Applications: 311–321. Amsterdam: Elsevier Science Publishers.

Hobbs, J. R. (1978): Resolving pronoun references. V B. J. Grosz, K. Spärck-Jones in B. L. Webber (ur.): Readings in Natural Language Processing: 339–352. Los Altos: Morgan Kaufmann Publishers.

Holozan, P. (2011): Samodejno izdelovanje besedilnih logičnih nalog v slovenščini: Magistrsko delo. Ljubljana: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko.

Holozan, P. (2014a): Piflar: sistem za učenje in odgovarjanje na vprašanja v naravnem jeziku. V M. Orel in S. Jurjevčič (ur.): Mednarodna konferenca InfoKomTeh: 350–358. Polhov Gradec: Eduvision.

Holozan, P. (2014b): Razreševanje sklicev pri analizi slovenskih besedil. V T. Erjavec in J. Žganec Gros (ur.): Zbornik 9. konference Jezikovne tehnologije: 135–140. Ljubljana: IJS.

Joshi, A. K., in Kuhn, S. (1979): Centered logic: the role of entity centered sentence representation in natural language inferencing. Proceedings of the International Joint Conference on Artificial Intelligence: 435–439. Tokyo.

Kocijančič Pokorn, N. (1997): A Slovene-English Contrastive Analysis of One. Slovene Linguistic Studies, 1 (1997): 17–34.

Korošec, T. (1981): Besediloslovna vprašanja slovenščine. XVII. seminar slovenskega jezika, literature in kulture: 173– 186. Ljubljana: Filozofska fakulteta.

Korošec, T. (2006): O besediloslovnih prvinah v slovenskem jezikoslovju. Slavistična revija, letnik 54 (posebna številka): 239–258.

Lappin, S., in Leass, H. J. (1994): An algorithm for pronominal anaphora resolution. Computatinal Linguistics, 20 (4): 535–561.

McShane, M., Beale, S., in Nirenburg, S. (2010): Reference Resolution Supporting Lexical Disambiguation. 2010 IEEE Fourth International Conference on Semantic Computing: 56-59. Los Alamitos: IEEE Computer Society.

Mitkov, R. (1999): Anaphora Resolution: The State Of The Art: Working

paper. University of Wolverhampton.

Mitkov, R., Evans, C., in Orasan, C. (2002): A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002): 17–23. Mexico City.

Němčík, V. (2006): Anaphora Resolution: Magistrsko delo. Brno: Masarykova universita, Fakulta informatiky.

Nevěřilová, Z., in Suchomel, V. (2014): Intelligent Search and Replace for Czech Phrases. Eighth Workshop on Recent Advances in Slavonic Natural Language Processing: 97 –105. Brno: Tribun EU.

Toporišič, J. (2004): Slovenska slovnica. Maribor: Obzorja.

Vicedo, J. L., in Ferrández, A. (2000): Importance of Pronominal Anaphora resolution in Question Answering systems. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL): 555–562. Morristown.

Zuljan Kumar, D. (2007): Narečni diskurz. Ljubljana: Založba ZRC.

Zuljan Kumar, D. (2010): Cohesive means in Slovenian spontaneous dialectal conversations. Slavia Centralis, 2010 (2): 17–34.

Žitnik, S. (2014): Iterativno pridobivanje semantičnih podatkov iz nestrukturiranih besedilnih virov: Doktorska disertacija. Ljubljana: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko.

THE SYSTEM FOR CO-REFERENCE RESOLUTION FOR SLOVENIAN TEXTS ANALYSIS AND POSSIBILITIES OF ITS USE

Co-reference resolution is an important part of language technologies, but has not yet been developed for Slovenian. There are various types of co-references and the paper focuses on anaphora resolution of personal pronouns. Seven methods, used in combination, were used; the most important one is based on activation. First results are promising, but for more extensive evaluation, Slovenian corpus with marked examples is needed.

Co-reference resolution was used in the question answering system Crammer, which can, as a result, answer more questions than before, because it can replace personal pronouns. At the same time, some other improvement were added to Crammer, e.g. answering to individual words and phrases and answering to declarative sentences. Added was also generation of long answers to questions with interrogative particles.

Co-reference resolution also improved working of Presis machine translation, especially for determining of gender of pronouns and for disambiguation of attributive subordinate clauses.

Keywords: co-reference resolution, coreference resolution, anaphora, question answering