Prvi leksikalni podatki o slovenskem znakovnem jeziku iz korpusa Signor

Špela Vintar, Boštjan Jerko, Marjetka Kulovec

Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, SI-1000 Ljubljana
spela.vintar@ff.uni-lj.si, {marjetka.kulovec, bostjan.jerko}@guest.arnes.si

Povzetek

O leksikalnih in slovničnih lastnostih slovenskega znakovnega jezika (SZJ) je bilo do nedavnega mogoče pisati zgolj na podlagi domnev in posamičnih opažanj. Prispevek predstavlja rezultate leksikalne analize, ki je bila izvedena s pomočjo korpusa Signor, prvega reprezentativnega korpusa SZJ. Po predstavitvi osnovnih korpusnih statistik se v prispevku posvetimo izbranim leksikalnosemantičnim elementom SZJ, med drugim tudi vlogi mašil in gest. Opažene pogostosti primerjamo s podatki, pridobljenimi v sorodnih raziskavah v tujini, predvsem za britanski BSL in avstralski Auslan. V zaključku razpravljamo o nadaljnjih raziskavah in možnostih uporabe korpusa Signor za posodobitev jezikovnih priročnikov in razvoj jezikovnih tehnologij.

First Lexical Analysis of the Slovene Sign Language based on the Signor Corpus

The lexical and grammatical properties of Slovene Sign Language (SZJ) have so far only been described on the basis of isolated observations or presumptions. This paper presents the results of a lexical analysis performed on the Signor corpus, the first representative corpus of SZJ. After presenting general corpus statistics we discuss selected lexical and semantic properties of SZJ, for example the role of fillers and gestures. The figures obtained are compared to related works, in particular corpus-based studies performed for BSL and Auslan. The paper concludes by outlining our plans for future research and the ways in which our corpus could help improve basic reference works for SZJ as well as serve as a basis of new technologies.

.

1. Uvod

Slovenski znakovni jezik (SZJ) je jezik gluhe skupnosti v Sloveniji in ga po aktualnih ocenah uporablja od 800 do 1600 oseb. Od leta 2002 je priznan kot eden od uradnih jezikov v Sloveniji, kar uporabnikom daje pravico do sporazumevanja v SZJ v vseh javnih in zasebnih situacijah. Ta pravica se običajno udejanja prek tolmačev SZJ. Gluhi pa se večkrat počutijo zapostavljene, ker je tolmačenje drago in državni sistem vaučerjev mnogim ne zadošča, tolmači pa tudi niso vedno na razpolago. To velja še posebej za izobraževanje.

Priročniki za učenje SZJ so zasnovani brez prave jezikoslovne osnove, saj je bilo doslej tovrstnih raziskav SZJ malo. Gluhi skupnosti sta na voljo dva multimedijska slovarja, vendar noben od njiju ne temelji na korpusnih podatkih o SZJ in oba težita k strogo normativni obravnavi kretenj. Prvi učbenik za učenje SZJ je bil objavljen leta 2010 in sicer zapolnil dotedanjo vrzel v izobraževanju SZJ, vendar se učbenik močno naslanja na slovenščino in zanemarja ne le lastnosti SZJ, ampak tudi edinstvene značilnosti znakovnih jezikov na splošno.

V dosedanjih znanstvenih objavah o SZJ se avtorji večinoma ukvarjajo s sociološkim vidikom ali temami (specialne) pedagogike (npr. Kuplenik, 1999; Globačnik, 2007). Bolj jezikoslovno usmerjeni članki obravnavajo temo standardizacije (Bauman 2007), primerjavo SZJ z govorjeno/pisno slovenščino (Globačnik 2001, Žele 2007) ali opisujejo izbrane in splošne lastnosti morfologije in fonologije SZJ (Žele in Bauman 2011). Glede na pomanjkanje znanstvene osnove se pojavljajo vprašanja o kakovosti poučevanja SZJ, pa tudi priprave in certificiranja za tolmače.

Predstavljena raziskava je pomemben korak v smeri izboljševanja stanja. S projektom, ki ga je financirala Javna agencija za raziskovanje RS v obdobju 2011-2014, je bil zgrajen prvi korpus SZJ (korpus Signor, spletna stran projekta http://lojze.lugos.si/signor/index.html). Zbrali smo posnetke prek 80 uporabnikov SZJ. Podatki so uravnoteženi po regijah, starosti in spolu. Korpus je ročno tokeniziran in lematiziran, nadaljujejo pa se označevanja kompleksnejših ravni. Korpus Signor predstavlja vir za opisne raziskave SZJ v obliki, kot se ta uporablja v realnih komunikacijskih situacijah gluhe skupnosti.

Namen prispevka je opisati nekatere leksikalne lastnosti SZJ, ki se izkazujejo na podlagi osnovne ravni označevanja, se pravi na ravni pripisovanja pomenskih oznak oz. lematizacije. Za ostale raziskave, predvsem skladenjske strukture in drugih slovničnih lastnosti, bo treba počakati na nadgradnjo označevalnih ravni.

2. Korpus SIGNOR

Gradnja korpusa se je pričela leta 2011 z namenom zagotovitve reprezentativnega in uravnoteženega korpusa izvirnih primerov besedil v SZJ (prim. Vintar in dr., 2012). Pred začetkom projekta smo pregledali podobne raziskave o znakovnih jezikih po svetu: ameriškega znakovnega jezika ASL (Lu in Huenerfauth, 2011), avstralskega Auslan (Johnston in dr., 2006), avstrijskega ÖGS (Krammer in dr., 2001, Dotter, 2011) in italijanskega LIS (Prinetto in dr., 2011), vendar smo se na koncu najbolj naslonili na projekt korpusa nemškega znakovnega jezika DGS. Kljub temu, da je slovenski projekt po trajanju in sredstvih skromen v primerjavi s podobnimi, smo uporabili podobno metodologijo za izbiro informantov in strukturo snemanja (Nishio in dr., 2010), pa tudi za strategijo segmentiranja (Hanke in dr., 2012) in označevanja lem ter kompleksnih struktur (Konrad in dr., 2012).

2.1. Gradnja korpusa

Korpus je reprezentativen glede na ocenjeno velikost gluhe skupnosti v Sloveniji. Zbrali smo posnetke 80 informantov, kar predstavlja 5 do 10 % celotne skupnosti uporabnikov SZJ. Informanti prihajajo iz vseh slovenskih

regij, vzorec pa je tudi dobro uravnotežen po spolu (37 žensk in 43 moških) in starosti (leta rojstva so enakomerno razporejena od 1932 do 1996). Vsakega informanta smo prosili še za nekaj osebnih podatkov, ki so shranjeni ločeno od posnetkov: kdaj se je pojavila gluhota in njena stopnja, primarna roka, stopnja izobrazbe, mesto in regija rojstva, mesto in regija izobraževalne ustanove in uporaba slušnega aparata (prim. Vintar in dr., 2012).

Glede na to, da so uporabniki SZJ znanje jezika pridobili na različne načine in v različnih starostih, je kompetenco SZJ težko oceniti. Podobno kot v mnogih drugih družbah se znakovni jezik v Sloveniji poučuje šele v zadnjem času. Tako se starejša generacija gluhih ni učila znakovnega jezika v šoli in so bili jezikovno zanemarjani, ali pa so jih učili govoriti in odgledovati. Tudi pri mlajši generaciji so razlike v znanju SZJ velike in so odvisne od mesta šolanja, saj je v Sloveniji še vedno edina šola, kjer sistematično poučujejo SZJ, Zavod za gluho in naglušno mladino Ljubljana. Seveda je pomemben podatek tudi stopnja gluhote.

Ker smo korpus Signor gradili z namenom jezikoslovnega opisovanja SZJ, smo vprašanje kompetence rešili pragmatično. Zavzeli smo stališče, da je uporabnik SZJ tisti, ki pogosto uporablja ta jezik za primarno komunikacijo z drugimi. Podatke, ki bi lahko vplivali na kompetenco in rabo, smo shranili kot meta podatke. Tako liberalen način se je izkazal za pretežno uspešnega, res pa je, da smo pri analizi posnetkov opazili enega informanta, ki je v večji meri uporabljal govor z minimalno uporabo SZJ.

Dogovarjanje za snemanja, komunikacijo z informanti in na koncu intervju ter snemanje so izvajali izključno gluhi študentje. Nekatera snemanja so potekala v društvih in nekatera na domovih informantov. Poleg tega smo v korpus vključili posnetke dijakov na Zavodu za gluhe in naglušne Ljubljana. Pred snemanjem mladoletnih oseb smo pridobili pisno privolitev njihovih staršev.

Snemanje je potekalo v treh delih. V prvem delu je informant v znakovnem jeziku pripovedoval o svojem življenju in družini. Namen tega dela je bila tudi vzpostavitev neformalnega dialoga med spraševalcem in informantom, ki slednjemu omogoči, da se sprosti in navadi na kamero. V drugem delu je informant pred snemanjem pogledal posnetek o splošni temi (npr. politika, telo, potovanje itd.). Zadnji del snemanja je bil namenjen zbiranju strokovnega besedišča in je lahko potekal kot pogovor med spraševalcem in informantom o priljubljeni temi slednjega (hobi, šport s katerim se ukvarja) ali pa je bil predvajan posnetek z bolj specializirano tematiko, o kateri je nato tekel pogovor. Posneti pogovori s posameznimi informanti so tako trajali od 10 do 20 minut.

2.2. Obdelava korpusa in označevanje

Vsi posnetki so pretvorjeni v enoličen format (.mov) in shranjeni na projektnem strežniku. Za označevanje korpusa smo uporabili orodje iLex (Hanke in Stolz, 2008), ki predstavlja prilagodljivo večuporabniško okolje za označevanje in shranjuje vse kretnje, lekseme in druge ravni oznak v bazo. Tako smo dosegli enolično uporabo oznak pri vseh označevalcih.

Označevalna shema v sedanji različici korpusa ima več ravni:

Tokenizacija. Posnetek dialoga v znakovnem jeziku je segmentiran v posamezne kretnje in ločen s časovnimi kodami. Pri segmentaciji smo se odločili za natančnejšo različico, ki zahteva več dela, a je za nadaljno analizo bolj natančna: prehodov nismo obravnavali kot dele kretenj, tako da je med dvema označenima kretnjama večinoma časovni presledek, ki predstavlja prehod.

Lematizacija. Označevanje posameznih kretenj z leksikalnimi oznakami ali glosi ustreza lematizaciji - vsaki kretnji dodamo enolično pomensko oznako. Označevalno okolje iLex uporablja leksikalno bazo, ki vsebuje vse poznane kretnje in njihove različice. Za osnovo leksikalne baze smo uporabili obstoječi slovar SZJ, ki so ga zgradili na Zvezi društev gluhih in naglušnih Slovenije,[1] ob vseh novih kretnjah ali pomenih pa smo leksikalno bazo dopolnili.

Izgovorjava. Artikulacija z glasom ali brez, ki spremlja kretnjo, lahko določa, potrjuje ali spreminja pomen kretnje.

Pomen. Vsaki kretnji je pripisan pomen glede na kontekst besedila.

Sestavljen pomen. Nekatere kretnje so sestavljene iz več delov, denimo DELAVKA iz kretenj DELATI in ŽENSKA. Te so označne v ločenem nivoju.

Grafični zapis v HamNoSys (Schmaling in Hanke 2001). Grafični zapis kretenj pomaga pri določanju različic kretenj in je pomemben korak h generiranju kretenj z animiranimi agenti.

Z označevanjem sta se ukvarjala dva raziskovalca: eden gluh od rojstva in drugi otrok gluhih staršev. Med označevanjem so se pojavljala številna vprašanja o segmentaciji in označevanju sestavljenih kretenj, nejasnih in nedokončanih kretnjah, razlikah med kretnjami in gestikulacijo in še mnoga druga. Reševali smo jih na najboljši možni način in se pogosto posvetovali s sodelavci iz Hamburga, ki se ukvarjajo z označevanjem korpusa nemškega znakovnega jezika DGS.

3. Korpusna analiza leksike SZJ

Razvoj korpusnega jezikoslovja temelji, vse od nastanka elektronskih korpusov v letih 1960 in 1970, na opazovanju in analizi reprezentativnih besedilnih vzorcev. V raziskovanju znakovnih jezikov so kvantitativne metode pomembne še iz enega razloga. Medtem ko pri govorjenih jezikih pojav zapisovanja privede do določene stopnje standardizacije ali vsaj soglasja o obliki zapisanih besed, je pri znakovnih jezikih celotno sporočilo v obliki vizuelne podobe, sestavljene iz gibov rok in telesa, obrazne mimike, artikulacije, gestikulacije in uporabe prostorskih elementov.

Ker znakovni jeziki nimajo standarda za zapisovanje, razen približnih grafičnih zapisov, namenjenih raziskovanju jezika in ne komunikaciji, je proces standardizacije jezika kljub želji mnogih uporabnikov težja naloga. Z analizo korpusnih podatkov si lahko pri standardizaciji SL pomagamo z različicami kretenj in primerjamo njihovo pojavnost.

Poznamo štiri sorodne korpusne raziskave za različne znakovne jezike. Morford in MacFarlane (2003) sta predstavila distribucijsko analizo ameriškega znakovnega

[1] http://www.zveza-gns.si/slovar-slovenskega-znakovnega-jezika/

[2] Ikonična kretnja je zasilni prevod angleškega izraza

jezika ASL z uporabo relativno majhnega korpusa (okoli 4000 kretenj). Veliko večjo zbirko sta uporabila McKee in Kennedy (2006), ki sta raziskala leksikalne značilnosti novozelandskega NZSL z uporabo korpusa Wellington, ki vsebuje več kot 100.000 pojavnic. V zadnjem času sta bili izvedeni še dve raziskavi. Prvo je izvedel Johnston (2011) za avstralski Auslan z uporabo označenega korpusa 63.436 pojavnic, drugo, za britanski BSL, pa Cormier s soavtorji (2011) na korpusu s 24.864 pojavnicami.

Naš pristop k leksikalni analizi je najbolj soroden Johnstonovi raziskavi (prav tam), saj uporabljamo tudi podoben način označevanja. Orodje iLex ima namreč tri pomembne lastnosti, ki omogočajo kvantitativno analizo: prvič, vse kretnje so shranjene v bazi in enolično povezane z glosi. Označevalec vedno izbere glos iz baze, razen če gre za novega, ki ga je potrebno v bazo dodati. Drugič, vsakemu leksemu dodajamo zapis HamNoSys, prav tako se ta zapis doda različicam leksema. Tako je vsak glos mogoče nedvoumno povezati z obliko kretnje. Tretjič pa ima označevalec dostop do video posnetkov vseh pojavitev določenega leksema in s tem možnost medsebojne primerjave za večjo doslednost.

iLex omogoča izvoz podatkov z uporabo ukazov SQL, zato smo za potrebe analize izvozili celotno bazo v Excel. Ker se nekateri deli analize nanašajo na semantične kategorije, ki jih sicer označevalna shema Signor ne vsebuje, smo morali nekatere dele ročno označiti.

3.1. Osnovna statistika korpusa

Celotna velikost označenega korpusa Signor je trenutno (junij 2014) 30.335 pojavnic in 2.976 različnic. 1.043 kretenj se v našem korpusu pojavi le enkrat. Najpogostejši kretnji po frekvenčni listi sta dve različici osebnega zaimka, JAZ1 in JAZ2, ki skupaj predstavljata 3,9 % celotnega zbira podatkov (glej Tabelo 1). Naslednji na spisku je POTEM, ki mu sledi kazalni zaimek TO. Skupna frekvenca prvih 10 kretenj je 10,8 %, kar pomeni, da deset najpogostejših kretenj predstavlja desetino celotnega korpusa. Pri prvih dvajsetih kretnjah je skupna frekvenca 17,4 % in pri stotih 38,9 %.

Če primerjamo naš vzorec SZJ z jezikoma Auslan in BSL, ni opaziti velikih razlik: korpus Auslana ima 55.859 pojavnic in vsebuje 6.171 različnic, od katerih je 3.606 enopojavnic, medtem ko je pri korpusu BSL 24.684 pojavnic z 2.507 različnicami, vendar število enopojavnic ni podano (Cormier 2011). Kazalna kretnja, ki predstavlja zaimek v prvi osebi, je najpogostejša tako v Auslanu kot v BSL-u s skupnima frekvencama 5 % in 6,9 % - v SZJ je frekvenca nekoliko nižja (3,9 %).

Spisek 20 najpogostejših kretenj v korpusu Signor vsebuje le štiri polnopomenske kretnje: DELATI, RAD, LETO in ŠOLA, medtem ko so preostale kretnje kazalne, kot so zaimki (JAZ, JAZ1, TO, MOJ, TAM), in ikonične, ki označujejo smer in/ali obliko.[2] Spisek vsebuje tudi glos za nejasne kretnje, saj v 184 primerih označevalca kljub sobesedilu nista mogla določiti kretnje ali pa sta s to oznako želela opozoriti na primer, o katerem se je potrebno posvetovati.

Višja frekvenca tako imenovanih funkcijskih kretenj, še posebej zaimkov in ikoničnih kretenj, se ujema z ugotovitvami pri Auslan in BSL.

Kot pričakovano je frekvenčni spisek korpusa Signor precej različen od pisane/govorjene slovenščine,[3] kjer se najpogosteje pojavlja pomožni glagol v tretji osebi *je*, ki mu sledijo vezniki (*in, da*), predlogi (*v, na, z, s*), povratnosvojilni zaimek (*se*) in preteklik *biti* (*bil*), prvi polnopomenski element pa se pojavi šele na 21. mestu (*leto*), prav tako je prvoosebni zaimek *jaz* šele na 26. mestu. To razliko lahko razložimo z dejstvom, da je slovenščina za razliko od SZJ tipični sintetični jezik, kjer sta osebni zaimek in povedek združena.

	Glos	**Pogostost**
1	JAZ	687
2	JAZ1	498
3	POTEM	354
4	TO	332
5	DELATI	247
6	PREJ	239
7	A	238
8	IKONIČNO-GIBANJE	229
9	IKONIČNO-OBLIKA	225
10	NE	225
11	JA	221
12	TAKO	218
13	MOJ	208
14	RAD	206
15	TAM	194
16	EN	192
17	NEJASNA KRETNJA	184
18	LETO	182
19	TUDI	177
20	ŠOLA	154

Tabela 1: Prvih 20 najpogostejših kretenj

3.2. Leksikalne in semantične lastnosti

V naslednjih korakih smo želeli raziskati leksikalne in semantične lastnosti besedišča SZJ. Za začetek smo želeli raziskati besednovrstno sestavo besedišča, zato smo na seznamu glosov uporabili oblikoskladenjski označevalnik za slovenščino ToTaLe (Erjavec in dr., 2010). Pri korpusih znakovnih jezikov je tak postopek iz več razlogov problematičen: prvič je glos kretnje zgolj pomenska oznaka, ki predstavlja približno preslikavo pomena v slovenščino, drugič je znano, da v znakovnih jezikih vlogo slovnice prevzemajo popolnoma drugačne

[2] Ikonična kretnja je zasilni prevod angleškega izraza *classifier*, ki se uporablja v tuji literaturi o znakovnih jezikih, pomeni pa poseben semantični razred kretenj, ki imajo atributivno vlogo in nakazujejo obliko, velikost, gibanje ipd. S tem niso mišljene polnopomenske ikonične kretnje, kot je npr. RIBA, ki oponaša gibanje ribe v vodi.

[3] Za primerjavo je uporabljen korpus Gigafida, http://www.gigafida.net.

strukture kot pri govorjenih/pisanih jezikih, in nenazadnje je kretnja s samostalniškim glosom lahko uporabljena v različnih kontekstih kot glagol, samostalnik ali določilo. Poleg tega pri samodejnem besednovrstnem označevanju označujemo osnovne kretnje in ne sestavljenih pomenov, čeprav nekateri sestavljeni samostalniki izhajajo iz glagola, ki mu dodamo kretnjo za osebo ali žensko (UČENEC = UČITI + OSEBA). Iz vseh teh razlogov gre rezultate v Tabeli 2, kjer vidimo distribucijo osnovnih besednih vrst v leksikonu SZJ, razumeti le kot zelo grob približek resničnemu stanju, saj smo pomenske oznake ali glose zgolj preslikali v besedne vrste, kot jih razumemo v slovenščini.

Daleč najpogostejša kategorija je samostalnik, ki ji sledijo glagol, prislov in na koncu pridevnik. Sumimo pa, da bi bilo število samostalnikov še višje, če bi posebej označevali tudi sestavljene kretnje.

Besedna vrsta	Št. različnic
samostalnik	1545
glagol	799
prislov	393
pridevnik	282

Tabela 2: Pogostost besednih vrst

Ročni pregled 300 najpogostejših pojavnic kretenj kaže na pomembnejše teme in semantične skupine, ki jih vsebuje naš korpus. Najpogostejši samostalniki so povezani s časom (LETO, MESEC, KONEC), družino (MAMA, BRAT, SIN), gluhoto (DRUŠTVO, ZAVOD), delom (SLUŽBA) in vsakdanjim življenjem (ŠPORT, FILM, ŠOLA, VRTEC, PRIJATELJ, RAČUNALNIK). Pogosti glagoli so povezani z delom in izobraževanjem (DELATI, UČITI SE, TISKATI, PISATI, ŠTUDIRATI), gibanjem (PRITI, HODITI, PRESELITI, POTOVATI), občutki (SLIŠATI, VIDETI, GLEDATI) in komunikacijo (KRETATI, GOVORITI, POGOVARJATI SE,). Veliko pogostih prislovov je načinovnih (RAD, LEPO, DOBRO, TEŽKO), medtem ko se pridevniki pogosto nanašajo na gluhoto (GLUH, SLIŠEČ, NAGLUŠEN), starost (STAR, MLAD, NOV), lastnost (LEP, VESEL) ali lastnino (MOJ, NJEGOV).

3.2.1. Mašila

Mašila so zanimiva skupina kretenj. Ko smo začeli z označevanjem korpusa Signor, nismo načrtovali ločevanja kretenj na semantične razrede ali pomenske skupine. Vendar pa nas je gluha označevalka kmalu opozorila, da so določene kretnje uporabljane predvsem kot mašila v toku pripovedi, ki pogosto nakazujejo fazo razmišljanja, kako formulirati preostanek izjave.

Skupna frekvenca teh kretenj je 647 pojavnic, kar je 2,1 % našega korpusa. Našli smo 33 različnic kretenj, ki jih lahko obravnavamo kot mašila. V Tabeli 3 je prikazanih deset najpogostejših.

Poglobljene analize mašil in njihove vloge v kretanem besedilu sicer še nismo izvedli, vendar iz naših vzorcev razberemo, da se nekatera mašila uporabljajo kot ločilo med pomenskimi deli kretanega besedila.

	Pogostost
TAKO	218
KAJ PA VEM	82
TO JE VSE	60
KAJ ŠE	55
TAKO JE	35
EH	34
KAJ ČEŠ	30
TA	20
TO JE TO	20
KAKO ŽE	13

Tabela 3: Deset najpogostejših mašil v SZJ

3.2.2. Geste

Cormier in dr. (2011) definirajo geste kot gestam podobne kretnje oziroma nize ponazoritvenih gibov. V korpusu Signor se geste pojavijo v skupni frekvenci 550 pojavnic in predstavljajo 1,8 % celotnega korpusa. Ker geste velikokrat ponazarjajo pomen, za katerega ni primerne kretnje, ali pa se uporabijo za poudarek določenega dogodka, je njihova variabilnost precej velika; preko 130 različnic je opredeljenih kot geste. Nekateri so po pomenu podobni mašilom, vendar je njihova vloga drugačna, spet drugi pa izražajo kompleksnejši pomen, kot npr. [vreči se na tla], [utripajoča luč] ali [sedeti na rami]. Takšni pantomimični gibi se običajno uporabljajo v spontanem kretanju in predstavljajo edinstveno lastnost znakovnega jezika, da je moč kompleksne ali sestavljene pomene izraziti izjemno gospodarno in ekspresivno.

Naša označevalca SZJ ločita med mašili in gestami s poudarkom, da je gesta vedno v funkciji nadomestila kretnje, medtem ko so mašila lahko kombinacija geste in kretnje v funkciji diskurza. Tako je mašilo NE VEM skomig z rameni, ki ga spremljajo navzgor obrnjene dlani, medtem ko gesta KAJ PA VEM predstavlja le skomig z rameni.

3.3. Variacije

Kot pri drugih znanih znakovnih jezikih je tudi v SZJ veliko sinonimov in različic kretenj. Sinonimi so definirani kot raba dveh ali več oblikovno nesorodnih kretenj z enakim pomenom, variacije pa kot raba dveh ali več oblikovno sorodnih kretenj z enakim pomenom. Eden dobro poznanih primerov sinonimije so tri različne kretnje za pomen [zdravnik], vsaka s svojo etimologijo in uporabo v različnih delih Slovenije, medtem ko je primer variacije med kar osmimi zabeleženimi variantami za prvoosebni zaimek JAZ1-JAZ8, kjer je razlika v obliki dlani in mestu telesa, kamor kaže. Žal trenutna označevalna shema korpusa Signor ne beleži razlike med sinonimi in različicami, tako da ne moremo podati kvantitativnih podatkov za posamezni frekvenci teh pojavov.

Od 2.976 različnic je 471 takšnih, ki imajo vsaj en sinonim ali različico, dve kretnji pa jih imata celo osem (JAZ in ITI). Največja variabilnost je pri kretnjah, ki

določajo količino ali obseg nečesa: kretnje za VELIKO, MALO, NIČ, VSE in KONEC. Te kretnje imajo po pet različic.

Pogostost teh različic je pomembna za morebitno leksikografsko obravnavo SZJ, kar je tudi razlog, da smo vsako različico ali sinonim označili z zapisom HamNoSys. Z uporabo tega zapisa lahko prikažemo posamezno kretnjo z animiranim agentom in leksikograf lahko vsako glos poveže z ustrezno kretnjo, ne da bi za to potreboval dostop do korpusa oziroma posnetkov v njem.

4. Zaključek

Projekt Signor predstavlja prvi poskus ustvarjanja označene, reprezentativne in avtentične zbirke besedil v SZJ. Določeni deli označevanja še potekajo, vendar lahko iz podatkov, ki so trenutno na voljo, dobimo prvi vpogled v leksikon SZJ in njegove kvantitativne lastnosti.

Predstavljene številke so do neke mere primerljive s podobnimi raziskavami, ki so bile narejene za BSL, ASL in Auslan, vendar pa je skupna pogostost kazalnih kretenj in gest v SZJ nekaj nižja kot denimo v BSL. Pri tem velja poudariti, da je primerjava pogostosti določenih jezikovnih pojavov med korpusi znakovnih jezikov približno tako nezanesljiva kot primerjava oblikoskladenjskih oznak korpusov dveh jezikov, ki uporabljata različna nabora oznak. Pomembna lekcija, ki smo se je naučili pri označevanju, je, da je razvrščanje kretenj v semantične in slovnične podrazrede subjektivno in zatorej v vsakem primeru pod vplivom osebnega jezikovnega občutka označevalca. Zavedamo se omejitev takega pristopa in v prihodnosti načrtujemo pregled vseh oznak, še posebej ikoničnih kretenj, gest in mašil.

Korpus Signor bo po zaključku primeren za analizo skladenjske strukture SZJ, kar bo predstavljalo tudi temelj za posodobitev gradiva za poučevanje SZJ. V ta namen imamo v načrtu dodajanje novih ravni označevanja, predvsem določanja meje med izjavami. Za poglobljeno analizo leksikalno semantičnih lastnosti načrtujemo poskuse z wordnetovimi sinseti. Razvijamo tudi spletni iskalnik, ki uporabniku omogoča vpogled v rabo posameznih kretenj s sobesedilom. Prek zapisa HamNoSys bo mogoče vsako kretnjo prikazati z animiranim agentom, izseki iz videoposnetkov pa bodo na voljo le za osebe, ki so dovolile objavo posnetkov na spletu.

Dolgoročno bi veljalo razmišljati tudi o razvoju sistema za strojno prevajanje med SZJ in slovenščino; takšni sistemi se že razvijajo za nekatere znakovne jezike (Schmidt in dr., 2011). Korpus Signor bi lahko uporabili kot učno množico za statistična orodja, iz oznak HamNoSys pa je mogoče generirati kretnje z animiranim agentom. Trenutno pa je ovira tudi majhnost korpusa, saj 30.000 pojavnic ni dovolj za izgradnjo jezikovnega in prevodnega modela.

Zahvala

Raziskava je nastala v okviru projekta, ki ga financira ARRS (koda projekta J6-4081). Zahvaljujemo se vsem informantom, ki so sodelovali pri projektu in prispevali posnetke za korpus Signor.

5. Literatura

Cormier, K., J. Fenlon, R. Rentelis in A. Schembri, 2011. Lexical frequency in British Sign Language conversation: A corpus-based approach. *Proceedings of the Conference on Language Documentation and Linguistic Theory 3*, uredili P.K. Austin, O. Bond, L. Marten in D. Nathan. London: School of Oriental and African Studies.

Erjavec, T., D. Fišer, S. Krek, N. Ledinek, 2010. The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Malta, 2010.

Globačnik, B., 2007. *Stališča Slovencev do slovenskega znakovnega jezika.* Magistrska naloga. Ljubljana: ISH.

Globačnik, B., 2001. Slovenski jezik in slovenski znakovni jezik. Defectologica slovenica 9/2. 19–28.

Hanke, Th. in J. Storz, 2008. iLex – A Database Tool For Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. *Proceedings of the Language Resources and Evaluation Conference 2008*, 28.-30. maj 2008.

Johnston, T., 2011. Lexical frequency in signed languages. *Journal of Deaf Studies and Deaf Education* 17:2. 163-193.

Hanke, Th., S. Matthes, A. Regen in S. Worseck, 2012. Where Does a Sign Start and End? Segmentation of Continuous Signing. In: *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon Language Resources and Evaluation Conference (LREC)* Istanbul, May 2012. 69-74.

Konrad, R., Th. Hanke, S. König, G. Langer, S. Matthes, R. Nishio in A. Regen, 2012. From form to function. A database approach to handle lexicon building and spotting token forms in sign languages. In: *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon Language Resources and Evaluation Conference (LREC)* Istanbul, May 2012. 87-94.

Kuplenik, N., 1999. O jezikovnih napakah pri pisnem izražanju gluhih srednješolcev. *Jezik in slovstvo* 44/5. 43–57.

Logar Berginc, N. in S. Krek, 2010. New Slovene corpora within the "Communication in Slovene" project. *Slavicorp conference.* Warsaw.

Lowenbraun, S., Appelman, K. in Callahan, J., 1980. *Teaching the hearing impaired through total communication*. Columbus, OH: Charles E. Merrill.

McKee, D. in G. Kennedy, 2006. The distribution of signs in New Zealand Sign Language. *Sign Language Studies* 6. 373-390.

Moderndorfer, M., 1989. Totalna komunikacija. V: Z. Juras (ur.): *O tematiki totalne komunikacije in organiziranje gluhih in naglušnih danes in jutri. Zbornik mednarodnega posveta.* Ljubljana: Zveza društev gluhih in naglšnih Slovenije. 123-129.

Morford, J. in J. MacFarlane, 2003. Frequency characteristics of American Sign Language. *Sign Language Studies* 3. 213-225.

Nishio, R., S. Hong, S. König, R. Konrad, G. Langer, Hanke, Th. in Ch. Rathmann, 2010. Elicitation methods in the DGS (German Sign Language) Corpus Project. Poster presented at the *4th Workshop on the Representation and Processing of Sign Languages:*

Corpora and Sign Language Technologies, following the 2010 LREC Conference in Malta, May 22.-23., 2010. 178-185.

Schmaling, C. in Th. Hanke, 2001. HamNoSys 4.0. Dostopno na: http://www.sign-lang.uni-hamburg.de/Projekte/HamNoSys/HNS4.0/englisch/HNS4.pdf.

Schmidt, Ch., D. Stein in H. Ney. 2011. Challenges in Statistical Sign Language Translation. *SLTAT 2011 - International Workshop on Sign Language Translation and Avatar Technology*, Berlin, Germany.

Vintar, Š., B. Jerko in M. Kulovec, 2012. Korpus slovenskega znakovnega jezika. *Zbornik 8. konference Jezikovne tehnologije, ISJT12*. 191-195.

Žele, A., 2007. Kako nevsiljivo povezovati znakovni jezik s pisnim sporočanjem. V: Jasna Bauman (ur.): *Standardizacija slovenskega znakovnega jezika v luči Resolucije o nacionalnem programu za jezikovno politiko 2007-2011*. Zbornik srečanja. Ljubljana: Združenje tolmačev SZJ. 13-17.

Žele, A. in J. Bauman, 2011. Slovenski znakovni jezik med normo in prakso. V: *Meddisciplinarnost v slovenistiki, 30. simpozij Obdobja.* Ljubljana: Univerza v Ljubljani. 557-582.