IVAN PANDŽIĆ i TOMISLAV STOJANOV

Čestoća riječi u hrvatskome jeziku

Neke od najčešćih riječi u kineskome jeziku jesu 人 (čovjek), 中 (sredina), 子 (dijete), 说 (govoriti) i 大 (velik), a u engleskome *time* (vrijeme), *person* (osoba), *year* (godina), *way* (put) i *day* (dan). Da bismo doznali koje su najčešće riječi u nekome jeziku, moramo se koristiti statistikom. Statistika se danas često primjenjuje u jezikoslovnim istraživanjima. Bez podataka o čestoći suvremeni se jezik ne može opisivati. Statistika je u pozadini i svih internetskih tražilica te drugih računalnojezikoslovnih alata (strojni prevodioci, sintetizatori govora, prepoznavači rukopisa itd.). Najtipičniji je jezičnostatistički podatak onaj o čestoći riječi (osobito su popularni popisi najčešćih imena i prezimena u Republici Hrvatskoj). Leksikografija je jedna od jezičnih disciplina u kojoj je podatak o čestoći riječi važan analitički podatak koji zamjenjuje metodologiju pisanja rječnika „iz glave". Tim su se načelom statističke relevantnosti, primjerice, povodili i autori *Hrvatskoga pravopisa* Instituta za hrvatski jezik i jezikoslovlje pri odabiru pravopisnih rješenja i sastavljanju pravopisnoga rječnika.

Da bi se došlo do statističkih podataka o jeziku, potrebno je prikupiti izvore i imati računalnu bazu tekstova. Oni se uobičajeno nazivaju *jezičnim korpusima*. Prikupljanjem tekstova i njihovom obradom stvorena je nova jezikoslovna disciplina, koja se naziva *korpusnim jezikoslovljem*. Tu disciplinu neki stručnjaci čak smatraju izvanjezikoslovnom i dijelom računalnih tehničkih znanosti.

Čestoća riječi danas se može točno izračunati, ali se prethodno mora postaviti temeljno pitanje njezine metodologije i određivanja kvalitativnih i kvantitativnih kriterija – koji se tekstovi, autori i žanrovi uključuju u istraživanje, je li zadovoljen kriterij relevantnoga uzorka itd.

Potraga za najčešćim hrvatskim riječima provedena je u trima najvećim hrvatskim korpusima: *Hrvatskome jezičnom korpusu* Instituta za hrvatski jezik i jezikoslovlje, http://riznica.ihjj.hr/, *Hrvatskome nacionalnom korpusu* Zavoda za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu, http://www.hnk.ffzg.hr/, te hrvatskome mrežnom korpusu pod nazivom *hrWaC*, koji je izrađen također na Filozofskome fakultetu Sveučilišta u Zagrebu, http://goo.gl/cT43jt. (Autori članka još jednom zahvaljuju dr. sc. Marku Tadiću i dr. sc. Nikoli Ljubešiću na ustupanju podataka o korpusima HNK v3.0 i hrWaC v2.0.)

	Hrvatski jezični korpus	Hrvatski nacionalni korpus	Hrvatski mrežni korpus
1.	i	i	i
2.	u	u	je
3.	je[1]	je	u
4.	se	se	se
5.	na	na	da
6.	da	za	na
7.	za	da	su
8.	su	su	za
9.	a	od	a
10.	od	o	od

Tablica 1. Najčešće riječi u trima hrvatskim korpusima

Od općega podatka o čestoći riječi mnogo je zanimljiviji podatak o čestoći punoznačnih riječi.

	Hrvatski jezični korpus	Hrvatski nacionalni korpus	Hrvatski mrežni korpus
1.	hrvatski	članak	htjeti
2.	godina	htjeti	moći
3.	moći	godina	godina
4.	jedan	što	imati
5.	Hrvatska	moći	velik
6.	reći	sav	dan
7.	predsjednik	hrvatski	trebati
8.	dan	Hrvatska	čovjek
9.	riječ	zakon	dobar
10.	posto	drugi	hrvatski

Tablica 2. Najčešće punoznačne riječi u trima hrvatskim korpusima

Uočava se da se korpusi minimalno razlikuju prema ukupnoj čestoći, dok se pokazatelji o punoznačnim riječima razlikuju zbog različitih korpusnih izvora. Tako je, primjerice,

[1] Pojavnica *je* u ovoj tablici označuje i zamjenicu i glagol.

za potrebe ovoga istraživanja *Hrvatski jezični korpus* zastupljen s 80-milijunskim novinskim potkorpusom, koji čine izdanja *Vjesnika*, *Vijenca* i *Sportskih novosti* u omjeru 35 % – 10 % – 55 % i koji se očekivano razlikuje od *Hrvatskoga nacionalnog korpusa* u kojemu su zastupljeniji pravni tekstovi i zakoni iz *Narodnih novina*.

Prvi rječnik izrađen na temelju korpusa tekstova engleski je *Collins COBUILD English Language Dictionary* iz 1987. Osim što su elektronički izvori postali važniji u leksikografskome radu, važnost statistike i jezičnih korpusa ogleda se i u tome što su leksikografske definicije u nekim rječnicima promijenile poredak te se na prvome mjestu navodi najčešće potvrđeno značenje koje riječi. Tako, primjerice, *Hrvatski jezični korpus* IHJJ-a u novinskome potkorpusu pokazuje veliku prevlast pojavnice *sapunica* u drugome, prenesenome značenju 'televizijska serija s mnogo nastavaka', što je obrnuto razmjerno u odnosu na knjižni potkorpus u kojemu je izvorno značenje 'pjena nastala topljenjem sapuna u vodi' bitno češće. Iz ovoga se podatka vidi u kojoj je mjeri važna uravnoteženost korpusnih tekstova za leksikografske potrebe.

I još nekoliko zanimljivih jezičnostatističkih podataka.

Riječi koje se pojavljuju samo jednom u promatranome tekstnom uzorku nazivaju se *jednopojavnice* ili *hapaksi*.

Podatcima o čestoći riječi koristio se i Ivan Mažuranić kad je 1844. dovršavao *Osmana* u XIV. i XV. pjevanju želeći upotrijebiti česte riječi koje je Ivan Gundulić dvjesto godina prije zapisao u preostalih osamnaest pjevanja.

Hrvatski jezik može se pohvaliti da je jedan od rijetkih svjetskih jezika koji ima tiskani čestotni rječnik. *Hrvatski čestotni rječnik* objavljen je 1999., a njegovi su autori Milan Moguš, Maja Bratanić i Marko Tadić.

Ipak, koliko god čestoća bila važna u proučavanju jezika, jezična normativistika ne može se isključivo povoditi kriterijem čestoće u donošenju pravila o suvremenome i standardnome hrvatskom jeziku. Jezik nije samo slika sadašnje uporabe, jezik je i tradicija i sustav koji se treba održavati.

Što vama govori podatak o najčešćim riječima u jeziku i što možemo zaključiti o društvu koje nas okružuje? U kojoj se mjeri čestoća određenih riječi mijenja s vremenom? Što vi mislite – trebaju li i hrvatski rječnici na prvo mjesto definicije stavljati statistički najčešće značenje ili i dalje slijediti tradicionalne jezikoslovne kriterije osnovnoga i prenesenoga značenja?

Za neke se hrvatske književnike popularno govori da imaju izrazito bogat rječnik (Miroslav Krleža, Vesna Krmpotić, Tomislav Ladan…). Hoće li komparativna računalna raščlamba to i dokazati, čitajte uskoro…

Zanima li vas neko specifično pitanje povezano s jezičnom čestoćom, pišite uredništvu časopisa *Hrvatski jezik* na e-adresu *hrjezik@ihjj.hr*.