NAREČNI FRAZEOLOŠKI SLOVAR – PRVI KORAKI

Jernej Vičič
Fakulteta za matematiko, naravoslovje in informacijske tehnologije, Koper

Karin Marc Bratina
Filozofska fakulteta, Ljubljana;
Scuola Superiore di Lingue Moderne per Interpreti e Traduttori, Trst

UDK 81'28'373.7'374.2:004.7

V prispevku Karin Marc Bratina (2009b) je bil podan osnutek gesla narečnega frazeološkega slovarja v e-obliki, ki bi poleg standardnih leksikografskih elementov podal tudi avdio izsek kontekstualne rabe ter prikaz umestitve frazema na karti narečnih frazemov. Slovarsko geslo je bilo zasnovano kot povezava standardov leksikografskega in dialektološkega dela s prednostmi obdelave s sodobnimi jezikovnimi tehnologijami. Logično nadaljevanje zasnove je nadgradnja gesla in realizacija e-slovarja narečne frazeologije na podlagi narečnega korpusa ter razširitev na vseslovenski narečni prostor. V prispevku predstavljamo spletno aplikacijo, ki bo skupnosti pomagala pri zbiranju narečnih frazemov. Aplikacija bo prosto dostopna.

narečna frazeologija, e-narečni frazeološki slovar, spletna aplikacija, množično zunanje izvajanje

The paper from Karin Marc Bratina (2009b) presented a draft entry from a dictionary of dialectal idioms in electronic format, which would include, in addition to the standard lexicographical elements, an audio clip of contextual use and geo-spatial information of the phraseme on the map of dialectal phrasemes. The dictionary entry is presented as a combination of the standards applied by previous lexicographic and dialectology work with the usage of modern language technologies. A further improvement is an upgrade of the entry as well as a realization of an e-dictionary of dialectal phraseology and expansion of Slovene dialects based on dialectal corpora. This paper presents a web application that will help a community collect dialectal phraseology. The application is freely accessible.

dialectal phraseology, e-dictionary of dialectal phraseology, web application, crowdsourcing

1 Uvod

Namen prispevka je predstaviti prva gesla interaktivnega e-slovarja slovenske narečne frazeologije; zasnova in zgradba gesla, oblikovanega na podlagi gradiva iz slovenskega istrskega narečja, sta bili predstavljeni v Marc Bratina 2009b. Takrat so bila predlagana teoretska izhodišča in ureditev slovarja na temeljih korpusne leksikografije, v okviru analize frazeoloških pomenov pa sta bila poudarjena pomen različnih kulturnosemantičnih pristopov (lingvokulturologija, etnolingvistika, etimologija,

kognitivna in konceptualna interpretacija sestavin frazema ipd.) ter pomen prve asociacije, izpričane s strani rojenih govorcev. Podan je bil osnutek slovarskega gesla, ki bi poleg standardnih leksikografskih elementov vseboval tudi avtentičen avdio izsek kontekstualne rabe frazema ter vizualno predstavitev s pomočjo sistema GIS.

Novost pričujočega prispevka je nadgradnja interaktivnega e-slovarja, v katerega bo lahko vsak s pridobljenim geslom vnašal primere iz svojega govora – lahko bi rekli, da gre za zametek vseslovenskega narečnega frazeološkega e-slovarja. Namesto sistema GIS smo za umeščanje frazema v prostor uporabili preprostejši Googlov sistem (Zemljevid). V prispevku natančno prikazujemo označevanje izsečkov kontekstualne rabe frazema, kot je to običajno pri gradnji govornega korpusa.[1]

2 Metodologija

Za temeljitejšo leksikografsko obdelavo slovarskega gesla (vseslovenskega) narečnega frazeološkega slovarja izhajamo iz posnetih govorjenih besedil, ki jih je za namen sekundarnih jezikoslovnih raziskav dobro urediti v narečni korpus. V prispevku predstavljamo način obdelave govornih jezikovnih virov za gradnjo korpusa, ki služi prvenstveno kot vir za raziskave narečne frazeologije, kot rečeno pa se lahko uporablja tudi za druge jezikoslovne raziskave. Gre za eno od pomembnejših faz pri urejanju gesla za narečni frazeološki slovar,[2] hkrati pa je to lahko zametek samostojnega projekta, tj. specializiranega narečnega korpusa, ki ga sicer lahko za svoje jezikoslovne namene uporabljajo tudi drugi raziskovalci.

2.1 Gradnja korpusa

Pri gradnji narečnega korpusa (oz. govornih zbirk) za namen narečnofrazeografske obdelave upoštevamo priporočila za gradnjo govornega korpusa iz Zemljarič Miklavčič 2008, smernice, ki so jim sledili sodelavci pri gradnji GOS (Verdonik, Zwitter Vitez 2011), ter seveda tudi predloge, upoštevane pri prvem narečnem korpusu GOKO, predstavljene v Šumenjak 2013. Naj poudarimo, da je narečno gradivo, ki ga dialektolog pridobi in posname v okviru terenskega dela (v našem primeru gre za vodene intervjuje), urejeno v korpus, dandanes zgolj logična posledica nadgradnje njegovega dela: že pred razvojem korpusnega jezikoslovja so namreč dialektologi pri svojih raziskavah izhajali iz govorjenih (fonetično transkribiranih) besedil, z razvojem korpusnega jezikoslovja, natančneje, priporočil za izgradnjo govorjenih korpusov pa se (sicer že fonetično transkribirana) govorjena besedila uredijo (označijo in dodatno transkribirajo) za gradnjo korpusa.

Gradivo za nastajajoči narečni korpus je nastalo na podlagi kontekstov, v katerih so narečni govorci ponazorili rabo določenega frazema. Za nadaljnjo obdelavo oz. urejanje je treba govorjena besedila ustrezno prekodirati/transkribirati. Odločili smo se

1 Natančneje, govorimo o t. i. govornih zbirkah (Gros idr. 2003; Gorjanc 2005; Gorjanc, Fišer 2010).

2 Najpomembnejša faza pri zbiranju narečnih frazemov je terensko delo s pomočjo narečne frazeološke vprašalnice. Nanj se je treba predhodno pripraviti s pregledom drugih virov, iz katerih npr. črpamo podatke o zunajjezikovni stvarnosti nekega narečja, ter študijem del s področja folkloristike in narečnega leposlovja (podrobneje o etapah pri zbiranju narečne frazeologije v Marc Bratina 2009a).

za trojno prekodiranje, in sicer 1. za fonetični zapis, saj so govorjene zbirke namenjene tudi ali predvsem dialektologom; 2. za t. i. poknjiženi/standardizirani zapis primerov rabe,[3] da lahko govorjene zbirke/korpus uporabljajo tudi drugi jezikoslovci; in 3. za prevod v slovenski knjižni jezik, ki pa zadeva le slovarsko obliko frazema.

Da bi ohranili narečno podobo in omogočili povezavo na avdio izsek, smo besedila zapisali fonetično, za kar smo uporabili vnašalni sistem ZRCola[4] (prim. tudi priporočila za računalniški simbolni fonetični zapis v Zemljak idr. 2002). Razmišljamo, da bi v okviru tega nivoja zapisa pripravili tudi zapis s fonetično transkripcijo s simboli IPA. Poknjiženi/standardizirani zapis razumemo kot prekodiranje na fonetični in morfološki ravni, medtem ko ostaja leksika narečna, dodani pa so prevodi; pri izpisu priporočamo nadpisane prevode.[5] Za dodane prevode smo se odločili 1. zaradi poenotenega iskanja primerov s pomočjo leksemov, 2. narečno leksiko s prevodom pa smo ohranili tudi zaradi izdelave dvojezičnega glosarija. Poknjiženi/standardizirani zapis je torej nujen zaradi lažjega iskanja po korpusu. Slovarska oblika frazema, ki nam v e-slovarju služi kot iztočnica, je prevedena (prim. Šumenjak 2013: 42), in sicer tudi na skladenjski ravni, kar omogoča iskanje po drugih jezikovnih virih, pomembna pa je tudi zaradi omogočanja uporabe že dostopnih jezikovnih tehnologij (lematizacija in MSD-označevanje).[6] Lematizacija in delno tudi oznake MSD omogočajo iskanje po korpusih, kjer lahko preverimo prisotnost enakih ali podobnih frazemov; v aplikaciji je implementirano iskanje po referenčnem korpusu Gigafida (glej razdelek Aplikacija). Drugi razlog pa je tudi, da želimo v nadaljevanju urediti tudi večjezični frazeološki glosarij. Prav tako nam prevedene slovarske oblike frazemov prek analize na pomenski ravni omogočajo lažjo določitev konceptualne metafore (gl. sestavo gesla v Marc Bratina 2009b), kar ugotavljamo tudi s pomočjo pomenske mreže slovenskih besed (Zupan 2013). Prevod torej potrebujemo tako zaradi semantične analize kot tudi za iskanje po referenčnih korpusih (Gigafida, GOS) in SSKJ (kontrastivna analiza). Za prevod zgolj slovarske oblike frazema smo se odločili iz praktičnega razloga: prevajanje vsega gradiva je namreč časovno zelo zamudno opravilo, obenem pa ne daje večjih raziskovalnih rezultatov.

2.2 Značilnosti besedil, zajetih v korpus

Glede na to, da smo govorce spodbudili, naj ponazorijo rabo nekega frazema v širšem kontekstu in pojasnijo njegov pomen, lahko v okviru stopnje spontanosti govorimo o deloma spontanem govoru (prim. Zemljarič Miklavčič 2009: 116). Druge značilnosti govorjenih besedil, zajetih v narečnem korpusu za frazeološke/frazeografske raziskave, so: prevladujejo monološka besedila neuradnega zasebnega značaja, ki

3 Prim. načela standardizacije v Verdonik, Zwitter Vitez 2011: 62–66.

4 Besedilo je bilo pripravljeno z vnašalnim sistemom ZRCola (http://ZRCola.zrc-sazu.si), ki ga je na ZRC SAZU v Ljubljani (http://www.zrc-sazu.si) razvil dr. Peter Weiss.

5 Prim. tudi predlog za vkodirani dodatni zapis besed za razširitev možnosti iskanja po korpusu pri fonemski transkripciji v Zemljarič Miklavčič 2009: 138–140.

6 Prim. t. i. standardni zapis pri Zwitter Vitez, Verdonik 2011: 58.

so glede na govorni položaj neformalna,[7] prenosnik pa je osebni stik (prim. Zemljarič Miklavčič 2009: 116; Zwitter Vitez idr. 2007).

2.3 Aplikacija

Ker je naš namen nadgradnja e-slovarja v vseslovenski frazeološki slovar, želimo pri sodelovalni skupnosti (*collaborative community – community*) spodbuditi zbiranje dialektološkega gradiva, ki pa mora biti nadzorovano in usmerjano. Na tak način lahko v veliko krajšem času zberemo večje količine gradiva, potrebujemo le še nadzor nad kakovostjo, ki pa je pri množičnem zunanjem izvajanju (*crowdsourcing*)[8] vedno pereč problem. Nadzor nad kakovostjo vnosov in možnost dopolnjevanja predvsem anonimnih vnosov sta natančneje predstavljena v razdelku Primeri uporabe.

Osnovne lastnosti, ki so bile vodilo že pri samem snovanju orodja, so:

- **dostopnost**: orodje omogoča vnos novih gradiv čim širšemu krogu uporabnikov;
- **enostavna uporaba**: uporaba orodja ne sme odganjati potencialnih uporabnikov – možnost anonimne uporabe, enostavne zahteve po strojni opremi, enostavne osnovne operacije, predvsem vnos novega gradiva;
- **podpora fonetičnemu zapisu**: podpora pisavi 00 ZRCola vnašalnega sistema ZRCola (Weiss 2004), podpora standardu IPA (Ladefoged 1990);
- **standardizirani vmesniki**: orodje omogoča standardizirani prenos podatkov, podporo sodobnim vmesnikom (JSON, XML, TEI-P5; TEI Consortium 2015).

Izdelali smo pilotno spletno aplikacijo Narečni frazem,[9] v kateri poskušamo doseči vse opisane cilje. Aplikacija implementira spletno različico modela MVC (Model, View, Controller; Moore idr. 2007). Temelji na preizkušenih in široko uporabljenih ogrodjih, ki v zadnjih letih pogosto služijo kot skupno ogrodje za snovanje spletnih aplikacij. CodeIgniter je ogrodje za izdelavo spletnih aplikacij, uporabljeno za strežniški del, AngularJS je ogrodje za izdelavo dinamičnih spletnih strani, uporabljeno za del aplikacije, ki se izvaja v brskalniku; ogrodje Bootstrap pa je bilo uporabljeno za osnovno podobo aplikacije.

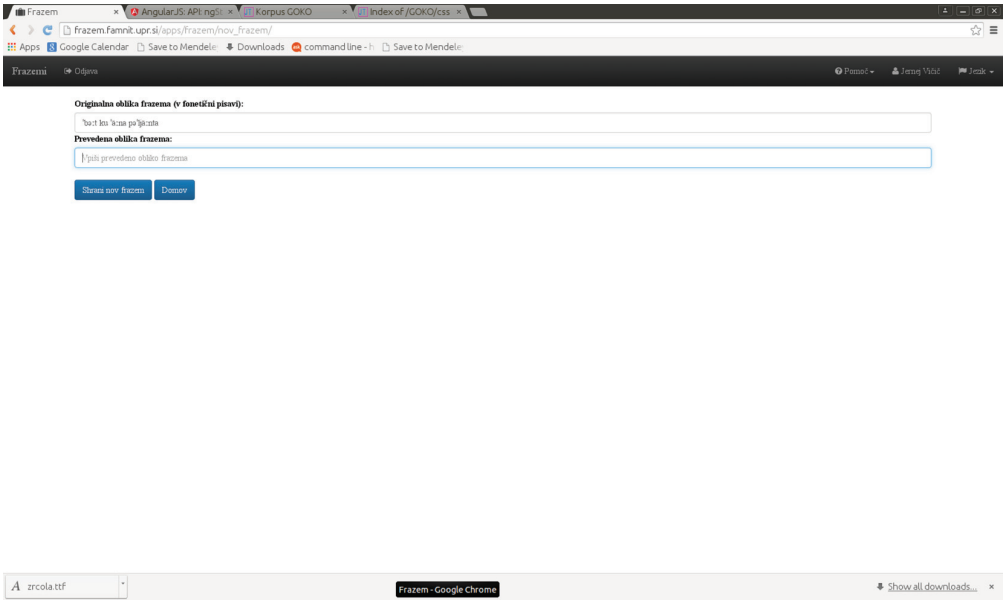
Tak način izdelave aplikacij omogoča kar največjo fleksibilnost in posledično olajša spremembe tudi v času uporabe, saj ne zahteva lokalne namestitve, preverjena ogrodja pa omogočajo hiter razvoj varnih aplikacij. Za uporabo je potreben le sodoben spletni brskalnik.

Pripravili smo tudi programski vmesnik API, tako da so vsi podatki dosegljivi s preprostimi http-klici v JSON-obliki, seveda ob ustrezni identifikaciji uporabnika.

7 Zavedamo se, da se pojma uradno/neuradno in javno/zasebno ne prekrivata (prim. Zemljarič Miklavčič 2009: 116), zato bi bilo treba na tem mestu natančneje opredeliti tudi situacijski kontekst, v katerem je potekalo zbiranje gradiva.

8 Islovar: Slovar informatike.

9 Narečni frazemi: http://frazem.famnit.upr.si



Slika 1: Spletna aplikacija za zbiranje narečnih frazemov; uporabniški vmesnik je zasnovan tako, da uporabnik čim enostavneje in hitro opravi svoj vnos

Aplikacija uporablja za poenostavitev obdelave frazemov prosto dostopne jezikovnotehnološke servise. Lematizacija in označevanje posameznih prevedenih oblik frazemov z oblikoskladenjskimi oznakami je mogoča s pomočjo spletnega servisa projekta JOS (Erjavec idr. 2010). Referenčni korpus slovenskega jezika Gigafida (Arhar Holdt idr. 2012) je uporabljen za iskanje pojavitev frazema v prevedeni osnovni obliki in tudi v posplošenih oblikah v obliki regularnih izrazov in lematiziranih oblik besed. Najdene frazeme lahko takoj opredelimo kot narečno nespecifične. Pomenska mreža slovenskih besed (Zupan 2013) je uporabljena za iskanje in opisovanje pomenske okolice posameznih besed (v lematizirani obliki) prevedenega frazema. To nam omogoča konceptualno analizo frazemov (pomensko podobne besede). Določanje geografske lokacije z uporabo programskega vmesnika Google Maps APIs (Hu, Dai 2013) omogoča geolociranje frazemov, uporabnik ob vnosu novega frazema definira tudi območje, kjer je naletel na njegovo uporabo. To predstavlja osnovo za analize, ki temeljijo na geoprostorski umestitvi frazemov.

Aplikacija predvideva tri vrste uporabnikov: anonimni uporabniki, preverjeni uporabniki, uredniški odbor.

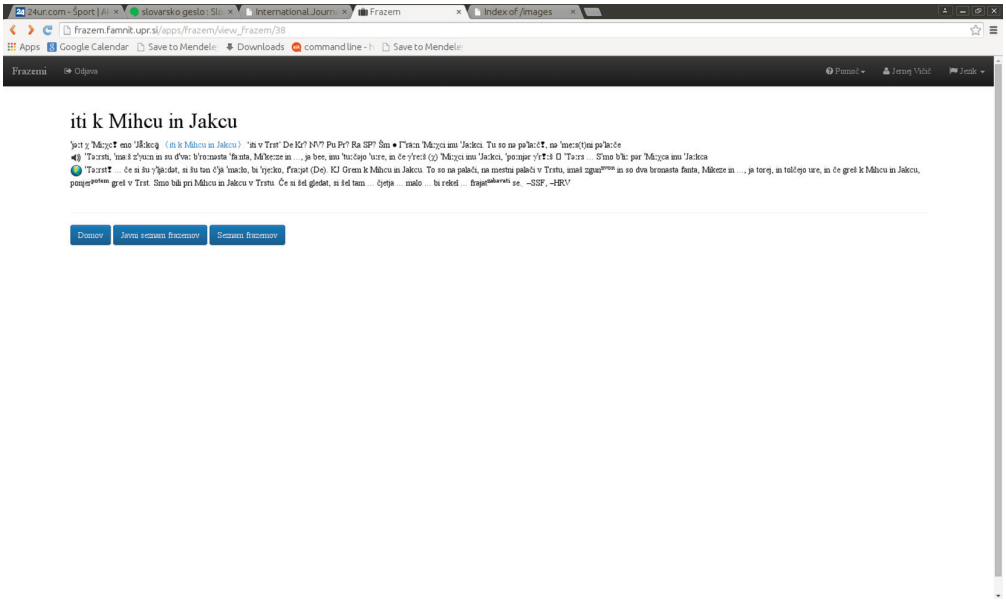
2.4 Primeri uporabe (use cases)

Pričakujemo štiri osnovne načine uporabe aplikacije. **Anonimni vnos frazemov**: navdušenci sami vpisujejo podatke o njim zanimivih narečnih frazemih, takšne vnose prevzamejo preverjeni uporabniki aplikacije, ki preverijo pristnost ter dopolnijo manjkajoče podatke. Pri tem si pomagajo s povezanimi jezikovnotehnološkimi servisi. **Vnos frazemov preverjenih uporabnikov**: preverjeni uporabniki, dialektologi, vna-

šajo frazeme, ki so jih opazili pri svojih raziskavah, in pregledujejo gradiva anonimnih uporabnikov. **Nadzor in urejanje gradiv**: uredniški odbor, sestavljen iz izbranih preverjenih uporabnikov, odobri frazeme. **Pregled slovarja**: iskanje in pregled celotnega slovarja.

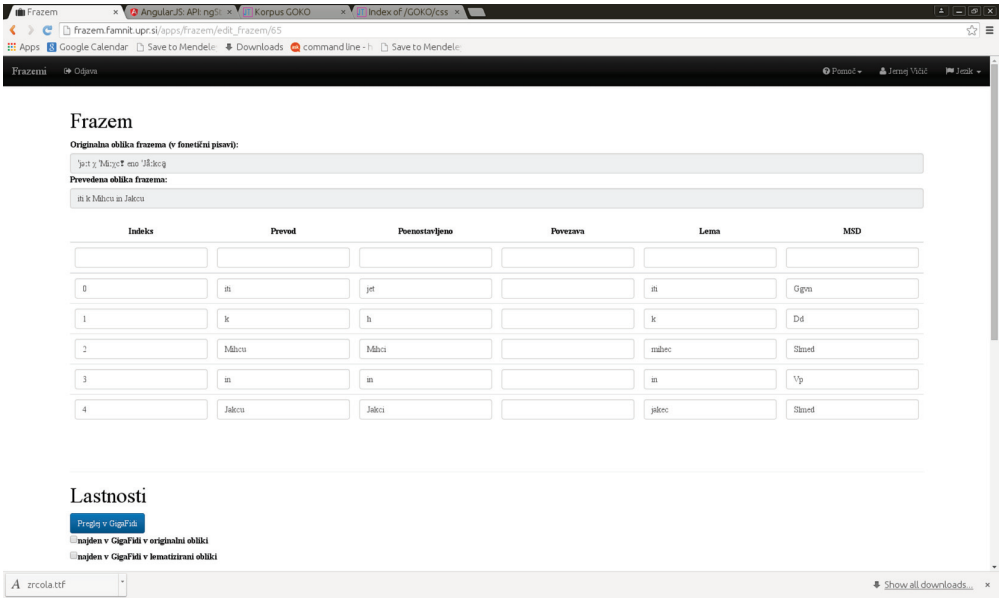
3 Rezultati in primeri

Oglejmo si primer vnosa novega naročnega frazema, ki je predstavljen na Sliki 2.



Slika 2: Frazem 1. ˈjəːt χ ˈMiːχceˌ eno ˈJåːkcaˌ <iti k Mihcu in Jakcu>; samodejno so dodane oznake MSD-projekta JOS.

Primeri vnesenih frazemov so predstavljeni v nadaljevanju.



Slika 3: Primeri vnesenih frazemov

4 Zaključek

V prispevku smo predstavili prosto dostopno spletno aplikacijo, ki bo skupnosti pomagala pri zbiranju narečnih frazemov. Opisali smo razloge za nastanek aplikacije in očrtali načine uporabe končnega izdelka – slovarja. Predstavili smo jezikovno-tehnološka orodja, s pomočjo katerih bo omogočeno hitrejše zbiranje gradiva – narečnih frazemov. Opisani so tudi najpomembnejši načini uporabe aplikacije.

S pomočjo sodobnih jezikovnih tehnologij jezikoslovci uporabnikom ponujamo interaktiven vpogled v jezikovno raznolikost slovenskega jezika in njegovih narečij, natančneje, v narečno frazeologijo. Prek slovenskih govorjenih (narečnih) frazemov lahko uporabniki na podlagi avtentičnih govorjenih besedil neposredno spoznavajo slovensko (narečno) frazeologijo, prek nje pa slovensko materialno, socialno in duhovno kulturo. Tako kot to počnejo etnologi in kulturni antropologi, lahko tudi jezikoslovci prispevamo k varovanju in ohranjanju nesnovne kulturne dediščine.

Literatura

ARHAR HOLDT, Špela, KOSEM, Iztok, LOGAR BERGINC, Nataša, 2012: Izdelava korpusa Gigafida in njegovega spletnega vmesnika. *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 16–21.

ERJAVEC, Tomaž, FIŠER, Darja, KREK, Simon, LEDINEK, Nina, 2010: The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10). 1806–1809.

Gigafida: Elektronska zbirka slovenskih besedil: http://www.gigafida.net

GORJANC, Vojko, 2005: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

GORJANC, Vojko, FIŠER, Darja, 2010: *Korpusna analiza*. Ljubljana: Znanstvena založba Filozofske fakultete.

GROS ŽGANEC, Jerneja, MIHELIČ, France, DOBRIŠEK, Simon, 2003: Govorne tehnologije: pridobivanje in pregled govornih zbirk za slovenski jezik. *Jezik in slovstvo* 48/3–4. 47–59.

HU, Shunfu, DAI, Ting, 2013: Online Map Application Development Using Google Maps API, SQL Database, and ASP.NET. *International Journal of Information and Communication Technology.* 102–111.

Islovar: Slovar informatike: http://www.islovar.org

LADEFOGED, Peter, 1990: The Revised International Phonetic Alphabet. *Language*. Linguistic Society of America. 550–552.

MARC BRATINA, Karin, 2009a: Etape zbiranja narečnega frazemskega gradiva. *Annales, anali za istrske in mediteranske študije. Series Historia et Sociologia* 19/1. Koper: Zgodovinsko društvo za južno Primorsko, *Znanstveno-raziskovalno središče*. 399–410.

MARC BRATINA, Karin, 2009b: Zasnova narečnega frazeološkega slovarja. Marko Stabej (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja 28*. Ljubljana: Znanstvena založba Filozofske fakultete. 233–238. http://www.centerslo.net/files/file/simpozij/simp28/Marc%20Bratina.pdf

MOORE, Data, BUDD, Raymond, BENSON, Edward, 2007: *Professional Rich Internet Applications: AJAX and Beyond*. Wiley.

ŠUMENJAK, Klara, 2013: *Opis govora Koprive na Krasu na osnovi dialektološkega korpusa. Doktorska disertacija*. Koper: Fakulteta za humanistične študije.

TEI, Consortium, 2015: *Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

VERDONIK, Darinka, ZWITTER VITEZ, Ana, 2011: *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

WEISS, Peter, 2004: ZRCola: vnašalni sistem za jezikoslovno rabo v programu Word. *Jezikoslovni zapiski*. 145–152.

ZEMLJAK, Melita, KAČIČ, Zdravko, DOBRIŠEK, Simon, GROS, Jerneja, WEISS, Peter, 2002: Računalniški simbolni fonetični zapis slovenskega govora. *Slavistična revija* 50/2. 159–169.

ZEMLJARIČ MIKLAVČIČ, Jana, 2008: *Govorni korpusi*. Ljubljana: ZIFF.

ZEMLJARIČ MIKLAVČIČ, Jana, 2009: Raziskovanje govorjenega jezika. Nike K. Pokorn (ur.): *Sodobne metode v prevodoslovnem raziskovanju*. Ljubljana: Znanstvena založba Filozofske fakultete. 110–130.

ZUPAN, Jure, 2013: *Pomenska mreža slovenskih glagolov.* Ljubljana: Znanstvena založba Filozofske fakultete; Založba ZRC, ZRC SAZU.

ZWITTER VITEZ, Ana, ZEMLJARIČ MIKLAVČIČ, Jana, STABEJ, Marko, KREK, Simon, 2009: Načela transkribiranja in označevanja posnetkov v referenčnem govornem korpusu slovenščine. Marko Stabej (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja 28*. Ljubljana: Znanstvena založba Filozofske fakultete. 437–442.
http://www.centerslo.net/files/file/simpozij/simp28/Zwitter_Zemlj_Stabej_Krek.pdf