FILOLOGIJA 58, Zagreb 2012

UDK 811.163.42'35
Izvorni znanstveni članak
Primljen 10.VI.2010.
Prihvaćen za tisak 20.VI.2011.

Šandor Dembitz
Fakultet elektrotehnike i računarstva
Zavod za osnove elektrotehnike i električka mjerenja
Unska 3, HR-10000 Zagreb
Sandor.Dembitz@fer.hr

FUNKCIONALNA LEKSIKOGRAFIJA MREŽNOGA PRAVOPISNOG PROVJERNIKA

Mrežni pravopisni provjernici nude jedinstvenu priliku za popravljanje vlastite jezične funkcionalnosti interakcijom sa svojim korisnicima. Navedenu mogućnost posebno je važno iskoristiti u jezičnotehnološki perifernim jezicima, kakav je hrvatski, radi prevladavanja jaza koji postoji u tehnologiji obrade prirodnoga jezika između njih i jezičnotehnološki središnjih jezika. Načini na koje je ta mogućnost iskorištena u slučaju hrvatskoga jezika bit će opisana na primjeru mrežnoga pravopisnog provjernika poznatog pod imenom *Hascheck*. *Hascheck* je prvi hrvatski javni pravopisni provjernik u uporabi od početka 1993. godine. Njegov je početni rječnik obasezao 100.000 različnica hrvatskoga općejezičnog fonda. Učenjem iz tekstova koji su mu pristizali na obradu opseg je njegova rječnika do svibnja 2010. godine narastao na 830 tisuća općejezičnih različnica i 600.000 različnica posebnojezičnoga fonda (osobna, vlastita i druga imena, kratice i tako dalje). To je rezultat obrade korpusa od 260 milijuna pojavnica ostvaren zahvaljujući ekspertnom sustavu za učenje inkorporiranom u programski sustav pravopisnoga provjernika. Iako je sustav za učenje visokoautomatiziran, nove se različnice ne uvrštavaju u leksičku bazu bez prethodnog ljudskog nadzora. Nadzor je potreban radi očuvanja točnosti rječnika. Tijekom nadzora posebno se vodi računa da u rječnik ne uđu različnice koje se vrlo rijetko javljaju u uporabi, a identične su pogreškama u pisanju mnogo učestalijih riječi hrvatskoga jezika. Velika količina podataka prikupljana godinama omogućuje i pouzdano matematičko modeliranje mnogih aspekata *Haschecko*va života, što će također biti iscrpno opisano u ovome radu.

1. Uvod

Ovaj je rad posvećen opisu *Hrvatskoga akademskog spelling checkera*, poznatijeg pod imenom *Hascheck*, što je akronim nastao iz njegova izvornoga naziva. *Hascheck* je mrežni pravopisni provjernik u pozadini kojega djeluje sustav za učenje. Zadaća je sustava funkcionalna leksikografija u smislu da se iz tekstova primljenih na obradu izdvajaju one različnice koje obogaćuju *Haschecko*v rječnik, a čime se popravlja njegova jezična funkcionalnost. Cjelokupni napor treba promatrati u okvirima nastojanja da se smanje hrvatski jezičnotehnološki deficiti. Nepostojanje zrelih hrvatskih sustava za strojnu tvorbu i strojno prepoznavanje govora, što je činjenica koja hrvatski nedvojbeno svrstava u skupinu jezičnotehnološki perifernih jezika[1], nerješiv je problem bez razvoja računalnoga pravogovornog rječnika, jezgrene sastavnice svake govorne tehnologije. Računalni pravopisni rječnik, osobito ako je korpusno utemeljen, dobra je podloga da se u takav razvoj krene.

Članak je organiziran u sedam poglavlja. Drugo poglavlje donosi opis *Haschecka* kao sustava s naglaskom na postupku učenja. Treće je poglavlje posvećeno prometnim podatcima i njihovoj analizi, te uvođenju matematičkih modela, uključujući logistički model kumulativne dinamike rasta *Haschecko*va prometa, koji govori o predvidivome vijeku *Haschecko*va trajanja kao internetske usluge u aktualnome obliku, potvrđujući da u Hrvatskoj nema računalnojezikoslovnog projekta koji bi se u skoroj budućnosti mogao uspoređivati s *Haschecko*m po opsegu obrađenoga korpusa. Matematičko modeliranje u četvrtome poglavlju bavi se ponašanjem triju osnovnih parametara pravopisnoga provjernika u vremenu. To poglavlje također donosi usporedbu *Haschecka* s Microsoftovim pravopisnim provjernikom za hrvatski jezik, te je u njemu demonstrirana primjenjivost Heapsova zakona kao sredstva za predviđanje opsega leksikografskoga rada u računalnoj leksikografiji. Peto poglavlje pokazuje kako učeći susta-

[1] Od blizu 7000 živih jezika navedenih u *Ethnologueu* (*http://www.ethnologue.com/*), samo 2% raspolaže kakvim-takvim jezičnotehnološkim proizvodima. I među tom manjinom uočljive su velike razlike. Sustavnost i izdašnost financiranja razvoja jezičnih tehnologija u pojedinim zemljama osnova su tih razlika, tako da je u jezičnotehnološkom smislu sredinom prošlog desetljeća uvedena podjela na središnje (engl. *central*) i periferne (engl. *non-central*) jezike (Streiter i dr. 2006) koju smo i mi prihvatili. U literaturi koja se bavi jezičnotehnološki perifernim jezicima često se rabi izraz »podfinanciran« (engl. *under-resourced*), iako hrvatski izraz semantički ne pokriva potpuno engleski naziv. Podfinanciranima držimo i sve žive jezike bez ikakve jezičnotehnološke skrbi, dakle 98% svjetskih jezika, sa svim pogubnim posljedicama takva stanja stvari za njihovu opstojnost u internetskoj eri.

vi implicitno u sebi nose i informaciju o svojoj prošlosti. U šestom je poglavlju iznijet prijedlog kako bi sustavi slični *Haschecku* mogli doprinijeti obavljanju puno ozbiljnijih leksikografskih zadataka koji stoje pred hrvatskom leksikografijom, dok sedmo poglavlje donosi zaključni osvrt.

2. O *Haschecku* i njegovu učenju

Hascheck je jedna od najstarijih internetskih usluga u Hrvatskoj. Usluga je zaživjela početkom 1993. godine u lokalnoj mreži Elektrotehničkoga fakulteta Sveučilišta u Zagrebu (danas Fakultet elektrotehnike i računarstva). Prva knjiga pisana hrvatskim jezikom koja je prošla sustavnu strojnu provjeru teksta bila je knjiga *Josip Lončar — život i djelo* (Bego i Butorac 1993). Od 21. ožujka 1994. godine *Hascheck* je besplatna javna usluga strojne pravopisne provjere teksta. Prvo razdoblje *Haschecko*va života, do ljeta 2003. godine, jest tzv. *e-mail* faza. U toj fazi tekst je pristizao u obliku poruke na poslužitelj koji bi ga obradio i rezultate obrade u obliku izvještajnoga popisa potom vraćao na adresu pošiljatelja. U *e-mail* fazi *Hascheck* je bio prilično ekskluzivna usluga sa svega nekoliko stotina korisnika u zemlji i inozemstvu, ali je u toj fazi ipak obrađen korpus od blizu 40 milijuna pojavnica, u što je uključena i obrada tri opsežna djela hrvatske leksikografije (Vujić 1996, 1997, Anić 2003, Jojić i Matasović 2004). U ljeto 2003. *Hascheck* dobiva *web*-sučelje dostupno na adresi *http://hascheck.tel.fer.hr/*. *Web*-faza službeno počinje 1. rujna 2003. godine i u njoj *Hascheck* prerasta u široko prihvaćenu uslugu, s desetcima tisuća korisnika u zemlji i inozemstvu. U toj je fazi do 1. svibnja 2010. godine obrađen korpus od 220 milijuna pojavnica.

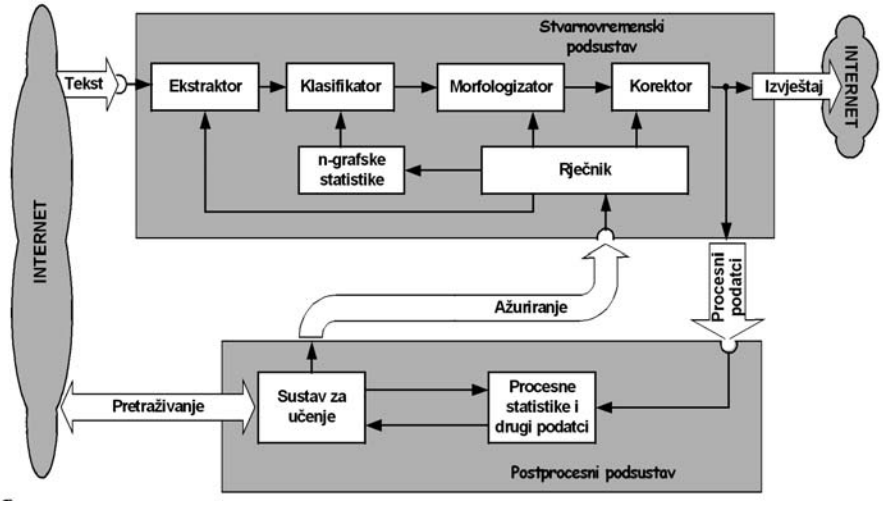
2.1 *Haschecko*va arhitektura

*Haschecko*va arhitektura bitno se razlikuje od arhitekture konvencionalnih pravopisnih provjernika, kakav je npr. Microsoftov pravopisni provjernik za hrvatski. Konvencionalni pravopisni provjernik u pravilu radi sa samo tri komponente (vidi sl. 1). To su:

1. rječnik;
2. ekstraktor, koji iz teksta za daljnju obradu izdvaja sve ono što u rječniku nije pronađeno;
3. korektor, koji nudi moguće ispravke za različnice koje nisu u rječniku.

*Haschecko*v stvarnovremenski podsustav, koji trenutačno reagira na tekstove zaprimljene na obradu, pored navedene tri komponente sadr-

žava još klasifikator i morfologizator. Te dvije komponente daju dodatnu »inteligenciju« sustavu. Da bi se njihova uloga objasnila, potrebno je osvrnuti se na strukturu *Hascheckova* rječnika.



Slika 1: *Hascheckova* arhitektura

Rječnik je najvažnija komponenta svakog pravopisnog provjernika jer o njemu ovisi jezična funkcionalnost sustava (Thompson 1992, 1994), (TEMAA Project 1997). U *Hascheckovu* slučaju rječnik se sastoji od tri popisa različnica:

1. Različnice hrvatskoga općejezičnog fonda (WT-popis); sadržava riječi koje podliježu hrvatskome pravopisu, a mogu se pisati malim ili velikim početnim slovom.
2. Različnice hrvatskoga posebnojezičnog fonda (NT-popis); sadržava osobna, vlastita i druga imena, zatim sve ostale elemente pisanja osjetljive na uporabu velikih i malih slova (kratice, mjerne jedinice i slično), riječi iz stranih jezika koja se u pisanju na hrvatskome rabe u izvornoj grafiji, itd.
3. Različnice engleskog općejezičnog fonda (EngT-popis).

Iako uporaba rječnika stranoga jezika nije uobičajena kod pravopisnih provjernika, dodavanje EngT-popisa *Hascheckovu* rječniku napravljeno je dok su se provjeravali samo tekstovi nastali na Elektrotehničkome fakultetu, koji, zbog naravi struke, često vrve engleskim izrazima. Danas se uključivanje EngT-popisa u *Hascheckov* rječnik može pravdati strukturom *Hascheckova* korpusa:

- 89,4% pojavnica pripada hrvatskomu općejezičnom fondu (WT-popisu),
- 5,9% pojavnica pripada posebnojezičnom fondu (NT-popisu),
- 1,7% pojavnica jesu zatipci (tipfeleri) ili pravopisne pogreške,
- 1,6% pojavnica pripada engleskome općejezičnom fondu (EngT-popisu),
- 1,4% pojavnica jesu čiste brojčane pojavnice koje nisu predmet pravopisne provjere teksta.

Čak i korpus komponiran znatno "konzervativnije" nego *Hascheckov*, *Hrvatska jezična riznica* Instituta za hrvatski jezik i jezikoslovlje, sadržava nezanemariv udio engleskih riječi pisanih u izvornoj grafiji. Prema stanju *Riznice* iz ljeta 2008. godine, dok je čestotnik korpusa još bio javno dostupan (CLC 2008), u njoj je pronađeno 4,2‰ pojavnica koje pripadaju engleskome općejezičnom fondu, a u *Riznicu* su ušle dominantno preko tiskovnoga potkorpusa, glavnina putem *Vjesnikova* dijela u njemu. Navedeni podatci u potpunosti opravdavaju uključivanje EngT-popisa u *Hascheckov* rječnik jer bi bez njega važan parametar jezične funkcionalnosti pravopisnoga provjernika, pokrivanje teksta, tj. udio pojavnica koje se ekstraktorom eliminiraju iz daljnje obrade, bio umanjen za 0,4 do 1,6%, ovisno o korpusu na koji se referenciramo. EngT-popis ne ugrožava ni druge parametre jezične funkcionalnosti provjernika, kao što je primjerice adekvatnost nuđenja ispravaka, jer je zbog udaljenosti dvaju jezika (Dembitz 1996), u smislu Damerau-Levenšteinove metrike (Damerau 1964, Levenshtein 1966), mala vjerojatnost da *misspelling* ili *typo* bude tretiran kao pravopisna pogreška, odnosno zatipak, i obrnuto.

Početni WT-popis izveden je iz desne strane *Englesko-hrvatskoga leksikografskog korpusa* (Bratanić 1975). To je bio radno vrlo intenzivan posao jer je svaku različnicu trebalo ručno provjeriti prema uporabi u *Korpusu*, a nerijetko i kolacionirati s drugim hrvatskim leksikografskim izvorima. Za dobivanje početnih 100.000 različnica hrvatskoga općejezičnog fonda utrošeno je između tisuću i dvije tisuće radnih sati. Zahvaljujući visokoautomatiziranom sustavu za učenje unutar provjernikova postprocesnoga podsustava, WT-popis je do svibnja 2010. godine narastao na 830.000 različnica, na što je ukupno utrošeno 3200 sati rada za nadgledanje učenja.

NT-popis započeo je rasti od nule da bi do svibnja 2010. godine narastao na 600.000 različnica. Za nadgledanje učenja elemenata NT-popisa utrošeno je ukupno 2100 sati rada.

Osnovica za EngT-popis izvedena je iz lijeve strane *Englesko-hrvatskoga*

leksikografskog korpusa. Dobivene različnice prošle su rigoroznu provjeru nekoliko američkih pravopisnih provjernika (Dembitz 1993a). Provedeni postupak dao je i vrijedne spoznaje za razvoj hrvatskoga pravopisnog provjernika. Rezultati provjere upotpunjeni su sadržajima iz rječnika *Unixova* programa `spell` (McIlroy 1982) i iz rječnika jednog starog UNIVAC-ova *spellcheckera* (Turba 1981), čime je dobiven popis oko 70.000 različnica engleskoga općejezičnog fonda. EngT-popis jedini je statični popis u *Haschecko-vu* rječniku jer sve nove engleske različnice koje nemaju potvrdu u njemu, a pojavljuju se u pisanju na hrvatskome, bivaju uvrštene u NT-popis. Pedeset radnih sati utrošenih na stvaranje EngT-popisa zanemarivo je vrijeme u odnosu na ono koje je utrošeno za dobivanje druge dvije sastavnice rječnika.

Popisi su najprimitivniji oblik stvaranja rječnika za potrebe pravopisnih provjernika. Većina suvremenih (konvencionalnih) pravopisnih provjernika koristi se slijednim razvojem različnica iz lema (Peterson 1980, Ispell 1996, Tórn i dr. 2005, Aspell 2008), ili dvorazinskom morfologijom (Arppe i dr. 2005). Lingvistički su razlozi višestruki i dobro su opisani u navedenim referencijama dok je tehnički razlog, barem u ranijim fazama razvoja pravopisnih provjernika, bilo ograničenje diskovnoga i memorijskoga prostora za pohranjivanje i brzo korištenje rječnika. Taj razlog, međutim, u slučaju mrežnoga pravopisnog provjernika otpada. Čak i vrlo skroman poslužitelj može bez problema obrađivati rječnik od 100 milijuna različnica budući da za njihovo pohranjivanje u nekomprimiranome obliku ne treba više od 1 GB prostora. Kao što je iz prije navedenih podataka razvidno, *Haschecko*v je rječnik 70 puta manjega opsega i s njegovim održavanjem i korištenjem nikada nismo imali problema, bez obzira na vrlo skromne uvjete u kojima se usluga razvijala. Popisi, zbog svoje jednostavnosti, donose i neke prednosti: lako ih je ažurirati, što je vrlo važno kod učećih sustava kakav je *Hascheck*, te je na njima lako primijeniti slobodnosoftverske alate (engl. *free and open-source software* — FOSS) za traženje sličnih nizova (engl. *approximate string matching*). U spektru takvih alata odlučili smo se za `nrgrep` (Navarro 2001), koji se pokazao izrazito podatnim za programiranje *Haschecko*va korektora (sl. 1).

Prvi pravopisni provjernik, koji je napisao Ralph Gorin sa Sveučilišta Stanford za računalo DEC-10 1971. godine (Peterson 1980:3), počivao je na rječniku s 10.000 engleskih različnica. Polazeći od razlika u flektivnosti između engleskoga i hrvatskoga jezika, naša je procjena bila da za prvi operativni hrvatski pravopisni provjernik treba oko 100.000 različnica (početni WT-popis). Ta je procjena bitno odstupala od Dolgopolovljeve koji je ustvrdio da za ruski pravopisni provjernik pune funkcionalnosti treba

rječnik od milijun do sto milijuna različnica (Dolgopolov 1986). Kako nije bilo razloga sumnjati u teorijsku procjenu za ruski jezik, koja se, ako je valjana, može primijeniti i na hrvatski jezik, problem flektivnosti hrvatskoga s malim početnim rječnikom trebalo je riješiti na inovativni način.

Problem je riješen uvođenjem stupnja zatipkanosti slovnoga niza koji različnici pridružuje klasifikator (sl. 1). Ideja potječe od pojma *index of peculiarity*, odnosno od informacijske mjere zatipkanosti slovnih nizova, koja je bila operacionalizirana u najstarijem *Unixovu* pravopisnom provjerniku zvanom `typo` (Morris i Cherry 1975). Program je na temelju učestalosti digrafa i trigrafa u obrađivanome tekstu za svaku različnicu izračunavao indeks zatipkanosti na temelju kojega je poslije izrađivao izvještaj. Različnice s najvećim indeksom zatipkanosti, u pravilu zatipci (engl. *typos*), nalazili su se na početku popisa. `Typo` je bio vrlo zanimljiv i relativno uspješan program, premda je radio bez ikakva rječnika, no početkom 80-ih godina nestao je iz *Unixovih* distribucija jer ga je sa svojom superiornijom jezičnom funkcionalnošću potisnuo McIlroyev `spell` (McIlroy 1982), koji je i danas standardni alat na svim *Unixovim* (*Linuxovim*) platformama. Ideja je iskorištena za uvođenje pojma težine slovnoga niza (Dembitz 1982), odnosno informacijske mjere koja se, za razliku od indeksa zatipkanosti, izračunavala na temelju statistike *n*-grafa, gdje je *n* mogao biti proizvoljno velik prirodni broj, izvedene iz rječnika i učestalosti pojavljivanja njegovih elemenata u korpusu nekoga jezika. Poslije je relativno složen izračun napušten i zamijenjen binarnom *n*-grafskom statistikom. Ona se u praksi pokazala robusnom, računalno jednostavnom pa time i brzom, i što je najvažnije, vrlo selektivnom.

Klasifikator (sl. 1) donosi odluku o stupnju zatipkanosti slovnoga niza na sljedeći način:

- izuzetno je zatipkan niz koji sadržava barem jedan trigraf bez potvrde u *n*-grafskoj bazi;
- vrlo je zatipkan niz koji sadržava barem jedan tetragraf bez potvrde u *n*-grafskoj bazi;
- umjereno je zatipkan niz koji sadržava barem jedan pentagraf bez potvrde u *n*-grafskoj bazi;
- nije zatipkan niz čiji su svi pentagrafi u *n*-grafskoj bazi.

Jasno je da je za stvaranje binarne *n*-grafske baze, koju možemo tumačiti kao svojevrstan statistički opis hrvatskoga pravopisnog sustava, dovoljan WT-popis, iz kojeg se ona izvodi i ažurira nakon svakoga učenja. Slovni nizovi obrađeni klasifikatorom u izvještaju dobivaju svoje boje, od crvene (izuzetno zatipkani) do zelene (nezatipkani). Odgovarajuća boja pri-

družuje se i nizovima kojima se ne može pridružiti stupanj zatipkanosti na prije opisani način: nizovi građeni od slova i znamenaka (npr. *A4*), nizovi s neuobičajenom uporabom velikih i malih slova (npr. *WordPerfect*) te nizovi s tri ili manje slova.

Rani eksperimentalni rezultati (Dembitz 1993b) pokazali su da klasifikator uspijeva svrstati 70% WT-popisu nepoznatih riječi u najniži stupanj zatipkanosti, dok je 25% nepoznatih riječi bilo razvrstano kao umjereno zatipkani nizovi. Obrnuta razdioba ostvarena je u razvrstavanju hrvatskih zatipaka i pravopisnih pogrešaka. Važno je napomenuti da u eksperimentima s engleskim jezikom nikada nismo uspjeli postići ni približno jednaku selektivnost. Najviše 45% EngT-popisu nepoznatih riječi bilo je svrstano u najniži stupanj zatipkanosti, dok su engleski zatipci i pravopisne pogreške bili gotovo jednoliko raspodijeljeni po svim stupnjevima zatipkanosti. Postignutu selektivnost u hrvatskome tumačimo fonetskim karakterom hrvatskoga pisma i pravopisa jer se ekonomija glasova, svojstvena svim prirodnim jezicima, pretočila u visoku predvidivost valjanih slovnih nizova u pisanim riječima. S porastom opsega WT-popisa selektivnost razvrstavanja još je porasla, tako da već dugo klasifikator svrstava preko 80% nepoznatih hrvatskih riječi u najniži stupanj zatipkanosti.

Opisana selektivnost bila je vrlo važna za dobro prihvaćanje usluge na samom početku njezina života. S WT-popisom od sto tisuća različnica *Hascheck* je imao pokrivanje teksta od svega 87% (Dembitz i dr. 1999). Otkrivanje i ispravljanje pogrešaka u tekstu bilo je olakšano time što se velika većina zatipaka i pravopisnih pogrešaka nalazila na početku izvještajnoga popisa.

Rezultate klasifikatora preuzima morfologizator (sl. 1). U njemu se na temelju statistike sufikasa u hrvatskome jeziku (Dembitz i dr. 1999) procjenjuje može li obrađivana različnica biti obličnicom hrvatskoga jezika ili ne može. Morfologizator u osnovi radi morfološko-statističko povezivanje različnica koje nisu u rječniku sa sadržajima WT-popisa, odnosno NT-popisa i EngT-popisa. Postupak je vrlo legitiman u računalnoj lingvistici što na globalnoj razini potvrđuje rad Johna Goldsmitha (Goldsmith 2001). Iz razloga točnosti samo različnice s najnižim stupnjem zatipkanosti mogu biti povezane sa sadržajem WT-popisa, dok je povezivanje sa sadržajima NT-popisa i EngT-popisa moguće bez obzira na stupanj zatipkanosti različnice. Povezivanje s EngT-popisom opravdava se činjenicom da brojne engleske imenice (pridjevi puno rjeđe) sklonidbom postaju obličnice hrvatskoga jezika, kao npr. *fitnessa*, koja se kao hrvatska izvedenica iz engleske riječi *fitness* pojavljuje na preko 60.000 hrvatskih *web*-stranica. Sprezanje glagola nije toliko učestalo, iako ima i toga: *uploadati* je hrvatski infini-

tiv izveden iz engleskog glagola *to upload* i u tom se obliku pojavljuje oko 15.000 tisuća hrvatskih *web*-stranica. Uspješno povezane različnice smatraju se oblič nicama hrvatskoga jezika i nad njima se ne obavlja postupak kojemu podliježu nepovezane različnice u korektoru (sl. 1).

Hascheckova preciznost povezivanja iznosi 91,3% za oblič nice hrvatskoga općejezičnog fonda, dok je ona u slučaju hrvatskih oblič nica posebnojezičnoga i engleskoga općejezičnog fonda jednaka 87,6%. Obje se vrijednosti mogu smatrati zadovoljavajućim s obzirom na statistički pristup morfološkoj analizi. Odziv morfologizatora, tj. mjera koliko različnica vrijednih učenja on uspijeva povezati s rječnikom, znatno je povoljniji u slučaju općejezičnih oblič nica. Dok je 81,9% različnica koje je *Hascheck* naučio bilo prije učenja povezano s WT-popisom, samo je 38,7% različnica koje su uvrštene u NT-popisu bilo prije učenje povezano s NT-popisom ili EngT-popisom. Ti podatci samo govore da je pojava »običnih« riječi i njihovih oblika u korpusu znatno predvidljivija od pojave imena, odnosno drugih posebnojezičnih elemenata i njihovih oblika.

Nepovezane različnice smatraju se pogreškama u pisanju i u korektoru se za njih (sl. 1) izračunavaju najvjerojatniji ispravci. Dok je pokrivanje teksta rječnikom bilo vrlo malo, nudili su se samo ispravci za dvije tipične hrvatske pravopisne pogreške: zamjena slovâ *č* i *ć*, te pogreške u pisanju dugoga i kratkoga diftonga /i̯e/. Kada se pokrivanje teksta stabiliziralo na razini od 95% i višoj, uvedeno je opće nuđenje ispravaka s Damerau-Levenšteinovom distancom 1 (jedno zamijenjeno slovo, jedno ispušteno slovo, jedno suvišno slovo ili transpozicija dvaju susjednih slova). Dobiveni kandidati za ispravke ponderiraju se sukladno hrvatskim pravopisnim pravilima, odnosno bliskosti zamjene slova na tipkovnici, čime se dobiva njihov redoslijed kod nuđenja ispravaka. Tako, primjerice, kandidati iz WT-popisa, kod kojih u odnosu na različnicu za koju se nude ispravci postoji zamjena *č*/*ć*, *ije*/*je* i obrnuto, ili je u različnici uočljivo nepoštivanje jednačenja suglasnika po zvučnosti i mjestu tvorbe, dobivaju najviši ponder. Jasno je da se hrvatska fonetska pravila pisanja ne primjenjuju za kandidate iz NT-popisa i EngT-popisa. Danas *Hascheck* nudi ispravke i s distancom 2, a u brojnim pravopisno ili jezično osjetljivim slučajevima ide se i na više vrijednosti. Tako će, na primjer, *Hascheck* za »uslov« i njegove brojne izvedenice kao ispravak ponuditi »uvjet« sa svojim izvedenicama, pri čemu je distanca između korespondentnih nizova najmanje 4 slovna znaka. Općenito govoreći, zbog fonetskoga karaktera hrvatskoga pravopisa, rizično bi bilo ići na distance veće od 2, jer bi se pojavio preveliki broj kandidata za ispravak koje bi bilo teško urediti u suvisli redoslijed prilikom nuđenja ispravaka. Osim toga, distanca 2 zadovoljava i najveći broj *misspe-*

llinga i *typoa* kada se nude ispravci u engleskim pravopisnim provjernicima (Kukich 1992). Također treba istaknuti da se ni Microsoftov pravopisni provjernik za hrvatski jezik za uredski paket *Office 2007* nije u nuđenju ispravaka odmakao dalje od distance 1.

U *Hascheckov* se rječnik može uvrstiti, uvjetno govoreći, i četvrti popis, označen kao ErrT-popis, u kojem se nalaze sve različnice koje su prošle kroz obradu, a da nikada nisu bile zapisane u ostala tri rječnička popisa. Tu se, dakle, radi o popisu svega onoga što *Hascheck* smatra pogreškom u pisanju. Korištenje ErrT-popisa znatno je ubrzalo učenje, o čemu će na kraju idućega potpoglavlja biti nešto više govora, ali on ima važnu ulogu i u stvarnovremenskom podsustavu. Različnice zapisane u ErrT-popisu kada se pojave u obradi nekoga teksta ne prolaze kroz morfologizator već idu izravno korektoru na nuđenje ispravaka. Na taj je način povećana točnost morfologizatora i smanjeno trajanje obrade. O vrijednosti ErrT-popisa najbolje govori njegov opseg: na dan 1. svibnja 2010. godine obasezao je preko 1,3 milijuna različnica i taj broj, jednako kao i broj različnica zapisanih u WT-popisu i NT-popisu, sa svakom obradom raste.

2.2 Postupak učenja

Stvarnovremenski podsustav opslužuje postprocesni podsustav s brojnim podatcima potrebnim za učenje, kao i s podatcima relevantnim za analizu rada cijeloga sustava. Svaka obrada popraćena je statističkim zapisom u kojem se nalazi opseg obrađenoga teksta mjeren brojem pojavnica u njemu, kao i razdioba pojavnica nepoznanica po stupnjevima njihove zatipkanosti. Zapis ima jedinstveni identifikator građen od adrese pošiljatelja, nadnevka i vremena obrade. Pod jednakim identifikatorom bilježe se i sve različnice nepoznanice zajedno s učestalošću njihova pojavljivanja u tekstu, oznakom stupnja zatipkanosti, te podatkom o povezivosti različnica s WT-popisom, odnosno NT-popisom ili EngT-popisom, a ako takve oznake nema, onda različnicu prati popis mogućih ispravaka, kada takav postoji. Konačno, za svaku pojavnicu nepoznanicu bilježi se njezin širi kontekst pojavljivanja (±5 pojavnica), ako i takav postoji.

Procesni podatci i statistike bilježe se zbog potreba sustava za učenje (sl. 1). Sustav za učenje onaj je *Hascheckov* dio koji ga po »inteligenciji« bitno izdvaja od drugih pravopisnih provjernika, bez obzira na to kojemu je jeziku riječ. Zahvaljujući sustavu za učenje jezična kompetencija *Haschec-kovih* korisnika postaje javno dostupno dobro. I sâm sustav za učenje generira statističke podatke, kao što je, primjerice, trajanje pojedinoga učenja.

Sustav za učenje aktivira nadgledatelj učenja kada procijeni da je za to

prikupljeno dovoljno novih procesnih podataka. To znači da sustav mora pamtiti kada je posljednje učenje obavljeno, što je uz jedinstveni identifikator `adresa_pošiljatelja-nadnevak-vrijeme_obrade` trivijalan programerski zadatak. Zbog asinkronosti između stvarnovremenskog podsustava i sustava za učenje, u smislu da se u novopristiglim stvarnovremenskim procesnim podatcima mogu ponoviti nepoznanice koje su upravo u procesu učenja, korpus nepoznanica pripremljen za učenje ponovno prolazi klasifikator, morfologizator i korektor, s tom razlikom da se povezivanje različnica s WT-popisom ne ograničava više samo na one koje su prethodno bile klasificirane kao najmanje zatipkane, već sve različnice iz korpusa pripremljenog za učenje, bez obzira bile one u ErrT-popisu ili ne, podliježu ocjeni morfologizatora. Jasno, one nepoznanice koje su u ErrT--popisu izdvajaju se u posebne datoteke jer je nad njima potrebna posebna ljudska provjera. Iz opisa stvarnovremenskih sastavnica sustava razvidno je da se opisnim postupkom dobiva građa koja je pripremljena za primjenu tehnika umjetne inteligencije kod učenja. One se i primjenjuju, ali uz ljudski nadzor radi očuvanja točnosti rječnika. Automatizaciju postupka učenja najbolje se dade opisati promjenama u brzini učenja. Dok je u čisto manualnoj fazi, prilikom stvaranja početnoga WT-popisa, ona bila između 50 i 100 različnica na sat, u travnju 2010. godine, kada je naučeno 42.108 novih različnica, 19.813 različnica općejezičnoga i 22.295 različnica posebnojezičnoga fonda, brzina učenja iznosila je 432 različnice na sat.

Važnost ljudskoga nadzora kod učenja bit će ilustrirana primjerima sa sl. 2. Na slici je prikazan odsječak korpusa pripremljenog za učenje u ljeto 2009. godine, s tim da nisu izdvojene različnice koje su prethodno bile zapisane u ErrT-popisu. Zapis započinje s oznakom stupnja zatipkanosti (*-ss-* oznaka je za različnice najnižeg stupnja zatipkanosti), nakon toga slijedi različnica nepoznanica, zatim broj njezina pojavljivanja u korpusu koji je obuhvaćen učenjem, da bi na kraju dolazila ili oznaka povezivosti (*p!* znači povezivost s WT-popisom, dok *P!* znači povezivost s NT-popisom ili EngT-popisom) ili ponuđeni ispravci.

1. -ss- Santaninog 6 => P!
2. -ss- Santaninom 3 => P!
3. -ss- Sonije 5 => Sonje? Sinije? Sočnije? Sofije? Sorije? Jonije?
4. -ss- Spiroskog 1 => Pirotskog? Sirskog? Spiljskog? Spinskog? Sportskog?
5. -ss- Sprem 1 => Šprem? Sprej? Sprema? Spreme? Spremi? Sperem?
6. -ss- Suveretu 1 => Suverenu?
7. -ss- UMIJESTO 1 => UMJESTO? UMIJESIO?
8. -ss- Valka 1 => Vlaka? Valja? Valla? Valma? Balka? Falka? Galka? Valjka?
9. -ss- Vartekstovog 1 => Varteksovog?
10. -ss- Vatana 1 => Vagana? Varana? Hvatana? Batana? Catana? Atana? Vataža?
11. -ss- Vidovic 1 => Vidović? Vidovica? Vidovice? Vidovi? Vidovac? Vidovec? Vidovit?
12. -ss- Zamijena 1 => Zamjena?
13. -ss- ŠTIČENIK 2 => ŠTIĆENIK?
14. -ss- amerikanci 1 => Amerikanci?
15. -ss- astronautovih 1 => p!
16. -ss- audi 1 => Audi? auri? sudi? audio? audit? gaudi? naudi? adi? auli? auti? žudi?

Slika 2: Uzorak pripremljen za učenje

Redci 1., 2. i 15. ukazuju da je povezivanje ispravno obavljeno i da se različnice mogu automatski naučiti. Ovdje možemo istaknuti i određene prednosti *Hascheckova* morfologizatora u odnosu na *Hrvatski morfološki leksikon i lematizacijski poslužitelj* (HML) Zavoda za lingvistiku Filozofskog fakulteta u Zagrebu (HML 2005). Morfologizator prepoznaje popridjevljene imenice i u njihovim kosim oblicima (Santana → Santanin → Santaninog, astronaut → astronautov → astronautovih), dok ih HML ne prepoznaje. HML zna sklanjati imenicu »astronaut« i prezime »Krleža« (model za popridjevljenje Santane) u svim padežima jednine i množine, dok su mu pridjevi »astronautov« i »Krležin« nepoznati i kao leme i kao oblici. O kolikom se »morfološkom manjku« HML-a tu radi govore podatci da samo pridjevskih lema koje završavaju na *-ov* u WT-popisu ima 2‰, dok pridjevskih lema koje završavaju na *-in* u NT-popisu ima 6‰.

Uvid u kontekstne zapise potvrdio je da su ispravci u redcima 5., 7., 9., 11., 12., 13. i 14. dobro ponuđeni i što se njihova redoslijeda tiče. »Sprem« (redak 5.) odnosio se na Katarinu Šprem, »Vartekstovog« (redak 9.) zatipak je pridjeva »Varteksovog«, »Vidovic« (redak 11.) se odnosio na Petru Vidović, dok su »amerikanci« (redak 14.) pravopisna pogreška kod pisanja Amerikanaca. Redci 7., 12. i 13. tipične su hrvatske pravopisne pogreške. U svim slučajevima prvorangirani ispravak, i onda kada je jedini, do-

bro je pogođen ispravak, što upućuje na visoku kvalitetu korektora. Može se činiti dvojbeno nuđenje ispravaka u retku 16. (npr. »moj audi«, a ne »moj Audi« kada se govori o vlastitom osobnom automobilu marke Audi). Međutim, kako iskustvo pokazuje da je u pisanju imena česta zamjena velikoga početnog slova malim, radije dižemo »lažnu uzbunu« nego što dopuštamo previđanje učestale pogreške. Da je u pripremu uzorka bio uključen ErrT-popis, svi do sada opisani primjeri pogrešaka ne bi se ni pojavili u uzorku za učenje jer su već prije bili registrirani kao pogreške.

Različnice u redcima 3., 4., 6., 8. i 10., za koje su također ponuđeni ispravci, prvi su se put pojavile u obradi. Uvid u kontekstualne zapise pokazuje da se redak 3. odnosi na Soniju Gandhi, redak 4. na osobu s makedonskim prezimenom Spiroski, redak 6. na mjesto Suvereto koje se nalazi kraj Livorna u Italiji, redak 8. na osobu s nizozemskim prezimenom Van der Valk, a da se redak 10. odnosi na etapu *Tour de Francea 2009* Vatan — Saint-Fargeau. Sve su različnice korektno napisane obličnice hrvatskoga jezika. Dok u slučajevima obličnica »Spiroskog«, »Suveretu« i »Vatana« nema nikakve dvojbe kod uvrštavanja u NT-popis jer ne interferiraju ni sa čim podudarnim u hrvatskome jeziku, obličnice »Sonije« i »Valka« traže promišljanje. Genitiv stranoga imena »Sonia« blizak je genitivu frekventnoga hrvatskog imena »Sonja«, koji je u retku 3. ponuđen kao prvi ispravak. Isto je tako genitiv nizozemskoga prezimena »Valk« blizak genitivu hrvatskoga prezimena »Vlak« koji je ponuđen kao prvi ispravak u retku 8. Učenje obličnica iz redaka 3. i 8. predstavlja rizik da zatipci frekventnih hrvatskih imena prođu neotkriveni. Pretraživanje interneta pokazuje da je »Sonia« vrlo često strano ime jer se pojavljuje na preko 50 milijuna *web*-stranica. »Van der Valk« je također često nizozemsko prezime jer se pojavljuje na milijun *web*-stranica. Međutim, jezična intuicija govori da je puno vjerojatnije da će se u budućnosti na hrvatskome pisati o različitim »Sonijama«, nego o »Van der Valkovima«. Stoga je obličnica »Sonije« uvrštena u NT-popis dok »Valka« nije; odluka je dijelom poduprta i velikom razlikom u učestalosti njihova pojavljivanja u lokalnome korpusu. Buduće obrade potvrdile su ispravnost intuitivne odluke. Danas se u NT-popisu, pored »Sonije«, nalaze i obličnice »Soniji«, »Sonijom« i »Soniju«, sve uredno povezane sa »Sonia« i kontekstualno potvrđene glede korektnosti uporabe, dok se o Van der Valku nije više pisalo.

Primjeri iz prethodnoga odlomka objašnjavaju zašto je prilikom učenja potreban ljudski nadzor. Iako kandidati za učenje stižu na nadzor razvrstani u različite datoteke, svaka datoteka mora proći kakav-takav ljudski nadzor prije nego što njezini kandidati budu uvršteni u odgovarajući popis. Danas učenje imena i drugih elemenata NT-popisa troši najviše ljud-

skoga vremena i zato su uzeti primjeri iz tog segmenta učenja. Učenje različnica hrvatskoga općejezičnog fonda postalo je, zahvaljujući opsežnosti WT-popisa i preciznosti morfologizatora, puno lakše nego što je to bilo na početku, premda je i tu ponekad potrebno pored lokalnoga konteksta provjeriti i širi *web*-kontekst u hrvatskoj domeni uporabe (naredbom `site:hr`). Sa sl. 1 je očito da se pored lokalnoga konteksta pojavljivanja neke različnice kandidata za učenje može po potrebi rabiti i globalni *web*-prostor kao najveći tekstualni repozitorij na svijetu. U tim se slučajevima pretraživanje često podešava naredbom `site:<ISO-3166 code>` prema domeni pretpostavljenoga jezika ili područja iz kojega se očekuje da kontekst izvorno dolazi (ISO-3166 2010).

Prilikom postupka učenja poštuju se u načelu rješenja svih triju sadašnjih hrvatskih pravopisa (Silić i Anić 2001, Badurina i dr. 2007, Babić i Moguš 2010) jer se smatra pravom pisca odabrati pravopis koji će rabiti. To znači da u WT-popisu postoje inačice pisanja pojedinih riječi za koje pravopisi nude različita rješenja, kao npr. *greška*/*grješka* i njihove izvedenice. Jasno je da učestalost pojedinih inačica ovisi o tome koliko se one često rabe u tekstovima koje je *Hascheck* obradio. Jedino nije prihvaćeno sklanjanje stranih imena (i riječi) kakvo je predloženo u poglavlju »Pisanje riječi iz stranih jezika« u (Badurina i dr. 2007:205—218) za onaj dio gdje bi po drugim pravopisima trebale nastupiti promjene na morfološkoj granici. Prema Badurina—Marković—Mićanovićevu pravopisu obličnice stranoga imena »Sonia« trebale bi se pisati ovako: »Soniama«, »Sonie«, »Sonii«, »Soniom« i »Soniu«. Takvo pisanje ne predstavlja tehnički problem za *Hascheck*, ali predstavlja problem za ono što neposredno za *Hascheckom* slijedi.

Usluga kojom se namjerava upotpuniti naš javni servis jest *HascheckVoice*, u probnoj inačici dostupan na adresi *http://hascheck.tel.fer.hr/voice2/voice.html*. Radi se o usluzi strojne tvorbe govora ponajprije namijenjene slijepim i slabovidnim osobama, ali i svima onima koji ne žele čitati, nego žele slušati izgovoren tekst. Kod strojne tvorbe govora postoji problem normalizacije teksta unutar kojega se kao potproblem javlja pitanje transkripcije stranih imena u fonološki sustav jezika za kojega se govor strojno generira (Dembitz i dr. 2010). U slučaju hrvatskoga jezika slovo »j« na morfološkoj granici upozorenje je normalizatoru da u izgovoru nastupaju glasovne promjene koje se razlikuju od neposredne (standardne) transkripcije slova ili skupine slovâ koje graničniku prethode. U konkretnom primjeru »Sonia« će u bazi normalizatora biti zapisna kao sljedeća lema: { Sonia: `Soni-a → [so][nj-a]` (s oznakom kratkosilaznog naglaska na prvom slogu)}. Zahvaljujući slovu »j« na morfološkoj granici kosih oblika

imena »Sonia« normalizator će lako prepoznati korijen i iz njega jednako tako lako (jer se dva glasa *j* ne mogu pojaviti zajedno u sufiksalnoj tvorbi u hrvatskome jeziku) izgenerirati izgovorljive oblike: `sonjama`, `sonje`, `sonji`, `sonjom` i `sonju`. Bez jednostavnosti u prepoznavanju morfološke granice sustav bi bio u opasnosti da, primjerice, »Sonie« prepozna kao francusku riječ *la sonie* = glasnost i da je tako i izgovori: `soni` (s kratkouzlaznim naglaskom na drugome slogu). Problematičnih bi se slučajeva moglo navesti još mnogo, no vjerujemo da i taj primjer dovoljno dobro ilustrira koliko bi teškoća strojnoj tvorbi govora na hrvatskome moglo stvoriti prihvaćanje rješenja iz Badurina—Marković—Mićanovićeva pravopisa u dijelu koji se odnosi na sklanjanje stranih imena i riječi.

Proširivanje rječnika pravopisnoga provjernika uvijek je povezano s pitanjem jezične točnosti sustava. Korisnici ne žele da im provjernik ne prijavi pogrešku koja u tekstu postoji. To se može dogoditi ako rječnik nije pročišćen, ali i onda ako korisnik zbog nepozornosti ili neznanja utipka pogrešnu riječ. Taj drugi slučaj rješava se kontekstnim pravopisnim provjernicima, no oni su za sada privilegija, uz ograničenu jezičnu funkcionalnost glede pronalaženja kontekstualnih pogrešaka, maloga broja središnjih jezika (Dembitz 2009).

Nadzirano učenje s provjerom konteksta u kojem se kandidat za učenje pojavljuje jedan je od načina očuvanja točnosti *Hascheckova* rječnika. Drugi je način ograničavanje sadržaja rječnika samo na one riječi različnice koje su potvrđene u pisanju na hrvatskome jeziku, što se dade interpretirati kao primjena Zipfova zakona na konstrukciju rječnika.

Zipfov zakon govori da se riječi u korpusu pojavljuju s izuzetnim razlikama u učestalosti pojavljivanja. Većina različnica u velikim korpusima su tzv. *hapax legomene*, riječi sa svega jednom potvrdom pojavljivanja. Korpus *Hrvatske jezične riznice*, koji je u ljeto 2008. godine brojio oko 85 milijuna pojavnica, sadržavao je 57% *hapax legomena*. Praktički isti udio *hapax legomena* (56,6%) sadržava i engleski korpus usporediva opsega (Kornai 2002), što samo govori da je taj statistički fenomen neovisan o jeziku. Zipfov zakon može se protegnuti i na obličnice visokoflektivnih jezika (Karlsson 1985, 1986 i 2000, Arppe 2006): dok se jedne pojavljuju učestalo, drugima je vjerojatnost pojavljivanja praktički jednaka nuli. Dobar se primjer može naći i u hrvatskome. Blizu 40% valjanih različnica hrvatskoga općejezičnog i imenskog fonda, koje je Denis Lacković razvio metodom slijednoga razvoja iz hrvatskih lema izvorno za potrebe `ispell`-a (Lacković 2003), nemaju potvrdu ni u *Riznici* ni u *Hascheckovu* korpusu. Rječnik je, inače, prilično skromnoga opsega od 200.000 različnica. S druge pak strane, *Hascheckov* rječnik pokriva korpus *Riznice* s 99,02%, uz visoku ko-

relaciju na razini učestalosti pojavljivanja riječi u dva korpusa. Nadalje, 0,33% sadržaja *Riznice Hascheck* smatra pogreškama u pisanju jer se nalaze u ErrT-popisu, pa prema tome one ne mogu davati potvrdu za dobre različnice u Lackovićevu rječniku. Prema procjeni morfologizatora 0,65% sadržaja *Riznice* bez potvrde u *Hascheckovim* popisima ima sljedeću strukturu: 0,11% pripada hrvatskome općejezičnom fondu, 0,23% jesu pogreške, a 0,31% pripada posebnojezičnome fondu. Kako je u Lackovićevu rječniku imenski, odnosno posebnojezični fond vrlo slabo zastupljen, 40% njegovih valjanih različnica trebalo bi pronaći potvrdu u jednom promilu sadržaja *Riznice*. Križanje Lackovićevih valjanih različnica bez korpusne potvrde i različnica iz nerazvrstanoga 0,65-postotnog sadržaja *Riznice* dalo je svega 317 zajedničkih različnica, što je praktički zanemarivo. Primjer lijepo ilustrira da rječnici razvijeni iz lema imaju za pravopisni provjernik mnogo neuporabljiva sadržaja koji samo donosi rizik za jezičnu točnost alata.

Prvo ispitivanje jezične funkcionalnosti hrvatskih konvencionalnih pravopisnih provjernika obavili smo prije desetak godina (Dembitz i Sokele 1997). Izmjerena netočnost varirala je između 2% i 6%, pri čemu je u mjerenju korišten reprezentativni uzorak pogrešaka uzet iz ErrT-popisa. Tada se najboljim pokazao tadašnji Microsoftov pravopisni provjernik za hrvatski jezik. Nedavno smo ponovili mjerenje točnosti Microsoftova provjernika u uredskome paketu *Office 2007*, pri čemu smo koristili uzorak od 2000 najučestalijih pogrešaka zajedničkih *Hascheckovu* korpusu i korpusu *Riznice*; neki od niže navedenih primjera jesu u *Riznici* ispravno pisane riječi, ali se u neispravnome obliku pisanja učestalo pojavljuju u ErrT-popisu. Izmjerena netočnost iznosila je 6,25%, što je porazan rezultat.

Mjerenje je ukazalo i na tri sustavne jezične slabosti Microsoftova alata:

1. Problem kapitalizacije — propuštaju se, primjerice, »vikica« i »icrc« zajedno s valjanim oblicima »Vikica« i »ICRC« (kratica za Međunarodni odbor Crvenoga križa). Od pogrešaka sa sl. 2 propušteni su »amerikanci« i »audi«.
2. Visoku toleranciju prema srbizmima — propuštaju se, primjerice, »stepen« i »milion« i svi njihovi oblici.
3. Toleranciju prema čestim pogreškama u pisanju, kao što su, na primjer, »presječe« ili »obuko«.

U trećem slučaju postoji jezikoslovno opravdanje tolerancije jer su navedeni oblici vokativi jednine imenica »presjek« i »obuka«. Vjerojatnost da se takvi vokativi pojave u pisanju na hrvatskome izvan gramatičkih priručnika i sličnih tekstova praktički je zanemariva, dok se »presječe«, kao tipična pravopisna pogreška u pisanju trećega lica jednine prezenta gla-

gola »presjeći«, pojavljuje 86 puta u ErrT-popisu, a oblik »obuko«, pretpostavljivo zatipak od »obukao« ili »obukom«, 31 put. Kako su rječnici svih konvencionalnih hrvatskih pravopisnih provjernika dobiveni slijednim razvojem različnica iz lema (neki su to radili s više, a neki s manje uspjeha), na temelju provedenih mjerenja procjenjujemo da postojanje vokativa sa zanemarivom vjerojatnošću pojavljivanja u pisanju doprinosi s barem 1% povećanju njihove netočnosti.

Hascheck, čiji je rječnik korpusno orijentiran, može sustavno upozoravati svoje korisnike na uporabu glagolskoga priloga i pridjeva »slijedeći« iz razloga česte zamjene s pridjevom »sljedeći«; dobar uvid u takve zamjene dobiva se pretraživanjem *Riznice*. Nadalje, u rječniku nema glagola »pokusati« ni većine njegovih oblika (zabilježen je samo oblik »pokusa« jer je istovremeno i genitiv česte imenice »pokus«) zbog njihove bliskosti s mnogo učestalijim glagolom »pokušati« i pogrešne prakse pisanja riječi hrvatskoga jezika engleskom abecedom nastale s pojavom elektroničke pošte. Također, neke obličnice, kao npr. »odječe« (vokativ jednine imenice »odjek« i treće lice množine prezenta glagola »odječati«) ili »odječi« (treće lice jednine prezenta glagola »odječati«) nisu u rječniku, iako njihove leme i mnogi drugi iz njih izvodivi oblici jesu, upravo zbog bliskosti s učestalim pravopisnim pogreškama, u konkretnome primjeru u pisanju odgovarajućih obličnica česte imenice »odjeća«. Takvih primjera moglo bi se navesti jako puno i u svim je slučajevima nepostojanje legitimnih hrvatskih riječi i oblika u WT-popisu motivirano njihovom bliskošću s učestalim pogreškama u pisanju na hrvatskome. Držimo da je opravdanije da pravopisni provjernik signalizira lažnu uzbunu u slučajevima rijetke ispravne uporabe ovakvih različnica nego da propušta učestale pogreške. Iz svega slijedi da korpusna orijentacija rječnika značajno doprinosi jezičnoj točnosti sustava.

Po okončanju nadgledanja učenja sustav stvara dvije datoteke, jednu za ažuriranje WT-popisa, drugu za ažuriranje NT-popisa. Nakon toga se obje datoteke pozorno pregledavaju prije samoga ažuriranja jer je ljudski čimbenik u učenju mogući razlog pojave pogrešnoga učenja. Ažuriranje popisa prati i ažuriranje frekvencijskih profila WT-popisa, NT-popisa, EngT-popisa i ErrT-popisa. Sustav također bilježi na bazi jedinstvenoga identifikatora `adresa_pošiljatelja-nadnevak-vrijeme_obrade` i tekst iz kojega je pojedina različnica naučena. Na koncu sustav bilježi i vrijeme utrošeno za nadgledanje novonaučenih različnica, iz čega je moguće izvesti parametar kojega nazivamo brzinom učenja.

Visokoautomatizirani postupak učenja i propusti u nadgledanju učenja mogući su uzroci uvrštavanja pogrešaka u *Hascheckov* rječnik. Posebni dio sustava za učenje vodi računa da se takvi propusti, kada se uoče, i

otklone. Tako je iz WT-popisa do svibnja 2010. godine izbrisano 5698 različnica ili blizu 0,7% njegova sadržaja. Sva se brisanja u WT-popisu bilježe jer ona utječu i na točnost klasifikatora, odnosno sadržaj *n*-grafske baze koja se pritom ažurira. Kako brisanja u NT-popisu ne utječu na ažuriranje *n*-grafske baze, ona se ni ne bilježe. Procjenjujemo da je iz NT-popisa izbrisano oko 1% njegova sadržaja, odnosno da je iz rječnika ukupno izbrisano oko 12.000 različnica. I korisnici pravopisnoga provjernika u mogućnosti su da nadziratelje učenja upozore na pogreške zapisane u rječniku, no to se do sada događalo iznimno rijetko.

U početcima učenja nije se puno marilo što se u datotekama pripremljenima za nadzor učenja ponavljaju i pogreške koje su već prošle višestruki nadzor. Međutim, s porastom prometa uvidjelo se da to predstavlja gubitak vremena, pa su se iz datoteka za učenje počele isključivati različnice zapisane u ErrT-popisu. Na taj način brzina je učenja povećana s 200 na 300 različnica na sat. Sljedeće povećanje brzine učenja ostvareno je isključivanjem iz učenja posebnojezičnih različnica sa svojstvom *hapax legomena*. Razlog za takvu radikalnu promjenu u strategiji učenja jest linearnost ovisnosti broja novih posebnojezičnih različnica o opsegu obrađenoga korpusa (vidi potpoglavlje 4.2), što je s gledišta nadzora učenja, uz znatno povećani promet provjernika, postalo radno neprihvatljivo. Isključivanje je tehnički ostvareno na način da su u datoteke za učenje uvrštavane samo povezive različnice, zatim različnice za koje nije ponuđen ispravak, dakle one s distancom 3 ili većom u odnosu na sadržaj triju rječničkih popisa, kao i sve različnice najnižega stupnja zatipkanosti bez obzira na ostale značajke, osim ako nisu u ErrT-popisu. To posljednje opravdava se željom da se iz učenja ne isključe općejezične *hapax legomene*. Radikalna promjena strategije učenja podigla je brzinu učenja na 350 različnica na sat jer je eliminiran velik broj pogrešaka s prvim pojavljivanjem koje su prije podlijegale nadzoru. Konačna brzina učenja od preko 400 različnica na sat posljedica je tehničkoga unapređenja sustava za učenje (Pavlek 2010).

3. Prometni podatci i analize

Brojni procesni podatci prikupljeni u postprocesnome podsustavu podloga su za analize ponašanja sustava, ali i za predviđanja što od sustava možemo očekivati. U ovome poglavlju, kao i u iduća dva, dat ćemo prikaz onih analiza koje se zasnivaju na uporabi *Excela*, standardnoga Microsoftova alata. Podloga za većinu analiza su podatci prikupljeni u *web*-fazi.

U tab. 1 dana je razdioba *Hascheckova* prometa po IP-domenama. U tablici su, istaknute oznakom iz standarda ISO-3166, prikazane samo one

domene iz kojih je *Haschecku* do 1. svibnja 2010. godine generirano 1‰ ili više ukupnoga prometa. Inače, *Hascheck* je registrirao korisnike iz 65 domena (zemalja) sa svih šest kontinenata[2]. U opisu prometa koriste se tri osnovna parametra pravopisnoga provjernika:

- pokrivanje teksta (engl. *text coverage* — TC), ili udio pojavnica koje je ekstraktor prepoznao kao valjane riječi (iskazuje se u postotcima);
- indeks učenja (engl. *learning index* — LI), ili prosječni broj novonaučenih različnica na stotinu obrađenih pojavnica;
- razina pogrešnosti (engl. *error rate* — ERR), tj. udio zatipkanih ili pravopisno pogrešno napisanih pojavnica u tekstu (također se iskazuje u postotcima).

Vrijednosti parametara prikazanih u tab. 1 srednje su vrijednosti za razdoblje od 1. rujna 2003. do 1. svibnja 2010.godine.

DOMENA	IP-adresa	KORPUS [pojavnica]	KORPUS [%]	TC	LI	ERR
AT	120	216.005	0,10	96,71	0,57	2,59
BA	1.832	3,578.603	1,62	96,81	0,39	2,60
CA	34	524.787	0,24	97,75	0.11	1,74
DE	277	654.403	0,30	96,44	0,50	2,52
GB	73	296.801	0,13	97,15	0,45	1,73
HR	83.578	208,831.860	94,62	97,87	0,37	1,58
HU	62	225.090	0,10	98,32	0,40	1,20
IT	223	666.775	0,30	97,75	0,30	1,82
ME	229	1,908.177	0,87	95,85	0,34	3,74
RS	285	535.734	0,24	93,87	0,46	5,25
SI	2.155	1,874.704	0,85	94,86	0,64	4,29
US	136	774.544	0,35	98,05	0,35	1,32
Drugi	552	620.763	0,28	96,37	0,39	2,82
UKUPNO	**89.556**	**220,708.246**	**100,00**	**97,79**	**0,37**	**1,65**

Tablica 1: Razdioba prometa po IP-domenama

[2] Do srpnja 2011. godine, kada je rukopis pripremljen za tisak, broj IP-domena narastao je na 78, broj IP-adresa na 200.000, od čega je HR-domeni pripadalo 185.000, što je tvorilo 9,2% hrvatskoga nominalnog internetskog prostora u to vrijeme, broj obrađenih tekstova na 2 milijuna, dok je obrađeni *web*-korpus obasezao pola milijarde pojavnica.

Pojavljivanje svih susjednih zemalja, Bosne i Hercegovina (BA), Mađarske (HU), Italije (IT), Crne Gore (ME), Srbije (RS) i Slovenije (SI), u tab. 1 bilo je očekivano. Pojavljivanje tri anglofonske zemlje, Kanade (CA), Ujedinjenoga Kraljevstva (GB) i Sjedinjenih Američkih Država (US), i dvije germanofonske zemlje, Austrije (AT) i Njemačke (DE), u tab. 1 tumačimo, pored intenziteta kontakata Hrvatske s navedenim zemljama, i željom korisnika iz tih zemalja naviknutih na visokofunkcionalne pravopisne provjernike za engleski, odnosno njemački jezik, da pronađu usporedivi alat i za hrvatski. Inače, 83.578 IP-adresa s kojih je *Haschеck* dosegnut iz Hrvatske (šesti redak tab. 1) tvore 4,7% hrvatskoga nominalnog internetskog prostora prema podatcima za travanj 2010. godine (MaxMind 2010). Realno je taj udio veći od 5% jer brojne IP-adrese dodijeljene Hrvatskoj nisu u uporabi već ih operateri čuvaju kao svojevrsnu pričuvu za budućnost.

Parametar razine pogrešnosti može se u slučaju tekstova pristiglih iz zemalja u kojima je za vremena Jugoslavije u uporabi bio »srpskohrvatski jezik« uzeti kao mjera udaljenosti jezikâ tih zemalja od hrvatskoga. Prema podatcima iz tab. 1, od hrvatskoga (ERR = 1,58%) najudaljeniji je srpski (5,25%), zatim slijedi crnogorski (3,74%), da bi po istim kriterijima hrvatskome najbliži bili jezici koji se rabe u Bosni i Hercegovini (2,60%), a gdje se pored hrvatskoga još rabe bošnjački i ijekavsko-jekavski srpski (bosanskosrpski). Za uzorke pisanja u Bosni i Hercegovini inicijalno su odabrana četiri potkorpusa iz sveučilišnih gradova: tuzlanski, kao uzorak bošnjačke pismenosti, opsega 157.133 pojavnica s ERR_{Tuzla} = 1,29%, banjalučki, kao uzorak bosanskosrpske pismenosti, opsega 44.036 pojavnica s $ERR_{Banja\ Luka}$ = 4,13%, mostarski, kao uzorak hrvatske pismenosti u BiH, opsega 980.608 pojavnica s ERR_{Mostar} = 1,71%, i sarajevski, kao uzorak svebosanske pismenosti, opsega 1,437.346 pojavnica s $ERR_{Sarajevo}$ = 3,27%. Kako postoji visoka vjerojatnost da u sarajevskom potkorpusu postoji nezanemariv udio bosanskosrpskih tekstova, kao dodatni uzorak bošnjačke pismenosti, jer iznimna čistoća tuzlanskoga potkorpusa traži objašnjenja koja nadilaze okvire ovoga rada, uzet je zenički potkorpus opsega 188.288 pojavnica s ERR_{Zenica} = 2,39%, za koji smatramo da najbolje odražava udaljenost između bošnjačkoga i hrvatskoga. Iz podataka prikazanih u tab. 1 razvidno je da hrvatski jezik putem *Haschecka* vrši utjecaj na jezike u BiH i Crnoj Gori, slično utjecaju engleskoga na hrvatski jezik.

Tablica 2 donosi razdiobu prometa prema opsegu tekstova zaprimljenih na obradu.

Opseg teksta	Tekstova	[%]	Korpus	[%]	Srednji opseg
[1, 10)	163.300	18,56	523.100	0,24	3,20
[10, 100)	248.889	28,29	11,343.202	5,14	45,58
[100, 1.000)	437.163	49,70	129,259.959	58,56	295,68
[1.000, 10.000)	29.544	3,36	62,235.873	28,20	2.106,55
[10.000, ∞)	748	0,09	17,346.112	7,86	23.189,99
UKUPNO	**879.644**	**100,00**	**220,708.246**	**100,00**	**250,91**

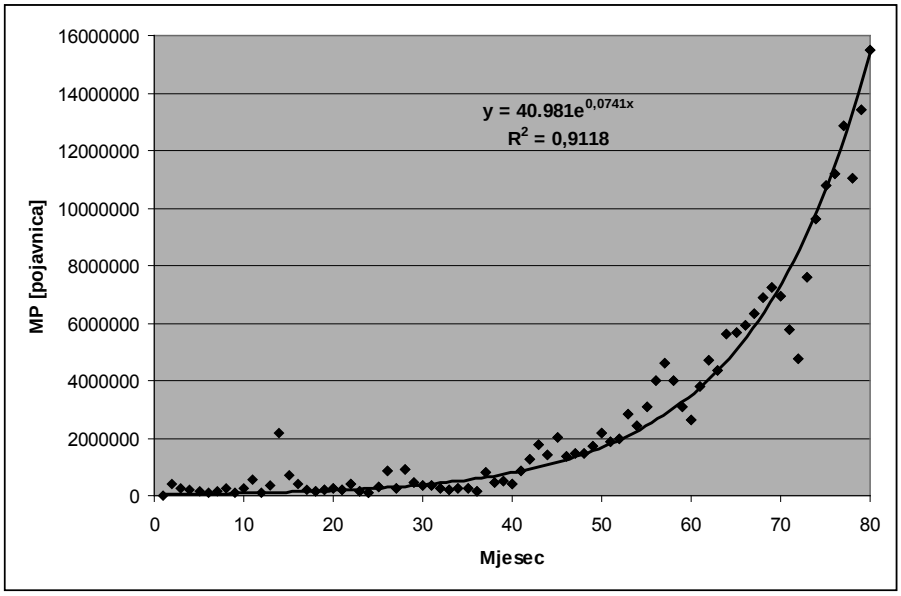
Tablica 2: Razdioba prometa prema opsegu obrađenih tekstova

Blizu 20% obrada, u kojima je srednja duljina teksta oko 3 pojavnice, zapravo su pravopisni upiti. *Hascheck* se pokazao pouzdanim pravopisnim savjetnikom jer pored riječi koje su u pravopisnim rječnicima sadržava i riječi koje tamo nisu zabilježene, a trebale bi biti, kao što je to npr. »pjeskaš«, riječ iskovana za označavanje igrača odbojke ili rukometa na pijesku čiju uporabnost u hrvatskome potvrđuju česte pojave u tekstovima *Vjesnika*, *Večernjeg lista* itd. Tekstovi opsega jedne stranice (oko 300 pojavnica) najučestaliji su u obradi, a za njima slijede tekstovi prosječnog opsega poruke u elektroničkoj pošti (45 pojavnica). Tekstovi opsega eseja (2000 pojavnica) ili novele (23.000 pojavnica) puno se rjeđe šalju na obradu, iako takvi tekstovi u ukupnome korpusu sudjeluju s 36%. Za obradu teksta opsega novele *Hascheck* u prosjeku troši 20 sekundi procesorskoga vremena.

Tablica 3 donosi statistički prikaz lojalnosti korisnika usluzi.

Broj obrada	IP-adr.	[%]	Tekstova	[%]	Korpus	[%]	PBO
[1, 10)	80.076	89,41	216.799	24,65	72,118.579	32,67	2,71
[10, 100)	9.187	10,26	184.752	21,00	48,597.588	22,02	20,11
[100, 1.000)	251	0,28	59.948	6,82	13,014.537	5,90	238,84
[1.000, 10.000)	33	0,04	99.064	11,26	25,282.564	11,46	3.001,94
[10.000,∞)	9	0,01	319.081	36,27	61,694.978	27,95	35.453,44
UKUPNO	**89.556**	**100.00**	**879.644**	**100,00**	**220,708.246**	**100,00**	**9,82**

Tablica 3: Lojalnost usluzi



Slika 3: Mjesečni promet (MP)

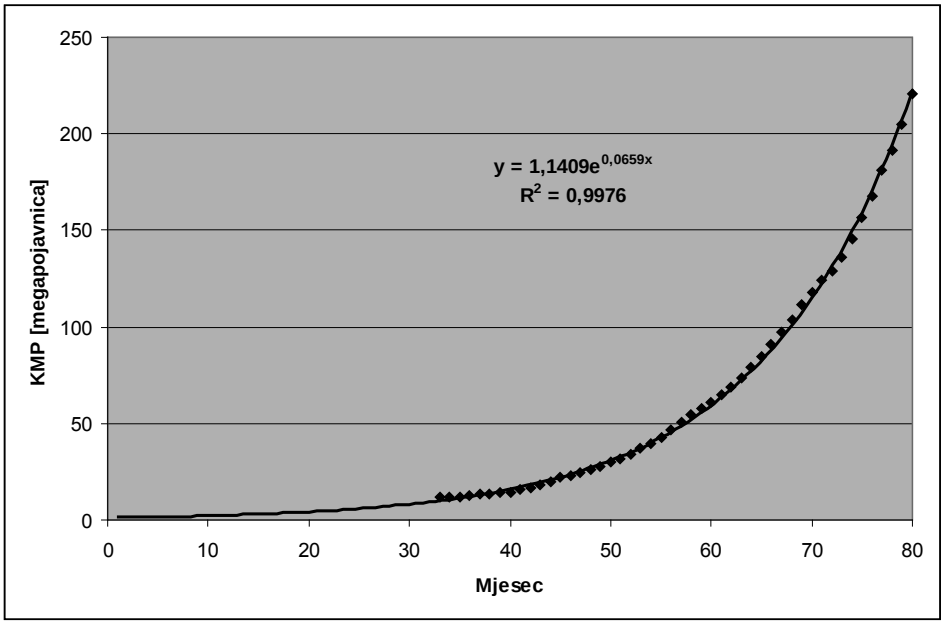
Činjenicu da je manje od 10 tekstova na obradu stiglo s nešto više od 80.000 IP-adresa tumačimo dinamičkim dodjeljivanjem IP-adresa velikom broju »malih« korisnika. Međutim, *Hascheck* ima i 42 velika korisnika sa statičkim IP-adresama i više od tisuću, odnosno deset tisuća obrada po adresi; stupac PBO donosi prosječni broj obrada po pojedinoj korisničkoj kategoriji. Među velikim korisnicima dva su slovenska, ostali su hrvatski, a većina pripada izdavačkome sektoru. Veliki korisnici sudjeluju u obradama s udjelom od 47%, dok njihov korpus tvori gotovo 40% ukupnoga korpusa.

Slika 3 prikazuje porast mjesečnoga prometa (MP) za razdoblje od 80 mjeseci, od rujna 2003. do uključivo travnja 2010. godine. Očito je da je promet u travnju 2010. godine dosegao opseg od blizu 16 milijuna pojavnica. Točke su empirijske vrijednosti, dok je krivulja ona matematička funkcija koja empirijske podatke najbolje opisuje. Porast prometa ima eksponencijalni karakter, a korelacijski koeficijent $R^2 = 0{,}9118$ ukazuje da empirijski podatci vrlo dobro koreliraju s funkcijom porasta prometa.

Još bolja korelacija ostvarena je u modeliranju kumulativnoga porasta mjesečnoga prometa (KMP):

$$KMP(i) = \sum_{j=1}^{i} MP(j) \qquad (6),$$

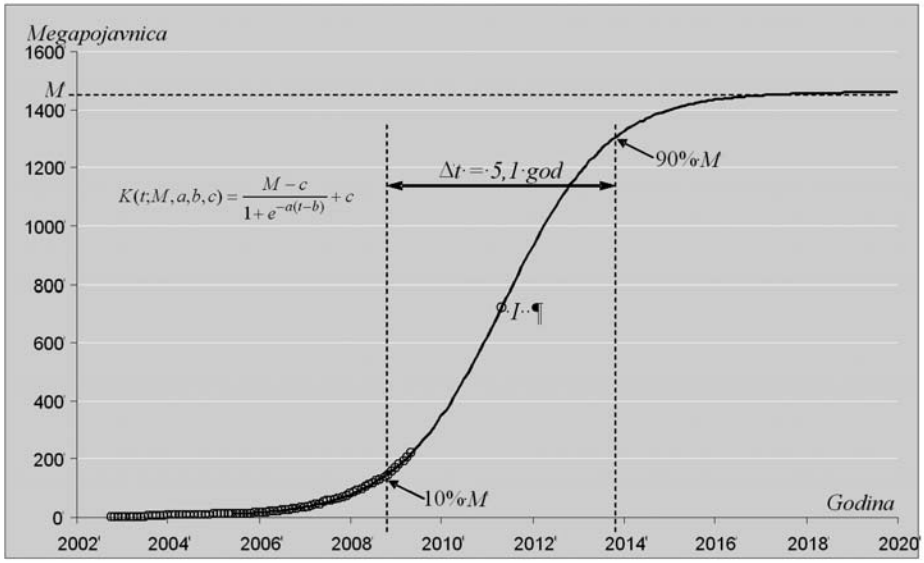
gdje je *MP(j)* mjesečni promet (točke sa sl. 3), a *i* broj mjeseci. Na sl. 4 prikazane su samo točke za posljednje četiri godine (48 mjeseci) u kojima je mjesečni promet imao vrijednosti od 200.000 pojavnica na više. Korelacija na razini $R^2 = 0{,}9976$, što praktički predstavlja vrijednost jedan, između empirijskih podataka i matematičke funkcije koja ponašanje tih podataka opisuje svojstvena je fizikalnim zakonima. To znači da funkciju sa sl. 4 možemo tretirati kao egzaktni zakon ponašanja *Hascheckova* korpusa u vremenu. Parametri funkcije govore da *Hascheckov* promet raste s prosječnom stopom prirasta od 6,81% mjesečno, što znači da se opseg obrađenoga korpusa udvostruči na godišnjoj razini.



Slika 4: Kumulativni mjesečni promet (KMP)

Podatci o kumulativnom prometu poslužili su Mladenu Sokeleu, stručnjaku za modeliranje internetskoga prometa (Sokele 2008, 2009), da izradi logistički model porasta *Haschecko*va korpusa (sl. 5), gdje su varijable korpus *K* u megapojavnicama i vrijeme *t* u godinama, s parametrima modela: *M* = 1.459,302 [megapojavnica], *a* = 0,864982 [1/godina], *b* = 2.011,344 [godina] i *c* = 1,579048 [megapojavnica]. Model ukazuje (parametar *b*) da će infleksija u rastu korpusa (točka *I* na sl. 5) nastupiti početkom svibnja 2012. godine do kada će biti obrađen korpus od 700 milijuna pojavnica (približno *M*/2). Iz analize dinamike rasta slijedi da *Haschecк* mora dobiti svoju napredniju zamjenu najkasnije do konca 2014. godine jer s 2015. godinom korpus ulazi u saturaciju (*K* > 90% *M*). Drugim riječima, konac životnoga vijeka *Haschecka* (u aktualnome obliku) nastupa po obradi od otprilike 1,314 milijarde pojavnica.

Logistički modeli primjereni su za analizu i predviđanje dinamike internetskih usluga (Meade i Islam 2006) jer se internetske usluge stalno mijenjaju kako bi opstale, a što očekuje i *Haschecк* u skoroj budućnosti.



Slika 5: Logistički model kumulativne dinamike prometa

Iz podataka prikazanih u ovom poglavlju razvidno je da u Hrvatskoj nema leksikografskoga pothvata koji bi se po osnovi korištenoga korpusa mogao mjeriti s *Hascheckovom* funkcionalnom leksikografijom. Takav se pothvat niti ne nazire jer tvorci drugih hrvatskih megakorpusa ne nude uz njih alate koji bi u sinergiji s korisnicima korpusa poslužili u leksikografske svrhe. Takvi alati nisu računalno pretjerano složeni, ali za njihovu implementaciju potrebno je znati što je u korpusu već leksikografski obrađeno, a što nije.

4. Ponašanje osnovnih *Hascheckovih* parametara u vremenu

Ovo je poglavlje posvećeno modeliranju ponašanja triju osnovnih parametara, pokrivanja teksta (TC), razine pogrešnosti tekstova (ERR) i indeksa učenja (LI), u vremenu. Za vremensku koordinatu u modeliranju uzima se opseg obrađenoga korpusa. Pri tome ćemo razlikovati dva »vremena«:

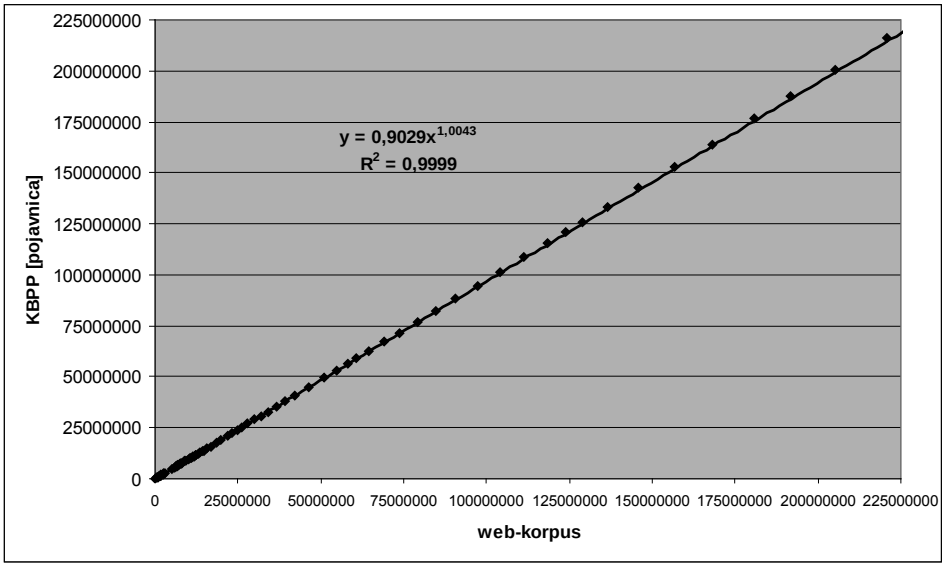
1. korpus iz *web*-faze, što je primjereno »vrijeme« za modeliranje TC-a i ERR-a;
2. ukupni korpus, jer se kod modeliranja LI-a mora voditi računa da je *Hascheckovo* učenje započelo s uvođenjem usluge u javnu uporabu.

Empirijski će podatci i nadalje biti prikazivani kao točke, dok je njihovo pozicioniranje kod prikaza TC-a, ERR-a i LI-a određeno sljedećom vremenskom koordinatom:

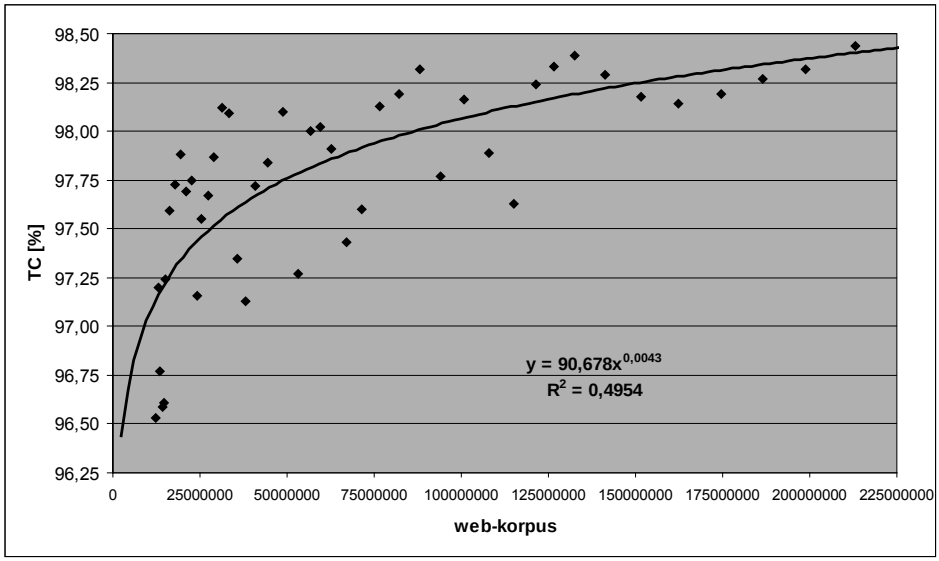
$$t_i = t_0 + KMP(i-1) + \frac{MP(i)}{2} \qquad (2),$$

gdje je t_0 = 0 za *web*-korpus, dok je u slučaju ukupnoga korpusa t_0 = 39,135.406 pojavnica, opseg korpusa koji je obrađen u *e-mail* fazi. Relacija (2) pokazuje da se empirijski podatci pozicioniraju u sredinu mjeseca na koji se odnose pri čemu mjeseci, zbog razlika u prometu, imaju različita »trajanja«.

4.1 Pokrivanje teksta i razina pogrešnosti



Slika 6: Kumulativni broj prepoznatih pojavnica
u obrađenim tekstovima (KBPP)



Slika 7: Ponašanje TC-a u vremenu — empirijski podatci i funkcija

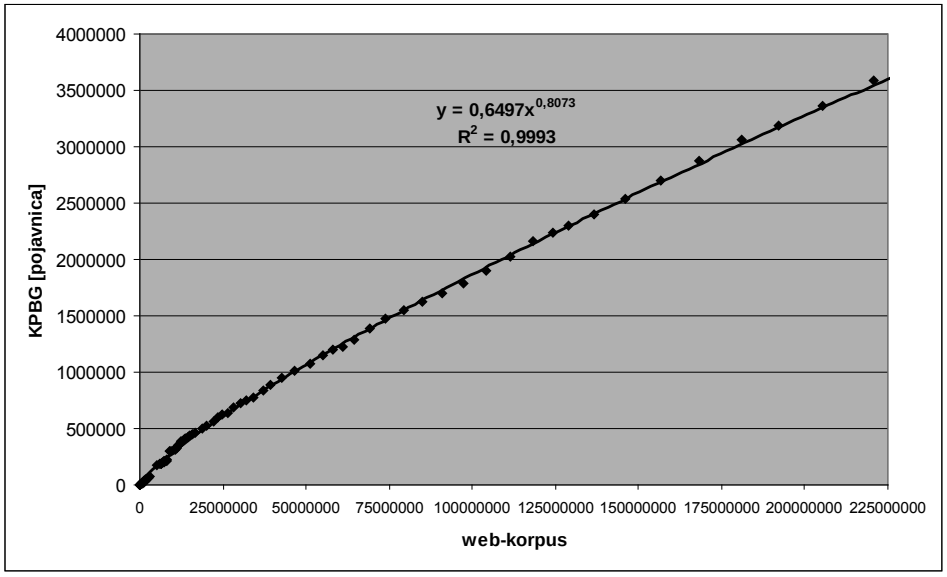
Slika 6 donosi empirijske podatke o kumulativnome broju prepoznatih pojavnica u obrađivanim tekstovima (KBPP) i funkciju oblika $a \cdot x^b$ koja empirijske podatke najbolje opisuje. Takve funkcije, uz $b > 1$, imaju monotono rastuću derivaciju. Kako je

$$TC(t) = 100 \cdot \frac{d}{dt} KBPP(t) \qquad (3),$$

slijedi da je pokrivanje teksta funkcija monotono rastuća u vremenu i to s velikom sigurnošću s obzirom na izuzetno visoku korelaciju empirijskih podataka i funkcije sa sl. 6.

Slika 7 donosi funkciju (3) i korespondentne empirijske podatke. Iako je na slici prikazano samo 45 empirijskih točaka, jer je na početku njihovo grupiranje pregusto pa time i nepregledno, korelacijski koeficijent R^2 = 0,4954 odnosi se na cijelo razdoblje od 80 mjeseci. Izračunali smo korelacijski koeficijent i za razdoblje od posljednje tri godine (R^2 = 0,6509), koji samo pokazuje da se s porastom opsega obrađenoga korpusa zakonomjernost ponašanja TC-a povećava.

Identičnu smo metodu primijenili i kod određivanja ponašanja razine pogrešnosti tekstova u vremenu.

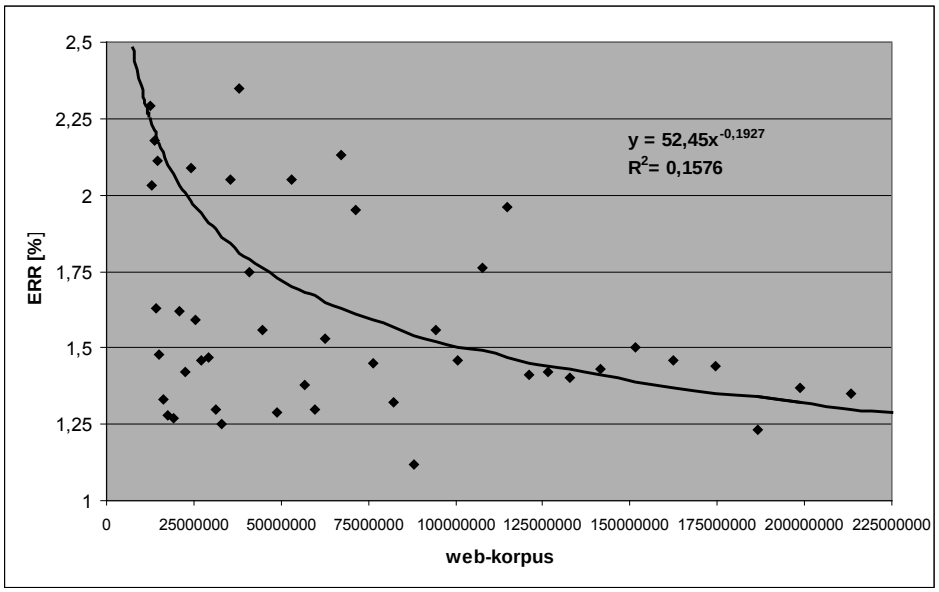


Slika 8: Kumulativni porast broja grešaka
u obrađenim tekstovima (KPBG)

Slika 8 prikazuje kumulativni porast broja grešaka (KPBG) u obrađenim tekstovima. Ponovno smo dobili funkciju oblika $a \cdot x^b$, samo ovaj put s parametrom $b < 1$, što znači da je funkcija

$$ERR(t) = 100 \cdot \frac{d}{dt} KPBG(t) \qquad (4)$$

monotono padajuća funkcija. Slika 9 donosi funkciju (4) i 45 točaka, korespondentnih empirijskih podataka. Ponovno je korelacijski koeficijent $R^2 = 0{,}1576$, koji je vrlo nizak jer je pogrešnost izrazito ljudska kategorija, izračunat za cjelokupno razdoblje od 80 mjeseci. Korelacijski koeficijenti za posljednje tri godine ($R^2 = 0{,}3154$), odnosno za posljednju godinu ($R^2 = 0{,}6751$), potvrđuju da se s porastom opsega obrađenoga korpusa zakonomjernost ponašanja ERR-a povećava.



Slika 9: Ponašanje ERR-a u vremenu — empirijski podatci i funkcija

Porast pokrivanja teksta u vremenu posljedica je učenja. Podatci sa sl. 7 kazuju da zadnjih deset mjeseci TC varira oko 98,25% s trendom porasta prema 98,5%. Da bi se te brojke stavile u kontekst strojne pravopisne provjere tekstova pisanih hrvatskim jezikom, izmjerili smo odziv i pokrivanje teksta Microsoftova pravopisnog provjernika za uredski paket *Office* 2007. Mjerenje je rađeno sukladno preporukama (TEMAA Project 1997) i u njima su korišteni reprezentativni uzorci od 8000 različnica hrvatskoga općejezičnog fonda, odnosno 2000 različnica hrvatskoga posebnojezičnog fonda. Dobiveni rezultati prikazani su u tab. 4.

	Općejezično (WT)	Posebnojezično (NT)
Odziv [%]	85,2	18,4
Pokrivanje teksta [%]	99,3	62,4

Tablica 4: Parametri Microsoftova pravopisnog provjernika za hrvatski

Polazeći od strukture *Hascheckova* korpusa opisane u potpoglavlju 2.1 lako je izračunati da bi pokrivanje *Hascheckova* korpusa od strane Microsoftova pravopisnog provjernika iznosila 93,86%. Kako je srednje *Haschecko*vo pokrivanje hrvatskoga potkorpusa u *web*-fazi 97,87% (tab. 1), dolazimo do razlike od četiri postotna poena u korist *Haschecka*. Takva je razlika primjerena kada se uspoređuju pravopisni provjernici jezičnotehnološki perifernih jezika s provjernicima jezičnotehnološki središnjih jezika (Dembitz i dr. 2009), ali nikako nije primjerena kada se oni uspoređuju unutar jednoga jezika. Navedena razlika posljedica je bogatstva posebnojezičnoga fonda u *Hascheckovu* rječniku. Ona, također, objašnjava stalni porast *Hascheckove* popularnosti.

Smanjenje razine pogrešnosti tekstova u vremenu neočekivani je rezultat modeliranja. Tumačimo ga sljedećim pretpostavkama:

1. *Hascheckovi* korisnici s vremenom su naučili izbjegavati standardne pravopisne pogreške u pisanju, a što je utjecalo na smanjenje ERR-a.
2. Ispravljanje pogrešaka u slučaju mrežnoga pravopisnog provjernika nešto je vremenski zahtjevnije nego u slučaju konvencionalnih pravopisnih provjernika. Kako se ljudsko ponašanje zasniva na načelu najmanjega napora, želja da se izbjegnu suvišna ispravljanja utječe na urednost pisanja, što smanjuje ERR.
3. Moguće je da su mnogi obrađeni tekstovi nastali u Microsoftovim alatima za pisanje i unutar njegova pravopisna provjernika, a uslijed čega su pristizali već prilično pročišćeni na dodatnu provjeru.
4. Dulje iskustvo s pravopisnim provjernicima uvjerit će svakoga korisnika da oni nisu »bezgrešni« alati. Naime, uz nemarno pisanje povezana je i vjerojatnost da se jedna riječ zatipka u drugu riječ koja ne odgovara kontekstu pisanja. Korisnička svijest o nesavršenosti pravopisnih provjernika također podiže urednost pisanja, odnosno smanjuje ERR.

Bez obzira jesu li i u kolikoj mjeri navedene pretpostavke točne, izvjesno je da *Hascheck* obavlja i društveno korisnu ulogu: tjera svoje korisnike da pišu urednije.

4.2 Heapsov zakon primijenjen na *Haschecku*

Heapsov zakon povezuje opseg korpusa s brojem različnica u njemu:

$$V(t) = \alpha \cdot t^{\beta} \qquad (5).$$

U tom izrazu V je vokabular (broj različnica u korpusu, u općem slučaju uključuje i pravopisne pogreške i zatipke), t je opseg korpusa, dok su α i β parametri, od kojih je parametar β uvijek manji od jedinice. Heapsov zakon dade se izvesti iz Zipfova zakona (Kornai 2002) i predstavlja njegov kumulativni oblik. Tipične vrijednosti parametara α i β za engleski jezik nalaze u granicama između 10 i 100, odnosno između 0,4 i 0,6.

Hascheck pruža jedinstvenu priliku da se pored porasta ukupnoga vokabulara njegova rječnika (V) modelira i porast općejezičnoga (V_{WT}), odnosno posebnojezičnoga (V_{NT}) vokabulara. Rezultati modeliranja prikazani su u tab. 5 za tri korpusna područja:

- početni desetomilijunski korpus (područje A);
- ukupni korpus u rasponu od 50 do 100 milijuna pojavnica (područje B);
- ukupni korpus iznad 100 milijuna pojavnica (područje C).

Prilikom modeliranja područja A morali smo uzeti u obzir da je *Hascheck* počeo raditi s početnim rječnikom izvedenim iz pretkorpusa (*Englesko-hrvatski leksikografski korpus*). Stoga smo u (5) morali dodati pomak P:

$$V(t) = \alpha \cdot (t - P)^{\beta} \qquad (6),$$

gdje P govori o veličini pretkorpusa. Negativne vrijednosti pomaka P u prvom i drugom retku tab. 5 vrlo su bliske i potvrđuju da je za dobivanje početnoga rječnika od 100.000 općejezičnih različnica upotrijebljen pretkorpus od 800.000 pojavnica, što odgovara veličini desne strane *Englesko-hrvatskoga leksikografskog korpusa*. Pozitivna vrijednost u slučaju V_{NT}-a pokazuje da je učenje posebnojezičnih različnica počelo brzo nakon puštanja usluge u javnu uporabu. Jasno je da je pomak u području A irelevantan za područja B i C. Kako je Heapsov zakon također kumulativni zakon, korelacijski koeficijenti dobiveni izračunom funkcija iz empirijskih podataka uvijek su bili veći od 0,99, isto kao i kod ostala tri kumulativna zakona primijenjena u ovome radu (sl. 4, 6 i 8).

	Heapsov zakon	α	β	P
Područje A	V	207,24	0,4578	-778.825
	V_{WT}	662,67	0,3716	-766.928
	V_{NT}	0,4310	0,7467	123.821
Područje B	V	145,13	0,4791	
	V_{WT}	4.281,1	0,2722	
	V_{NT}	0,006014	0,9701	
Područje C	V	819,9	0,3852	
	V_{WT}	5.398,1	0,2592	
	V_{NT}	6,8616	0,5885	

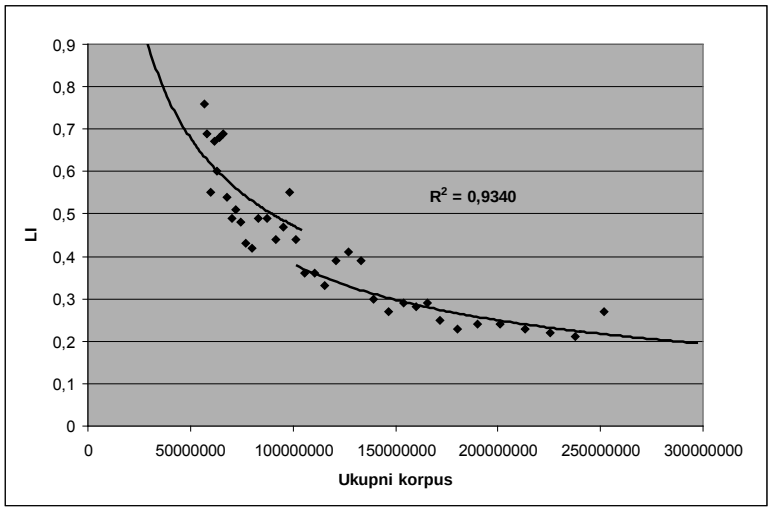
Tablica 5: Parametri Heapsovog zakona za *Hascheckov* korpus

Parametri funkcije V za područja A i B malo se razlikuju, što potvrđuje i činjenica da se opsezi rječnika za t = 100,000.000, izračunati njihovom primjenom, razlikuju za svega 3%. Međutim, parametri funkcija V_{WT} i V_{NT} se u istim područjima primjetno razlikuju. Parametar α funkcije V_{WT} znatno je porastao, dok se isti parametar funkcije V_{NT} znatno smanjio. Obrnuto ponašanje u područjima A i B pokazuje parametar β, s tim da je on u području B za V_{NT} postao vrlo blizak vrijednosti 1 (linearni rast). Kako je parametar β dominantan za ponašanje Heapsova zakona, iz ovih promjena slijedi da porastom korpusa vokabular općejezičnoga fonda teži ka (kvazi)saturaciji, što potvrđuje i smanjenje parametra β u području C, dok vokabular posebnojezičnoga fonda teži prema beskonačnosti. Ovim se daje i jezikoslovno objašnjenje Kornaijeve matematičke tvrdnje (Kornai 2002) da su vokabulari svih prirodnih jezika potencijalno beskonačni. Beskonačnost vokabulara posljedica je stalnoga uvoza novih imena i drugih posebnojezičnih oblika iz stranih jezika, ili njihova stvaranja u vlastitome jeziku, bez izgleda da u tom jezičnom segmentu brzo nastupi saturacija.

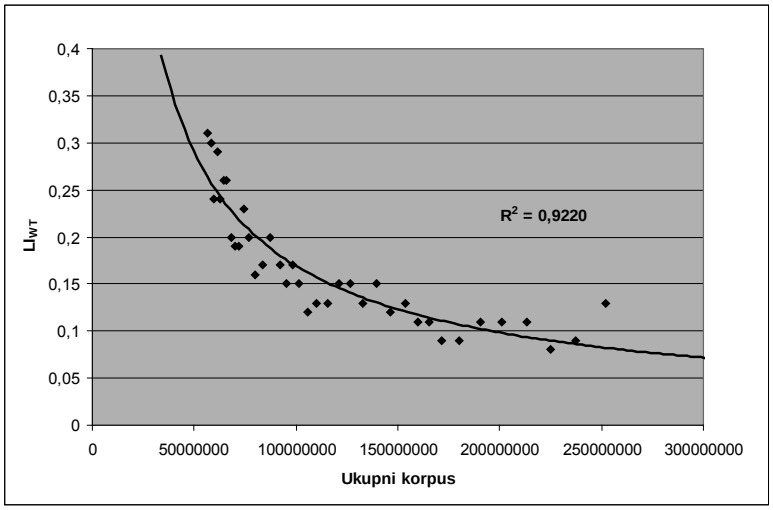
Spoznaja o gotovo linearnoj karakteristici funkcije V_{NT} kod stomilijunskih korpusa utjecala je na promjenu strategije učenja opisane na kraju 2. poglavlja. Naime, s eksponencijalnim rastom korpusa u vremenu, opisanim u 3. poglavlju, učenje novih posebnojezičnih elemenata, koji bi po broju bili srazmjerni opsegu novoobrađenoga korpusa, postalo bi radno neprihvatljivo. Iz

$$LI\ (t) = 100 \cdot \frac{dV}{dt} \qquad (7),$$

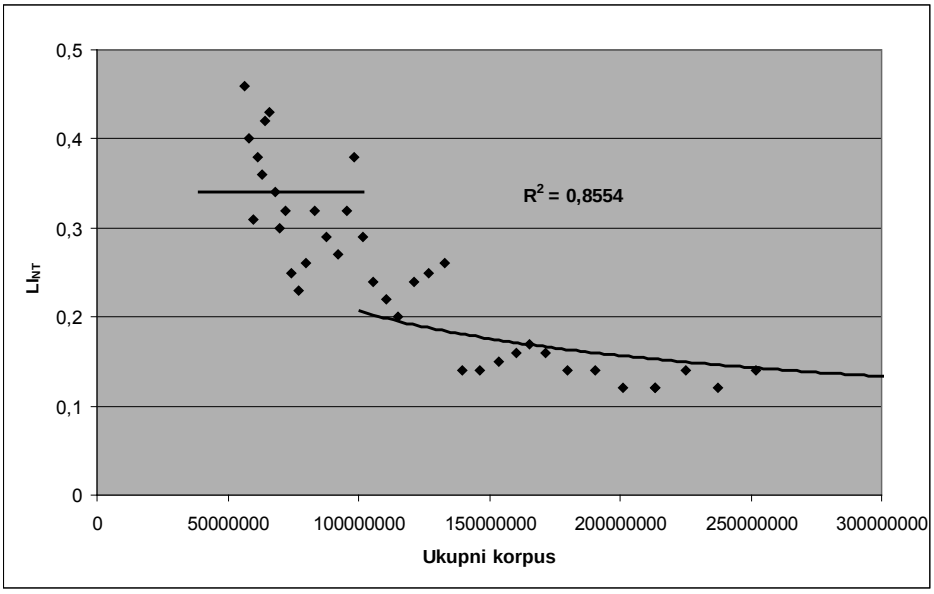
s tim da se izraz (7) može primijeniti i za izračunavanje funkcija indeksa učenja, kako za općejezični (LI_{WT}), tako i za posebnojezični fond (LI_{NT}), slijedi po parametrima iz područja B za raspon korpusa od sto milijuna do milijardu pojavnica da je LI_{NT} u granicama od 0,336 do 0,314, što je praktički identično. Mijenjanje strategije učenja nametnulo se kao nužda.



Slika 10: Ukupni indeks učenja — empirijski podatci i funkcija



Slika 11: Indeks učenja za općejezični fond — empirijski podatci i funkcija



Slika 12: Indeks učenja za posebnojezični fond — empirijski podatci i funkcija

Sl. 10, 11 i 12 donose empirijske podatke i funkcije indekasa učenja. Diskontinuiteti funkcija na sl. 10 i 12 posljedica je promjene u strategiji učenja, koja je izravno utjecala da parametri funkcije V_{NT} u području C postanu »normalniji« (vidi tab. 5), što je doprinijelo da se promijene i parametri funkcije V kojoj je funkcija V_{NT} sastavnica. Činjenica da funkcija LI_{WT} nema diskontinuitet (sl. 11) potvrđuje da smo mijenjajući strategiju učenja uspjeli postići da općejezični fond ostane nezakinut u učenju. Valja naglasiti da promjena strategije učenja nije utjecala na zakonitost mijenjanja funkcija *TC* i *ERR* (sl. 7 i 9) jer 0,1 do 0,2% *hapax legomena*, koliko ih u stomilijunskim korpusima otprilike postoji na razini posebnojezičnih pojavnica, ni ne može pokvariti relativno nisko korelirajuće funkcije čije se temeljne vrijednosti iskazuju u postotcima.

Na sl. 10, 11 i 12 prikazano je samo po 38 empirijskih točaka, koje prikazuju odgovarajuće indekse učenja ostvarene nad korpusima čiji je mjesečni opseg premašivao milijun pojavnica, tj. od mjeseca ožujka 2007. godine. Kada se mjesečni promet ustalio na razini od milijun i više pojavnica, nametnula se potreba planiranja rada kojega će trebati uložiti u nadgledanje učenja. Parametri potrebni da se predvidi koliko će radnih sati trebati uložiti po osnovi nadzora učenja u nekom vremenskom razdoblju izvedivi su iz funkcije *KMP* sa sl. 4 i funkcije *LI* sa sl. 10, te iz podataka o brzini učenja s kraja potpoglavlja 2.2. Program za predviđanje opsega učenja sastavni je

dio postprocesnoga podsustava (sl. 1) i prilično je točan: korelacija između predviđenih i ostvarenih opsega učenja u rasponu duljem od tri godine iznosi $R^2 = 0{,}8938$. Visoku preciznost predviđanja opsega učenja tumačimo vrlo visokom točnošću funkcija *KMP* i *LI*, iskazane njihovim korelacijskim koeficijentima većim od 0,9.

Nešto niži korelacijski koeficijent u slučaj funkcije LI_{NT} u odnosu na koeficijent dobiven za funkciju LI_{WT} statistička je potvrda konstatacije da je ponašanje posebnojezičnih sastavnica nekoga korpusa manje predvidivo od ponašanja općejezičnih sastavnica istoga korpusa. Inače, razlike u korelacijskim koeficijentima dobivene za indekse učenja u odnosu na korelacijske koeficijente dobivene za pokrivanje teksta i razinu pogrešnosti tekstova (sl. 7 i 9), koji su znatno niži, tumačimo razlikama u utjecaju pojedinca na odgovarajući parametar. Jasno je da parametar ERR podliježe jakom individualnom utjecaju unutar ukupnih kolektivnih svojstava pisanja, što se statistički iskazuje ukupno niskim korelacijskim koeficijentom. Taj je utjecaj puno manji u slučaju parametra TC jer on umnogome ovisi o učenju. Najmanji pojedinačni utjecaj postoji kod funkcija indekasa učenja jer one opisuju ponašanje jezika i njegovih sastavnica, a jezik pripada kolektivitetu sa statističkom zanemarivošću individualnoga djelovanja unutar njega.

5. Kognoelektrička analogija

Zipfov zakon na početku korpusa ima zakon eksponencijalnoga pada (Kornai 1999) koji se znatno razlikuje od linearnih zakona u logaritamskom mjerilu prikazanih na sl. 10, 11 i 12. Takav smo zakon i mi dobili za početni desetomilijunski korpus:

$$li(t) = a + (1 - a) \cdot e^{-\frac{t-T}{\tau}}, \quad li = LI/100 \qquad (8).$$

Sada je indeks učenja normiran (LI koji smo do sada koristili podijeljen je sa 100), dok parametri funkcije imaju sljedeće vrijednosti: $a = 0{,}025855$ [različnica/pojavnica], $\tau = 2{,}035.042$ [pojavnica] i $T = -4{,}774.792$ [pojavnica]. Fiksirajući parametre τ i T potražili smo funkciju koja bi odgovarala pokrivanju teksta u istom rasponu korpusa. Dobili smo zakon eksponencijalnoga rasta:

$$tc(t) = b \cdot \left(1 - e^{-\frac{t-T}{\tau}}\right), \qquad tc = TC/100 \qquad (9),$$

gdje je pokrivanje također normirano, a parametar $b = 0{,}952595$ bezdimenzionalni je parametar. Korelacija je kod oba modela bila iznad 0,9.

Identične funkcije opisuju nabijanje realnoga kondenzatora priključenog na istosmjerni naponski izvor (Lončar 2006):

$$i(t) = k \cdot \left[a' + (1 - a') \cdot e^{-\frac{t}{\tau}} \right] \qquad (10),$$

$$u_C(t) = k \cdot b' \cdot \left(1 - e^{-\frac{t}{\tau}} \right) \qquad (11),$$

gdje je *i(t)* struja koju daje izvor, $u_C(t)$ napon na kondenzatoru, dok su parametri funkcija izvedenice iz fizikalnih svojstava sustava: E je napon izvora, R_s je njegov unutarnji otpor, C je kapacitet kondenzatora, a R_C otpor paralelno spojen na kapacitet, iz čega slijedi:

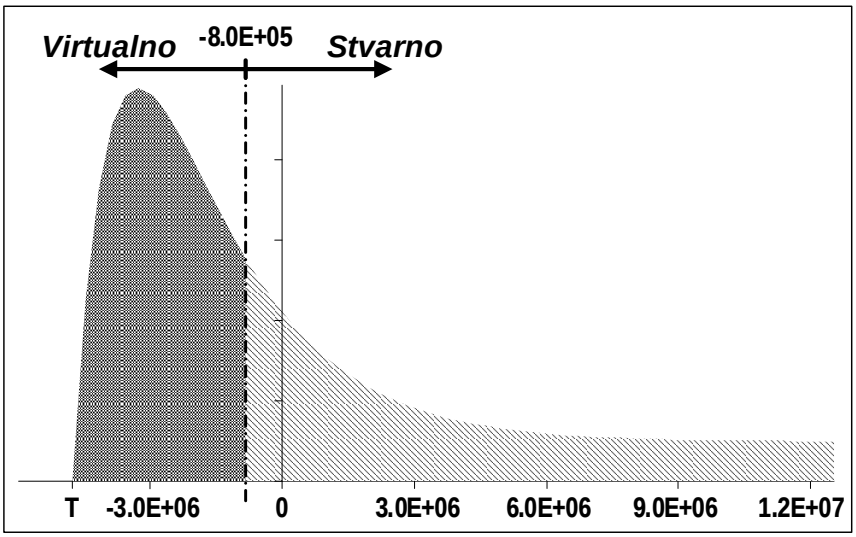
$$k = \frac{E}{R_s}, \quad a' = \frac{R_s}{R_C + R_s}, \quad b' = \frac{R_C \cdot R_s}{R_C + R_s}, \quad \tau = C \cdot \frac{R_C \cdot R_s}{R_C + R_s} \qquad (12),$$

s time da realni sustavi uvijek zadovoljavaju uvjet $R_C >> R_s$.

Identičnost *Hascheckovih* funkcija s fizikalnim zakonom, uz činjenicu da se njegovo učenje dade interpretirati kao struja riječi koja puni rječnik s ciljem podizanja stupnja pokrivanja teksta (napona), potakla nas je da ustvrdimo da se *Hascheck* ponaša kao kondenzator priključen na izvor, pri čemu se pod izvorom podrazumijevaju njegovi korisnici, te da uvedemo pojam snage učenja (*PoL*, *Power of Learning*):

$$PoL(t) = \left[a + (1 - a) \cdot e^{-\frac{t-T}{\tau}} \right] \cdot b \cdot \left(1 - e^{-\frac{t-T}{\tau}} \right) \qquad (13),$$

kao umnoška funkcija indeksa učenja i pokrivanje teksta, po analogiji na fizikalni sustav u kojemu je snaga umnožak struje i napona. Funkcija (13) okosnica je naše kognoelektričke analogije (Dembitz i dr. 1998).



Slika 13: *Hascheckova* funkcija snage učenja (*PoL*)

Na sl. 13 prikazana je funkcija snage učenja te su naznačena dva razdoblja u *Hascheckovu* životu, virtualno i stvarno. Naime, pomak T u funkcijama (8) i (9) puno je veći od opsega pretkorpusa upotrijebljenoga za stvaranje početnoga *Hascheckova* rječnika. Stoga je »vrijeme« prije pretkorpusa (800.000 pojavnica) proglašeno virtualnim vremenom *Hascheckova* života, vremenom dok je on bio samo ideja za čije je oživotvorenje trebalo uložiti istraživački rad. Svojstva funkcije (13) istaknuta na slici odgovaraju stvarnosti u razvoju i eksploataciji računalnih učećih sustava: istraživački rad intenzitetom je puno zahtjevniji (tamnija površina na sl. 13) od rada potrebnoga za održavanja i unapređivanje dobro dizajniranoga učećeg sustava.

Kako je integral snage u vremenu energija (rad), poznavanje funkcije *Hascheckove* snage učenja omogućilo nam je da izračunamo koliko je nominalno vrijedan rad utrošen u njegov razvoj:

$$W(t)=\int_{T}^{t} PoL(t)\cdot dt=\tau\cdot b\cdot\left[p\cdot a+(1-2a)\cdot(1-e^{-p})-\frac{1-a}{2}\cdot(1-e^{-2p})\right] \qquad (14),$$

gdje je $p = (t - T)/\tau$, dok W ima dimenziju [različnica], što i jest jedinica rada u stvarnome *Hascheckovu* vremenu. S uvrštavanjem granice $t = -800.000$ dobili smo da je rad uložen u razvoj *Haschecka* ekvivalentan radu potrebnom da se nauči 750.273 različnica. Provjera valjanosti računa na-

pravljena je uvrštavanjem $t = 0$, pri čemu smo dobili radni ekvivalent od 844.384 različnica. Razlika od 844.384 − 750.273 = 94.111 različnica blizu je vrijednosti od 100.000 različnica, koliko je otprilike brojao početni *Hascheckov* rječnik, i govori da energetski račun prije podcjenjuje nego li što precjenjuje rad uložen u istraživanje i razvoj.

Podijeli li se iznos od 750.273 različnice s 200 različnica na sat, kolika je bila brzina učenja prije bilo kakvih promjena u strategiji učenja, dobiva se vrijednost od 3751 sati rada. Pridoda li se tomu iznosu još 1500 sati kao srednja vrijednost rada uloženoga u ručnu kontrolu sadržaja početnoga *Hascheckova* rječnika, koji je procijenjen na tisuću do dvije tisuće sati, dolazimo do 5251 radnih sati utrošenih u ostvarivanju *Haschecka* kao javne usluge ili do 2,78 čovjek-godina. *Hascheck* je rezultat doktorske disertacije (Dembitz 1993b), a rezultat izračuna precizno se poklapa s nominalnim trajanjem doktorskoga studija od tri godine.

Kognoelektrička analogija primjer je da učeći sustavi u sebi nose podatke ne samo o svojoj aktualnosti nego i o svojoj prošlosti, dok su bili tek predmetom istraživanja i razvoja. To ne treba čuditi jer učeći sustavi na započinju život kao *tabulae rasae*, već se njihovo inicijalno znanje oplemenjuje genetikom, odnosno u računalnim sustavima programima koji znanje opredmećuju, obogaćuju i čine upotrebljivim u budućnosti. Rad potreban za osmišljavanje i izradu tih programa dobiva svoj nominalni radni ekvivalent kroz parametre izvedive iz rada sustava u stvarnome vremenu, jedino je potrebno domisliti se kako iz njih izvući priču o prošlosti.

Ovim se poglavljem željelo pokazati da je u računalnoj leksikografiji sve mjerljivo, ne samo rad da se računalni leksikon održava i obogaćuje, već i rad potreban da se on osmisli. Tu spoznaju držimo važnom zato što jezičnotehnološki projekti nisu jeftini (*Hascheck* je do svibnja 2010. godine konzumirao oko 6 čovjek-godina) i investitoru se moraju moći pravdati troškovi u svim fazama života projekta.

6. Kako dalje?

U ovom poglavlju ne namjeravamo razmatrati budućnost *Haschecka*, koja je manje-više predvidiva, nego puteve kojima bi trebala ići računalna leksikografija u Hrvatskoj.

Dvadeset i prvo stoljeće bit će stoljeće računalne leksikografije, što već potvrđuje *Wikipedia* na globalnoj razini. Smatramo da Hrvatska treba *Croatopædiju*, suvremenu inačicu Akademijina rječnika na internetu, upotpunjenu općom, tematskom i dvojezičnom (prijevodnom) leksikografi-

jom. Za razliku od tiskanih izdanja, internetski leksikoni moraju biti živi i stalno se dopunjavati novim sadržajima jer jezik živi i donosi nove pojmove koji traže leksikografsku obradu. Pri tome treba voditi računa da je internetsko doba vrlo ubrzano, što potvrđuje pojava engleskoga neologizma *googlezoic* kojim se opisuje internet sadašnjega doba. Pitanje je vremena kada će se engleski neologizam pretočiti u hrvatsku riječ guglzoički ili paleoguglzoički, pridjev koji dobro opisuje *Hascheck* u ukupnosti njegova trajanja.

Prva ideja koja je mogla dovesti do nečega sličnog *Croatopædiji* iznjedrila se u vrijeme dok je autor surađivao s »Novim Liberom« za vrijeme priprema *Hrvatskoga jezičnog portala* (HJP 2006). Kako je *Hascheck* već bio obradio većinu građe na kojoj se *Portal* zasnivao, razmišljalo se da se *Hascheckova* usluga inkorporira u *Portal*, s namjerom da se na taj način prikupljaju pojmovi (natuknice) koji još nisu u njemu. Ideja nije prihvaćena i tako je *Hascheck* ostao samostalna usluga, a *Portal* statični leksikon.

Za korpus *Hrvatske jezične riznice* izrijekom se navodi da se stvara kao »podloga za izradu Velikoga rječnika hrvatskoga jezika« (*http://riznica.ihjj.hr/dokumentacija/index.hr.html*). Vjerojatno slične namjere imaju i tvorci *Hrvatskoga nacionalnog korpusa*, *http://www.hnk.ffzg.hr/*. Iz leksikografskoga iskustva prikazanoga u ovome radu slijedi da treba izraditi alat sličan *Haschecku* koji bi sadržavao sve natuknice i obličnice pojmova koje je leksikon obuhvatio, te ga ponuditi javnosti na besplatnu uporabu, a u cilju prikupljanja nove građe za ažuriranje i upotpunjavanje. Kroz isti alat trebalo bi propuštati i sve nove sastavnice što se korpusu pridodaju. Na internoj razini takav bi alat bio također uporabljiv, sadržavajući ono što su leksikografi već obradili tijekom izrade leksikografskoga djela, a još nisu ponudili javnosti. Na taj bi se način u vrlo konačnome vremenu, promatrano u odnosu na vrijeme utrošeno da se *Croatopædijin* hrvatski uzor zgotovi, i uz mjerljiv rad moglo doći do natukničkog opsega Akademijina rječnika u formi koja korespondira današnjem razdoblju stvaranja. Višejezična komponenta takvoga leksikona, barem što se međunarodnih jezika tiče, mogla bi se riješiti s poveznicama na brojne javno dostupne rječnike dotičnih jezika.

7. Zaključak

Hascheck je primjer izvorno hrvatskoga jezičnotehnološkog projekta koji je rezultirao javnim sustavom opće namjene, iznimno dobro prihvaćenim od strane brojnih korisnika u zemlji i svijetu. Takvih je projekata u Hrvatskoj malo te je stoga bilo potrebno *Hascheck* iscrpno prikazati, i po-

datkovno i metodološki. *Hascheck* je, također, dobar primjer kako inovativnost i upornost mogu nadomjestiti novac. Microsoftov pravopisni provjernik za hrvatski jezik, iza kojega stoji novac i tehnološka izvrsnost jedne od najvećih svjetskih korporacija, pokazuje znatno nižu jezičnu funkcionalnost od *Hascheckove*. I to je iskustvo vrijedno isticanja jer su jezične tehnologije u Hrvatskoj još uvijek podfinancirani tehnološki sektor.

Hascheck valja promatrati u okvirima napora da se smanje hrvatski jezičnotehnološki deficiti. Al Goreova *Digitalna deklaracija međuovisnosti* (Gore 1998) postulira strojnu govornu prevodivost (engl. *speech-to-speech translation*) kao drugu po važnosti od pet globalnih informacijsko-tehnoloških zadaća našega vremena. To od hrvatskih istraživača traži da pored bavljenja strojnim prevođenjem rješavaju i izazove govornih tehnologija za hrvatski jezik. Do kvalitetnih govornih tehnologija ne možemo doći bez računalnih pravogovornih rječnika. *Hascheckov* rječnik, posebice zbog postojanja čestotnikâ za sve njegove rječničke sastavnice, dobra je podloga da se započne s postupnim razvojem hrvatskoga računalnog pravogovornog rječnika, gdje će svaka pojedina etapa imati cilj da se pravogovornim rječnikom pokrije neki udio govornoga hrvatskog jezika. Bavljenje govornim tehnologijama imat će i povratni učinak na *Hascheck*. Strojna tvorba govora na hrvatskome neminovno će se suočiti s problemom heteronima budući da je hrvatski izrazito bogat istopisnicama raznozvučnicama. Morfosintaktički algoritmi nužni za prepoznavanje izgovora heteronima upotrebljivi su i u kontekstualnoj pravopisnoj provjeri teksta. Istraživanja usmjerena prema kontekstualnome *Haschecku* već su započela pa je nužno napore objediniti radi što efikasnijega trošenja vremena i novca. Iskustva s *Hascheckom* i *HascheckVoiceom* pokazuju da bavljenje jezičnotehnološkim sustavima javne namjene proizvodi sinergijske učinke koji pridonose kvalitetnijim rezultatima istraživanja i uspješnijem ostvarivanju zadanih ciljeva. Takav je učinak dokazan u ovome radu kroz odnos *Haschecka* i njegovih korisnika.

Literatura

Arppe, Antti, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mikael Suominen, Martti Vainio, Hanna Westerlund, Anssi Yli-Jyrä (eds.). 2005. *Inquiries into Words, Constraints and Contexts: Festschrift for Kimmo Koskenniemi on his 60th Birthday*. Stanford (CA) : CSLI Publications. 330 str.

Arppe, Antti. 2006. Frequency Considerations in Morphology, Revisited — Finnish Verbs Differ, Too. A *Man of Measure. Festschrift in Honour of Fred Karlsson* (Mickael Suominen, Antti Arppe i dr., eds.). Turku : Special Supplement to SKY Journal of Linguistics, Linguistic Association of Finland, Vol. 19/2006:175—189.

Aspell. 2008. *GNU Aspell*. *http://aspell.net/*, preuzeto 19.V.2010.

Anić, Vladimir. 2003. *Veliki rječnik hrvatskoga jezika*. Zagreb : Novi Liber. 1881 str.

Babić, Stjepan, Milan Moguš. 2010. *Hrvatski pravopis*. Zagreb : Školska knjiga. 454 str.

Badurina, Lada, Ivan Marković, Krešimir Mićanović. 2007. *Hrvatski pravopis*. Zagreb : Matica hrvatska. 662 str.

Bego, Vojislav, Josip Butorac (ur.). 1993. *Josip Lončar — život i djelo.* Zagreb : Hrvatska akademija znanosti i umjetnosti — Elektrotehnički fakultet Sveučilišta u Zagrebu. 205 str.

Bratanić, Maja. 1975. Englesko-hrvatski leksikografski korpus. *Bilten Instituta za lingvistiku*, 1(1):71—73.

CLC. 2008. *Hrvatska jezična riznica*. Zagreb: Institut za hrvatski jezik i jezikoslovlje. *http://riznica.ihjj.hr/*, čestotnik preuziman od 21.VI. do 30.IX.2008.

Damerau, Fred J. 1994. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171—176.

Dembitz, Šandor. 1982. Word generator and its applicability. U: *Proceedings of the International Zurich Seminar on Digital Communications: Man-Machine Interaction*. Zürich : IEEE, 59—64.

Dembitz, Šandor. 1993a. Rezultati usporedbe nekih spelling checkera. *Acta Graphica*, 5(1):29—43.

Dembitz, Šandor. 1993b. *Automatizacija postupka otkrivanja grešaka u tekstu u novim telekomunikacijskim službama* (doktorska disertacija). Zagreb : ETF. 142 str.

Dembitz, Šandor. 1996. Udaljenost između jezika. U: *Zbornik radova SoftCOM 96*. Split : FESB, 219—226.

Dembitz, Šandor, Mladen Sokele. 1997. Usporedba hrvatskih spelling checkera. *Zbornik radova SoftCOM 97*. Split—Bari—Dubrovnik : FESB, 191—197.

Dembitz, Šandor, Petar Knežević, Mladen Sokele. 1998. Learning Words — A Cognoelectrical Analogy, *Proceedings of the 9th Artificial Intelligence/Cognitive Science Conference — AISC'98* (John Dunnion, Gregory O'Hare i dr., eds.). Dublin : UCD, 47—54.

Dembitz,Šandor, Peter Knežević, Mladen Sokele. 1999. Hascheck — The Croatian Academic Spelling Checker. *Applications and Innovations in Expert Systems VI* (Robert Milne, Ann Macintosh i dr., eds). London : Springer, 184—197.

Dembitz, Šandor, Gordan Gledec, Mirko Randić. 2009. Spellchecker. *Wiley Encyclopedia of Computer Science and Engineering* (Benjamin W. Wah, ed.). Hoboken (NJ) : John Wiley & Sons, Vol. 5, 2793—2804.

Dembitz, Šandor, Jakov Pavlek, Dejan Stupar. 2010. Problem stranih imena u strojnoj tvorbi govora na hrvatskome. *Proizvodnja i percepcija govora* (Vesna Mildner i Marko Liker, eds.). Zagreb : FF-press, 406—417.

Dolgopolov, A. S. 1986. Automatic spelling correction. *Cybernetics*, 22(3):332—339.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language**.** *Computational Linguistics***,** 27(2):153—198.

Gore, Al. 1998. Digital Declaration of Interdependence, in *Remarks from Vice President Al Gore,* 15th International ITU Plenipotentiary Conference, Minneapolis, MN, October 12, 1998, *www.itu.int/newsarchive/press/PP98/Documents/Statement_Gore.html*, preuzeto 26.V.2010.

HJP. 2006. *Hrvatski jezični portal***.** Novi Liber i SRCE. *http://hjp.srce.hr/***,** preuzeto 25.V.2010.

HML. 2005. *Hrvatski morfološki leksikon i lematizacijski poslužitelj*. Zavod za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu. *http://hml.ffzg.hr/hml/unos.php*, preuzeto 20.V.2010.

Ispell. 1996. *International Ispell*. http://www.lasr.cs.ucla.edu/geoff/ispell.html, preuzeto 19.V.2010.

ISO-3166. 2010. *English Country Names and Code Elements*. http://www.iso.org/iso/english country_names_and_code_elements, preuzeto 21.V.2010.

Jojić, Ljiljana, Ranko Matasović (ur.). 2004. *Hrvatski enciklopedijski rječnik*. II. dopunjeno izdanje u 12 svezaka. Zagreb: Novi Liber — Europapress Holding. 4.203 str. ukupno.

Karlsson, Fred. 1985. Paradigms and Word Forms. *Studia Gramatyczne VII*, Ossolineum, 135—154.

Karlsson, Fred. 1986. Frequency Considerations in Morphology. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 39(1):1986, 19—28.

Karlsson, Fred. 2000. Defectivity. *Morphology: An International Handbook on Inflection and Word Formation* (Geert Booij, Christian Lehmann i dr., eds.). Berlin & New York: Mouton de Gruyter, Vol. 17.1:647—654.

Kornai, András. 1999. Zipf's law outside the middle range. *Proceedings of the Sixth Meeting on Mathematics of Language* (James Rogers, ed.). Orlando (FL): University of Central Florida, 347—356.

Kornai, András. 2002. How many words are there? *Glottometrics*, 2002/4:61—86.

Kukich, Karen. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377—439.

Lacković, Denis. 2003. *Croatian dictionary and affix file. http://www.lasr.cs.ucla.edu/geoff/-ispell-dictionaries.html#Croatian-dicts*, preuzeto 24.I.2004.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707—710.

Lončar, Josip. 2006. *Osnove elektrotehnike — knjiga prva i druga* (pretisak 4. izdanja prve knjige iz 1956. i 4. izdanja druge knjige iz 1958.). Zagreb : Graphis. 359+347 str.

MaxMind. 2010. *Allocation of IP addresses by Country*. *http://www.maxmind.com/app/techinfo*, preuzeto 15.IV.2010.

McIlroy, M. Douglas. 1982. Development of a spelling list. *IEEE Transactions on Communications.* COM-30(1):91—99.

Meade, Nigel, Towhidul Islam. 2006. Modelling and forecasting the diffusion of innovation — A 25-year review. *International Journal of Forecasting*, 22(2006):519—545.

Morris, Robert, Lorinda L. Cherry. 1975. Computer detection of typographical errors. *IEEE Transactions on Professional Communications*, PC-18(1):54—64

Navarro, Gonzalo. 2001. NR-grep: a fast and flexible pattern matching tool. *Software: Practice and Experience*, SPE-31(13):1265—1312.

Pavlek, Jakov. 2010. *Proširenje leksičke baze mrežnog pravopisnog provjernika* (magistarski rad). Zagreb: FER. 105 str.

Peterson, James L. 1980. *Computer Programs for Spelling Corrections: An Experiment in Program Design*; Lecture Notes in Computer Science, Vol. 96. Berlin : Springer. 213 str.

Silić, Josip, Vladimir Anić. 2001. *Pravopis hrvatskoga jezika*. Zagreb : Školska knjiga-Novi Liber. 970 str.

Sokele, Mladen. 2008. Growth models for the forecasting of new product market adoption. *Telektronikk*, 3/4(2008):144—154.

Sokele, Mladen. 2009. *Analytical Method for Forecasting of Telecommunications Service Life-Cycle Quantitative Factors* (doctoral thesis). Zagreb : FER. 130 str.

Streiter, Oliver, Kevin P. Scannell, Mathias Stuflesser. 2006. Implementing NLP projects for noncentral languages: instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4):267—289.

TEMAA Project. 1997. *A model for NLP evaluation: Spelling checkers*. http://www.cst.dk/temaa/ D16/d16exp-3.html#Heading14, preuzeto 18.V.2010.

Thompson, Henry. 1992. The strategic role of evaluation in natural language processing and speech technology. *Record of the ESPRIT/DANDI/ELSNET/HCRC Workshop*. Edinburgh : Human Communication Research Centre, University of Edinburgh.

Thompson, Henry. 1994. TEMAA : A testbed study of evaluation methodologies: Authoring aids. *Proceedings of the Language Engineering Convention*. Paris : ELSNET, 147—148.

Trón, Viktor, György Gyepesi, Péter Halácsy, András Kornai, László Németh, Dániel Varga. 2005. Hunmorph: open source word analysis. *Proceedings of the ACL 2005 Workshop on Software*. Ann Arbor (MI) : ACL, 77—85.

Turba, Thomas N. 1981. Checking for spelling and typographical errors in computer-based text. *ACM SIGPLAN Notices*, 16(6):51—60.

Vujić, Antun (ur.). 1996. *Hrvatski leksikon*. I. svezak. Zagreb : Naklada Leksikon. 669 str.

Vujić, Antun (ur.). 1997. *Hrvatski leksikon*. II. svezak. Zagreb : Naklada Leksikon. 738 str.

Functional Lexicography of an Online Spellchecker

Abstract

Online spellchecking offers a unique possibility of permanent improving of spellchecker linguistic functionality through an interaction with the community of spellchecker users. Such a possibility is crucial for spellchecking in NLP no*n*-central languages, like Croatian, in order to overcome gaps in natural language processing (NLP) tools between them and NLP central languages (English, Japanese, German, French, Russian, Mandarin Chinese etc.). The possibility will be discussed based on *Hascheck* example. *Hascheck* started as the first Croatian public spellchecker, operating with a very modest dictionary of 100,000 Croatian common word-types. Due to the learning the dictionary increased to 830,000 common word-types and 600,000 name-types, acronyms, abbreviations etc. It is a result of processing of a corpus which amounts to 260 millions tokens. *Hascheck's* corpus is the biggest corpus ever processed in Croatia with a lexicographic aim. All those happened because of Learning System incorporated into spellchecker software environment, which converts individual user language competence into collective value. The Learning System is highly automated, but its results do not enter into *Hascheck's* dictionary without human supervision. The supervision is needed because of precision reasons. The supervision takes a special care about potentially valid words which might be close to frequent or potentially frequent misspellings or typos. Abundance of collected data allows mathematical modeling of many aspects of *Hascheck's* life, which are also presented in the paper.

Ključne riječi: pravopisni provjernik, korpus, indeks učenja, pokrivanje teksta, Heapsov zakon

Key words: spellchecker, language corpus, learning index, text coverage, Heaps' law