Simona Klemenčič
Inštitut za slovenski jezik Frana Ramovša ZRC SAZU

Polne etimološke osvetlitve v novem slovarju

Etimološke osvetlitve bodo predstavljale pomemben del novega slovarja slovenskega jezika. Na več plasteh besedja je prikazan pomen razvoja jezikovnih tehnologij za raziskave izvora besed, pa tudi njihove trenutne omejitve.

Full etymological explanations in the new Dictionary

Etymological explanations will constitute an important part of the new dictionary of Slovene language. Importance of development of language technologies for research of word origins is shown on different layers of a language's lexicon, as well as their present limitations.

Ključne besede: etimologija, slovaropisje, slovar, primerjalna metoda, verjetnostna metoda

Keywords: etymology, lexicography, dictionary, comparative method, probabilistic method

Etimološke osvetlitve v slovarju se delijo na sklicevalne (sklic na besedo, ki je obravnavana na drugem mestu v slovarju) in polne. Polne etimološke osvetlitve zajemajo: a) plast besedja, ki je nastalo v relativno nedavnem razvoju jezika – tvorjenke iz besednih zvez in predložnih zvez, poenobesedene besedne zveze, besede, nastale iz lastnih imen, in razlage kratic, b) etimološke osvetlitve izposojenk in tujk ter c) polne etimološke osvetlitve besed, podedovanih iz starejših jezikovnih plasti.

Načela pisanja etimoloških osvetlitev v novem slovarju slovenskega jezika so bila natančneje opredeljena v Snoj 2009 in v SNB (*Slovarju novejšega besedja*) 2012: 45–49 ter uporabljena pri pisanju 4670 etimoloških osvetlitev v SNB.

V SNB je zaradi specifike slovarja, ki zajema novejše besedje, večji poudarek na pisanju sklicevalnih etimoloških osvetlitev v primerjavi z novim slovarjem slovenskega jezika, ki bo zajel večji del prevzetega in predvsem podedovanega besedja. Pred začetkom pisanja etimoloških osvetlitev za novi slovar bi bilo zato koristno še enkrat soočiti poglede na zapisovanje iz starejših jezikovnih plasti podedovanega besedja. V tem se kot najbolj pereča

kažejo vprašanje pogledov na praslovanske jezikovne plasti v diahroni perspektivi, vprašanje zapisovanja laringalov (barvajoči ali nebarvajoči; h_1 do h_4 ali simbol za laringal, ki pokriva vse štiri?), pogledi na praindoevropski *šva* ter interdentalni pripornik ter vprašanje zapisa posebnih znakov, kot so palatalizirani, labializirani in pridihnjeni konzonanti, pri čemer predlagam, da ohranimo zapis, ki je v skladu s tradicijo sodobne slovenske in zahodnoevropske indoevropeistike.

Ko govorimo o etimoloških osvetlitvah, se poraja vprašanje, v kolikšni meri lahko jezikovne tehnologije pohitrijo delo etimologa. Pomoč jezikovnih tehnologij je seveda dobrodošla in tudi nujna. Primerjalno jezikoslovje je kljub natančnim postopkom induktivna metoda, pri kateri imajo sprejeti argumenti (v tem primeru etimološke osvetlitve) sicer praviloma zelo veliko logično moč, še vedno pa so občutljivi za nove podatke (prim. Klemenčič 2012: 19–34). Če primerjamo npr. etimologijo slovenske besede *kašelj* v Snoj 2003: 263 in njene hrvaške sorodnice *kašalj* v HER (*Hrvatski enciklopedijski rječnik*) 2003: 561, vidimo, da se indoevropski rekonstrukciji do neke mere razlikujeta. V prvem slovarju se beseda izvaja iz korena *$k^{\mu}\bar{a}s$- (*$k^{\mu}ah_2s$-), v drugem pa iz korena *keh_2s-. Etimolog se mora odločati med različnimi možnimi etimologijami v skladu s svojim strokovnim znanjem in materialom, ki ga ima na razpolago. Zahteva po upoštevanju celotne evidence je področje, na katerem je računalniška podpora nujna. »Nabiranje češnjic« ni zaželeno pri nobenem sklepanju, jezikovnega materiala za osvetlitev posameznega problema pa je zelo veliko, medtem ko je časa za pregledovanje in ovrednotenje le-tega relativno malo. To je razlog, zaradi katerega je pri etimologovem delu nujno potrebna pomoč tehnologije. Iskanje po digitaliziranih bazah je neprimerljivo hitrejše kot po slovarjih v knjižni obliki. Idealno bi bilo, ko bi digitalizirali vse etimološke, enojezične in historične slovarje, ki so na voljo, predvsem slovanske, in čim večje število ustreznih člankov in monografij. Potrebovali bi aplikacijo, ki bi pregledovala vse te baze in v njih poiskala določen niz znakov v določenem jeziku (ne pa tudi v drugih), in sicer tako, da bi upoštevala le tiste besede, ki so v besedilu posebej obravnavane v kontekstu etimologije (ležeči tisk), ne pa tudi ostalih pojavitev tega niza v besedilu. Težava s tem ni toliko tehnična, saj bi bilo to dokaj preprosto izvedljivo, kot se je pokazalo že pri izdelavi pete knjige *Etimološkega slovarja slovenskega jezika*. Ideja o takšni aplikaciji se kljub temu ustavi že pri slovenščini, kjer niti digitalizirana različica *Etimološkega slovarja slovenskega jezika* ni objavljena zaradi težav z avtorskimi pravicami.

Poleg besed, za katere lahko iščemo ključ v katerem od obstoječih etimoloških slovarjev, ostaja veliko število slovenskih besed, ki še nimajo etimološke osvetlitve. Te bodo

predstavljale največji del etimologovega dela pri novem slovarju. Ali bi si bilo pri tem mogoče pomagati z ustrezno računalniško aplikacijo, ki bi na podlagi ustreznih algoritmov ter slovenskega in tujega besedja v bazi ponudila najboljšo možno etimologijo? Bi, a ne še prav kmalu in zato žal še ne v okviru tega projekta.

Prva plast besed, kjer vidimo možnost pomoči tehnologije, so **iz praslovanščine podedovane besede**. Računalniški programi, ki aplicirajo fonetične zakone na vnesene besede, že obstajajo. Primer takšne aplikacije je *The Sound Change Applier*. Deluje tako, da v eni datoteki vnesemo besede izvornega jezika in v drugi nabor pravil za delovanje jezikovnih zakonov v razvoju iz izvornega v ciljni jezika. Aplikacija sprocesira obe datoteki in v tretji prikaže, kako se vnesene besede v ciljnem jeziku izoblikujejo v skladu s podatki iz prve in druge datoteke. Takole:[1]

latin.lex	**port.sc**	**port.out**	
lector	V=aeiou	leitor	[lector]
doctor	C=ptcqbdgmnlrhs	doutor	[doctor]
focus	F=ie	fogo	[focus]
jocus	B=ou	jogo	[jocus]
districtus	S=ptc	distrito	[districtus]
civitatem	Z=bdg	cidade	[civitatem]
adoptare	s//_#	adotar	[adoptare]
opera	m//_#	obra	[opera]
secundus	e//Vr_#	segundo	[secundus]
	v//V_V		
	u/o/_#		
	gn/nh/_		
	S/Z/V_V		
	c/i/F_t		
	c/u/B_t		
	p//V_t		
	ii/i/_		
	e//C_rV		

Prikaza fonetičnega razvoja jezika se uspešno lotevajo tudi aplikacije *IpaZounds*, *Wordcorr* in *Phonix*. Predstavljajo zametek nujno potrebne velike baze za etimologijo in zgodovino

[1] Vir: *The Sound Change Applier*, http://www.zompist.com/sounds.htm, 24. februar 2014.

svetovnih jezikov. Za zdaj so v pomoč študentu primerjalnega jezikoslovja, ne pa tudi etimologu, ki jezikovne zakone že pozna. Poleg tega so za prikaz delovanja takšnih aplikacij izbrani neproblematični primeri. V razvoju jezika prihaja do nesistemskih sprememb, kot je denimo delovanje analogije. Take spremembe lahko pojasnimo, ko se zgodijo, ne moremo pa jih predvideti.

Druga plast besed, kjer vidimo možnost podpore jezikovnih tehnologij pri ugotavljanju izvora besede, so **tvorjenke**. Lahko bi napisali program, ki bi tvorjenko povezal z najverjetnejšo motivirajočo besedo. Pustimo ob strani vprašanje, ali bi bilo to smiselno, saj etimolog načeloma pozna slovensko in slovansko besedotvorje. Če pomislimo na težave s povezavami tipa *ujeda* k *jesti*, *okajen* h *kaditi* ali *ogorek* h *goreti*, vidimo, da bi se stvari zapletle. Niso nerešljive, zahtevalo pa bi dosti dela, da bi popisali nabor vseh pravil. Vendar pa se prave težave začnejo drugje.

Prva težava je ločevanje med tvorjenko, nastalo v slovenščini, in med praslovansko tvorjenko. Pri prvi iščemo motivirajočo besedo v okviru slovenščine, pri drugi pa ustreznice v drugih slovanskih jezikih in motivirajočo besedo v okviru praslovanščine. Iz tvorbe besede pogosto ni razvidno, za katero plast besedja gre. Bo računalnik besedo *ogorek* povezal z *goreti*? Če bo pritegnil material iz drugih slovanskih jezikov, nas bo morda prej napotil na poljski *ogórek* ʻkumaricaʼ. Tu se pokaže druga težava: aplikacija bi bila koristna samo, če bi upoštevala tudi pomene besed. Nesmiselno je namreč, da izdelamo program, ki bi npr. besedo *sence* povezal s *senca* ali *seno*, ne le s *sen*, in ki bi v *šalici* videl pomanjševalnico od *šala*, v *vilici* pa pomanjševalnico od *vila* in ne od *vile*. Pomen besede se po drugi strani lahko iz najrazličnejših razlogov hitro spremeni tudi pri sorodnih besedah, včasih celo v svoje nasprotje. Primer: slovensko *zal* ʻlep, postavenʼ je etimološko sorodno z *zel* ʻslab, hudobenʼ, *sladek* pa je etimološko sorodno s *slan*. Uporabiti bi bilo potrebno postopke, ki bi preprečili, da že iz semantičnih razlogov ne bi prihajalo do nesmiselnega povezovanja nesorodnih besed, obenem pa bi program zaznal semantično ustrezanje med besedami, katerih pomeni so le na videz oddaljeni. To je gotovo mogoče, čeprav kompleksno, in gotovo bi zahtevalo veliko časa in ljudi.

Nekaj primerov povezav, ki bi jih lahko ponudil računalnik na podlagi pravil za tvorbo besed v slovenščini, če bi upošteval tudi semantično povezljivost:

ravnatelj gl. *ravnati*

stranka gl. *stran*

krožek gl. *krog*

bivak gl. *bivati*

Videti je, kot da bi to olajšalo etimologovo delo. Vendar pa so besede *ravnatelj*, *stranka* in *krožek* prevzete iz drugih slovanskih jezikov in niso nastale kot slovenske tvorjenke, *bivak* pa je iz spodnjenemškega *bīwake* in nima nobene zveze s slovenskim glagolom *bivati*. Tu imamo opravka s tretjo plastjo besed – s **prevzetimi besedami**. Zgornje rešitve bi bile torej povsem napačne. Prav tako ne bi bilo v pomoč, če bi računalnik izvor besede *guglati* iskal v praslovanski besedi **gug(ъ)la* ʻstorž; pokrivaloʼ ali besedo *tabela* po tvorbi primerjal s *tamlada*. Če aplikacija ne bi znala ločevati med prevzetimi in podedovanimi besedami in če ne bi upoštevala semantike, bi morala ponuditi prav vse teoretično mogoče možnosti izvora, med temi tudi neizpričane, a mogoče (prim. *obvod* iz neizpričanega **obvesti* iz *vesti*). To pa pomeni dve stvari: pisanje teh pravil in preverjanje delovanja aplikacije bi predstavljalo projekt zase, ki bi zahteval kar nekaj let dela, in drugič, na koncu bi imeli za etimološko osvetlitev vsake besede na voljo množico strojno generiranih predlogov možnih izhodišč, od tega večino takih, ki bi povzročali le zgago in zastranitev, ker bi bili nesmiselni že iz semantičnih razlogov. Odločitev za najboljšo možno etimološko osvetlitev bi bila odvisna od etimologovega znanja in materiala, ki ga ima na razpolago: z »mukotrpnim ročnim delom« (gl. spodaj) bi moral poiskati, katero izhodišče je pravo. To pa etimolog počne že brez podpore jezikovnih tehnologij. Bližnjice tu za zdaj še ni. Računalniški postopki so znani že dovolj dolgo, da bi primerjalni jezikoslovci takšen program že pred nekaj desetletji izdelali sami, če bi bilo to smiselno.

Kratkovidno bi bilo reči, da bo tako tudi ostalo. Pred desetletji smo bili enako skeptični glede strojnega prevajanja. Vendar pa bi bil projekt, ki bi se ustrezno lotil računalniško podprtega pisanja etimoloških osvetlitev, po vloženem delu in številu podatkov primerljiv z gigantskim projektom, kakršen je *Google Translate* z vsemi vključenimi jeziki. Slovenska beseda *knjiga*, ki je v slovenščino verjetno prišla iz stare kitajščine, je primer, ki prikaže, kako daleč besede prehajajo iz jezika v jezik – pa ne gre za eno od sodobnih kulturnospecifičnih besed, ki bi jih našli tudi v obstoječih zahodnoevropskih slovarjih. Za potrebe našega jezika bi takšen projekt zahteval sodelovanje diahronih jezikoslovcev za vse slovanske jezike in za vse relevantne

indoevropske in sosednje jezike. Za vse te jezike bi bilo treba popisati vse znane zakonitosti fonetičnega in morfološkega razvoja skozi čas,[2] vse zabeleženo besedje v sinhronih prerezih in vsa doslej znana rekonstruirana stanja. Šele v okviru tako velike baze in dobro premišljenih algoritmov bi slovenščina lahko črpala relevantne podatke in predloge možnih izvorov posameznih besed, na katere slovenski etimolog morda še ni pomislil.

Kolikor mi je znano, se o takšnem projektu nismo še niti začeli pogovarjati, vsekakor pa bi bil nujno potreben. Računalnik bi tako lahko zadostil zahtevi po zadostni evidenci, o logični moči induktivnega argumenta oziroma sklepanju na najboljšo razlago pa za zdaj še ne more presojati.

Vprašanje laringala v vzglasju rekonstruiranih praindoevropskih besed ilustrira težave s presojo logične moči argumenta. Hipoteza, da se v indoevropskem prajeziku nobena beseda ni začenjala na vokal, je prepričljiva in danes široko sprejeta. V vzglasju rekonstruiranih besed, ki ne izkazujejo konzonanta pred vokalom, postuliramo laringal, torej glas, za katerega v nobenem jeziku ni videti neposrednega refleksa na tem mestu. Hipoteza je v takih primerih podprta samo s teorijo. Primerjalni jezikoslovec bi v skladu z veljavno teorijo v aplikacijo vnesel **Hes*- za indoevropski koren ʻbitiʼ, medtem ko bi računalnik v najboljšem primeru ponudil rekonstrukcijo **es*-. Od tega, katero obliko vzame za izhodišče, pa je odvisno, katere nize bo iskal v drugih jezikih, da jih poveže z rekonstruiranim korenom. Kaj pa drugi faktorji, ki jih upoštevamo pri določanju sorodnosti dveh jezikov, kot so ujemanja v tipologiji, lastna imena ali morda številski sistem? Kako bi strojno rešili denimo problem substrata v pragermanščini ali v grščini, ko pa o tem substratu ne vemo kaj dosti, poleg tega pa za grščino velja, da izkazuje več substratnih jezikov, ne le enega? Zaradi teh težav je um etimologa, ki k argumentaciji pritegne vedno nova dejstva, še zmeraj bistveno zanesljivejši od računalnikovega. Res pa je, da človek s svojim umom lahko seže le do določene časovne globine, saj postane količina jezikovnih dejstev, ki jih je treba obdelati, za bolj oddaljena časovna obdobja sčasoma neobvladljiva tudi za najboljšega primerjalnega jezikoslovca. Za našo jezikovno družino postanejo dognanja nezanesljiva nekje v petem tisočletju pred našim štetjem, ko moramo k rekonstruiranemu materialu za indoevropski prajezik pritegniti druge rekonstruirane prajezike, predvsem uralskega. Jezikoslovčeve intelektualne sposobnosti tu odpovejo, ker zaradi obsega materiala ni strokovnjakov za primerjalno jezikoslovje več kot ene jezikovne družine. Za osvetlitev jezikovnih stanj, starejših od indoevropskega prajezika,

[2] Poskus popisa jezikovnih zakonov, ki nastaja na naših tleh, je *Aplikacija za vnos in prikaz glasoslovnih razvojev* avtorja Roberta Jakomina (http://lgm.fri.uni-lj.si/hf/).

bo dobrodošla pomoč računalnika. Čeravno ji iz zgoraj opisanih razlogov ne bo mogoče povsem zaupati, bo to najboljše, kar bomo kdaj imeli na razpolago.

V lanskem letu je bilo primerjalno jezikoslovje za kratek čas deležno pozornosti svetovne javnosti zaradi članka z naslovom *Avtomatizirana rekonstrukcija starih jezikov s pomočjo verjetnostnih modelov glasovnih sprememb* (Bouchard-Côté idr. 2013). Avtorji so v članku opisali pristop k rekonstrukciji prajezika z verjetnostno metodo, ki naj bi bila uspešnejša pri rekonstrukciji izumrlih jezikov kot »mukotrpna ročna procedura, ki ji pravijo primerjalna metoda« (Bouchard-Côté idr. 2013). Članek je zbudil veliko pozornosti in je v časopisju sprožil val novic z obetajočimi naslovi kot *Znanstveniki so ustvarili »časovni stroj« za rekonstrukcijo starih jezikov* (Anwar 2013).

Avtorji so primerjali besedje 637 avstronezijskih jezikov, da bi ugotovili skupni izvor obravnavanih besed in s tem jezikov. Iz članka je razvidno, da so primerjali majhen nabor besed iz baze *Austronesian Basic Vocabulary Database*. Gre za osnovne glagole kot ʻhoditiʼ in ʻleteti', za poimenovanja delov telesa, barv, številkov in bližnjih sorodnikov. To besedje je praviloma podedovano; le izjemoma se prevzema iz drugih jezikov. Kljub temu da so se avtorji z izborom te plasti besed izognili težavam, ki jih prinaša prevzeto besedje, je ujemanje z izsledki jezikoslovcev glede na navedbe v članku le 85 odstotkov. To je zelo nizka številka, ki pove, kako težavne so računalniške rekonstrukcije celo na majhnem vzorcu izbranih besed, ki bi morale biti dokaj neproblematične.

Probabilistična metoda je tako kot Swadeshev seznam (prim. Klemenčič 2013: 15–16) pomembna za ugotavljanje eventualne sorodnosti pri primerjanju velikega števila jezikov, ki so slabo zabeleženi in raziskani. Povsem zavajajoče pa je trditi, da »medtem ko je ročna rekonstrukcija mukotrpen proces, ki traja tudi po več let, lahko ta sistem izvede rekonstrukcijo na veliki količini materiala v nekaj dneh ali celo urah« (Anwar 2013). Avstronezijska jezikovna družina ni primerljiva z indoevropsko po raziskanosti ali izpričanosti vej. Kar je naredil ta program s komaj 85-odstotno natančnostjo na relativno neproblematičnem vzorcu, je za indoevropske jezike narejeno že več kot sto let, in to mnogo bolj natančno in zanesljivo. Verjetnostna metoda nam ne bo dala etimologije besede *bivak*, zato je nesmiselno govoriti o primerjalni in verjetnostni metodi v istem kontekstu. Gre za različne cilje, ki jih znanstveniki poskušajo doseči z različnimi postopki.

Do podpore jezikovnih tehnologij pri izdelavi etimoloških osvetlitev je še dolga pot in v tem trenutku možnosti, ki se kažejo, še niso relevantne za slovenščino in za novi slovar

slovenskega jezika. V prihodnosti pa bo nujno razmišljati o mednarodnih interdisciplinarnih projektih za primerjalno jezikoslovje, ki bodo vključevali tudi slovenščino.

Literatura

ANWAR, Yasmin, 2013: *Scientists create automated 'time machine' to reconstruct ancient languages*. http://newscenter.berkeley.edu/2013/02/11/ancientlanguages.

Austronesian Basic Vocabulary Database. http://language.psy.auckland.ac.nz/austronesian/.

BOUCHARD-CÔTÉ, Alexandre, idr., 2013: Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Science* 110/11. www.pnas.org/cgi/doi/10.1073/pnas.1204678110.

HER, 2003: ANIĆ, Vladimir, BROZOVIĆ RONČEVIĆ Dunja, idr.: *Hrvatski enciklopedijski rječnik*. Zagreb: Novi Liber.

IpaZounds. http://zounds.artefact.org.nz/.

KLEMENČIČ, Simona, 2011: *Spisovnik primerjalnega jezikoslovca*. Ljubljana: Znanstvena založba Filozofske fakultete.

KLEMENČIČ, Simona, 2013: *Pregled indoevropskih jezikov*. Drugi natis. Ljubljana: Znanstvena založba Filozofske fakultete.

Phonix. https://code.google.com/p/phonix/.

SNB 2012: BIZJAK KONČAR, Aleksandra, SNOJ, Marko (ur.): *Slovar novejšega besedja slovenskega jezika*, Ljubljana: Založba ZRC, ZRC SAZU.

SNOJ, Marko, 2003: *Slovenski etimološki slovar*. Druga, pregledana in dopolnjena izdaja. Ljubljana: Modrijan.

SNOJ, Marko, 2009: Etimološke osvetlitve v novem slovarju slovenskega jezika. A. Perdih (ur.): *Strokovni posvet o novem slovarju slovenskega jezika*, 23. in 24. oktober 2008. Ljubljana: Založba ZRC, ZRC SAZU. 83–93.

The Sound Change Applier. http://www.zompist.com/sca2.html.

Wordcorr. http://www.wordcorr.org/index.htm.