NORMATIVNE ZADREGE – EMPIRIČNI PRISTOP

Helena Dobrovoljc
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU, Ljubljana

Simon Krek
Amebis, d.o.o., Kamnik; Institut »Jožef Stefan«, Ljubljana

UDK 811.163.6'26:81'322.2

Namen prispevka je predstaviti možnosti povezovanja jezikoslovja in informatike pri oblikovanju metodologije za ugotavljanje problematičnih jezikovnih vprašanj, ki so pri opisovanju norme slovenskega standardnega jezika najpogosteje v ospredju. Prikazan bo empirični pristop k oblikovanju nabora kategorij normativnih zadreg: postopek oblikovanja nabora, dopolnjevanje nabora in namen ter cilj oblikovanja ontologije.

standardizacija jezika, podatkovno rudarjenje, korpusno jezikoslovje

The aim of the paper is to present new approaches in creating a methodology for detecting difficult areas in language codification, singled out in normative descriptions of the standard or literary Slovene language, by linking the fields of linguistics and computer science. An empirical method in creating a set of categories concerning normative dilemmas will be presented: the process of creating the list of categories, its revision and the creation of the final ontology.

language standardization, data mining, corpus linguistics

Meddisciplinarnost, humanistika in jezikoslovje ter informacijske tehnologije

Osrednja tematika simpozija Obdobja 30 je meddisciplinarnost v slovenistiki, kar napoveduje razpravo o tem, kako se slovenistična raziskovalna in jezikovna skupnost ob jezikovnih vprašanjih povezuje in oplaja z ugotovitvami drugih znanstvenih disciplin. Vprašanja, na katera danes pričakujemo kar konkretne odgovore, so si v našem humanističnem prostoru približno pred dvema desetletjema in več začeli zastavljati najprej tisti slovenisti in raziskovalci drugih humanističnih disciplin, ki jih je pot vodila na raziskovalna polja zunaj slovenskih meja. Sredi 90. let prejšnjega stoletja so začeli odkrivati, da hitro napredujoč razvoj civilizacije privilegira izključno tiste znanosti, ki so zmožne avtonomnih raziskav in ki prispevajo več k »splošnemu družbenemu blagostanju« (Hladnik 1994). Pomembno je tudi spoznanje, da se humanistične discipline ne uvrščajo med te izbrane znanosti in da jim tega statusa ne podeljuje niti njihov nacionalni predznak. Rešitev za eksistenčno krizo humanističnih disciplin se je ponujala v meddisciplinarnem povezovanju s tehnološko-naravoslovnimi vedami ali pa z usmeritvami v aplikativnost. Obe možnosti je jezikoslovje v zadnjih desetletjih izkusilo in že preizkusilo in – kot je bilo rečeno v povabilu k udeležbi na simpoziju – »zavidljivim rezultatom lahko sledimo v sociolingvistiki, korpusnem jezikoslovju, kognitivnem jezikoslovju in psiholingvistiki«, in to kljub odsotnosti sistemskih[1] spodbud (McCarty 2004). Zdi se, da je jezikoslovje s svojimi meddisciplinarnimi povezavami zavzelo posebno mesto, saj že v svoji osnovni dejavnosti skuša odkrivati zakonitosti, podobno kot naravoslovne znanosti – iz

raziskovalne želje, da bi preučevali jezik in njegove pojavne oblike čim bolj objektivno. Že leta 1861 je jezikoslovec orientalist Max Müller zapisal, da je jezikoslovje v nasprotju s humanističnimi znanostmi (kot so zgodovina, literatura in pravo) pravzaprav naravoslovna znanost – tako kot biologija in geografija. »Jezik sam postane izključni predmet znanstvenega spraševanja« (Milroy, Milroy 1999). Po drugi strani je jezikoslovje med tistimi kognitivnimi znanostmi, ki so upoštevale, da z empiričnimi pristopi pridobivajo nove perspektive, ki jih tradicionalne metode niso omogočale. Interesno polje, na katerem sta se obe vedi srečali, je korpusno jezikoslovje. To jezikoslovcem ponuja obsežne in uravnotežene zbirke raznovrstnih besedil – besedilne korpuse, ki so v raziskave jezika prinesli možnosti posploševanja, predvidevanja in napovedovanja jezikovnih vzorcev,[2] seveda pod pogojem, da so ti gradivski viri avtentični. Korpusi so jezikoslovcem omogočili, da se znebijo mnogokrat nezaželene pristranskosti in subjektivnosti, kakršno so jim očitali v preteklosti, ko so se bili prisiljeni zanašati na svoj jezikovni čut.

Povezovanje jezikoslovja in informacijskih tehnologij je pomembno tako za aplikativne kot za teoretične jezikoslovne raziskave, po drugi strani pa lahko raziskovalnemu sodelovanju obeh ved pripišemo več obojestranskih koristi.[3] Medtem ko je uporaba korpusov za leksikografske, terminološke in že omenjene raziskave, ki gradijo teoretične predpostavke na pogostnosti pojavljanja posameznih enot, očitna, pa se v okviru standardizacije jezika srečujemo s tako poimenovanimi »normativnimi zadregami«, ki jih lahko le deloma rešujemo s pomočjo korpusnih pristopov, zato bomo v nadaljevanju pokazali tudi drugačne možnosti sinergične povezave obeh ved.

Normativne zadrege in standardizacija

Nabor problematičnih vprašanj oz. normativnih zadreg slovenskega jezika smo oblikovali z empiričnim pristopom. Z izrazom »normativne zadrege« smo namreč poimenovali vsa tista mesta v jezikovnem sistemu, ob katerih je uporabnik jezika oz. govorec negotov. Občutek negotovosti in nekompetentnosti, ki ga pogosto pripisujemo »škrbinam« v posameznikovem jezikovnem znanju, je lahko tudi posledica jezikovne neustaljenosti, ki jo v procesu jezikovnega opisa, standardizacije oziroma normiranja zaznajo tudi jezikoslovci. Normativne zadrege so torej jezikovni problemi tako govorcev kot jezikoslovcev, in če pri prvih vzbujajo splošni občutek jezikovne negotovosti in nezmožnosti, dajejo pri drugih vtis neučinkovitosti. Kljub našemu zavedanju, da je jezikovna ustaljenost mogoča le v idealnih pogojih, si s kontinuiranim posodabljanjem jezikovnega standarda želimo doseči stanje, ki ga s pomočjo praškega izraznega aparata opisujemo kot »prožno ustaljenost« ali v novejši praški (korpusnojezikoslovni) tradiciji »koncept minimalne intervencije« (Cvrček 2008). To pomeni, da v okviru zavestnega prizadevanja za vsestransko razvit in ustaljen nadregionalni knjižnojezikovni standard evidentiramo jezikovne spremembe, ki so

1 Willard McCarty (*A brief history of humanities computing*; 1964–1970, 2004) opozarja na to, da je pri povezovanju humanistike in informacijskih tehnologij na eni strani opazna abstinenca vodilnih ustanov v tradicionalnih akademskih okoljih in odsotnost motivacije tistih, ki bi morali o tehnologizaciji odločati.

2 V slovenski teoriji jezikovne naravnosti je pogostnost jezikovnega pojava, vzorca, zgradbe ali enote tesno povezana s t. i. naravnostjo oz. zaznamovanostjo. Prim. Dobrovoljc (*Slovenska teorija jezikovne naravnosti*, 2005: 46) in Fenk - Oczlon (*Familiarity, information flow, and linguistic structure*, 2001), ki trdi, da bi namesto izraza *nezaznamovanost* lahko uporabljali kar *pogostnost*.

3 Erjavec (*Računalniške zbirke besedil*, 1996/97) opozarja na pomen korpusnih raziskav tudi za področja jezikovnih tehnologij. Hladnik v prispevku *Količinske in empirične raziskave literature* (1994) poudarja, da se humanistične in znanstvene discipline »zlivajo med seboj« zlasti na področju empiričnih raziskav in da je interes obojestranski – »vplivi grejo v obe strani«.

posledica jezikovnega razvoja in ki jih prinašajo med drugim tudi pogovorni substandardi, in na ta način poskrbimo, da se knjižna in govorna norma čim manj razhajata. V procesu standardizacije jezika se v štirih fazah odvrti krožni postopek:

(1) ugotavljanje jezikovne rabe,
(2) ugotavljanje jezikovne norme,
(3) zapis jezikovne norme,
(4) preverjanje ustreznosti kodifikacije.

Četrta faza, torej preverjanje ustreznosti kodifikacije, poteka na različnih ravneh; eden od načinov preverbe je tudi ugotavljanje jezikovnih zadreg pri uporabnikih oz. usklajenost njihove jezikovne rabe z aktualnimi kodifikacijskimi določili. Prvi pogoj za verodostojno razčlenitev dejanskega stanja v rabi je ustrezna gradivska osnova. Zaradi ugotovitev, da je nelektoriranih besedil v korpusih manj (Erjavec 2008: 11), kot bi si za naše potrebe želeli, smo pri postopku zbiranja podatkov o normativnih zadregah, ki bodo neodvisni od vseh dosedanjih ugibanj, izbrali drug pristop. Ugotovili smo namreč, da nam v elektronski dobi dovolj relevantno zbirko jezikovnih problemov ponujajo že spletni forumi ali svetovalnice za jezikovne probleme, kjer se obiskovalci »oglašajo« brez posrednika in prosijo za mnenje strokovnjaka jezikoslovca, lektorja, ponekod pa so deležni tudi komentarjev in mnenj drugih jezikovnih uporabnikov. Zato smo se odločili, da empirično ugotovimo, glede katerih vprašanj so obiskovalci spletnih forumov najpogosteje negotovi.

Empirični pristop k normativnim zadregam

Pri izbiri gradiva za potrebe analiz pri četrti fazi, tj. preverjanje ustreznosti kodifikacije, so bili tako izbor gradiva kot postopki vnašanja morebitnih metainformacij usmerjeni predvsem v identifikacijo tistih delov kodifikacijskega sistema, pri katerih se kaže razkorak med odločitvami tvorcev besedil in sprejeto normo, ne glede na dejstvo, ali se sami tvorci tega razkoraka zavedajo ali ne. Postopek identifikacije normativnih težav je osnovan na dveh temeljnih virih. Prvi vir je večja količina gradiv, ki bi jih lahko imenovali »avtorefleksivna«[4] – to so zbirke vprašanj, odgovorov ali debat na tistih spletnih forumih ali spletnih straneh, ki so izrecno posvečeni jezikovnim zadregam pri slovenščini. Forumi in druge oblike komunikacije na spletu (klepetalnice, blogi, socialna omrežja itn.) v času velike rasti svetovnega spleta v zadnjih letih nudijo povsem novo dimenzijo spremljanja tematik, ki so v središču zanimanja določene skupnosti. Ta dejavnost je v svetu spletnega podatkovnega rudarjenja znana pod imenom *opinion mining*, *sentiment analysis* ali *opinion extraction*, torej rudarjenje oz. luščenje mnenj ali analiza odnosa oz. naravnanosti do določene teme. Čeprav v našem primeru ne gre za povsem enake postopke, so predvsem viri za spletno podatkovno rudarjenje in analize normativnih težav enaki.

Forume in spletne strani s postopki spletnega pajkanja (ang. *web crawling*) shranimo in obdelamo za potrebe kasnejše kategorizacije. Spletno pajkanje v grobem zajema tri postopke, ki so povezani v enoten program:

a. uporaba spletnega pajka (ang. *web crawler*): na podlagi danih parametrov,[5] ki jih določimo sami, išče spletne strani in jih shranjuje na lokalne nosilce podatkov;

b. odstranjevanje elementov formata HTML (ang. *boilerplate removal*): ker za analizo potrebujemo zgolj besedilo spletnih strani, ne pa tudi vseh metainformacij, ki jih

4 Kot »avtorefleksivno« pojmujemo tisto dejavnost govorcev, ki je povezana z iskanjem izhoda iz lastnih jezikovnih zadreg brez pomoči strokovnjaka, torej posrednika.

5 To so lahko konkretne spletne strani s podstranmi, druga in pogostejša možnost pa je pajkanje s pomočjo t. i. semenskih besed (ang. *seed words*) in mehanizmov, ki jih ponujajo splošni iskalniki (Google, Yahoo itn.). Primer uporabe takih mehanizmov za gradnjo korpusov je denimo WebBootCat (Baroni idr. 2006).

te vsebujejo in so zakodirane v formatu HTML, je treba te prvine odstraniti, kar ni povsem trivialen postopek;

c. odstranjevanje podvojenih delov besedila (ang. *duplicate removal*, *de-duplication*): ker je narava spleta taka, da se informacije velikokrat ponavljajo (ista novica, objavljena na različnih spletnih straneh, objava enakih spiskov, sporedov ipd.), sta identifikacija in odstranjevanje odvečnih ponavljajočih se delov besedil nujna.

Tako obdelana besedila lahko potem nadalje obdelujemo z različnimi jezikovnotehnološkimi postopki (oblikoskladenjsko označevanje, skladenjsko razčlenjevanje, prepoznava imenskih entitet itn.) in s tem izboljšamo možnosti nadaljnjih analiz.

Pri vzpostavljanju sistema izgradnje gradivne baze za potrebe ugotavljanja jezikovne rabe smo na začetku sami izbrali dve tipični strani, kjer poteka tovrstna spletna komunikacija. Prva je spletna stran ŠUSS – *Odgovori na jezikovna vprašanja* (http://www2.arnes.si/~lmarus/suss/), ki nima forumske, torej debatne narave, temveč gre za klasični sistem odgovorov na vprašanja. Izbran je bil med drugim tudi zato, ker je vsebinsko bližji načrtovanemu jezikovnotehnološkemu orodju, ki ga navadno imenujemo sistem odgovorov na vprašanja (ang. *question answering system*). Drugi vir je klasični spletni forum, kjer poteka debata na določeno temo in je tako bližji jezikovnotehnološkemu orodju, ki ga imenujemo virtualni agent ali klepetalnik (ang. *chatbot*, *virtual agent*). Za te potrebe smo izbrali spletni forum *med.over.net* oz. njegov del, ki je namenjen vprašanjem v zvezi z jezikovnimi zadregami »Al' prav se piše ...?«. Predvideno je, da bo po preizkusnih postopkih kasneje potekalo tudi pajkanje celotnega spleta s pomočjo semenskih besed, ki bodo izbrane iz testne zbirke besedil.

Po strojni obdelavi pajkanih besedil smo te za potrebe ročne obdelave oz. kategorizacije uvozili v program Excel in začeli s postopki ročnega kategoriziranja vprašanj.[6] Pred tem pa smo morali oblikovati začasni nabor, ki smo ga poimenovali prva kategorizacija in ki je nastal ob pregledovanju obstoječih normativnih priročnikov. V tem postopku smo skušali ugotoviti, pri katerih vprašanjih bo zaradi razhajanja med dejansko rabo in zapisanimi določili največ normativnih zadreg. Predvidevali smo, da bodo uporabnike k vprašanju spodbudile tiste jezikovne težave, pri katerih si zaradi različnih razlogov s priročniki ne morejo pomagati. Ugotovili smo, da gre za več vrst tovrstne problematike:

1. Pogosto je aktualna rešitev (govorimo o t. i. obstoječi dogovorni rešitvi) v razkoraku z jezikovno rabo in določili, ki urejajo rabo teh jezikovnih prvin na drugih področjih. V ta sklop smo uvrstili npr. pisanje imen blagovnih znamk in industrijskih izdelkov ter zdravil, pa tudi rabo velike oz. male začetnice pri izlastnoimenskih pridevnikih s priponskimi obrazili *-ov*, *-ev*, *-in* (*ahilova*/*Ahilova tetiva*) ter imenih praznikov (*dan*/*Dan državnosti*) in posebnih nagrad (*Zlati*/*zlati znak ZRC SAZU*). Vzorčni primeri za neskladje med normativnimi določili in rabo so tudi stava pristavčne vejice in pa nekatera določila, povezana z oblikoslovnim in besedotvornim pregibanjem posameznih tujih imen (*Franz – z Franzom*/*z Franzem*; *Keats – Keatsov*/*Keatsev*; *Jean Paul – Jeanpaulov*/*Jean Paulov*).

2. Drugi sklop predstavljajo prezapletena normativna določila. Zadrege so pogoste zlasti pri jezikovnih prvinah, ki so v jeziku nove, še nenormirane, saj uporabnik nima dovolj podatkov, da bi po analogiji z že standardiziranimi primeri znal prvino ustrezno uporabiti. Ta sklop vključuje problem zapisovanja predložnih zemljepisnih imen in večbesednih ledinskih imen, kjer je pravilo zasnovano tako, da mora uporabnik poznati specifično lokalno situacijo (npr. *Pod Hribom*/*Pod hribom*).

6 Postopek je izvajala delovna skupina za *Slogovni priročnik* spomladi 2010.

3. Probleme, ki smo jih uvrstili v tretji sklop, povezuje dejstvo, da aktualna določila v pravopisu oziroma opisi v slovnici ne zajemajo vseh mogočih načinov zapisa, izjem in podrobnosti. V ta sklop sodi velika večina vprašanj, povezanih s prevzemanjem besed in zvez (ohranjanje citatnosti in stopnja podomačitve v odvisnosti od področja rabe posamezne besede; prevzemanje stvarnih lastnih imen), ter položaji, v katerih je raba tujih jezikovnih prvin citatna (*Rimski-Korsakov* proti *Rimskega - Korsakova*). Precej je tudi problemov, povezanih z zapisom skupaj ali narazen pri različnih zloženkah, npr. *evro-atlantski* proti *avstro-ogrski* ali *format A4* proti *2. b-razred*.

4. V posebno skupino smo vključili vse tiste jezikovne probleme, za katere v sodobnih normativnih priročnikih ne najdemo rešitve, saj ti problema (še) niso evidentirali. Največ teh je povezanih s posebnostmi jezikovne rabe v informacijskih tehnologijah (ločila, njihova stičnost, zapis malih in velikih črk, uporaba velikih črk sredi besede …), precej pa je tudi novih predmetnosti (imena plezalnih smeri, pešpoti, znamenitosti, parkov, prireditvenih prostorov, plinovodov …). Neevidentirano je tudi pisanje prirednih zloženk brez vezaja (*arteriovenski*), položaji za stavo vejice, kadar ta upade na besednozvezno raven (*to je bilo reci piši v petih minutah*; *na koncu je hočeš nočeš popustil*), posamezni primeri stave vezaja in pomišljaja (*razmerje posameznik – družba*; *Mojca Koželj - Pevec*), pisanje svetniških (ne)naselbinskih imen (*sv./Sv. Primož*) itn.

Na podlagi analize različnih normativnih virov smo torej izdelali nabor predvidenih težav jezikovnih uporabnikov in ga imenovali prva kategorizacija; s to smo si pomagali pri ročnem kategoriziranju.

Dobljeno kategorizacijo smo v želji po celostni obravnavi problemov dopolnjevali z aktualnimi in priložnostnimi vprašanji, ki jih jezikovni uporabniki zastavljajo posameznim jezikoslovcem javno ali zasebno, tako z vidika nizanja novih kategorij kot tudi gradivsko. Gradivsko smo izpopolnjevali kategorizacijo

2.12	**POSEBNOSTI 1. MOŠKE SKLANJATVE**		
2.12	Sklanjanje srbohrvaških imen	C	*Kragujevac*
2.12.1	Sklanjanje imen na nemi -e	C	*George*
2.12.2	Daljšanje osnove z j		*Heidegger* *Ranciere* *Baudelaire* *Shakespeare*
2.12.3	Daljšanje osnove s t	C	*Jaka (pog.)*
2.12.4	Daljšanje osnove z n	C	*nagelj*
2.12.5	Pregibanje angleških in francoskih imen, ki se končujejo na *-re*	A	*Moliere* *Ranciere* *Baudelaire* *Shakespeare*
2.12.6	Pregibanje francoskih imen na nemi soglasnik	A	*Dumas*
2.12.7	Orodnik in preglas	A	*bruc, z brucom/z brucem* *Franz* *Fritz* *Bush*
2.12.8	Pregibanje tujih imen na samoglasnik	C	*Galileo Galilei*

Slika1: Poskus prve kategorizacije jezikovnih problemov oziroma normativnih zadreg

KAT.	FREK.	VSEBINA	ZGLEDI
E3a	91	**Pomen**	*mediokritet, vizažist, skaz, ergološki; iti rakom žvižgat*
E1c	46	**Izbira med domačimi sopomenkami**	*bolnica – bolnišnica, višek – presežek, vezica – skoba, pijan – vinjen, treba – potrebno; komuniciranje-komunikacija; prati – umivati; dalje – naprej; utoniti – utopiti se; vzrok – razlog; poleg – zraven*
E1a	43	**Ali je s to besedo res vse v redu?**	*trošarina (je »grdo«), homoseksualec, permutabilen, le-ta; postoriti prekršek, ali lahko uporabim* glagol *pripustiti*, kadar nismo na področju veterine?
F2b	39	**Druga vezljivost (predlogi, vezniki)**	*čez/prek(o) ceste, delati v/na vrtu; v spomin Jožetu Kastelicu/Jožeta Kastelica; grem na/v Ptuj; zaščita pred/proti sevanjem; lahko sodelujejo vsi, razen zaposlenih/zaposleni*
E1b	35	**Ali lahko napišem/rečem, da …?**	*vlogo podajamo/predajamo/dajemo*
E3b	29	**Izvor**	*kodrlajsast <?, domačija Pužman, kiper <?*
E1d	28	**Izbira med sopomenkami: domače – prevzeto**	*tabela – preglednica, slika – fotografija, kakovost – kvaliteta, online – spletni/na internetu, e-naslov – e-mail*
A1b7	27	**Vrstna poimenovanja društev, organizacij ipd., strank ipd. ter organov, služb, podjetij**	*občina – Občina Škofja Loka; ustava – Ustava Republike Slovenije*
A2a1	21	**Vejica pri pristavku**	*Nemška kanclerka Angela Merkel je dejala, da bo okrepila nemško gospodarstvo*
G2	19	**Oblika dopisa: glave, ogovori, pozdravi, podpisi ipd.**	*Kdo je naveden prvi: tisti, ki piše, ali tisti, komur pišemo?; dr., g.; direktor …, Lepo pozdravljeni/S spoštovanje!; poravnava, kje je datum …;*

Slika 2: Vzorčni pogostnostni seznam kategorij po prvi kategorizaciji

tudi z različnimi korpusi, ki so nastali v okviru projekta *Sporazumevanje v slovenskem jeziku*, med njimi predvsem korpus *Gigafida* z 1,15 milijarde besedami in korpus *Šolar* z milijon besedami. Korpus *Gigafida* je trenutno največja in najbolj raznovrstna zbirka besedil v slovenščini, ki je jezikoslovno označena in pripravljena za jezikoslovno podatkovno rudarjenje. Korpus *Šolar* (Rozman idr. 2011) je korpus šolskih pisnih izdelkov, ki so jih učenci slovenskih osnovnih in srednjih šol samostojno tvorili pri pouku, in omogoča predvsem identifikacijo kodifikacijskih težav šolajoče se mladine. Tretji način dopolnjevanja kategorizacije je potekal vzporedno z ročno kategorizacijo in s pomočjo rezultatov le-te: obstoječi nabor namreč še ni vseboval vseh vprašanj, ki so se pojavljala v obeh izbranih forumih.

V prvi fazi ročnega označevanja s pomočjo prve kategorizacije smo označili nekaj več kot 2500 vprašanj s forumov *med.over.net* in ŠUSS. Nato smo naredili vzorčni pogostnostni seznam označenih kategorij. V nasprotju s predvidevanji ob pregledu obstoječih priročnikov, torej s prvo kategorizacijo, je pogostnostna razvrstitev najpogosteje zastavljenih vprašanj pokazala, da normativnih zadreg ne moremo pripisati le slabostim v kodifikaciji, torej nejasnosti ali prezapletenosti jezikovnih priročnikov. Najpogosteje zastavljena vprašanja se namreč ne nanašajo na izrazno podobo besede, njen zapis ali izreko, temveč gre večinoma za probleme,

povezane z uporabo jezikovnih sredstev v njihovem kontekstu:

- *kakšen je pomen besede A?*
- *s katero (domačo) sopomenko lahko nadomestim besedo A oziroma zvezo B C?*
- *ali lahko uporabim besedo A v besedilu B?*
- *ali lahko napišem/rečem, da …?*
- *kdaj izbrati domačo, kdaj »tujo« besedo?*

Na probleme, ki jih imajo uporabniki pri izbiri leksikalnih enot in pri izbiri njihovih besedilnih funkcij tradicionalni normativni priročniki, kakršen je na primer pravopis, ne dajejo celovitih odgovorov. Prva tipologija na spletu zastavljenih vprašanj je tako pokazala, da so »normativne zadrege« v manjši meri povezane z neustreznostjo obstoječe kodifikacije in v večji meri z recepcijo le-te, torej z nedostopnostjo in pomanjkanjem jezikovnih priročnikov, ki bi bili usmerjeni v razumljivo in praktično razlago najbolj perečih vprašanj. Ugotovitev potrjuje tudi primerjava dobljenega seznama kategorij z rezultati ankete, izvedene med jezikovnimi uporabniki poleti 2010 v okviru projekta Sodobni pravopisni priročnik …,[5] saj so anketiranci na vprašanje, katere podatke najpogosteje iščejo v slovarju, najpogosteje odgovorili, da iščejo: podatek o zapisu besede, o pomenu besede in o rabi besede v besedni zvezi ali rabi oblike, predloga. Kljub temu pa je treba upoštevati, da so predstavljeni rezultati vzorčni in da bomo dobili celovitejšo sliko ob analizi večjega nabora jezikovnih vprašanj.

Od vprašanj k ontologiji problemov

O smiselnosti ugotavljanja normativnih zadreg nas ne prepričujejo le potrebe, nastale ob samem postopku standardizacije, temveč bo nabor problematičnih mest jezikovne rabe uporabljen tudi kot ogrodje t. i. spletnega slogovnega priročnika.[6] V okviru tega bo uporabniku omogočeno, da dobi informacijo normativnega značaja v treh pojavnih oblikah oz. formalno v treh rubrikah: *Kratko in malo*, *Na dolgo in široko*, *Za navdušence*. Metodologija, potrebna za oblikovanje mikrostrukture odgovorov na vprašanja, ki jih zastavlja tipologija, se izdeluje vzporedno z ročnim označevanjem vprašanj in bo predmet nadaljnjih razprav, tu naj omenimo le, da bomo za celovitejši prikaz problematike posamezne kategorije v odgovorih izpopolnili s podatki iz leksikalnih, korpusnih in bibliografskih virov, in sicer s pomočjo metode podatkovnega rudarjenja. Kot primer lahko navedemo vprašanje izločanja frekvenčnih podatkov o svojilnih pridevnikih, izpeljanih iz tujih lastnih imen, ki se končajo na govorjeni *r* in imajo za seboj nemi *-e*. V spodnji preglednici so navedeni podatki za najpogostejše tovrstne pridevnike v korpusu *Gigafida*.

Z opisanim izpopolnjevanjem bo naša kategorizacija postala ontologija problematičnih vprašanj izrazne ravnine jezika. Posamezna enota v tej ontologiji bo oblikovana tako, da bo zajela čim bolj zaokrožen pojav, ki je s stališča norme ali standarda dojet kot težaven, in da bo na vprašanje, ki ga odpira, mogoče odgovoriti z zgoraj omenjenim tripartitnim sistemom (*Kratko in malo – Na dolgo in široko – Za navdušence*). Bistvenega pomena je tudi, da bo osnovna enota ontologije čim bolj konsistentno povezljiva s podatki v leksikalnih in korpusnih virih in čim bolje uvrstljiva v sprejeti sistem (oz. ontologijo). S takim kategoriziranjem gradiva sicer ne bomo kar takoj rešili vseh

5 Sklicujemo se na anketo med več kot 300 jezikovnimi uporabniki, ki smo jo poleti 2010 izvedli v okviru aplikativnega raziskovalnega projekta Slovenski pravopisni priročnik v knjižni, elektronski in spletni različici (L6-0166), ki sta ga sofinancirala ARRS in SAZU v obdobju od 1. 2. 2008 do 30. 1. 2011. Rezultati ankete še niso bili javno objavljeni.

6 V okviru projekta Sporazumevanje v slovenskem jeziku (2008–2013) bodo nastali referenčni korpus in leksikalna baza slovenskega jezika s slovničnim analizatorjem ter nova didaktika poučevanja slovenskega jezika, pedagoška korpusna slovnica in slogovni priročnik.

		-rejev	*-reov*	*-rjev*	*-rov*	**Skupna vsota**
Shakespea(re)	*-a*	3	2	64	253	322
	-e	1	5	130	564	700
	-ega	6	7	111	436	560
	-em		1	44	257	302
	-emu			1	24	25
	-i	2	2	44	214	262
	-ih	5	5	166	494	670
	-im	1		13	71	85
	-ima			1	2	3
	-imi	1		4	24	29
	-o	3	3	69	313	388
	(prazen)	5	7	52	326	390
Vsota *Shakespea(re)*		27	32	**699**	**2978**	3736
Moo(re)	*-a*		5		102	107
	-e	1	5		24	30
	-ega		6		93	99
	-em		4		30	34
	-emu				16	16
	-i				20	20
	-ih				26	26
	-im		4		17	21
	-ima				1	1
	-imi				3	3
	-o				44	44
	(prazen)	2	14		249	265
Vsota *Moo(re)*		3	38		**625**	666
Pha(re)	*-a*				45	45
	-e		1		39	40
	-ega	2	2		179	183
	-em	1			20	21
	-emu				1	1
	-i				16	16
	-ih				113	113
	-im				12	12
	-imi				13	13
	-o	1			17	18
	(prazen)	1	2		97	100
Vsota *Pha(re)*		5	5		**552**	562

Preglednica 1: Primer izločanja frekvenčnih podatkov za potrebe *Slogovnega priročnika*

»normativnih zadreg« jezikovnih uporabnikov, vendar pa bomo z njihovim evidentiranjem pripomogli h koncipiranju priročnikov, ki v svoji obvestilnosti sledijo potrebam uporabnikov.

Literatura

BARONI, Marco, KILGARRIFF, Adam, POMIKÁLEK, Jan, RYCHLÝ, Pavel, 2006: *WebBootCaT: instant domain-specific corpora to support human translators. EAMT-2006: 11th Annual Conference of the European Association for Machine Translation*. Oslo: Norway. Proceedings. 247–252.

CVRČEK, Vàclav, 2008: *Regulace jazyka a Koncept minimální intervence*. Praga: Nakladatelství Lidové Noviny.

DOBROVOLJC, Helena, 2005: *Slovenska teorija jezikovne naravnosti*. Ljubljana: Založba ZRC, ZRC SAZU.

ERJAVEC, Tomaž, 1996/97: Računalniške zbirke besedil. *Jezik in slovstvo* 42/2–3. 81–96.

ERJAVEC, Tomaž, 2008: *Analiza metapodatkov korpusa FidaPLUS. Institut Jozef Stefan*. http://nl.ijs.si/et/project/Fida/Quant/Analiza_FidaPLUS.pdf

FENK - OCZLON, Gertraud, 2001: Familiarity, information flow, and linguistic form. John Bybee in Paul Hopper (ur.): *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins. 431–448.

HLADNIK, Miran, 1995: Količinske in empirične raziskave literature. *Slavistična revija* 43/3. 319–340.

McCARTY, Wilard, 2004: *A brief history of humanities computing*; 1964–1970. http://lists.village.virginia.edu/lists_archive/Humanist/v17/0771.html

MILROY, James, MILROY, Lesley, 1999: *Authority in language: investigating standard English*. London, New York: Routledge.

ROZMAN, Tadeja, STRITAR, Mojca, KRAPŠ VODOPIVEC, Irena, KOSEM, Iztok, KREK, Simon, 2010: *Nova didaktika poučevanja slovenskega jezika: sporazumevanje v slovenskem jeziku*. Ljubljana: Ministrstvo za šolstvo in šport: Amebis.