KAKO "OSLOBODITI" HRVATSKI JEZIK

Premda nikada ranije nismo mislili da ćemo imati neke veze s lingvistikom i premda smo vrlo slabo upućeni u problematku (hrvatskog) jezika i u sve ono što je s njom povezano, stjecajem okolnosti morali smo se pozabaviti nekima od pitanja koja ulaze u ovu domenu. Ukratko, radilo se (i radi) o projektu tj. programu Udruge za slobodne informacije "SINBAD" pod nazivom mki (Make Index), koji je nastao iz naše namjere da izradimo slobodni softver pomoću kojega bi se automatski ili barem "poluautomatski" mogla kreirati kazala za knjige (pojmova, imena, geografskih naziva i sl.).[1] Rad na programu započeo je prije nešto više od godinu dana, no on ne može biti završen, ne barem na odgovarajući način, zbog zapreka o kojima će biti govora u ovom tekstu.

Zasigurno najveći problem u ovom programiranju proizlazi iz okolnosti da riječi koje se uključuju u kazalo u pravilu mogu imati različite gramatičke oblike. Program (koji kazalo slaže na osnovi pripremljene liste pojmova), mora dakle "prepoznati" sve moguće oblike za sve riječi koje ulaze u kazalo, pa prema tome mora posjedovati neki oblik "inteligencije" potrebne za ovo prepoznavanje.

Da bismo pronašli pravi način kreiranja te "inteligencije", konzultirali smo se s ljudima upućenim u jezična pitanja. Zaključak je bio da je problem teško ili nemoguće riješiti samo programski tj. upotrebom određenih (jezičnih) algoritama, nego da bi bilo najbolje koristiti se rječnikom hrvatskog jezika u kojemu su popisane sve hrvatske riječi i svi njihovi oblici (ili onaj njihov broj koji je dovoljan da bi jedan ovakav program davao "dovoljno dobre" rezultate). Dakako, program bi mogao raditi i bez rječnika - on zapravo u ovoj fazi tako i radi - no u ovom slučaju korisnik treba, s jedne strane, biti upućen u neke "tehničke" pojedinosti (kao što je recimo rad s "regularnim izrazima" koje program koristi u pretraživanju teksta), dok, s druge, mora voditi računa o svim gramatičkim oblicima riječi koje ulaze u kazalo, što znatno komplicira cijeli postupak njegove "strojne" izrade (može se doduše raditi i s korijenima riječi, no tada je pretraga u dosta slučajeva prilično neprecizna). Također, u ovom je slučaju ograničena i funkcionalnost programa - zgodno bi recimo bilo da program sam može pronaći sva osobna imena u tekstu i onda automatski kreirati odgovarajuće kazalo imena (dakle bez da ih iz teksta najprije mora "izvlačiti" sam korisnik), što je sada nemoguće.

Uglavnom, nakon zaključka da nam za realizaciju projekta treba "digitalizirani" rječnik hrvatskog jezika, i to onaj s "morfologijom", tj svim gramatičkim oblicima riječi, krenuli smo u potragu za njim. No pokazalo se da jednu takvu stvar, koja bi bila "slobodna", što znači da je dostupna i da se može koristiti bez ograničenja za bilo koju svrhu (pa tako i u raznim "slobodnim" informatičkim projektima), nije baš jednostavno pronaći ili da se ona u nekoj zadovoljavajućoj formi, danas, prema svemu sudeći, ne može pronaći uopće.

Htjeli bismo na ovom mjestu izniijeti neka zapažanja nastala tokom ove naše "potrage", jer nam se čine zanimljivima i za širu javnost. Priča će možda nekome biti dosadna, jer se izgleda svodi na notorne činjenice koje "svi znaju", no one, barem za autora ovoga teksta, koji je na (hrvatski) "jezični teren" banuo potpuno neobaviješten i nepripremljen, predstavljaju prilično iznenađenje.

Prvo što nam je s ovim u vezi našega problema palo na um, bio je Anićev rječnik uz čija novija izdanja uključuju i CD na koji je rječnik spremljen u elektronskom formatu. Međutim, pokazalo se da je normalan rad s njim nemoguć, s obzirom da je elektronska forma rječnika neprimjerenog "formata", a k tomu je i "kriptirana", pa se CD ne može kopirati. Kriptiranje je zasigurno učinjeno u svrhu zaštite "autorskih prava".

Međutim, sad se postavlja pitanje - ne predstavljaju li "autorska prava" u ovom konkretnom slučaju, u neku ruku, "prava" na riječi hrvatskog jezika? Doduše, može se reći kako nisu u pitanju riječi nego rječnik kao cjelina, skup riječi koje je netko prikupio i uredio. No, ne bi li bilo logično i normalno da se i rječnik, kao i same riječi od kojih je složen, koristi slobodno, bez ikakvih zapreka. Bi li itko išta izgubio da se rječnik može slobodno distribuirati i koristiti? Možda i bi, ali sigurno ne onaj tko je za nastanak toga rječnika zaslužan. Pa zar to onda nije nepravedno, da po osnovi zaštite autorskih prava iz jednog rječnika korist izvlači netko tko za njegov nastanak nije zaslužan? Na koncu, sam je izdavač Anićevog rječnika, u suradnji sa "Srcem", pokrenuo internetski "Hrvatski jezični portal", gdje su svi podaci iz Anićevog rječnika (a i iz nekih drugih) dostupni. Zašto se onda rječnik ne može slobodno distribuirati i na CD-u i to u normalnom formatu kojega je moguće koristiti pod bilo kojim operativnim sustavom? Zašto ga se ne može koristiti u slobodnim informatičkim projektima? Zašto je, čak štoviše, rad s rječnikom onemogućen nekome tko "nema" ni internet ni Windowse i/ili nekome tko ne voli takav način pretrage i kojekakva (grafička) sučelja, nego bi rječnik pretraživao grepom?

---

1 "Testna" verzija programa **mki** može se pronaći na web stranicama Udruge za slobodne informacije "SINBAD" (**http://sinbad.bplaced.net/projekti.html/**).

Inače, na ovakve i slične apsurde, do kojih dovodi zaštita "autorskih prava", takva kakva se primjenjuje danas, te uopće koncept "vlasništva" primijenjen u intelektualnoj sferi, nailazimo na bezbrojnim drugim primjerima i u raznim drugim područjima našeg duhovnog stvaralaštva. Zagovaratelji ovih koncepata imaju neke svoje argumente, no teško da ikoji od njih može "držati vodu". Kako se argumenti ove vrste mogu pobiti na jednostavan način demonstrirao je recimo Richard Stallmanovom u svom "GNU Manifestu".

No, vratimo se našim dogodovštinama pri potrazi za "slobodnim" hrvatskim rječnikom. Nakon spomenute male "istrage" oko Anićevog rječnika, obratili smo se onima za koje je bilo sasvim sigurno da imaju i znaju sve ono što nas zanima - stručnjacima sa Zavoda za lingvistiku zagrebačkog Filozofskog fakulteta. Prije toga smo pogledali njihov portal "Jezične tehnologije za hrvatski jezik" iz kojega se vidi da posjeduju zaista značajne "jezične resurse" (premda je sam portal pomalo zapušten). Međutim, kako smo saznali u razgovoru, ništa od tih resursa nije dostupno za "slobodnu " upotrebu, već je potrebno sa Fakultetom sklopiti ugovor o njihovom korištenju, s tim da se svaki pristup njihovim bazama plaća po određenoj tarifi.

Na pitanje - Zašto se plaća? - sigurno bi se mogli navesti brojni razlozi, neki zacijelo i opravdani. Možda je, s obzirom na prilike u kojima živimo, uistinu opravdano naplaćivati pristup određenim podacima tj. resursima koji se mogu koristiti, a vjerojatno se i koriste u "komercijalne" svrhe. No, zar ne bi bilo logično i normalno da se barem ono osnovno, **rječnik hrvatskog jezika u elektronskom obliku**, prepusti društvu na slobodno korištenje, kao temeljno "javno dobro" hrvatske kulture, hrvatskog naroda i svih hrvatskih građana, koje su stvorili danas živući naraštaji hrvatskih lingvista. Uostalom, upravo to društvo, tj. porezni obveznici financiraju (u najvećoj mjeri) njihov rad, pa bi nam se oni mogli "odužiti" barem jednom ovakvom gestom.

Jasno, za ovakvo stanje stvari nisu "krivi" ljudi koji rade u lingvističkoj znanosti, problem je u "sustavu" u kojemu se izgubila svaka odgovornost za dobrobit i čovjeka-pojedinca i društva u cjelini, pa tako (uz sve ostalo) omogućava i da prisvajanja intelektualnih i svih ostalih dobara, koja su se sve do pred neku godinu ili desetljeće smatrala "javnim vlasništvom", dosegne upravo neshvatljive razmjere. O opasnostima i štetama koje na taj način nastaju, mnogo se piše i govori, pa to ne treba ponavljati.[2]

Inače, u vezi s tendencijama prisvajanja na "području duha" vrlo je zanimljivo pogledati i tendencije upotrebe znaka kopirajta na literaturu koja je izlazila u našoj zemlji. Danas se ta literatura, posebice ona tiskana pod kopirajtom izdaje u pravilu i gotovo je nemoguće pronaći bilo što relevantnije, a da nije označeno ovim znakom ili da je izdano pod nekom od "slobodnih licenci" (kopileft). No ono čega mnogi možda nisu svjesni, jest da se ranije, u razdoblju prije "demokratskih promjena", u Hrvatskoj većina knjiga izdavala bez ove oznake, pa da tako pod kopirajtom nisu bili ni razni rječnici, kako oni hrvatskog jezika (poput onog Anićevog), tako i oni stranih. Ova tendencija širenja " prava vlasništva" nad našom duhovnom baštinom, nije nešto neočekivano, no postavlja se pitanja - tko je tu tendenciju nametnuo i iz kojeg razloga. Da se zaista u toj priči radi o "zaštiti" autora, sve bi se to moglo i razumjeti, pa se takvim tendencijama ne bismo puno protivili, no svima je jasno da se uopće ne radi o tome. Poznato je naime da u većem broju slučajeva autorska prava preuzimaju izdavačke kuće, a da je korist koju autor izvlači na osnovi svoga rada, barem što se tiče naše literature, mala ili nikakva (osim ako nije "umrežen" sa stanovitim "skupinama" koje od toga uistinu profitiraju). Sve u svemu treba reći da se ovdje radi o interesima određenih krugova moći koji i na taj način provode svoju vlast nad "duhovnom sferom" našega društva.

Treća, a ujedno i posljednja epizoda naše priče o potrazi za "slobodnim" hrvatskim rječnikom posve nam je vratila nadu u uspjeh našeg projekta, premda još uvijek ne možemo reći da je došlo do sretnog raspleta. Naime, u nastavku smo malo "prošvrljali" po internetu i naišli na jedan za nas vrlo zanimljiv web site, na kojemu je predstavljen projekt "Rječnika hrvatskih jezika" (**http://www.igaly.org/rjecnik-hrvatskih-jezika/**). Projekt je pokrenuo dr. Goran Igaly, matematičar s PMF-a, a kako je napisano u njegovom opisu, cilj projekta je obuhvatiti sve riječi koje se pojavljuju u hrvatskom jeziku, pa i one "dvojbene" za koje se mišljenja stručnjaka o tome jesu li uistinu hrvatske riječi ili nisu, razilaze. Rječnik (koji trenutačno ima 342.002 riječi tj. njihovih oblika) se može slobodno skinuti s interneta, a oko korištenja u komercijalne svrhe potrebno je dogovoriti se s autorom.

Rječnik smo "skinuli", isprobali, no učinio nam se ipak nedovoljno obimnim za potrebe našeg projekta. Kontaktirali smo i dr. Igalyja koji nam je rado dopustio da ga koristimo kao bazu na osnovi koje **mki** postaje

2 O toj temi pisali smo i u članku posvećenom Aaronu Swartzu, koji je objavljen nedavno na portalu zg-magazin (vidi niže). Okolnosti tragične smrti ovog briljantnog softverskog stručnjaka i borca za slobodu informacija pokazuju koliko je daleko taj "sustav" spreman ići u zaštiti interesa onog "jednog postotka" ljudske populacije (a možda ih je i manje) što prisvaja javna dobra koja pripadaju svima nama - jer su nastala zahvaljujući znanjima i radu cijele ljudske zajednice, i to ne samo današnjih naraštaja, nego i svih onih koji su nam prethodili. Dakako "prisvajanje" u ovom kontekstu nije baš prikladan termin - prava riječ je ona kojom je ovakvu praksu okarakterizirao primjerice Eben Moglen u svom "The dotCommunist Manifestu", a ta riječ je - "krađa".

dovoljno "inteligentan". Pritom nas je obavijestio da je u pripremi nova verzija rječnika koja će obuhvaćati preko 600.000 riječi, i mi se nadamo da će ona biti adekvatna za naše potrebe. Inače, možemo napomenuti kako će se zasigurno moći uspostaviti i svojevrsna interakcija našeg programa i "Rječnika hrvatskih jezika" - **mki** pri obradi teksta može detektirati eventualne riječi kojih u rječniku nema, tako da će se rječnik moći nadopunjavati i na ovaj način. Dakle, korist može biti i obostrana. Doduše, pomalo je nejasno kakav je "status" Igalyjevog Rječnika (vjerojatno zato što još nije završen), no nadamo se da će on biti stavljen pod neku od "slobodnih" licenci.

Premda, kao što rekosmo, sve još "nije gotovo", i ovaj primjer pokazuje kako se radom entuzijasta, koji čak i ne moraju biti neki veliki stručnjaci u poslu kojim se uz svoj redovni posao bave, mogu, posebice u ovom današnjem "digitalnom i informacijskom dobu", stvoriti razna korisna "intelektualna dobra", dostupna za slobodnu distribuciju i korištenje, koja će riješiti razne probleme s kojima se susrećemo u životu i radu. Radom entuzijasta stvara se tako slobodni softver i Wikipedija, stvara se slobodno znanje i slobodna kultura. Slobodni projekti bilježe se i na drugim područjima, recimo na području hardvera, robotiziranih sustava i sl. Putem svojih projekata i propagandnih aktivnosti ovom trendu želi pridonijeti i naša Udruga za slobodne informacije "SINBAD".

Vjerujemo da je kroz ovakve projekte i nastojanja moguće suprotstaviti se mnogobrojnim negativnim tendencijama, i onima koje smo spominjali u ovom tekstu, kao i raznim drugim, koje ugrožavaju ne samo naš normalan razvoj i napredak, nego i neke od temelja opstanka naše civilizacije. Možda je na taj način moguće potaći i službene institucije, pa čak i institucije vlasti, da napokon počnu funkcionirati na primjeren način, odnosno voditi brigu o interesima cijelog društva. A kad bi se o tim interesima na tom "nivou" zaista vodila briga, onda naša "situacija" ne bi bila takva kakva je danas, a peripetije poput ovih naših, oko rječnika hrvatskog jezika koji bi se mogao koristiti u slobodnim informatičkim projektima, teško da bi se mogle i zamisliti.

Zagreb, veljača 2013.