Hascheck neprestano u naponu snage

Hrvatski akademski pravopisni provjernik, ili *Hascheck*, važan je kotač u razvoju hrvatskih prirodnojezičnih tehnologija

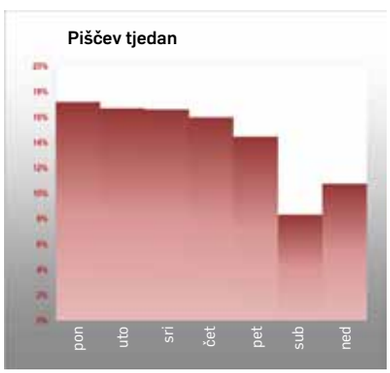
Hrvatski akademski *spelling checker*, poznatiji pod akronimom *Hascheck*, jedna je od najstarijih internetskih usluga u Hrvatskoj. Zaživjela je početkom 1993. godine u lokalnoj mreži Elektrotehničkoga fakulteta Sveučilišta u Zagrebu (danas Fakultet elektrotehnike i računarstva - FER), da bi 21. ožujka 1994. godine postala javnom uslugom strojne provjere teksta pisanoga hrvatskim jezikom.

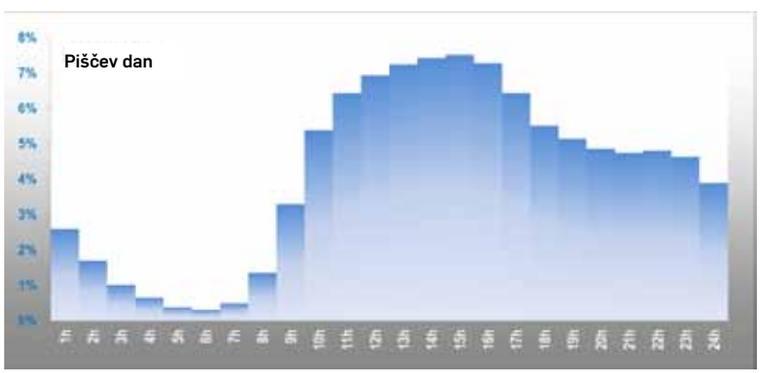
U dvadeset godina djelovanja *Hascheck* je obradio osam milijuna tekstova, koji tvore korpus od dvije milijarde pojavnica. Za razliku od konvencionalnih pravopisnih provjernika, *Hascheck* je ekspertni sustav, jer stalno uči nove riječi iz tekstova svojih korisnika. Učenje je visoko automatizirano, no ono podliježe ljudskome nadzoru radi očuvanja visoke preciznosti pravopisnoga rječnika. Rječnik je na početku brojao oko 100.000 različnica, ali je tijekom dvadesetogodišnjega učenja narastao na više od dva milijuna različnica, od kojih polovica pripada hrvatskom općejezičnom fondu, dok drugu polovicu tvore posebnojezične, dominantno imenske različnice.

Za razliku od konvencionalnih pravopisnih provjernika, *Hascheck* je ekspertni sustav, jer stalno uči nove riječi iz tekstova svojih korisnika. Učenje je visoko automatizirano, no ono podliježe ljudskome nadzoru radi očuvanja visoke preciznosti pravopisnoga rječnika

***Hascheck* je do sada** opslužio blizu pola milijuna korisnika u zemlji i svijetu. Ukupno 88 posto prometa inicirano je iz HR-domene, koja trenutačno broji oko 2,3 milijuna IP-adresa, od kojih je 560.000 zabilježeno kao izvorište domaćega prometa, dok ostatak tvori inozemni promet s izvorištima u 123 IP-domene diljem svijeta. Pored Hrvatske, *Hascheck* se učestalo koristi u Bosni i Hercegovini, Sjedinjenim Američkim Državama, Njemačkoj, Ujedinjenome Kraljevstvu i Crnoj Gori. Svaka od navedenih zemalja sudjeluje u ukupnome prometu s korpusom većim od 10 milijuna pojavnica.

Kako se zbog učenja funkcionalnost sustava stalno poboljšava, tako raste i intenzitet njegova korištenja. U prva četiri mjeseca 2014. godine dnevno se u prosjeku obrađivalo 7.000 tekstova, ili korpus od 1,8 milijuna pojavnica. *Haschecko*v se posao može uspoređivati s lektorskim, s tim da se godišnja norma za lektoriranje kreće u korpusnom rasponu od 600 do 900 tisuća pojavnica. Dakle, *Hascheck* danas dnevno odradi dvije do tri godišnje lektorske norme, radeći i svetkom i petkom, i danju i noću, i ne tražeći nikakvu naknadu za svoj rad.

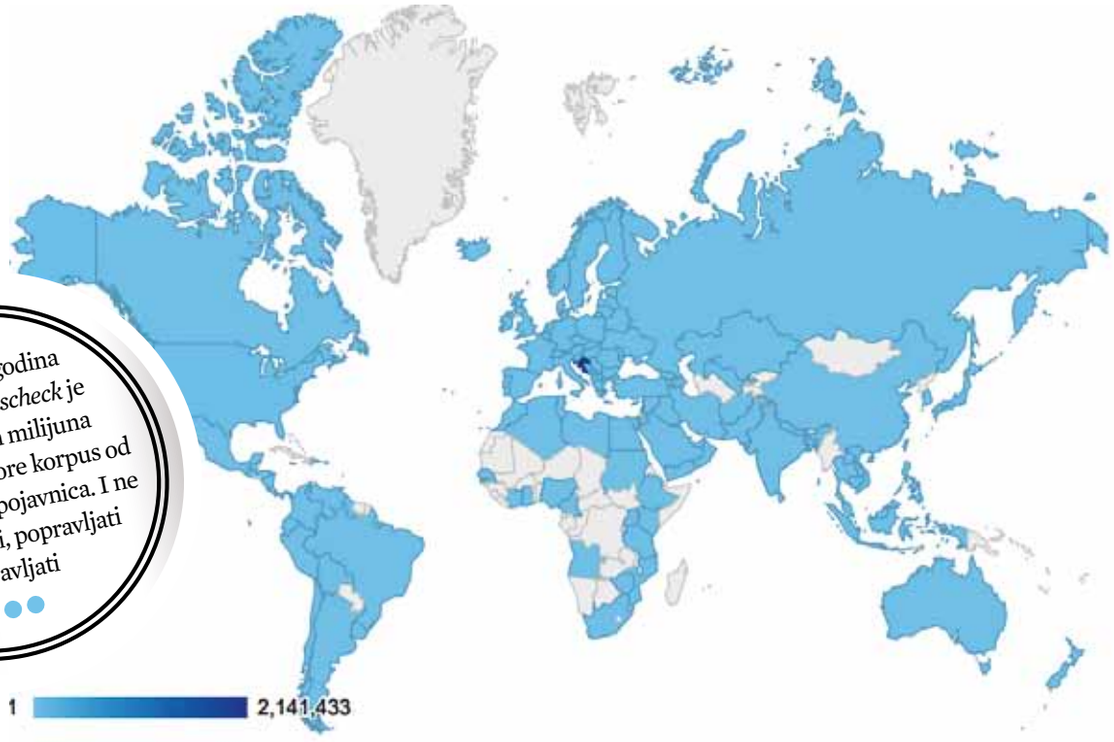




Hascheck danas dnevno odradi dvije do tri godišnje lektorske norme, radeći i svetkom i petkom, i danju i noću, i ne tražeći nikakvu naknadu za svoj rad

Što *Hascheck* znači onima kojima je stalo da im tekstovi budu dobro i brzo uređeni, lapidarno iskazuju riječi Sande Ham, sveučilišne profesorice kroatistike i suautorice jednog od hrvatskih pravopisâ. Ona je održavateljima usluge, nakon obrade vrlo zahtjevnoga teksta opsega 60.000 pojavnica, poručila sljedeće: *Hascheck je više nego koristan! Ono što bih ja radila danima, bilo je gotovo za desetak minuta i to još s mojom dodatnom provjerom!* U iskazu nema nimalo pretjerivanja, jer samo za pročitati tekst navedenoga opsega sa slovkanjem, radi otkrivanja pogrešaka koje neminovno prate pisanje, treba barem tridesetak sati rada. Stoga nije ni pretjerivanje kada se kaže da se dosadašnji doprinos *Haschecka* porastu produktivnosti u proizvodnji tekstova pisanih hrvatskim jezikom mjeri stotinama tisuća, ako već ne i milijunima uštođenih radnih sati.

U dvadeset godina djelovanja *Hascheck* je obradio osam milijuna tekstova, koji tvore korpus od dvije milijarde pojavnica. I ne prestaje učiti, popravljati i ispravljati



***Hascheck* nikada** nije bio sâm sebi svrhom. Recentni razvoj prirodnojezičnih tehnologija u svijetu obilježila je krilatica: *Više podataka, manje lingvističkog označavanja!* Rezultat novoga trenda jest pojava leksičkih n-gramskih baza podataka s milijardama zapisa, do sada izvedenih za desetak svjetskih jezika. Dobivene *big data* infrastrukture namijenjene su sofisticiranom jezičnom modeliranju, sa svrhom poboljšavanja performansi postojećih i razvoja novih jezičnotehnoloških aplikacija. U svibnju 2007. godine, niti godinu dana nakon što je *Google* objavio postojanje n-gramskog sustava za engleski jezik, prve leksičke *big data* infrastrukture izvedene iz korpusa opsega 1025 milijardi pojavnica (1025 Gpojavnica), započelo je, s osloncem na *Haschecku*, kreiranje usporedive infrastrukture za hrvatski jezik.

Engleski, kineski i hrvatski n-gramski sustav u osnovi su vrlo slične *big data* infrastrukture. Brojnost n-grama u sva tri sustava ravna se po gotovo identičnom zakonu (korelacijski koeficijenti veći su od 0,995 u svim usporedbama), a nagibi pravaca linearne regresije vrlo su blizu omjeru opsegâ sustava. Očito je da je hrvatski, zahvaljujući *Haschecku*, dobio ozbiljnu podatkovnu podlogu za poticanje razvoja niza novih jezičnotehnoloških aplikacija, ciljanih prema smanjivanju deficitâ koji ga danas u ovome području karakteriziraju u odnosu na brojne europske jezike.

Hrvatski n-gramski sustav iskorišten je da se *Hascheck* pretvori u kontekstualni provjernik. Već je dosegnut omjer 4:1 u prijavljivanju pravopisno-zatipkovnih i gramatičkih, odnosno stilskih pogrešaka, što *Haschecka* svrstava u red vrhunskih proizvoda ove vrste u svijetu. Jezični modeli izvedeni iz hrvatskih n-grama poslužili su za razvoj uporabljivih sustava za strojnu tvorbu, odnosno strojno prepoznavanje hrvatskoga govora, dok je u tijeku razvoj sustava za visokokvalitetno strojno prevođenje s hrvatskoga na francuski i obrnuto. Međutim, potencijali n-gramskog sustava nisu ni izdaleka ovim iscrpljeni.

▲ 88 posto prometa inicirano je iz HR-domene, koja trenutačno broji oko 2,3 milijuna adresa. Njih 560.000 zabilježeno je kao izvorište domaćega prometa. Ostatak tvori inozemni promet s izvorištima u 123 IP-domene diljem svijeta

Ono što je u proteklih sedam godina napravljeno na razvoju hrvatske n-gramske infrastrukture prikazano je putem usporedbe s najvećima ▼

Razmjena znanja

'Ferovci' dijele novostvorenu vrijednost

Svi zainteresirani za korištenje hrvatskog n-gramskog sustava u istraživačko-razvojne svrhe pozivaju se da se obrate na adrese sandor.dembitz@fer.hr i gordan.gledec@fer.hr. FER ne namjerava čuvati novostvorenu vrijednost, relevantnu za očuvanje identiteta Hrvata i Hrvatske u internetskoj eri, samo za vlastite uporabe.

Usporedba hrvatskog n-gramskog sustava s dva najveća svjetska n-gramska sustava

	Engleski 1025 Gpojavnica	Kineski 883 Gpojavnica	Hrvatski 1,9 Gpojavnica
1-grama	13 588 391	1 616 150	2 814 198
2-grama	314 843 401	281 107 315	103 786 725
3-grama	977 069 902	1 024 642 142	298 840 686
4-grama	1 313 818 354	1 348 990 533	401 083 297
5-grama	1 176 470 663	1 256 043 325	393 462 025
Ukupno	**3 795 790 711**	**3 912 399 465**	**1 199 986 931**

PROMO

E-adresar http://hascheck.tel.fer.hr; http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison;