View metadata, citation and similar papers at core.ac.uk
brought to you by CORE

Iza profesora Rosandića ostat će njegova djela u kojima je riječju, ali i vlastitim primjerom pokazao kako poučavati, proučavati, učiti i voljeti hrvatski jezik i književnost.[1] Ogleda se to i u njegovim stihovima Učitelju stvaraocu:

„Tebi koji svakoga dana za radnim stolom stvaraš sutrašnji dan
u učionici i nestrpljivo ga očekuješ.
...
Tebi koji si osvjedočen da radiš najljudskiji posao,
tebi koji u sebi nosiš stvaralački zanos i nemir.
Neka tvoje djelo bude uzor drugima i trajno nadahnuće
onima koji dolaze i koji će doći koračati stazama
kojima si i ti koračao.
Neka tvoje djelo bude iskra iz koje će nastajati nova svjetlost.“

Počivajte u miru, dragi profesore!

Dunja Pavličević-Franić

[1] Kronološka bibliografija profesora Rosandića objavljena je 2010. u časopisu Metodika (god. 11., br. 2., str. 264. – 312.) u povodu njegovih obljetnica – 80. obljetnice života i 60. obljetnice pisanog stvaralaštva.

25 GODINA HAŠEKA

Šandor Dembitz

Uvod

Ime iz naslova čitatelja vjerojatno najprije podsjeća na Dobrog vojaka Švejka a ponekog, možda, i na Ljudevita Jonkea, prvog urednika Jezika, prevoditelja romana na hrvatski. Za razliku od Čeha Jaroslava Hašeka, koji je svoju svjetski poznatu satiru pisao tijekom i nakon Velikoga rata, hrvatski je vojnik Švejk – pridjev „dobar“ namjerno je izostavljen – svoj Hašek počeo pisati tijekom Domovinskoga rata te ga i dandanas dopisuje.

Hašek je pohrvaćeni oblik akronima *Hascheck*, izvedenog iz naziva Hrvatski akademski *spelling checker*, i označava jezgrenu sastavnicu mrežnoga pravopisnog provjernika koji u različitim oblicima, danas na adresi https://ispravi.me/, od 21. ožujka 1994. stoji na raspolaganju svima koji žele da im se tekst prije objavljivanja strojno provjeri.

Danas, u guglzoiku, *spellchecking* nije posebno atraktivno područje prirodnojezičnih tehnologija, što u domaćim okvirima potvrđuje spominjanje Hašeka u knjizi Hrvatski jezik u digitalnom dobu, u kojoj mu je posvećena jedna jedina rečenica: „*On-line Hrvatski akademski spelling checker* (Hascheck) postoji od 1994. i još je uvijek u uporabi.“[1] U citiranoj se monografiji njezini autori, svi odreda barem jednom izabrani za člana-suradnika HAZU-a, iscrpno bave temama danas opredmećenima u *Google Translateu* ili *Google Dictateu* itd. Jedino im je promakla činjenica da je Hašek davna hrvatska anticipacija istih, ali što se tu može.

Čemu uopće *on-line spellchecking*? U paleoguglzoiku, dok su se Amerikanci još intenzivno bavili pravopisnim provjernicima, o problemu je napisano i ovo

> „Recept za izradu gulaša od slona započinje s: prvo ulovi slona. Ako vaš recept za izradu pravopisnog provjernika započinje s: prvo pronađi sve valjane riječi-različnice u engleskom jeziku, vjerojatno ćete brzo uvidjeti da je puno lakše napraviti ukusni gulaš od slona.“[2]

Lako je predočiv američki lovac, opremljen puškom za uspavljivanje, kako lovi svoga slona. Što da radi njegov hrvatski parnjak, oboružan kamenom sjekirom, ako slučajno uspije ošamutiti svoga mamuta? „Na internet s njime, jer inače gulaša nema!“ Da je to paleolitičko razmišljanje bilo ispravno, potvrđuje činjenica da danas, osim Microsoftova pravopisnoga provjernika za hrvatski, korisnicima hrvatskoga u stvarnosti za te svrhe još jedino Hašek stoji na raspolaganju. Prije dvadesetak godina konvencionalnih hrvatskih pravopisnih provjernika bilo je na lopate bacati,[3] ali nisu preživjeli. Međunarodnim veletvrtkama šaka jada ne može konkurirati po modelu: „vidjela žaba kako potkivaju konja pa i sama digla nogu“. Za takve izazove ipak treba malo soli u glavi. Da je izazivač strancima na koncu pokazao tko je tko na domaćem bunjištu, potvrđuje i nedavna usporedba.[4]

Što je napravljeno?

Kako je Hašek nastao, čemu sve služi, kako radi i još puno toga zainteresirani čitatelj Jezika može pronaći u Kolu[5] i Filologiji.[6] Stoga će ovdje ukratko biti prikazano samo ono što je u 25 godina napravljeno, a da ima neku vrijednost.

---

[1] Tadić, M., Brozović-Rončević, D., Kapetanović, A., 2012., Hrvatski jezik u digitalnom dobu, META-NET White Paper Series, Springer, str. 26., dostupno na http://www.meta-net.eu/whitepapers/e-book/croatian.pdf

[2] Bentley, J., 1985., A Spelling Checker, Communications of the ACM, god. 28., br. 5., str. 460.

[3] Sokele, M., 1997., Hrvatski spelling-checkeri, WIN.INI, 3./97., str. 38. – 49.

[4] Pilić, L., 2017., Jezične tehnologije, VIDI, br. 252., str. 100. – 105.

[5] Dembitz, Š., 2017., Strojna obrada hrvatskog jezika – mađarski doprinosi, Kolo, 4. / 2017., str. 108. – 122.; dostupno na http://www.matica.hr/kolo/539/strojna-obrada-hrvatskog-jezika-maarski-doprinosi-27748/

[6] Dembitz, Š., 2012., Funkcionalna leksikografija mrežnoga pravopisnog provjernika, Filologija, god. 58., str. 55. – 98.

Hašekov je rječnik od početnih 100 000 različnica hrvatskog općejezičnog fonda u 25 godina strogo nadziranog učenja, nadziranog radi očuvanja preciznosti rječnika, narastao:

- na 1 051 189 različnica hrvatskog općejezičnog fonda;
- na 957 620 različnica hrvatskog posebnojezičnog, pretežito imenskog fonda;
- na 70 528 različnica engleskog općejezičnog fonda, u kojemu nema onih riječi koje se jednako pišu u engleskome i hrvatskome, npr. atom ili zebra.

Engleski leksik uključen je u Hašekov rječnik jer je engleski jezik današnja *lingua franca*. Čak se i u Hrvatskoj jezičnoj riznici,[7] stomilijunskom dijakronijskom korpusu sa stoljetnim rasponom tekstova, koji su sastavili kroatisti, potvrđuje 13 175 različnica iz engleskog dijela Hašekova rječnika (najučestaliji je određeni član *the* s ukupno 7 988 pojavljivanja), koje tvore 0,4 % cjelovitoga korpusa Riznice. Uzimajući u obzir i ukošene oblike engleskih riječi tipa *rolla*, *rollu* itd., udio engleštine u Riznici penje se do 0,8 %, što odgovara razini zatipkovno-pravopisnih pogrješaka u njoj. Inače, Hašekov bi rječnik, kada bi ga tko želio tiskati, tražio najmanje 3 standardna leksikografska sveska.

U 25 godina usluzi je pristupljeno s 1.368.702 IP-adrese iz 177 vršnih internetskih domena, pretežito zemalja. Prikaz opsega pružene usluge po vršnim domenama dan je u Dodatku ovomu radu. Prema evidenciji HTTP kolačića, tj. tragu koji svaki korisnik ostavlja za sobom nakon obavljene obrade, uslugu je koristilo oko milijun osoba. U Tablici 1. prikazana je ukupnost 25-godišnjeg Hašekovog usluživanja najvažnijih vršnih domena s nekoliko bitnih parametara.

Izvorišta prometa	Obrađeni korpus [pojavnica]	Udio po izvorištima [%]	Prosječno prekrivanje korpusa rječnikom [%]	Prosječni udio zatipkovno-pravopisnih pogrješaka u korpusu [%]
Hrvatska	6 313 123 913	87,26	98,47	1,50
BiH	460 404 455	6,36	97,17	2,81
Srbija[8]	58 941 003	0,81	97,31	2,67
Njemačka	58 714 427	0,81	98,13	1,83
SAD	54 830 162	0,76	98,67	1,31
Ostala	289 082 052	4,00	97,68	2,29
Ukupno	7 235 096 012	100,00	98,34	1,62

Tablica 1.

[7] Dostupno na http://riznica.ihjj.hr/index.hr.html

[8] Uključuje i promet pokrenut iz Republike Kosovo. Premda je po ISO-3166-1 standardu Kosovu već dodijeljena vršna domena KO (vidjeti https://hr.wikipedia.org/wiki/ISO_3166-1), razdvajanje vršnih domena Kosova i Srbije još nije obavljeno.

Obrađeni korpus od 7,2 Gpojavnica (gigapojavnica) odgovara korpusu od 30 milijuna autorskih kartica teksta i šest je puta veći od „najvećeg hrvatskog korpusa hrWaC“, kojim se diči uvodno citirana monografija.[9] To je samo još jedna potvrda da kod malih primjereno osmišljeni pristupi znaju polučiti bolje rezultate od nekritičkog slijeđenja velikih po žabljem modelu.

Ono što zabrinjava jest podatak koji upućuje da se hrvatski urednije piše u SAD-u negoli u samoj Hrvatskoj (posljednji stupac Tablica 1.), ali to je pitanje kojim bi se morale pozabaviti hrvatske obrazovne vlasti. Poziv se opravdava činjenicom da su unatrag nekoliko posljednjih godina one bile vrlo izdašne u dodjeljivanju nagrade „Ivan Filipović“ za značajna ostvarenja u odgojno-obrazovnoj djelatnosti hrvatskim normativistima,[10] kojima je zadaća hrvatske učenike uputiti kako treba uredno pisati na hrvatskom jeziku. Nas sretnima čine priznanja sljedeće vrste:

> *Poštovani, pohvala za vašu stranicu https://ispravi.me/! Nisam izvorna govornica hrvatskog jezika i teško mi pada pohvatati sve gramatičke cake. Vaša stranica mi daje samopouzdanja jer učim pri svakom pisanju. Hvala puno i samo naprijed! Lp, Tena*[11]

Hašek je odavno prestao biti konvencionalni pravopisni provjernik. Ispravljanje gramatičkih pogrješaka započelo je mijenjanjem nepostojećeg glagolskog priloga prošlog, primjerice „slijedivši“, u valjani glagolski prilog sadašnji, tj. „slijedeći“, i obrnuto, „proslijedeći“ u „proslijedivši“. Čak ni pismeni korisnici hrvatskoga nisu više sasvim sigurni, vjerojatno zbog gubitka aorista, odnosno imperfekta u svakodnevnoj uporabi, koji su hrvatski glagoli svršeni, a koji nesvršeni. Bavljenje „nekonvencionalnim pogrješkama“ nastavljeno je stvaranjem hrvatskog n-gramskog sustava, koji je omogućio da se kontekstno prepoznaju, po potrebi i isprave, učestale gramatičke i stilske pogrješke u pisanju na hrvatskome.

Skupljanje i uređivanje hrvatskih n-grama započelo je, potaknuto projektom *Google Translate*,[12] sredinom 2007. godine. N-gramski je sustav nužna podatkovna podloga za suočavanje s izazovima kao što su strojno prevođenje, strojna pretvorba govora u tekst itd. U Tablica 2. nalazi se usporedni prikaz hrvatskoga s dva najveća Googleova n-gramska sustava s početka rečenoga projekta.

---

[9] Vidjeti prvu referenciju, str. 35.

[10] Vidjeti http://ihjj.hr/stranica/nagrade-i-priznanja/25/

[11] Citiranu je poruku 27. siječnja 2019. Hašeku (hascheck@fer.hr) uputila Tena Ćorić, osoba rođena i odrasla u Švicarskoj.

[12] Vidjeti https://en.wikipedia.org/wiki/Google_Translate

	Engleski[13] WaC 1,025 Tpojavnica	Kineski[14] WaC 883 Gpojavnica	Hrvatski Hašekov korpus 7,2 Gpojavnica
1-grami	13 588 391	1 616 150	5 757 442
2-grami	314 843 401	281 107 315	265 171 603
3-grami	977 069 902	1 024 642 142	918 083 221
4-grami	1 313 818 354	1 348 990 533	1 390 001 665
5-grami	1 176 470 663	1 256 043 325	1 463 796 046
Ukupno	3 795 790 711	3 912 399 465	4 042 809 977

Tablica 2.

Google se poslužio cjelokupnim WWW kao tekstovnim repozitorijem, odnosno tzv. pristup *Web as Corpus* (WaC) – isti je poslužio i za dobivanje spomenutoga „najvećeg hrvatskog korpusa" – i čestotnost n-grama, ili tzv. *cut-off* kriterij, da bi dobio gore prikazane sustave. To u hrvatskom slučaju ne može voditi do usporedivih rezultata, ali do usporedivih se rezultata dolazi ako se iskoriste Hašekove obrade i leksičnost kao kriterij za uvrštavanje n-grama u bazu, tj. da su konstituenti svih n-grama riječi s potvrdom u Hašekovom rječniku. Valja napomenuti da preko 50 % unigrama u hrvatskom slučaju tvore različnice – brojevi, no već s $n \geq 2$ udio n-grama s takvim konstituentima pada ispod 2 %.

Hašekov 25-godišnji društveni doprinos može se sažeti u sljedećem:

1. Uštađeno je oko 10 000 radnih godina sričućega čitanja, koje bi se bez usluge potrošile radi otkrivanja i otklanjanja pogrješaka, neizostavnih pratiteljica nastajanja novoga teksta.
2. Stvoren je hrvatski n-gramski sustav, podatkovna podloga nužna za uspješno suočavanje s izazovima koji stoje pred hrvatskim jezičnim tehnolozima, čiji je opseg veći od opsega svih knjiga koje su od Gutenberga do danas tiskane na hrvatskom jeziku.

Kako je usluga https://ispravi.me/ zapravo predlektoriranje, osmišljena da bi se uređivaču teksta olakšao i skratio najnekreativniji, a vrlo zamorni dio posla, izračun prvoga doprinosa polazi od:

- davna lektorska norma kretala se između 10 i 20 autorskih kartica teksta dnevno;
- radna godina prema europskom standardu broji 1.720 radnih sati, odnosno 215 radnih dana.

Hašek je obradio 30 000 000 autorskih kartica teksta, pa računajte.

[13] Vidjeti https://catalog.ldc.upenn.edu/LDC2006T13
[14] Vidjeti https://catalog.ldc.upenn.edu/LDC2010T06

Opseg korpusa svih knjiga tiskanih od Gutenberga do 2010. godine broji 18,2 Tpojavnica[15], iz čega slijedi procjena da sve knjige ikada tiskane na hrvatskome tvore korpus čiji opseg ne premašuje 20 Gpojavnica. Opseg hrvatskoga n-gramskog sustava, mjeren pojavnicama, računa se iz podataka posljednjega stupca Tablice 2. na sljedeći način:

$$\sum_{i=1}^{5}(broj\_i\_grama) \cdot (i+1) = 20{,}2\ Gpojavnica$$

i na tome se temelji navedena veličina drugoga doprinosa.

Hašek je opstao zahvaljujući uplatama manje od jednog promila njegovih korisnika, koji ga rabe ili su ga rabili u profesionalne svrhe. Skrb o usluzi počiva na leđima aktualnoga dekana FER-a i njegovog umirovljenika, čije je zdravlje dobrano narušeno. Srećom, obojica još dišu.

Što nije napravljeno?

Vijest o postojanju hrvatskoga n-gramskog sustava potaknula je Francuze, koji rade na sustavu Ariane,[16] da predlože da se njihov francusko-ruski par, razvijan od vremena kada je Francuska pod de Gaulleom napustila NATO, metodom samonadopunjavanja (engl. bootstrapping) pretvori u francusko-hrvatski par za strojno prevođenje. Prijedlog je djelovao zdravo, jer je nudio mogućnost da se u razumnom roku s malim ulaganjima dođe do visokokvalitetnoga sustava za strojno prevođenje s francuskoga na hrvatski i obrnuto. O kakvoj se kvaliteti prevođenja razmišljalo dovoljno govori podatak da je za *benchmarking*, tj. usporedbu pokazatelja kakvoće prevođenja, odabran Saint-Exupéryjev *Le Petit Prince*, kod nas davno preveden od strane jedne Splićanke kao Mali princ, potom u izdanju iz 2011. preimenovan u Malog kraljevića. Međutim, od zamisli se nije daleko stiglo, jer ni tražena sredstva za pokrivanje materijalnih troškova projekta nisu odobrena. Zašto?

Hrvatska politika, bilo koje vrste, nikada nije ozbiljno shvaćala Digitalnu deklaraciju međuovisnosti,[17] političku najavu guglzoika napisanu od strane osobe koja je dobila Nobelovu nagradu za mir 2007. godine. Posebno je njezinu drugu točku:

> „Moramo prevladati naše jezične barijere razvijajući stvarnovremenske sustave za strojno govorno prevođenje, tako da svatko na svijetu može razgovarati s bilo kim drugim“

ona doživljavala kao *science fiction*. Izravni dokazi s početka guglzoika za potkrjepu ove tvrdnje trebali bi se nalaziti u arhivima MZO-a, HAZU-a i IHJJ-a. Nešto svježiji, premda neizravni dokaz slijedi:

---

[15] Michel, J.-B., et al., 2011., Quantitative Analysis of Culture Using Millions of Digitized Books, Science, god. 331., br. 6014., str. 176. – 182.

[16] Dostupno na https://www.liglab.fr/fr/la-recherche/plates-formes-du-lig/plate-forme-pour-le-traitement-automatique-des-langues

[17] Gore, A., 1998., Digital Declaration of Interdependence, Remarks from Vice President Al Gore, 15th International Plenipotentiary ITU Conference, Minneapolis, MN, USA; October 12, dostupno na https://www.itu.int/newsarchive/press/PP98/Documents/Statement_Gore.html

- iz adresnih raspona Hrvatskoga sabora (IP-adrese 194.152.219.0 – 194.152.219.255, odnosno 195.29.174.0 – 195.29.175.255) u 25 godina obrađena su 2 872 teksta koji su tvorili korpus od 864.479 pojavnica, od čega je 99,94 % prometa ostvareno u posljednjih 15 mjeseci, od početka 2018. do konca ožujka 2019.;
- iz adresnoga raspona Europskog parlamenta (IP-adrese 136.173.0.0 – 136.173.255.255) Hašek je od početka 2013. do konca ožujka 2019. zaprimio na obradu 14 522 teksta koji su tvorili korpus od 2 122 054 pojavnice, s manje-više jednolikom razdiobom prometa u vremenu.

Dostatno.

U govornotehnološkom dijelu (strojna tvorba govora, odnosno strojno pretvaranje govora u tekst) jednostavnija rješenja (strojna tvorba govora, upravljanje govorom) na hrvatskom tržištu nude slovenske i srpske tvrtke, jer hrvatskih tvrtki, koje bi im konkurirale, jednostavno nema. No, pravo vrhnje u ovom području bere Newton Technologies Adria, lokalna podružnica češke tvrtke, koja je nedavno Ministarstvu pravosuđa RH prodala sustav za pretvorbu kontinuiranoga govora u tekst „s pripadajućim specijaliziranim uređajima za diktiranje za 800 korisnika“ za 33,5 milijuna kuna.[18] Uzalud svi prijedlozi davno upućeni Hrvatskoj zakladi za znanost da je nastupilo vrijeme za pokretanje projekata usmjerenih prema razvoju hrvatskih govornotehnoloških proizvoda. Uzalud dokazivanja da se uporabljivi prototipovi sustava, kako za strojnu tvorbu govora,[19] tako i za pretvaranje kontinuiranoga govora u tekst,[20] dadu brzo napraviti, i to bez ikakvih financijskih ulaganja, samo temeljeno na dobrim domaćim podatkovnim podlogama i radu ne doktoranada, već diplomanata. Izgleda da je u Hrvatskoj isplativije sufinancirati tuđi nego poticati vlastiti tehnološki razvoj, čak i kada je u pitanju jezik bez kojega bi Hrvatska bila tek zemljopisna odrednica. Valja napomenuti da su prije 25 godina Česi i Hrvati dijelili istu razinu razvijenosti prirodnojezičnih tehnologija.[21]

Zaključak

Prije 150 godina pokrenuta je izrada tzv. Akademijina rječnika, grandioznoga projekta koji je trajao više od 100 godina, da bi se pokazalo kako je hrvatski rav-

[18] Dostupno na https://pravosudje.gov.hr/vijesti/potpisan-ugovor-o-nabavi-programskog-rjesenja-za-pretvaranje-govora-u-tekst-s-pripadajucim-specijaliziranim-uredjajima/19861 , nadnevak vijesti je 18.07.2018.

[19] Šoić, R., 2010., Sinteza hrvatskog govora uporabom sustava Festival, diplomski rad br. 74., FER, Zagreb.

[20] Bajo, D., Turković, D., Dembitz, Š., 2014., Rapid Prototyping of a Croatian Large Vocabulary Continuous Speech Recognition System, Proceedings of the IARIA, str. 13. – 18., Curran Associates, Red Hook, NY.

[21] Dembitz, Š., 1993., Automatizacija postupka otkrivanja pogrješaka u tekstu u novim telekomunikacijskim službama, doktorska disertacija, ETF-Zagreb, str. 5.

nopravan svim drugim europskim jezicima. U današnjoj su Europi svi jezici nazivno ravnopravni, no u stvarnosti su neki nešto ravnopravniji, kao u onoj poznatoj životinjskoj farmi. Za male narode, njihovu kulturu i identitet, nužno je stoga da u 21. stoljeću izbore, i putem jezičnih tehnologija, svoje mjesto pod suncem ravnopravnosti. Malo je područja nad kojima danas mali narod može iskazivati potpuni suverenitet kao što je to njegov jezik.

Jasno je da se od suvereniteta uvijek može odustajati, ako za to postoje valjani razlozi. Takva odustajanja imaju svoju cijenu i u pravilu počivaju na političkim procjenama. O cijenama je ovdje bilo nešto riječi, a za političke procjene Hašekov autor nije mjerodavan. Može samo izraziti svoju bojazan da će se hrvatskom jeziku do kraja 21. stoljeća nametnuti status *Küchensprachea* odustanu li Hrvati od razvoja jezičnih tehnologija za vlastiti jezik. Ovaj rad upućuje da je takav scenarij, na autorovu veliku žalost, danas već na djelu. Čemu su se onda Strossmayer i toliki nakon njega uopće trudili, neki i ginuli?

DODATAK

Prikaz opsega pružene usluge po vršnim domenama

Budući da su nazivi vršnih domena uzeti iz američke baze, prikaz je pisan engleskim pravopisom.

	IP-domains (countries)	#IP-addresses	#Texts	Corpus [tokens]
1.	Afghanistan	14	128	10,907
2.	Albania	665	3,808	652,605
3.	Algeria	20	40	7,319
4.	Andorra	6	22	5,172
5.	Angola	2	5	194
6.	Anonymous Proxy	20	1,646	330,606
7.	Argentina	104	492	168,571
8.	Armenia	7	41	13,557
9.	Asia/Pacific Region	11	67	13,578
10.	Australia	738	7,590	1,869,227
11.	Austria	7,019	129,741	25,148,812
12.	Azerbaijan	13	26	2,868
13.	Bahrain	4	9	279
14.	Bangladesh	7	18	14,873
15.	Barbados	5	40	2,865
16.	Belarus	32	78	24,734
17.	Belgium	1,608	25,464	5,409,281
18.	Belize	7	292	41,935
19.	Bermuda	1	1	41
20.	Bolivia	10	98	47,783

21.	Bosnia and Herzegovina	108,122	1,491,045	460,404,455
22.	Botswana	1	15	10,887
23.	Bouvet Island	1	7	42,037
24.	Brazil	212	975	196,390
25.	British Virgin Islands	3	13	2,784
26.	Brunei	1	1	928
27.	Bulgaria	306	12,359	1,272,561
28.	Burkina Faso	1	1	19
29.	Burundi	3	16	695
30.	Cambodia	115	695	91,950
31.	Cameroon	14	22	83,891
32.	Canada	1,190	43,247	9,996,040
33.	Cape Verde	2	10	63
34.	Chile	58	309	124,996
35.	China	371	5,498	1,344,131
36.	Colombia	53	428	85,234
37.	Congo - Brazzaville	1	8	717
38.	Congo - Kinshasa	4	30	4,419
39.	Costa Rica	22	70	12,677
40.	Côte d'Ivoire	6	78	26,673
41.	Croatia	1,155,346	23,142,519	6,313,123,913
42.	Cuba	4	4	50
43.	Curaçao	1	1	125
44.	Cyprus	47	236	51,506
45.	Czech Republic	890	35,282	7,002,622
46.	Denmark	564	11,565	1,799,119
47.	Dominican Republic	4	31	1,114
48.	Ecuador	17	83	10,697
49.	Egypt	83	652	26,395
50.	El Salvador	5	99	15,377
51.	Estonia	1,503	12,057	3,123,082
52.	Ethiopia	15	116	28,273
53.	Europe	1,398	96,952	15,193,772
54.	Faroe Islands	3	11	848
55.	Finland	248	4,546	962,307
56.	France	2,027	109,255	20,372,694
57.	French Polynesia	4	14	6,946
58.	Gambia	1	1	1
59.	Georgia	35	156	27,077
60.	Germany	17,675	293,479	58,714,427
61.	Ghana	4	5	1,000
62.	Gibraltar	1	2	444
63.	Greece	357	1,533	477,706
64.	Grenada	13	40	9,442

65.	Guadeloupe	2	2	3,010
66.	Guatemala	5	49	7,369
67.	Guernsey	1	3	662
68.	Haiti	1	1	163
69.	Honduras	1	1	45
70.	Hong Kong SAR China	175	1,239	215,751
71.	Hungary	1,601	18,159	4,801,973
72.	Iceland	62	299	118,901
73.	India	329	1,116	334,232
74.	Indonesia	158	522	157,330
75.	Iran	30	117	21,279
76.	Iraq	73	151	20,819
77.	Ireland	2,098	18,091	4,936,897
78.	Isle of Man	5	59	18,481
79.	Israel	133	430	137,631
80.	Italy	3,050	49,308	8,844,232
81.	Jamaica	11	37	12,695
82.	Japan	216	1,792	322,026
83.	Jersey	1	2	190
84.	Jordan	27	66	104,807
85.	Kazakhstan	32	167	21,420
86.	Kenya	34	798	101,094
87.	Kuwait	37	122	55,197
88.	Kyrgyzstan	6	12	5,744
89.	Laos	19	62	12,999
90.	Latvia	123	1,118	261,875
91.	Lebanon	12	34	4,674
92.	Liberia	1	1,029	284,667
93.	Libya	5	12	4,655
94.	Liechtenstein	12	2,489	366,166
95.	Lithuania	2,236	12,556	2,950,112
96.	Luxembourg	539	4,412	1,231,743
97.	Macau SAR China	3	8	1,206
98.	Madagascar	5	8	833
99.	Malawi	14	171	692,180
100.	Malaysia	98	335	68,028
101.	Maldives	6	8	357
102.	Malta	102	924	142,161
103.	Martinique	1	1	1,310
104.	Mauritania	2	3	2,790
105.	Mauritius	20	51	6,205
106.	Mexico	171	1,320	358,737
107.	Moldova	106	1,763	499,313
108.	Monaco	22	390	44,370

109.	Mongolia	2	9	204
110.	Montenegro	5,921	74,412	26,743,505
111.	Morocco	59	226	42,278
112.	Mozambique	4	22	6,768
113.	Myanmar (Burma)	31	625	62,308
114.	Nepal	21	103	26,308
115.	Netherlands	2,299	59,282	15,222,549
116.	New Zealand	104	988	188,779
117.	Nicaragua	10	18	12,338
118.	Nigeria	33	2,015	232,345
119.	North Macedonia	1,653	18,334	4,433,953
120.	Norway	360	5,474	1,982,203
121.	Oman	115	628	48,591
122.	Pakistan	17	79	4,449
123.	Palestinian Territories	1	1	7
124.	Panama	19	231	91,467
125.	Paraguay	1	1	5
126.	Peru	35	224	23,228
127.	Philippines	95	382	51,338
128.	Pitcairn Islands	1	2	249
129.	Poland	2,358	45,167	12,304,620
130.	Portugal	419	3,151	778,821
131.	Puerto Rico	5	40	12,507
132.	Qatar	93	1,815	494,898
133.	Réunion	2	23	1,959
134.	Romania	567	19,195	3,749,730
135.	Russia	512	8,487	1,759,307
136.	Rwanda	2	2	90
137.	Saint Kitts and Nevis	3	40	29,618
138.	Saint Lucia	2	2	182
139.	Satellite Provider	4	11	665
140.	Saudi Arabia	53	439	67,714
141.	Senegal	10	38	57,054
142.	Serbia	9,676	88,909	58,941,003
143.	Seychelles	62	42,806	6,526,067
144.	Sierra Leone	1	2	6
145.	Singapore	158	956	707,089
146.	Slovakia	466	9,813	1,946,409
147.	Slovenia	12,774	246,846	33,146,688
148.	South Africa	78	803	225,069
149.	South Korea	85	323	52,287
150.	South Sudan	1	6	2,200
151.	Spain	1,384	13,014	6,896,783
152.	Sri Lanka	31	46	7,340

153.	Sudan	6	12	1,249
154.	Suriname	1	1	73
155.	Sweden	1,829	50,094	7,935,319
156.	Switzerland	1,647	27,318	8,642,473
157.	Syria	5	9	302
158.	Taiwan	56	214	64,589
159.	Tajikistan	2	2	108
160.	Tanzania	37	96	33,855
161.	Thailand	809	3,378	1,151,445
162.	Timor-Leste	11	57	9,244
163.	Togo	1	1	694
164.	Tunisia	13	73	16,137
165.	Turkey	631	3,990	2,102,011
166.	Uganda	7	17	3,342
167.	Ukraine	337	4,731	2,499,272
168.	United Arab Emirates	336	1,600	375,077
169.	United Kingdom	3,992	142,487	24,480,273
170.	United States	6,467	266,984	54,830,162
171.	Uruguay	6	13	3,681
172.	Uzbekistan	8	19	573
173.	Vatican City	6	18	2,570
174.	Venezuela	3	5	461
175.	Vietnam	347	2,903	465,490
176.	Zambia	8	42	4,609
177.	Zimbabwe	1	2	5
TOTAL:		1,368,702	26,701,365	7,235,096,012

Last update: Mon Apr 1 08:19:41 CEST 2019

Prema dostupnim MaxMindovim GeoIP podatcima (https://dev.maxmind.com/geoip), hrvatska vršna domena raspolaže s ukupno 2.818.597 IP-adresa, od kojih dobar dio nije izravno dostupan krajnjim korisnicima interneta. Prema podatcima iz gornjega prikaza proizlazi da je 41 % hrvatskih IP-adresa koristilo Hašekovu uslugu. Iz toga slijedi da je Hašek nedvojbeno infrastrukturna usluga u Hrvatskoj. Uzimajući u obzir udio Hrvata u populaciji BiH te činjenicu da je 13 % bosansko-hercegovačkih IP-adresa koristilo istu uslugu, zaključak se može protegnuti i na tu zemlju. Specifičnost Hašeka kao hrvatske infrastrukturne usluge jest ta da nikada nikakve veze nije imao, unatoč svim nastojanjima da se takav status promijeni, sa zaduženima za skrb o nacionalnim interesima. Sigurno je da to tako ne može ići do u nedogled, ako ni zbog čega drugoga onda zbog smrtnosti njegova održavatelja.

Sažetak

Šandor Dembitz, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
UDK 81'27:811.163.42, izvorni znanstveni rad
primljen 2. travnja 2019., prihvaćen za tisak 7. listopada 2019.

25 years of Hašek

Hašek is a Croatian on-line spellchecker that continuously operates since March 21, 1994, nowadays at the address https://ispravi.me/. In 25 years of functioning Hašek processed nearly 30 million texts, which build a corpus of more than 7 billion tokens. By comparison, all books ever published in Croatian form a corpus with less than 20 billion tokens. As a WWW-embedded tool, Hašek took advantage of many web-based services including learning. Thanks to Hašek's learning capability, its dictionary increased from initial 100 thousand to more than 2 million word-types. Another aspect of learning was the creating and regular updating of the Croatian n-gram system. Unlike Google, whose n-gram systems are based on the WaC (Web as Corpus) approach and cut-off criteria, Croatian n-grams were extracted from processed texts by a lexical criterion: each n-gram constituent must be proven by the spellchecker as valid in Croatian spelling. The difference in approaches made Croatian n-gram system comparable in size to the largest Google n-gram systems. Unfortunately, the advantages of on-line spellchecking for rapid breakthroughs into much more sophisticated language technology areas were not recognized by Croatian decision makers, with some consequences mentioned in the paper.

Key words: Hašek, spellchecking, learning, Google, n-gram systems

UPORABA NEODREĐENIH ZAMJENIČNIH RIJEČI U REČENICAMA POPUT
Je li me netko/tko/tkogod tražio?

Mate Milas

U radu se istražuje gramatičko-semantičko tumačenje jezičnoga savjeta o pravilnoj/nepravilnoj uporabi neodređenih zamjeničnih riječi u rečenicama poput *Je li me netko/tko/tkogod tražio?* u hrvatskom standardnom jeziku. U jezičnim se savjetima razlika između neodređenih zamjeničnih riječi s predmetkom *ne-* (*netko*, *nešto*...) u odnosu na zamjenične riječi bez tvorbenih dodataka (*tko*, *što*...) te s dometkom *-god* (*tkogod*, *štogod*...) zasniva na dvama razlikovnim obilježjima: postojanje/nepostojanje entiteta na koji zamjenica upućuje i nepoznatost entiteta. Autor smatra da je prvi semantem nedovoljno precizan, a drugi dijelom pogrješan. Razlika se između tih neodređenih zamjeničnih riječi može precizno opisati gramatičko-semantičkim konceptima *specificiranosti* ili *referencijalnosti*, koji se u jezikoslovlju počinju rabiti od 60-ih godina 20. st. Zanimljivo je da je te koncepte još krajem 19. st. opisao hrvatski jezikoslovac August