# Računalna obrada i kategorizacija hrvatskih čestica za POS tagging

**Matić, Katharina**

SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA LINGVISTIKU

DIPLOMSKI STUDIJ

RAČUNALNA LINGVISTIKA

**Katharina Matić**

# Računalna obrada i kategorizacija hrvatskih čestica za *POS tagging*

Diplomski rad

Mentor:

Doc. dr. sc. Božo Bekavac

Zagreb, 2021.

UNIVERSITY OF ZAGREB

FACULTY OF HUMANITIES AND SOCIAL SCIENCES

DEPARTMENT OF LINGUISTICS

GRADUATE PROGRAMME

COMPUTATIONAL LINGUISTICS

**Katharina Matić**

# Computer processing and categorization of Croatian particles for POS tagging

Master's thesis

Supervisor:

Božo Bekavac, PhD

Zagreb, 2021

**Zahvala**

Prije svega, zahvaljujem doc. dr. sc. Boži Bekavcu, koji mi je pristao biti mentor unatoč pretrpanom rasporedu.

Posebno sam zahvalna i šefici dr. sc. Nini Tuđman Vuković, koja me uvijek sa strpljenjem poticala da pišem.

Napokon, zahvalna sam roditeljima na bezuvjetnoj potpori tijekom oduženog trajanja studija.

**Sadržaj**

**Sažetak**

Čestice nisu jednako i sustavno kategorizirane u svim tradicionalnim gramatikama hrvatskog jezika jer za tu kategoriju ne postoji jedinstvena precizna definicija kojom bi se svi autori mogli voditi bez prostora za različite interpretacije. Budući da se status nekog leksema kao čestice uglavnom utvrđuje prema njegovoj značenjskoj funkciji u rečenici, a priroda jezične produkcije podrazumijeva i visoku razinu jezične kreativnosti te je broj rečeničnih konstrukcija i kombinacija pojedinih leksema u njima nemjerljiv, u tradicionalnim se gramatikama u kategoriju čestica u pravilu uključuje i niz drugih vrsta riječi, kao što su prilozi, ovisno o njihovoj funkciji u pojedinom kontekstu. To je osobito slučaj kad se uzme u obzir semantička dimenzija svakog iskaza, bilo u pogledu namjeravanog značenja, bilo u pogledu značenjskih odnosa među leksemima. Međutim, takva kategorizacija čestica nije praktična ni produktivna kad je riječ o računalnoj obradi jezika, u ovom slučaju konkretno označavanju vrsta riječi u korpusima hrvatskog jezika, jer se u takvoj vrsti obrade uglavnom gube semantičke i pragmatičke nijanse jezične upotrebe te je fokus na oblicima leksema i njihovim sintaktičkim ulogama.

U ovom se radu popis hrvatskih čestica nastoji ograničiti na što manji mogući broj nepromjenjivih riječi koje se neupitno mogu smatrati česticama neovisno o kontekstu. Takav se pristup čini najproduktivnijim rješenjem za potrebe računalne obrade jezika zbog njegove jednostavnosti i činjenice da se preostali leksemi koji se tradicionalno dvojno kategoriziraju kao čestice, iako prvenstveno pripadaju drugoj vrsti riječi, na taj način i dalje mogu bez poteškoća kategorizirati u skladu sa svojim primarnim kategorijama bez uplitanja semantičkih kriterija koji računalu nisu dostupni na istoj razini kao čovjeku. Pritom se izbjegava i problem relativnosti semantičkih tumačenja u vrsti obrade za koju su prije svega potrebna jasna razgraničenja i preciznost.

**Ključne riječi**

POS označavanje, čestice, računalna lingvistika

**Abstract**

Particles are not equally and systematically categorized in traditional Croatian grammars, as there is no single precise definition for this category that all authors could use as a guide with no room for different interpretations. Since the status of a lexeme as a particle is mainly determined by its semantic function in a sentence, and since language production by its nature contains a high level of linguistic creativity, as well as considering the fact that the number of sentence constructions and combinations of individual lexemes within them is immeasurable, traditional grammars as a rule include a number of other types of speech, such as adverbs, in the category of articles depending on their function in a particular context. This is especially the case when considering the semantic aspect of each utterance, either in terms of intended meaning or in terms of semantic relations between lexemes. However, such categorization of particles is neither practical nor productive when it comes to computer processing, in this case specifically part-of-speech tagging for Croatian language corpora, because in this kind of processing the semantic and pragmatic nuances of language use are mostly lost and the focus is on the forms and syntactic roles of lexemes.

The aim of this paper is to limit the list of Croatian particles to the smallest possible number of invariable words that can undoubtedly be considered particles regardless of context. Such an approach seems to be the most productive solution for part-of-speech tagging due to its simplicity and the fact that the remaining lexemes that are traditionally doubly categorized as particles, despite primarily belonging to another part of speech, can still be easily categorized according to their primary categories without the interference of semantic criteria that are not available to computers at the same level as to humans. At the same time, this approach avoids the issue of varying semantic interpretations in a type of language processing that primarily requires clear distinctions and precision.

**Key words**

POS tagging, particles, computational linguistics

## 1. Uvod

Označavanje vrsta riječi u korpusu temeljni je postupak za brojne druge razine označavanja. Takvo se označavanje prvo obavlja ručno te se zatim na temelju tako označenog korpusa strojno trenira sustav za daljnje automatsko označavanje. Automatizacijom tog postupka znatno se ubrzava cijeli proces, no bitno je da korpus na kojem sustav temelji svoj rad bude što točnije označen i da za njega postoje što precizniji kriteriji.

Čestice su u hrvatskom jeziku problematična vrsta riječi jer se njihova kategorizacija među autorima gramatika donekle razlikuje (vidi poglavlje 3.). U načelu je riječ o istim nepromjenjivim riječima, no u tu se kategoriju svrstava i niz leksema koji već pripadaju drugim vrstama riječi, najčešće prilozima, kad se njihova upotreba smatra atipičnom.

Takva se dvojna kategorizacija obično temelji na semantičkim kriterijima, što otežava sustavan opis kategorije čestica na način koji bi bio razumljiv računalu. Naime, za spomenute semantičke kriterije najčešće ne postoje odgovarajući sintaktički obrasci koji bi nedvojbeno upućivali na to da je u samom tekstu riječ o čestici, a ne prilogu ili nekoj drugoj vrsti riječi. Stoga ćemo u ovom radu nastojati jasno razgraničiti što jest, a što nije čestica kako bi računalo neupitno moglo ispravno označiti njihovu vrstu riječi u korpusu.

Budući da su čestice u užem smislu nepromjenjive riječi koje već ne pripadaju nijednoj drugoj vrsti riječi, kao moguće rješenje nameće se sastavljanje ograničenog popisa čestica koji isključuje sve druge riječi koje već pripadaju nekoj drugoj kategoriji. Prilozi bi se u tom slučaju uvijek kategorizirali kao prilozi, bez obzira na to je li njihova upotreba atipična ili ne, a isto bi vrijedilo i za ostale vrste riječi. U pravilu, ako leksem u funkciji čestice već pripada drugoj vrsti riječi i zamjenjiv je njome, ne bi ga trebalo smatrati česticom.

U ovom se radu nastoji pronaći pragmatično rješenje za što jednostavnije označavanje čestica u korpusima hrvatskoga jezika bez ugrožavanja ukupne točnosti označavanja korpusa. Drugim riječima, cilj je ograničiti popis čestica samo na one lekseme koji su neupitno čestice kako bi se smanjili problemi koje može uzrokovati dvojna kategorizacija određenih leksema te tako povećala točnost POS označavanja čestica u hrvatskome jeziku.

## 2. POS označavanje i kategorizacija vrsta riječi

*Part-of-speech tagging* (u daljnjem tekstu „POS označavanje") je „pridruživanje gramatičke kategorije svakoj pojavnici u tekstu" koje se ponekad naziva „gramatičko označavanje ili morfosintaktičko obilježavanje" i jedna je od osnovnih vrsta lingvističkog označavanja. Nadalje, „POS oznake prvi su korak u razrješavanju istopisnica, tj. pojavnica koje imaju isti lik a različite gramatičke kategorije i/ili značenje" (Bekavac, 2002: 177). Alat kojim se obavlja POS označavanje zove se *POS tagger* (ili označivač). Označivači se dijele u dvije vrste: vjerojatnosni označivači (*probabilistic*) i označivači zasnovani na pravilima (*rule-based*). Prva se vrsta temelji na statistici, a druga na ručno pisanim lingvističkim pravilima. Iako većina suvremenih označivača koristi prvi pristup, on se i dalje primjenjuje u kombinaciji s drugim pristupom kao prvim korakom, i koji je fokus ovog rada. Ti su alati najčešće vrlo precizni, no nije moguća potpuna točnost jer ne postoji nijedan računalnolingvistički alat koji radi s potpunom točnošću. Primjenom pravila na rezultate vjerojatnosnog označivača povećava se točnost rezultata (Bekavac, 2002: 173-177).

Ma i suradnici (2011: 57) pokazuju važnost točnog označavanja vrsta riječi na primjeru engleske rečenice *They might at any time turn against their masters*, u kojoj leksem *against* ima funkciju čestice, a ne prijedloga kako bi se inače očekivalo na temelju njegova oblika. Dakle, računalo bi trebalo prepoznati da je *against* u toj rečenici dio frazalnog glagola *turn against*, a ne imenske fraze *against their masters*. U suprotnom bi moglo doći do poteškoća u daljnjoj obradi te rečenice, primjerice u slučaju strojnog prevođenja, te dovesti do netočnog krajnjeg rezultata. Naravno, čestice na engleskom jeziku ne podudaraju se u doslovnom prijevodu s česticama na hrvatskom, no isto načelo vrijedi neovisno o jeziku. Segmentacija rečenica i označavanje vrsta riječi u početnim fazama obrade korpusa temelji su za druge vrste označavanja korpusa te stoga moraju biti što točniji i precizniji.

Razlike između tradicionalne kategorizacije vrsta riječi i kategorizacija za POS označavanje mogu se izložiti i na primjeru Pranjkovićeve (2008: 238-246) analize leksema *što* i njegovih brojnih funkcija. Primjer koji Pranjković nudi za *što* kao uzvik zapravo se može bez problema svrstati u priloge jer je u njemu *što* zamjenjiv prilogom *kako*, stoga u tom slučaju nije potrebna zasebna kategorizacija leksema *što* kao uzvika ili, kako on predlaže, čestice:

> (Uh) *što* se ona danas pravi važna! (Pranjković, 2008: 238)

Za računalo bi se leksem *što* mogao tretirati kao prilog i kada se koristi umjesto *zašto*, i u slučajevima gdje bismo ga inače smatrali česticom, poput *što više* ili *što bolje* (slično kao *sve više* i *sve bolje*), odnosno kad se javlja kao intenzifikator. Na taj se način samo jednom kategorizacijom pokrivaju dvije upotrebe (prilog i čestica).

Pranjković (2008: 240) za *što* kao prilog navodi primjere u kojima se taj leksem upotrebljava umjesto upitnog priloga *zašto*:

> *Što* nas mučite? *Što* i njih ne pozoveš? *Što* si tako blijed?

Navodi i upotrebu u obliku okamenjenoga dativa, npr. *čemu si se udavala, čemu uopće pokušavati* i sl. Takve slučajeve s okamenjenim dativom vjerojatno možemo za potrebe računala ignorirati jer nisu dovoljno česti da bi bilo opravdano njihovo uključivanje, pa se možda mogu svrstati u zamjenice, a kod priloga je jedini promjenjivi dio zapravo gradacija, koja nije primjenjiva na leksem *što*.

Jednako tako, *što* se upotrebljava i kao i zamjenica i kao veznik. Međutim, za POS označavanje upotreba leksema *što* kao veznika može biti pokrivena i samo kategorizacijom kao zamjenice, s obzirom na to da doista i dalje jest riječ o zamjenici. Time se rješava i problem odvajanja tih dviju upotreba kontekstualnim pravilima koja nisu nužno uvijek dovoljna za precizno razgraničavanje. Takvim se odlukama općenito pojednostavljuje cjelokupan postupak POS označavanja bez ugrožavanja točnosti izlaznih podataka. Mogla bi se kritizirati činjenica da je takvo označavanje korpusa u tom smislu manje precizno jer se ne uzima u obzir nijansiranost stvarne jezične upotrebe, ali se u strogom smislu zadržava točnost kategorizacije samih vrsta riječi, što je u ovom slučaju prioritet.

Dakle, ne čini se potrebnim zamjenice u službi veznika kategorizirati kao veznike, nego samo kao zamjenice, uključujući slučajeve kad uvode zavisne rečenice. Drugim riječima, prema takvim kriterijima *što* se u svim slučajevima kad je veznik može kategorizirati kao zamjenica.

Kao čestica *što* se najčešće upotrebljava kao svojevrsni intenzifikator uz komparative i superlative i može se opisati kao poredbena čestica kojom se pojačavaju poredbena značenja, npr.

> *Što* više, to bolje. Dođi *što* možeš prije. Obuci *što* ljepšu haljinu. Došli smo najbrže *što* smo mogli. Uvijek radi najbolje *što* zna. Pjeva najjače *što* mu grlo može podnijeti. (Pranjković, 2008: 242)

U tim primjerima leksem *što* malo više naginje čestici nego u drugima, ali ne nužno. Ako ga zbilja odlučimo kategorizirati kao česticu, potrebna je detaljna razrada konteksta u kojima to jest, a u kojima nije, a tijekom pretraživanja korpusa teško je naći primjere u kojima je neophodno označiti ga kao česticu. Treba li nam uistinu *što* kao čestica samo u slučajevima gdje se javlja kao intenzifikator? Pragmatičniji pristup bi, čini se, bio tretirati ga kao prilog i u tim slučajevima kako bi se smanjile komplikacije pri određivanju sintaktičkih pravila.

Kao mogući problem nameću se sljedeće rečenice koje navodi Pranjković (2008: 242):

> *Što* misliš da ga nisam pitao? A *što* ih ti sama ne umiješ uzeti? A *što* ti za to uopće nisi čuo?

U takvim kontekstima leksem *što* zamjenjiv je česticom *zar* te se stoga čini kao „prava" čestica, ali ne čini se opravdanim uključiti taj leksem u konačan popis čestica na temelju tako ograničene upotrebe, pogotovo ako se uzme u obzir nedostatak karakterističnog sintaktičkog obrasca u kojem bi nedvojbeno bila riječ o čestici. Za svaki od navedenih sintaktičkih obrazaca možemo zamisliti rečenice u kojima je konstrukcija jednaka, ali upotreba leksema različita. Tako usporedno s prvom rečenicom, na primjer, možemo imati rečenicu *Što misliš da mu se dogodilo?*, a za preostale dvije *A što ih ti pitaš na predavanjima?* i *A što ti za njega uopće radiš?*

Zaključno, leksem *što* na temelju navedenoga ne bi trebalo kategorizirati kao česticu za potrebe POS označavanja. To je najjednostavnije rješenje – da *što* bude samo zamjenica i prilog, odnosno da kategorizacija kao zamjenice pokriva i upotrebu kao veznika, a prilog upotrebu kao čestice – umjesto izdvajanja svih mogućih interpretacija koje računalu nisu nužno toliko bitne.

## 3. Čestice u tradicionalnim gramatikama

Čestice su nepromjenjiva vrsta riječi čije se definicije, iako se u načelu podudaraju, donekle razlikuju među autorima gramatika. Karlić i Tušek (2013: 210-211) navode da se u gramatikama južnoslavenskih jezika definicije čestica mogu razvrstati u tri kategorije: 1. čestice kao zasebna vrsta riječi, pri čemu se definicija ne zadržava samo na njihovim morfološkim obilježjima, nego uključuje i njihove semantičke, sintaktičke, a nekad i pragmatičke funkcije, 2. čestice kao zasebna vrsta riječi, ali isključivo kao sintaktička pojava, i 3. klasifikacije prema kojima čestice nisu zasebna vrsta riječi.

Čestice su se u hrvatskim gramatikama uglavnom svrstavale među priloge sve do 1979., kad je objavljena *Priručna gramatika hrvatskoga književnog jezika*, u kojoj se kategorija čestica uvodi na sljedeći način:

> „Posebnu vrstu priloga čine riječi koje su po obliku prilozi, ali se ne prilažu pojedinim riječima ili dijelovima rečenice, nego cijeloj rečenici. One pokazuju stav govornika prema onome što se u rečenici govori i ne vrše službu nijednog njezina dijela, pa ih odvajamo kao posebnu vrstu riječi pod imenom čestice". (Barić i sur., 1979: 139)

Čestice su u toj gramatici dalje definirane kao „riječi koje iskazuju stav govornika prema onome o čemu govori, s obzirom na njegovo znanje, želje i osjećanja" (Isto: 214). Upotrebljavaju se u poricanju neke tvrdnje (*ne*), u pitanju je li tvrdnja istinita (*li, zar*), za pojačanje tvrdnje ili poricanja (*da, jest, dabome, dakako, svakako, ne, nikako*), za izricanje nestrpljenja, želje, zadovoljstva i sl. onim što znači riječ uz koju se čestica stavlja (*bar, baš, čak, i, jedva, još, ni, niti, opet, samo, tek, već*), za izricanje ravnodušnosti, dopuštenja (*ma, makar, bilo, god*), za izricanje dojma o onome o čemu se govori ili svoje ocjene toga (*doista, gotovo, istina, možda, naravno, nekako, potpuno, sasvim, sigurno, skoro, vjerojatno, veoma, vrlo, zaista* i sl.), odnosno „riječi koje su po postanju prilozi načina, ali se ne odnose na glagole, ni na pridjeve, ni na priloge, niti ih određuju, već se odnose na smisao čitave rečenice" (Isto: 215). Leksemi iz posljednje skupine čine se nezavisnima od rečenice te se nekad od nje odvajaju zarezima, npr.

> *Gotovo* svi su uskliknuli od radosti.
>
> To je *sigurno* najbolje rješenje.
>
> Njihove su riječi, *dakako*, mnogo pomogle da se stvar razjasni.
>
> *Istina*, on o svemu tome nije imao ni pojma.

Medu čestice se ubrajaju i poštapalice, odnosno „riječi koje neki govore bez ikakve veze s njihovim smislom, obično kad u govoru žele dobiti vremena da nađu potrebne riječi za ono što žele reći" (Isto) kao što su *ovaj, onaj, čuj, kaže, vele, dragi moj* i sl. Npr.

> Bio sam tamo pa, *ovaj*, nisam vidio ništa, ali, *ovaj*, čuo se, *ovaj*, neki potmuo glas.

U poglavlju o česticama spominju se i rečenice i skupovi riječi koji imaju modalnu službu te „gube svoj puni smisao i služe samo da održe pažnju slušalaca ili da daju oduška osjećajima koji obuzimaju govornika" (Isto) kao što su *tako reći, općenito govoreći, sve u svemu, hoćeš-nećeš, bolje rekavši, kao što znate, oprostite* i sl. Npr.

Sve je to, *tako reći*, izvitopereno.

On je na sve to, *oprostite*, jednostavno šutio.

Vi to, *na sreću*, niste osjetili, ali ja, *kažem vam*, i te kako jesam.

Srž takve razdiobe zadržala se i u suvremenim gramatikama, no razrada kategorizacije primjetno se razlikuje među njima.

Babić i suradnici (1991: 734) u svojoj gramatici čestice definiraju kao nepromjenjive riječi „koje služe za oblikovanje ili preoblikovanje rečeničnog ustrojstva, za isticanje ili davanje drugačijega značenja pojedinim riječima, a služe i za subjektivno-modalnu ocjenu rečenice kao cjeline". Čestice su podijeljene u dvije skupine – jedne služe za preoblikovanje rečenica ili značenja pojedinih riječi, a druge su jednakovrijedne rečenici. U prvu skupinu ubrajaju se čestice *li, zar* (upitne), *ne* (niječna), *bar, baš, čak, ma, također, upravo, ta* i sl. (za posebno isticanje), *evo, eto, eno* (upozorenje na ono što je blizu), *god, put, puta, neka* i *da*. U drugu skupinu svrstane su nepromjenjive riječi kao *sporno, sigurno, jamačno, vjerojatno, zacijelo, naravno, navodno, možda, valjda, međutim, naprotiv, jednostavno, štoviše* i sl. kad se upotrebljavaju kao jednakovrijedne rečenicama. U skladu s time predlaže se podjela čestica u dvije vrste riječi, odnosno čestice i modalne (načinske) riječi. Ta gramatika, za razliku od nekih drugih, u čestice ne ubraja poštapalice i slične riječi, već ih obrađuje zajedno s usklicima.

Raguž (1997: 277) u svojoj gramatici navodi da se među čestice ubrajaju „različite riječi koje ostaju izvan uobičajenih skupina vrsta riječi". Pobliže su opisane kao riječi kojima se „modificira značenje drugih riječi ili čitava iskaza" te se ističu i njihove gramatičke funkcije. Kao dodatan semantički kriterij navodi se da se njima izražava stav govornika, te se na temelju tih kriterija u čestice ubrajaju i prilozi i poštapalice, pa čak i neke umetnute rečenice. Međutim, autor ističe kako je na taj način definirana kategorija preopširna za detaljnu analizu i stoga obrađuje samo čestice „koje su bitne za gramatičke funkcije i koje nisu ni u jednoj drugoj skupini riječi", odnosno *zar, li, ne, da, jest, ma, makar, god, bilo, bar, barem, de/hajde, neka, daj, put/puta, pa, ta, evo, eto* i *eno*.

Silić i Pranjković (2005: 253) u svojoj gramatici čestice definiraju kao „suznačne […] i nepromjenjive riječi kojima se izražava stav govornika prema sadržaju cijeloga iskaza ili prema njegovu dijelu, odnosno riječi koje na bilo koji drugi način modificiraju dijelove rečenice, rečenicu, odnosno iskaz ili sudjeluju u oblikovanju njihova gramatičkog ustrojstva". S obzirom na sintaktički status podijeljene su u dvije skupine: „nesamostalne gramatikalizirane riječi koje modificiraju značenja drugih riječi, spojeva riječi ili rečenica" (*li, zar, barem, god* i sl.) i

samostalne „čestice koje se odnose na čitavu rečenicu i modificiraju po čemu njezin sadržaj", odnosno modifikatore (npr. *nažalost*). Kao što vidimo iz primjera koji navode za drugu kategoriju, zbog sintaktičke prirode navedene definicije čestica u tu kategoriju uključuju i priloge, ovisno o njihovoj upotrebi.

U nesamostalne čestice ubrajaju upitne čestice (*li, zar, da*), intenzifikatore (*i, ni, također, ta, pa, samo, bar, barem, god, ma, makar, bilo*), usporedne čestice (*mnogo, puno, malo, vrlo, veoma, dosta, gotovo, skoro, jedva, još, nešto, previše, odviše, posve, potpuno, sasvim, skroz, posebno, osobito, naročito, prilično, neobično* itd.), poticajne čestice (*neka, hajde, daj, de, deder, da*), jesno-niječne čestice (*da, ne, jest*) i prezentative (*evo, eto, eno, gle*).

Ako sagledamo širu sliku kategorizacije čestica u navedenim gramatikama, možemo uočiti nekoliko stvari. Definicija iz *Priručne gramatike* iz 1979. nije detaljno razrađena i temelji se na semantičkim kriterijima te bi se kao takva mogla primijeniti na niz leksema ovisno o interpretaciji pojedinca. Osim toga su u kategoriju čestica uključeni prilozi koji se prilažu rečenici u cjelini. Nijedan od tih kriterija nije nužno relevantan u kontekstu POS označavanja, prvenstveno zato što ih je teško primijeniti. Čak i ako kao kriterij za prilaganje rečenici uzmemo odvojenost zarezima, on nije primjenjiv u svim slučajevima, a izvan toga je takva kategorizacija čestica gotovo nemoguća ako se uzme u obzir slobodan redoslijed riječi koji je svojstven hrvatskom jeziku. Ukratko, svaki kriterij koji se prvenstveno temelji na semantičkim nijansama u ovom konkretnom kontekstu nije dovoljno učinkovit pristup.

Definicija iz *Praktične hrvatske gramatike* ukazuje na to da se čestice obično kategoriziraju na temelju toga što ne pripadaju nijednoj drugoj vrsti riječi. Čestice su tako u gramatikama obično podijeljene u dvije kategorije, koje bismo mogli nazvati česticama u „užem" smislu i česticama koje izvorno pripadaju drugim vrstama riječi, najčešće prilozima. Dakle, u slučaju druge kategorije, takve se čestice ne kategoriziraju prvenstveno prema zajedničkim obilježjima, već im se ta obilježja pripisuju retroaktivno na temelju atipičnog sintaktičkog i semantičkog ponašanja. Raguž u toj gramatici također ističe da bi obrađivanje druge kategorije bilo preopširno te ograničava svoju analizu na prvu kategoriju.

Babić i suradnici kategoriziraju čestice na temelju sintaktičko-semantičkih kriterija te ih, kako je već napomenuto, dijele u dvije skupine. Prva skupina sadrži čestice kojima se preoblikuju drugi elementi iskaza, a druga čestice koje su jednakovrijedne rečenici. U drugu skupinu svrstavaju uglavnom priloge i u nju ne ubrajaju poštapalice. Primjećujemo da Silić i Pranjković također dijele čestice u dvije kategorije. Čestice u „užem" smislu opisuju kao „nesamostalne", a čestice u „širem" smislu kao „samostalne". Kao primjer samostalne čestice navode prilog, što

je u skladu s podjelama u drugim gramatikama, ali isto tako priloge svrstavaju i u neke potkategorije nesamostalnih čestica, npr. „usporedne čestice". Takva se podjela ponajprije temelji na semantičkom odnosu između dotične „čestice" i iskaza u kojem se ona javlja.

Navedene gramatike u „širu" kategoriju čestica uključuju naizgled neograničen popis priloga, što je u kontekstu tradicionalne gramatičke analize opravdano semantičkim kriterijima. No budući da je i dalje u načelu riječ o prilozima, za potrebe POS označavanja njihovo uključivanje nije produktivno. Rezultat POS označavanja će bez njihova uključivanja u kategoriju čestica i dalje biti njihovo točno razvrstavanje u priloge jer oni doista izvorno pripadaju toj vrsti riječi.

## 4. Problematika kategorizacije hrvatskih čestica za POS označavanje

Čestice su kao kategorija u hrvatskom jeziku općenito problematične jer se u njih svrstava gotovo nepresušan niz leksema koje gramatičari na sintaktičkoj i semantičkoj razini u određenim slučajevima ne žele svrstati u njihove primarne kategorije zbog njihove atipične upotrebe u usporedbi s pravilima koja su predodređena za njih. Odgovor bi, naime, mogao ležati u prilagodbi pravila primarnih kategorija, odnosno u povećanju njihove fleksibilnosti, umjesto postojeće prakse dvostruke kategorizacije kao čestica za sve slučajeve u kojima se upotreba čini „neuobičajenom".

Kao što Josić (2011: 7) navodi, „[u] čestice se mogu ubrojiti i one riječi koje na obličnoj razini već pripadaju nekima od […] devet vrsta, odnosno uzvicima, prilozima, veznicima, zamjenicama i glagolima. U tim slučajevima riječ je o upotrebi jedne vrste riječi u službi druge, a u službi čestica najčešće su prilozi". U takvim slučajevima, odnosno u konstrukcijama u kojima leksem koji već pripada drugoj vrsti riječi tek vrši funkciju čestice, nema smisla označavati sve te lekseme kao čestice jer bi se time bespotrebno otežao postupak pripreme kriterija za računalo. Povrh toga bilo bi nemoguće na isključivo sintaktičkoj razini dovoljno precizno razgraničiti sve moguće upotrebe drugih vrsta riječi kao čestica, što bi zasigurno dovelo i do veće učestalosti netočnog označavanja korpusa.

Još je jedan od čestih problema pri označavanju korpusa podudaranje pisanih oblika različitih leksema koji pripadaju različitim vrstama riječi, odnosno homografija. Hudeček i Mihaljević (2009: 159) opisuju homonimiju i homografiju na sljedeći način: „Homonimija se može ostvariti kao homofonija (istozvučnost – isti izgovorni oblik), homografija (istopisnost – isti pisani oblik) ili pak oboje istodobno (potpuna homonimija). Osim toga homonimne mogu biti cijele riječi (leksemi) ili samo pojedini oblici riječi." Kako bi se taj problem riješio, potrebna su sintaktička odnosno ko-tekstualna pravila na temelju kojih računalo može razlikovati i

ispravno kategorizirati homografne lekseme odnosno njihove homografne oblike. Budući da je to potrebno prvenstveno za razlikovanje među vrstama riječi, koje se u načelu različito ponašaju i javljaju u različitim neposrednim sintaktičkim okruženjima, prikladnim se pravilima uspješno mogu razgraničiti.

No što ako je u pitanju razlikovanje, primjerice, priloga u funkciji priloga i priloga u funkciji čestice? U takvim se slučajevima, barem u hrvatskom jeziku, ne bi moglo reći da su u pitanju samo homonimi jer je doista riječ o istom leksemu. Ovisno o konkretnoj rečenici ili vrsti upotrebe, moguće je da će različiti gramatičari različito kategorizirati isti leksem u istom okruženju kao česticu ili kao prilog. Nadalje, velik dio te kategorizacije ovisi o njegovoj semantičkoj funkciji, koja može biti predmet rasprava i uzrok nesustavnosti kategorizacije.

Takvim nedoumicama u računalnoj lingvistici nema jednako puno mjesta kao u tradicionalnim gramatikama, a posebno kad je u pitanju POS označavanje, koje je konkretno usmjereno na vrstu riječi. Stoga se čini da nema smisla za takve potrebe u popis čestica uključivati dug i neograničen niz priloga kad oni već imaju vlastitu kategoriju kojoj pripadaju. Ako kao jedno od okvirnih pravila uzmemo da čestica nije nijedan leksem koji je u svim kontekstima zamjenjiv prilogom, kako bismo opravdali kategorizaciju samog priloga kao čestice? Naravno, na razini jezične teorije uvijek se može dalje ući u detalje s obzirom na to da je jezik po svojoj prirodi podložan interpretaciji, pa tako i njegova pravila. Međutim, računalu takve nedoumice ne samo da nisu važne, nego, štoviše, mogu dovesti do niže kvalitete rezultata zbog neusklađenosti i maglovitosti kriterija koji su temelji za daljnji rad. Stoga takva teorijska načela nisu nužno mjerodavna za praktičnu primjenu, osobito ako u teoriji ne postoji jedinstven i sustavan dogovor u pogledu kategorizacije na kojoj se rad računala temelji.

Rezultati pretraživanja korpusa hrWaC pokazuju da trenutačna kategorizacija čestica u njemu uključuje lekseme kao što su *sve, što, naime, i, to, tek, pak, baš, ni* i *niti*, koji se zapravo mogu svrstati u druge vrste riječi bez da ih se uključi u popis čestica. Tako možemo lekseme *i, pak, ni* i *niti* kategorizirati kao veznike, *to* kao zamjenicu, *sve* i *što* kao zamjenice i eventualno priloge te *naime, tek* i *baš* kao priloge.

Kad isključimo sve lekseme koji se neupitno mogu kategorizirati kao druge vrste riječi, preostaje nam ograničen popis leksema koje možemo pojedinačno analizirati kako bismo procijenili treba li ih svrstati u čestice ili u neku drugu kategoriju, odnosno analizom utvrditi koje od odabranih „tradicionalnih" čestica za potrebe POS označavanja nije moguće izolirati kao takve jer ne postoje specifična sintaktička okruženja po kojima bi se diferencirale.

Na temelju dosad navedenog, popis čestica trebao bi biti što kraći i sve lekseme koji se mogu svrstati u druge vrste riječi trebalo bi tako i kategorizirati radi ekonomičnosti opisa. Računala ne mogu jednako kao čovjek razlikovati nijanse u značenjima izvan sintaktičkog okruženja, koje je u nekim slučajevima naprosto nemoguće izolirati na način koji bi obuhvatio isključivo upotrebu određenih leksema kao čestica ako se one inače kategoriziraju i kao neka druga vrsta riječi. Naime, sintaktička okruženja koja su za računalo jednaka mogu imati različite semantičke nijanse koje bi stvarnom govorniku ukazale na razliku u upotrebi.

Može se postaviti i pitanje je li kategorija čestica kao takva neophodna za POS označavanje ako se većina njih može svrstati u druge vrste riječi, no leksemi kao što su *zar, li, da, ne* i sl., odnosno čestice u užem smislu, pokazuju da to nije uvijek moguće. Usto, činjenica da se čestice uopće mogu podijeliti na lekseme koji su isključivo čestice i one koji pripadaju i nekoj drugoj vrsti riječi ukazuje na to da ipak postoje čestice koje su, jednostavno rečeno, „više čestice" od drugih. Stoga nije teško zamisliti razgraničenje prema kojem se samo takvi leksemi kategoriziraju kao čestice.

Pristup koji se predlaže u ovom radu usmjeren je na pronalaženje pragmatičnog rješenja kojim se nastoji eliminirati što više semantičkih kriterija te bitno umanjiti broj kompliciranih sintaktičkih pravila. To se može postići tako da se riječi koje se u gramatikama sekundarno kategoriziraju kao čestice jednostavno u svim slučajevima svrstaju u njihovu primarnu vrstu riječi.

Drugim riječima, predlaže se ograničavanje čestica za potrebe POS označavanja konkretno na čestice u užem smislu, odnosno na lekseme koji se ne mogu svrstati ni u jednu drugu postojeću vrstu riječi, kao što su primjeri koje Pranjković navodi za „nesamostalne" čestice. U pravilu, ako leksem u funkciji čestice već pripada drugoj vrsti riječi i uvijek je zamjenjiv njome, ne bi ga trebalo smatrati česticom.

## 5. Odabrane čestice

Kad sagledamo prethodno obrađene gramatike kao cjelinu, vidimo da se u njima u „užoj" kategoriji čestica ponavljaju većinom isti leksemi: *da (jest), ne, li, zar, god, ma, makar, bar, barem, čak, pa, ta, bilo, evo, eto, eno te de, deder, daj, neka* i *put (puta)*.

Iako je taj popis već poprilično kratak, ako se vodimo kriterijem da su čestice riječi koje su gramatikalizirane i već ne pripadaju drugoj vrsti riječi, možemo eliminirati riječi koje su zapravo glagoli (*jest, de, deder, daj*). U tom slučaju kao čestice za potrebe ovog rada preostaju *da, ne, li, zar, god, ma, makar, bar, barem, čak, pa, ta, bilo, neka, put (puta), evo, eto* i *eno*. Među njima postoje leksemi koji su homonimni s veznicima (*da, pa*), zamjenicama (*ta, neka*), imenicama (*bar, god, bilo, put*) i glagolima (*bilo*), no ipak je riječ o različitim vrstama riječi, što se pobliže analizira u nastavku na primjerima iz korpusa hrWaC (http://nlp.ffzg.hr/resources/corpora/hrwac/).

### 5.1. Li

***Li*** je upitna čestica koja u svim prethodno obrađenim gramatikama spada u čestice u užem smislu. Odgovara svim relevantnim kriterijima, odnosno nepromjenjiva je, nesamostalna i gramatikalizirana riječ koja sudjeluje u oblikovanju gramatičkog ustrojstva iskaza i ne pripada nijednoj drugoj vrsti riječi. Stoga za taj leksem nema previše sumnje po pitanju njegove kategorizacije kao čestice neovisno o položaju u rečenici i leksemima uz koje se javlja, kao što možemo vidjeti u sljedećim primjerima:

1. Vjeruje *li* on da je normalno uvoziti ugljen s drugog kraja svijeta, dok ekološki pogodniji plin teče plinovodom nedaleko Plomina?
2. Na trenutak smo zaboravile nervozu kada smo vidjele poznata lica ljudi koji su doputovali iz različitih zemalja ne bi *li* se okupili tu.
3. Jesi *li* ikad pomislila da je postavljanje pitanja samo sebi svrhom?
4. Potkrijepi svoje dokaze svakako i pokojom fotkom i otkrij jesi *li* nas uspio uvjeriti da jedne od 8 pari jedinstvenih tenki završe na tvojim nogama.
5. Nije *li* krajnje vrijeme da shvatimo pravu istinu?
6. Svakog dana brojio je svoje kovčege da vidi nije *li* uz put što izgubio.

### 5.2. Zar

Leksem ***zar***, baš kao i leksem *li*, ispunjava sve navedene kriterije i stoga je primjer čestice u užem smislu, odnosno „prave" čestice, kao što možemo vidjeti u sljedećim primjerima:

7. Sve to će Vam omogućiti da je vrijeme na Vašoj strani, a to ste i htjeli, *zar* ne?
8. *Zar* su oni bili zločinci koje je trebalo kazniti?
9. *Zar* dvije žene mogu biti blesave a on koji laže pošten?
10. *Zar* ćemo ih ostaviti da čame u tom mračnom zatvoru?
11. Pa *zar* ne vidiš da se glavni grad vise ne zove Titograd a država ne pripada Jugoslaviji?
12. I *zar* netko ima potrebu pisati pjesme i u sretnim razdobljima?

### 5.3. Ne

Leksem ***ne*** još je jedan primjer „prave" čestice. Na Hrvatskom jezičnom portalu već se kao prva definicija navodi da je on „riječ za odricanje, negira glagolsku radnju ili značenje drugih punoznačnih riječi". Kao i prethodne dvije čestice, odgovara svim relevantnim kriterijima. Pretraživanjem korpusa hrWaC nalazimo sljedeće primjere:

13. Zašto *ne*?
14. Ta na njemu je da nas uči, pokaže primjerom i vodi, zar *ne*?
15. *Ne*, neću dopustiti da zbog nekog ili nečeg padnem u napast.
16. *Ne* radi se o tome da veliki kreativni geniji pomažu običnom lijenom puku, nego su obični poreznici ti koji pomažu propalim kreativnim genijima.
17. *Ne* ignorirajte osjećaj hladnoće kao ni drhtavicu koji nas upozoravaju da se trebamo utopliti ili prekinuti boravak na hladnoći.
18. Naravno, *ne* smije nigdje ulaziti voda ili vlaga, isto vrijedi i za slamu.
19. Naime, ona je obvezni dokument, bez kojeg dijete *ne* može polaziti dječji vrtić.
20. Vjerovali ili *ne*, i takve stvari su se događale.

Može se nalaziti u bilo kojem dijelu rečenice, npr. na kraju rečenice (primjeri 13. i 14.) ili surečenice (primjer 20.), na početku rečenice (primjeri 15., 16. i 17.), ispred rečeničnih znakova kao što je zarez (15. i 20.) ili iza njih (18.) ili bez rečeničnih znakova oko nje (19.). Često se

javlja neposredno ispred glagola, koje negira (16., 17., 18. i 19.), te se upotrebljava za negiranje imperativa kako bi se naznačilo zabranu ili upozorenje (primjer 17.). Osim toga, u korpusu se vrlo često javlja iza čestice *zar* (primjer 14.) za oblikovanje pitanja na koje se očekuje potvrdan odgovor. Kao što vidimo iz navedenih primjera, položaj u rečenici, interpunkcija, vrste i oblici riječi uz koje se javlja ili bilo koji drugi sintaktički kriteriji ne utječu na njegov status čestice.

### 5.4. Da

I leksem ***da*** je čestica u užem smislu u svim obrađenim gramatikama, no javlja se i kao veznik. Te dvije funkcije svakako treba odvojiti za POS označavanje jer je razlika između njih prevelika da bi se zanemarila – doista je riječ o različitim vrstama riječi. Usto, ima i homograf koji je glagol (3. l. jd. pz. glagola *dati*). Pretraživanjem korpusa hrWaC kao često kontekstualno pravilo za leksem *da* pokazuje se da je kao čestica u pravilu okružen pravopisnim znakovima jer se često nalazi na početku rečenice:

21. *Da*, ja sam vjernik.
22. *Da*, točan i pravilan naziv je ta umanjenica.
23. *Da*, postoje trenuci snažnog kolektivnog sudjelovanja kada lokalne zajednice raspravljaju i odlučuju…

Isto vrijedi i za rečenice u kojima mu na početku rečenice prethodi uzvik ili veznik *i*, a slijedi ga rečenični znak:

24. Ah *da*, UN dijeta, sad si negdi mršav i treperiš na svakom povjetarcu…
25. Istražili smo. I *da*, istina je.

Dakle, na početku rečenice riječ je o čestici ako se iza nje nalaze zarez ili drugi rečenični znak. Osim toga, ako se nalazi na početku rečenice, to znači da su obično ispred nje točka, upitnik ili uskličnik koji naznačuju kraj prethodne rečenice. Stoga jedan od kriterija može biti okruženost rečeničnim znakovima s obiju strana, jer se inače radi o vezniku ako rečeničnih znakova nema ili ako se nalaze samo s jedne strane, s obzirom na to da se kao veznik u pravilu ne javlja u istom okruženju:

26. *Da* sam jučer imao sjekiru, sada bi kuhali fini ručak.

27. Dogradonačelnik Željko Sikirica napominje kako i dalje ostaje obveza Grada *da* naručivanjem poslova pomogne DES-u te *da*, iako su i u tom pogledu poduzeti koraci, treba puno bolje raditi.
28. Budući *da* se radilo o hitnom slučaju jer začepljenje mokraćovoda vrlo brzo može teško oštetiti bubrege, pas je bez odlaganja operiran…
29. Ta procjena je bila *da* je branša izuzetno dinamična, osjetljiva na pogreške, *da* se globalno prate svi parametri na domaćem tržištu i *da* se donose strateške odluke za razvoj regije s naglaskom na brigu oko kvalitete proizvoda i svih procesa.

U službi veznika često se nalazi u sintaktičkom okruženju bez pravopisnih znakova, što ima smisla s obzirom na to da povezuje surečenice. Postoje i rečenice kao što je *Rekao je da*., što je temelj za sljedeće pravilo: ako se nalazi na kraju rečenice, jednako se tako ne radi o vezniku, nego o čestici. Kad povrh toga uzmemo u obzir slučajeve u kojima se rečenica nastavlja, npr. *Rekao je da, ali je dodao da nije siguran.*, možemo pravilu dodati da se radi o čestici ne samo na kraju rečenica, nego i surečenica, koje su obično odvojene rečeničnim znakovima ili veznicima.

No i za ovo pravilo naravno postoje iznimke, primjerice u slučaju nizanja više surečenica ili u rečenicama s umetnutim rečenicama:

30. …nema neke zagonetke u tome što je Lippmann tvrdio, radi se o činjenici; tajna leži u tome *da*, svjesni nje, igramo igru.

Moguće rješenje mogao bi biti uvjet da se iza *da* i zareza nalazi suprotni veznik, kao u primjeru *Rekao je da, ali je dodao da nije siguran*. Za taj bi se kriterij problematičnim mogao pokazati suprotni veznik *dok* jer ovisno o kontekstu može označavati i zavisnu vremensku surečenicu koja nosi značenje istovremenosti (npr. *Često se može čuti da, dok su mladi, ljudi ne brinu o zdravlju*.).

No treba uzeti u obzir i objektivnu učestalost takvih rečenica u jeziku. Budući da hrvatski jezik ima slobodan redoslijed riječi, nekad je gotovo nemoguće sintaktičkim pravilima obuhvatiti sve slučajeve. Cilj bi trebao biti obuhvatiti barem većinu kako bi se s pragmatičkog stajališta postigla najveća moguća učinkovitost kriterija.

Posljednji problem koji se nameće je homograf 3. l. jd. pz. glagola *dati*. Na primjer, *da* se javlja i kao veznik i kao glagol u rečenici *Tri i pol godine čekam da on **da** gol, ali je zaista milina raditi s ovim momcima*. Općenito je teško zaključiti u kojim se sintaktičkim okruženjima sigurno radi o glagolu. Primjerice, ako kriterij želimo temeljiti na činjenici da je riječ o prijelaznom glagolu, kriterij bi mogao biti da se (neposredno) iza njega nalazi imenica ili zamjenica u dativu (za neizravni objekt) ili akuzativu (za izravni objekt). No možemo zamisliti rečenice kao što su *Ne želim **da** psu daješ čokoladu.* i *Jasno mi je **da** ga ne voliš*. U oba je slučaja riječ o vezniku. Dodatan bi kriterij mogao biti i da se (neposredno) ispred njega nalazi imenica ili zamjenica u nominativu (kao u prethodno navedenom primjeru *Tri i pol godine čekam da **on** **da** gol*) te, u slučaju da je zamjenica ili imenica izbačena, da se homograf *da* ponavlja dva puta za redom (*Tri i pol godine čekam **da da** gol*).

Na koncu, ako su pravila za *da* kao česticu i *da* kao veznik dovoljno detaljno razrađena, *da* bi se kao glagol jednostavno mogao označiti u svim preostalim slučajevima. Tako u prethodnim primjerima možemo vidjeti da se *da* kao veznik javlja na početku surečenice, ali i neposredno iza glagola. U takvom okruženju u načelu možemo znati da nije riječ o glagolu, već o vezniku. Detaljnijom analizom korpusa u opsegu većeg rada ili projekta bilo bi moguće pronaći još sintaktičkih obrazaca koji bi poslužili kao temelj za uspješno razgraničavanje svih upotreba.

### 5.5. God

Čestica ***god*** još je jedan primjer čestice u užem smislu. Ima homograf u značenju prstena u deblu stabla, no on se u pravilu javlja u množini (*godovi*) te stoga ne bi trebao predstavljati prevelik problem za POS označavanje. Kao čestica se obično kategorizira kao intenzifikator, iako semantički nema jednaku funkciju kao primjerice *čak*. U korpusu se najčešće javlja iza leksema *kad(a), kako, koliko, kakav, koji, kud(a), kamo, što, tko, dok, dokle* i sl. (kao što vidimo u primjerima u nastavku), iako se između njih može naći pomoćni glagol ili (povratna) zamjenica (kao u primjerima 39. i 40.):

31. Teritorij mačke ne mijenja se dok *god* ona dominira njime.
32. Kad *god* dođete, atmosfera koja ih obavija vrlo je ugodna, opuštena, ali i radna u svim mogućim smislovima.
33. Vozi se biciklom kud *god* možeš; idi na posao ili u školu s biciklom umjesto javnim prijevozom ili autom.

34. Što *god* odabrali kao svoju životnu misiju bolje ćemo to činiti ako to radimo punim prsima.
35. U koju *god* kuću uđu neka najprije kažu: «Mir kući ovoj».
36. Tko *god* prihvati da nečiju zemlju zasadi lozom, mora je dva puta godišnje okopavati i jednom godišnje obrezivati, pod kaznom od četiri libre.
37. Ne fascinira sposobnost Samsunga da izvodi kopije, koliko fascinira sposobnost Applea da je uvijek prvi u nametanju trendova, koliko *god* oni kasnije izgledali „očigledno" i „banalno".
38. Imenujte emocije, kakve *god* one bile, i ohrabrujte svoje dijete da čini isto.
39. Koliko je *god* to moguće s obzirom na trenutnu (ne) zakonsku situaciju nastojat ćemo da se u njemu nađu kvalificirani stručnjaci te ćemo ga maksimalno promovirati putem javnih medija.
40. Koliko se *god* odgovor činio teškim i slojevitim, on se može ipak sažeti u jednu jedinu riječ, a to je – sloboda.

### 5.6. Pa

Upotrebu leksema ***pa*** kao čestice možemo prikazati na primjerima od 41. do 45. Leksem *pa* čestica je upravo u takvim kontekstima, na početku rečenice, dok se inače smatra veznikom:

41. *Pa* kako to da si se baš ti odmetnuo od crkve?
42. *Pa* nemamo mi vremena za to.
43. *Pa* pobila sam pet ljudi, kako ne bih prihvatila takvu ulogu?
44. *Pa* moguće je, zašto ne, ali sam ovdje pred svima vama prije svega jer želim demantirati priču po kojoj me navodno nikakvi milijuni ne mogu zadržati u Splitu i Hajduku.
45. Molim? *Pa* nisam ja to izmislila. Pogledajte u rječnik.

Ovdje se kao problem nameće činjenica što jednako tako možemo zamisliti te iste rečenice s veznikom *ali* umjesto *pa*, što znači da je zapravo zamjenjiv veznikom. Ipak, istina je da leksem *pa* kao veznik i *pa* kao čestica nisu zapravo ista riječ, što je svakako argument u prilog njegovu

uključivanju u popis čestica, a javljanje na početku rečenice čini se kao dovoljan kriterij. To možemo vidjeti i u sljedećim primjerima u kojima se ne javlja na početku rečenice i riječ je o vezniku:

46. Izvorni je tekst toga svjetskog dokumenta opsežan, ima članaka, pisali su ga pravnici, *pa* je mnogima „težak".
47. Nakon završenog dokaznog postupka te govora tužitelja i oštećenika prvo govori branitelj pravne osobe, *pa* predstavnik pravne osobe, a zatim branitelj odgovorne osobe *pa* odgovorna osoba.
48. Planinarskim putem smo se spustili do ceste i nekim starim planinarskim putem došli do mjesta Trstenik, *pa* onda cestom do sela Račja vas.

Budući da se kao veznik inače ne javlja na početku rečenice, ako se vodimo ovim kriterijem, vjerojatno možemo točno označiti veliku većinu slučajeva.

### 5.7. Ta

Ovaj je slučaj problematičniji od prethodnog jer se riječ ***ta*** na početku rečenice može javiti i kao čestica i kao zamjenica. Kontekstualno gledajući također je teško odrediti koja se vrsta riječi javlja iza koje kategorije jer se iza obaju leksema može nalaziti više vrsta riječi. Razmotrimo sljedeće primjere čestica:

49. *Ta* nisam je namjerno zgazio ako je zgazim.
50. *Ta* ona mu je i ostavila otvorena vrata od kuće.
51. *Ta* svećenik je naš duhovni primjer.
52. Ali tko će po toj magli, mrazu; *ta* psi ne mogu baš ništa za takova dana, a neće mi se uzalud po bregovima hodati.
53. *Ta* nije Crkva bez razloga uzela pacijente po bolnicama živim blagom Crkve.
54. …još uvijek ne razumijem zašto me poslao, *ta* nije li rekao da imam važnog posla…
55. *Ta* bila je to žena od devedeset godina koja je često bila samo teret i teško sam nalazio za nju ljubavi već samo neku obvezu pomaganja.

56. *Ta* vidjela sam ja i dosad gospode.
57. *Ta* na njemu je da nas uči, pokaže primjerom i vodi.

Usporedimo prethodne primjere s primjerima zamjenica:

58. Ja sam vodio jednu Zagrepčanku prije par godina i *ta* nije prestajala slikat džamije i klanjanja.
59. Nije joj *ta* bila ni do koljena.
60. Ovo nije šala, ja doslovno ne znam što da radim. *Ta* nije normalna.
61. *Ta* žena je zaista napravila čudo.
62. *Ta* su prava službeno zapisana u zajedničkom svjetskom dokumentu koji se zove „KONVENCIJA UJEDINJENIH NARODA O PRAVIMA DJETETA".
63. *Ta* je ljubav stara koliko i njihovo postojanje i slavimo je svaki dan uz zalogaj sira i kapljicu vina.

Na temelju primjera iz korpusa hrWaC primjećujemo nekoliko obrazaca. Ako iza *ta* slijedi imenica koja nije u nominativu ženskog roda jednine, sigurno je riječ o čestici – možemo općenito uključiti kriterij neslaganja u rodu, broju i padežu. U tom slučaju za rečenice kao što je primjer 51. ne bi bilo dvojbe da je riječ o čestici. Naravno, vrlo lako možemo zamisliti i rečenicu jednaku navedenoj u kojoj umjesto *svećenik* stoji *svećenica*, no takvi se slučajevi mogu isključiti na temelju toga što bi i čovjeku u tekstualnom obliku, bez pomoći intonacije, mogli biti dvosmisleni zbog slaganja u rodu, broju i padežu.

Još jedan obrazac koji vidimo u navedenim primjerima jest da je sigurno riječ o zamjenici ako se neposredno iza *ta* nalazi pomoćni glagol u potvrdnom obliku (primjeri 62. i 63.). Obrnuto pravilo ne možemo s jednakom sigurnošću primijeniti jer se pomoćni glagol u niječnom obliku može javiti i iza zamjenica (primjeri 58. i 60.) i iza čestica (primjeri 53. i 54.). No ako neposredno nakon tog niječnog pomoćnog glagola slijedi imenica ili zamjenica u nominativu (kao u primjeru 53.) uglavnom će biti riječ o čestici, a uvijek je riječ o čestici ako je taj glagol u množini (zbog neslaganja) ili ako ga slijedi čestica *li* (primjer 54.). Jednako tako, uvijek je riječ o čestici ako neposredno iza *ta* slijedi zamjenica u nominativu, neovisno o rodu i broju zamjenice (primjer 50.).

Naravno, za sve ovo postoje i iznimke. Razgraničenje između tih dviju kategorija zapravo se najviše oslanja na semantiku i intonaciju, što otežava njihovo razlikovanje na temelju sintaktičkih pravila. No problem koji se nameće ako ne uključimo *ta* kao česticu jest to što će se onda u svim navedenim slučajevima na temelju oblika analizirati kao pokazna zamjenica, što je naprosto netočno. Stoga je svakako potrebno uključiti *ta* u popis čestica.

### 5.8. Barem

Leksem ***barem*** može se ponašati i kao prilog i kao čestica, ali radi jednostavnosti kategorizacije bilo bi bolje označiti ga samo kao česticu. Razmotrimo sljedeće primjere:

64. Ako nakon čitanja ovih savjeta promijenite način života i ponašanja u *barem* jednome detalju, ona je postigla svoju svrhu.
65. *Barem* na jedan dan zaboravite brige, ponesite kupaći kostim i otputujte s nama na obližnje otoke oko otoka Lošinja.
66. Zato u desetom kolu nećemo fantazirati, *barem* ne onako kako mislite.
67. Situaciju dodatno pogoršava veliki uvoz namještaja koji permanentno raste, a domaći potrošači u nedostatku hrvatskog namještaja koji bi *barem* približno slijedio trendove u stanovanju izabiru uvozne proizvode.
68. Svi volimo *barem* otprilike znati kako će se osoba s kojom smo u dodiru postaviti u nekoj situaciji, kako će reagirati i koliko na tu osobu možemo računati.
69. Ja to mislim iz čistog prijateljstva i želje da ti pomognem da ti se i jedna osoba koja *barem* od oka poznaje situaciju prestane smijati…
70. Nadam se da ćemo u prvoj godini uspjeti *barem* okvirno snimiti stanje i iz te slike izvući upute za ono što bi trebalo činiti dalje.

U primjerima 64. i 65. leksem *barem* zamjenjiv je prilogom *najmanje*, odnosno ponaša se kao prilog u tim rečenicama. Međutim, takva zamjena nije moguća u preostalim primjerima i teško je u njima zamisliti bilo koju drugu vrstu riječi na njegovu mjestu. Sintaktički gledano ne nameće se obrazac koji bi dosljedno mogao razgraničiti te dvije upotrebe, te je stoga s pragmatičkog stajališta vjerojatno najbolje odlučiti se samo za jednu od tih vrsta riječi. U ovom je slučaju odabrana kategorija čestica jer se leksem prečesto javlja u konstrukcijama u kojima nije zamjenjiv prilozima.

### 5.9. Bar

Čestica ***bar*** ponaša se isto kao čestica *barem* i uvijek je zamjenjiva njome. Glavna je razlika među njima to što leksem *bar* ima nekoliko homografa, pa je za potrebe POS označavanja potrebno razgraničiti upotrebe tih leksema kontekstualnim pravilima. Kao čestica se javlja u sljedećim primjerima:

71. Pitala sam majku u savjetovalištu, koja nije bila zadovoljna uspjesima svoga sina, da se dosjeti *bar* jedne njegove dobre osobine ili uspješnog ponašanja.
72. Mačka je posebna, *bar* što se tiče prehrambenih potreba.
73. Osamdeset i sedma minuta je i tu je priči kraj, *bar* što se tiče regularnog dijela.
74. Često vlasnici pitaju da li je važno da mačka ima jedno leglo, ali ne postoji niti jedan zdravstveni razlog zbog kojeg bi bilo bolje da vaša mačka ima *bar* jedno leglo prije nego što se sterilizira.
75. *Bar* nam tuđi ne otimaju novce.

Njegovi su homografi *bar* u značenju *pub* i *bar* kao mjerna jedinica te izrazi *mini bar* i *bar kod*:

76. U listopadu, točnije 10. listopada 2009. g. na sjevernoj strani dvorane otvoren je caffe *bar* Golden VIP koji je potpuno u vlasništvu Športskog centra Višnjik d.o.o.
77. Novouređeni Maraschino *Bar* Varaždin i Lounge *Bar* TTS Sport Centra postali su glavna točka okupljanja svih uzrasta stanovnika grada Varaždina.
78. …morska površina = 1 *bar*, 10 m = 2 bara, 20 m = 3 bara…
79. Na obje etaže nalaze se SAT TV, mini *bar*, telefon, klimatizacijski uređaj, sef te je omogućen pristup internetu.
80. Kontinuirani način rada prenosi podatke svaki puta kad *bar* kod uđe u polje očitanja.
81. Aplikacija funkcionira tako da korisnik snimi *bar* kod proizvoda mobitelom te aplikacija daje korisniku sadržaj proizvoda i njegovo objašnjenje.
82. Idemo u *bar*!
83. Sporna reklama prikazuje tri žene koje ulaze u *bar* i naručuju Jägermeister.

Homografe u izrazima *lounge bar* i *caffe bar* vrlo je lako riješiti kontekstualnim pravilom koje nalaže da je riječ o imenici ako se nalazi iza leksema *lounge* i *caffe* (i ostalih varijacija tog leksema). No kao što vidimo u primjeru 77., pravilo bi se moglo proširiti na strane riječi općenito jer su imena barova često upravo takve riječi. Za slučajeve u kojima je ime bara hrvatska riječ opcije su poprilično ograničene zbog elementa nepredvidljivosti i činjenice da se čestica *bar* može naći i iza vlastite imenice, pa je najbolje rješenje kriterij prema kojem obje riječi imaju veliko početno slovo (*Imenica u nominativu* + *Bar*).

Napokon, leksem *bar* u značenju *pub*, kada nije riječ o imenima ugostiteljskih objekata, ne javlja se vrlo često u istom obliku kao čestica. Naime, češće ćemo naići na rečenice koje sadrže oblike *baru/barovima*, no za slučajeve u kojima se u tom značenju javlja u obliku *bar* isto je moguće odrediti sintaktičko pravilo koje bi u većini slučajeva ispravno razlikovalo predmetne homografe. Primjerice, ako se ispred njega nalazi prijedlog *u*, a iza njega rečenični znak ili veznik (kao u primjerima 82. i 83.).

Za niz *bar kod* na prvi se pogled čini da bi se mogao riješiti istom vrstom pravila kao za *caffe bar*, no on se može javiti i u rečenicama kao što je *Bar kod njega uvijek ima hrane*. Za takve slučajeve moglo bi se uključiti sintaktičko pravilo koje upućuje na kontekste u kojima *kod* iza leksema *bar* nije imenica, kao što je ovdje padež imenice ili zamjenice koja ga slijedi (genitiv). Međutim, takvim se pravilom ne isključuju slučajevi kao što je rečenica u primjeru 81. Detaljnijim pretraživanjem korpusa bilo bi moguće sastaviti popis leksema u genitivu koji se često javljaju iza niza *bar kod* kad je riječ o imenici, a ne čestici, kao što su leksemi *proizvod* i *artikl*. Takvim bi se pravilom uspješno obuhvatila većina slučajeva.

Napokon, što se tiče homografa *bar* u značenju mjerne jedinice, većina se slučajeva može obuhvatiti pravilom da se javlja iza znamenki.

### 5.10. Makar

Leksem ***makar*** je ovisno o kontekstu zamjenjiv česticom *bar/barem*, kao u primjerima 84. i 85. ili veznikom *iako* (ili mu je značenjem i funkcijom sličan), kao u primjerima 86. i 87.:

84. Kad bi uspjeli spasiti *makar* pola bačene hrane, učinak na klimu bio bi isti kao da uklonimo četvrtinu svih automobila.
85. Natječaj je valjan ako pristigne *makar* i samo jedna valjana ponuda.
86. *Makar* za to i dobio gore spomenutih 100 milijuna kuna, to nema smisla.

87. Trebaju gledati i svijet oko sebe, *makar* su još uvijek male.

Kao što vidimo iz navedenih primjera, riječ je o funkcijama koje se dovoljno razlikuju da ih je potrebno razgraničiti. U načelu, ako se *makar* nalazi na početku rečenice ili surečenice, riječ je o vezniku, a u preostalim slučajevima u pitanju je čestica. Daljnjim pretraživanjem korpusa mogle bi se naći moguće iznimke i na temelju njih izraditi detaljnija pravila.

### 5.11. Čak

Na Hrvatskom jezičnom portalu leksem ***čak*** definiran je kao prilog i čestica „za isticanje i za dodavanje onome što je rečeno; dapače, štoviše". Razmotrimo sljedeće primjere:

88. *Čak* i brojni Sudanci ne znaju za njih jer se u školi povijest uči od dolaska islama na dalje.
89. Od ostalih golova, *čak* dvaput se na listi našao Oscar.
90. Zakon o prebivalištu je potrebno realizirati jer je već 20 godina prošlo od osnutka Hrvatske države, a još uvijek postoje mogućnosti manipulacije koje *čak* u nekim slučajevima mogu utjecati na izborne rezultate.
91. Ustvari, *čak* se ne biste trebali kupati niti tuširati u vodi koja dolazi direktno iz slavine.
92. *Čak* su i članovi i/ili male grupe u međunarodnim humanitarnim organizacijama i snagama UN bile optužene za sudjelovanje u takvim nečasnim aktivnostima, a u nekim su slučajevima i proglašeni krivima.
93. *Čak* ni kad bismo razgovarali u istoj prostoriji, ne bismo imali isti trenutak življenja jer smo prožeti brojnim i različitim prošlim zbivanjima.

U nekim od navedenih primjera zamjećujemo da, iako je moguće leksem *čak* zamijeniti drugim prilozima, pritom dolazi do izmjene semantičkih i pragmatičkih aspekata tih rečenica, odnosno redoslijeda elemenata i značenja cjeline.

Činjenica da su u definiciji na Hrvatskom jezičnom portalu kao parasinonimi tog leksema navedeni prilozi *dapače* i *štoviše* ide u prilog njegovoj kategorizaciji kao priloga. No je li on doista dosljedno zamjenjiv takvim prilozima? To nije slučaj u sljedećim primjerima:

94. Za nastup na današnjem treningu prijavljena su bila *čak* 84 skijaša.
95. Dakle, *čak* i kada bismo otkrili da postoji nekakav vanjski razlog našeg postojanja, to ne bi mogao biti dobar kandidat za odgovor o smislu života.
96. Maja živi blizu Sesveta i danas svi za nju kažu da je druga osoba, *čak* i izgledom.
97. Popuštanje napetosti se može obavljati na velikim zavarenim i drugim konstrukcijama teškim *čak* do 18.000 kg.

Za ovaj je leksem konkretno teže odlučiti u koju ga kategoriju svrstati. Svakako se čini suvišnim kategorizirati ga i kao prilog i kao česticu jer razgraničenje njegove upotrebe između tih dvaju slučajeva obično ovisi o interpretaciji (npr. u rečenicama kao u primjeru 91. i sl. u kojima se može zamisliti prilog na njegovu mjestu, doduše ne s istim semantičkim ishodom), a ne nužno o nekim sintaktičkim obrascima koji bi sustavno i neupitno mogli uputiti računalo u njihovo dosljedno razlikovanje. Radi jednostavnosti kategorizacije ćemo ga za potrebe ovog rada kategorizirati samo kao česticu.

### 5.12. Bilo

Čestica ***bilo*** problematična je jer je homografna s jednim od oblika glagola *biti* (3. l. jd. sr. r. prf.) i imenicom *bilo (puls)*. Osim toga postoji i (korelativan) veznik *bilo...bilo*. Etimološki gledano riječ je izvedena iz glagola *biti* pa bi se potencijalno mogla i tretirati kao njegov oblik za potrebe POS označavanja. Takvu bi upotrebu doduše svakako trebalo razgraničiti od korelativnog veznika, ali detalji takvog razgraničenja nisu obuhvaćeni opsegom ovog rada.

Što se tiče konkretne upotrebe kao čestice, u konstrukcijama kao što je *bilo kakav* ponaša se kao čestica i ne čini se točnim opisati ga kao prilog. Naime, možemo zamijetiti da se ponaša slično kao čestica *god*, a glavna je razlika u redoslijedu elemenata (*bilo* kakav / kakav *god* i sl.). Shodno tome, može se uvesti pravilo prema kojem se označava kao čestica isključivo neposredno ispred konkretnog popisa leksema (*koji, kakav, čiji, što, tko, gdje, kamo, kud(a), kad(a)* itd.) bez rečeničnih znakova između njih, kao u sljedećim primjerima:

98. Dijete ima pravo tražiti da država spriječi njegovo nezakonito preseljenje i zadržavanje u inozemstvu od strane jednoga roditelja ili *bilo* koga drugoga.

99. Dijete koje je žrtva *bilo* kakva nehaja, izrabljivanja, zlouporabe ili oružanog sukoba ima pravo na odgovarajući tjelesni i duševni oporavak i ponovno uklapanje u društvo.
100. Klijent može definirati projekt suradnje po želji na *bilo* kojem procesu.
101. Tako nudimo kompletnu uslugu uz montažu na terenu, te vlastitu dostavu *bilo* gdje u Istri, a spremni smo pronaći rješenje i za udaljenije destinacije.
102. Tele2 mobilni internet na bonove omogućuje ti jednostavan i brz pristup internetu s računala *bilo* gdje i *bilo* kada, bez sklapanja ugovornog odnosa.
103. Glas ljubavi je na svojoj probi pronašao dva auta koji bi ih vozili negdje da nastupaju. Pitanje je *bilo* gdje?

Kao što vidimo, predloženo pravilo obuhvatilo bi veliku većinu slučajeva. Naravno, i za njega postoje neke iznimke, kao što je druga rečenica u primjeru 103. Iako je u pitanju niz *bilo gdje*, nije riječ o čestici, nego o obliku glagola *biti*. Međutim, možemo postaviti kontekstualno pravilo kojim bi se i takvi slučajevi točno označili, primjerice da je riječ o glagolu, a ne o čestici, ako se ispred niza *bilo* + *gdje, koji, kakav* i sl. nalazi niz *pitanje je*. Detaljnijom analizom i pretraživanjem korpusa taj bi se popis mogao dopuniti drugim odgovarajućim nizovima ako postoje te tako riješiti problem.

### 5.13. Ma

Upotreba leksema ***ma*** također je često slična upotrebi čestice *god*, utoliko što nose slično značenje i javljaju se u sličnom ko-tekstualnom okruženju, konkretno uz lekseme *koliko, kako, gdje* i sl. U takvim je slučajevima, kao i u slučaju čestice *bilo*, glavna razlika to što se *ma* javlja ispred, a *god* iza tih leksema:

104. Moguće ozljede nastale nestručnim obrezivanjem, *ma* koliko sitne bile i oku nevidljive mogu biti ulaz za mnoge patogene bakterije…
105. *Ma* kako bilo, na pravom ste mjestu da – po prvi puta ili ponovno – otkrijete svoj glas.
106. Kristov vjernik nosi vjeru *ma* gdje bio, unosi je u sve što čini.

U drugim se slučajevima koristi za naglašavanje ili kao znak odbijanja ili negiranja, obično na početku rečenica ili surečenica:

107. *Ma* nemoj!
108. *Ma* to je neobrazovano do kosti.
109. *Ma* bolje da ne kažem ništa i svakom prepustim na mašti nastavak ove započete rečenice.
110. Bio sam razigran, bahat, nepromišljen, *ma* bolje reći – pokvarenjak.

U takvim se slučajevima može smatrati i uzvikom, ali je za potrebe POS označavanja vjerojatno dovoljno kategorizirati taj leksem samo kao česticu.

### 5.14. Put(a)

Čestica ***put(a)*** ima dva homografa, a oba su imenice – imenica ženskog roda, najčešće u odnosu na ljudsko tijelo ili kožu, i imenica muškog roda u značenju putanje, staze i sl. Imenica je homografna samo u oblicima *put* i *puta* jer se čestica morfološki ne mijenja na isti način kao imenica. Stoga odmah možemo zanemariti oblike kao *puti, putevi*, *puteva* i sl.

Općenito se kao dobro početno pravilo za razgraničenje imenice i čestice čini javljanje čestice neposredno iza brojeva i rednih brojeva ili znamenki (primjeri 111., 112. i 113.) te priloga (primjeri 116. i 117.). Između njih se mogu naći i pomoćni glagol ili povratna zamjenica, kao u primjerima 114., 115. i 118.

111. Dva *puta* sam ih radila i oba *puta* su završili u mrvicama, ja u plaču.
112. U Rijeci nisam bio omiljen od početka karijere, a u Split su me prvi *put* pozvali tek prošle godine.
113. Ponovite vježbu po 15 *puta* na svakoj nozi.
114. Institucije Europske unije prvi su *puta* objavile natječaj za hrvatske konferencijske prevoditelje…
115. Ime Vodice prvi se *put* spominje 1402. godine u dokumentu izdanom u Šibeniku.
116. Kako bi uspjeli, moramo više *puta* pogriješiti.
117. Otac ga je vadio nekoliko *puta* iz zatvora, koristeći svoj ugled i svoja poznanstva sa ljudima na položaju…
118. Više se *puta* čula ta brojka kao informacija i iz ministarstva poljoprivrede.

Naravno, i za to se mogu naći iznimke:

119. I bez mene je svijet pun loših pjesnika pa sam odabrao drugi *put*.

Za takve bi se primjere također moglo analizom uzoraka iz korpusa utvrditi postoje li leksemi koji se opetovano javljaju u kontekstualnom okruženju kad je riječ o imenici, kao što je u ovom primjeru glagol *odabrati*, te na temelju toga sastaviti popis takvih leksema i konkretnih konstrukcija u kojima se javljaju.

Slaganje u broju za određene brojeve nije isto za imenicu i česticu. Usporedimo, na primjer, nizove *pet puta* i *pet puteva*. U prvom je slučaju riječ o čestici, a u drugom o imenici. Tako bi još jedan kriterij za čestice mogao biti upravo ta razlika u obliku kad se javljaju uz određene brojeve (brojeve veće od četiri sve do 20, a zatim brojeve veće od 24 do 30, pa veće od 34 do 40 itd.). Naravno, usto bi trebalo uzeti u obzir i vjerojatnost da će se u korpusu bilo koji veći broj odnositi na puteve u značenju staza, pa bi se u tom pogledu analizom uzoraka iz korpusa mogao pronaći i broj koji bi se mogao primijeniti kao gornja granica za upotrebu uz imenicu.

Kad se javlja nakon zamjenica može biti riječ i o imenici i o čestici, kao što vidimo u sljedećim primjerima:

120. Iako dečki ovaj *put* nisu osvojili medalje, vjerujem da će se sav njihov trud ubrzo isplatiti.
121. Najbolje vrijeme za krenuti na taj *put* je krajem lipnja.

I ove bi se upotrebe mogle razgraničiti usporedbom učestalosti zamjenica koje se javljaju uz *put* kao imenicu i kao česticu. Tako se primjerice u korpusu zamjenica *ovaj* pretežno javlja uz česticu, a *taj* uz imenicu.

### 5.15. Nek(a)

Čestica ***nek(a)*** homografna je sa zamjenicom *neka* i veznikom *nek(a)*. Kao čestica se obično upotrebljava za tvorbu imperativa (primjeri 122., 123., 124., 125., 131. i 132.).

122. *Neka* se ne uznemiruje vaše srce i *neka* se ne straši.
123. U ovoj godini vjere *neka* nas sve prati blagoslov svemogućega Boga Oca i Sina i Duha Svetoga.
124. *Neka* Gospodin blagoslovi sve koji rade u ovom djelu naše pastve kao i one koji nas čitaju i prate.
125. *Neka* ti ni jedan dan ne prođe da nekome ne kažeš hvala.

126. Danas nastojim sve savršeno raditi kako se ne bi dogodila *neka* od tih strašnih posljedica o kojima je uvijek govorila.
127. Tko zna, možda nas uskoro očekuje i *neka* pjesma na francuskom.
128. S takvim pogledom često ćeš pomisliti da *neka* tvoja voljena osoba više voli nešto drugo ili nekog drugog od tebe, ako tome pridaje u nekom trenutku više pažnje ili vremena.
129. Zaboravi prošla loša iskustva i stare ljubavi, i ne odgađaj ljubav za *neka* druga, bolja vremena.
130. *Neka* žena je došla i dodirnula mi lice...
131. *Neka* žena svih naroda koja je nekoć bila Marija bude naša zagovornica…
132. Dakle, ako neko voli ovo *neka* jede ovo, ako neko voli ono *neka* jede ono.
133. Svi muškarci bi bili u katalogu i to po par slika za svakog muškarca pa tako ako *neka* voli recimo nabildanog Šveđana ona bi samo stavila kvačicu…

Ako se neposredno iza leksema *neka* javlja osobna zamjenica, u pravilu je riječ o čestici (primjeri 123. i 125.). Ako se iza njega javlja imenica (primjeri 124., 127., 130. i 131.), situacija je nešto složenija. Zasigurno možemo reći da je riječ o čestici ako se imenica ne slaže sa zamjenicom *neka* u rodu, broj i padežu (primjer 124.). Međutim, isto pravilo ne možemo primijeniti u slučaju kad se rod, broj i padež slažu jer tada može biti riječ i o čestici i o zamjenici (primjeri 130. i 131.). Ipak, u tom je slučaju ipak češće u pitanju zamjenica, osim ako je riječ o vlastitoj imenici, a čestice bi se mogle naknadno dodatno razgraničiti pravilima nakon opsežnije analize obrazaca iz korpusa. Ispred prijedloga je također najčešće riječ o zamjenici (primjer 126.). Što se tiče glagola, iako se može naći neposredno ispred njih kao zamjenica ako se slažu u rodu i broju (primjer 133.), pretraga korpusa pokazuje da to nije toliko često kao čestica (primjer 132.), te se stoga i to može uzeti kao okvirno pravilo za česticu.

Nadalje, leksem *neka* je kao veznik obično zamjenjiv veznicima *iako* i *da*:

134. I *neka* sam ja šljakerija bar pošteno zaradim i znam da je moje i hvala bogu nisam nikome dužan...
135. Dajem još jedan kabel kojem je izolacija oštećena *neka* se nađe.

Prva je upotreba u korpusu prilično rijetka te se na temelju toga potencijalno može zanemariti za potrebe POS označavanja. Druga je nešto češća od nje, iako nije vrlo česta, ali bi se možda mogla obuhvatiti kategorijom čestice na temelju toga što se upotrebom ustvari i ne razlikuje znatno od čestice. Doduše, opsežnija analiza korpusa mogla bi ukazati na postojanje dosljednih obrazaca po kojima bi se te dvije upotrebe mogle razlikovati.

Općenito je nešto teže pravilima izolirati upotrebu leksema *neka* kao čestice, ali je ipak bitno uvrstiti ga među čestice zbog činjenice da doista nije riječ o istoj vrsti riječi.

### 5.16. Evo, eto, eno

Napokon, prezentativi ***evo***, ***eto*** i ***eno*** čestice su koje se razlikuju prema kategoriji lica s kojom su povezani (*evo* s prvim licem, *eto* s drugim i *eno* s trećim) (Silić i Pranjković, 2005: 257), no sintaktički među njima nema bitnih razlika:

136. *Evo* ima svega nekoliko konkurentnih modela.
137. Ako među onima koji se bune netko može bolje, *evo* bacam rukavicu.
138. Ma *evo* ti i dozvola, imam ja još tri dozvole kod kuće.
139. I svi, *eto*, jako vole Hrvatsku.
140. *Eto* ljudi moji, samo malo više pameti u glavu.
141. Ali *eto* vidiš, kažu da su ljudi koji su preživjeli bankrote bolji i sigurniji jamci…
142. Jest, utopio se Potjeh, *eno* ga, leži na dnu vode, bijel kao vosak.
143. *Eno* ti tamo kanta za smeće pa baci.
144. Pa *eno* Kinezi ti raspršuju neke aerosole koji izazivaju kišu, znači da neka tehnologija ipak postoji.

Ne pripadaju nijednoj drugoj vrsti riječi i uvijek su čestice neovisno o položaju u rečenici, stoga i njih svakako treba uključiti u konačan popis čestica za POS označavanje.

## 5. Zaključak

Iako je s razvojem tehnologije i upotrebom neuronskih mreža omogućena točnija računalna interpretacija ljudskog jezika za potrebe, primjerice, strojnih prijevoda, u ranoj fazi obrade jezika koja uključuje POS označavanje neposredna ljudska uputa i dalje može značajno povećati točnost. Te se upute uglavnom temelje na popisima leksema, a kasnije i kontekstualnim pravilima. Budući da se kategorizacija mnogih tradicionalno-gramatičkih čestica temelji na semantičkim kriterijima koje nije uvijek moguće izraziti u obliku računalu čitljivih obrazaca i budući da postoji niz leksema koji se ovisno o interpretaciji smatraju „rubnima" u toj kategoriji, potrebno je postaviti čvrste kriterije za njihovu kategorizaciju za potrebe računalne obrade.

Naravno, semantička dimenzija jezika neizostavan je dio lingvističke analize, osobito zato što je primarna funkcija jezika upravo komunikacija. No pri računalnoj obradi jezika važno je uzeti u obzir ograničenja računala u usporedbi s ljudskim mozgom. Računalo može dosljedno i ispravno razlikovati vrste riječi samo na temelju pravila i obrazaca koje je moguće izraziti u formatu koji mu je čitljiv, kao što su popis i pravila navedeni u ovom radu.

U hrvatskim gramatikama ne postoji jedinstvena i usklađena kategorizacija čestica kojom bi se računalo moglo neupitno voditi, a treba uzeti u obzir i slobodan redoslijed riječi u hrvatskome jeziku koji otežava primjenu kontekstualnih pravila za razgraničavanje čestica od drugih homografnih pojavnica. Stoga se najpragmatičnijim pristupom čini skraćivanje i ograničavanje popisa čestica na samo one nepromjenjive riječi koje se mogu smatrati česticama neovisno o kontekstu.

Zaključno, na temelju kriterija, analize i argumenata izloženih u ovom radu kao konačan popis čestica za potrebe POS označavanja predlažem sljedeće: *da, ne, li, zar, god, ma, makar, bar, barem, čak, pa, ta, bilo, neka, put (puta), evo, eto i eno*. Svi su leksemi s popisa nepromjenjive riječi koje se ne mogu jednoznačno svrstati u druge vrste riječi. Za većinu tih leksema nisu potrebna složena sintaktička pravila kako bi u korpusu bili označeni kao čestice, osim u slučaju homografa, te ih je u usporedbi s drugim tradicionalnim česticama lakše ispravno označiti u većini slučajeva.

Zbog ograničenog opsega rada nisu obrađeni svi detalji potrebni za izravnu primjenu, te ostaje prostora za dodatna istraživanja na ovu temu, no nadam se da će prijedlog pojednostavljene kategorizacije čestica izložen u ovom radu biti koristan u daljnjoj računalnoj obradi hrvatskog jezika.

**Popis literature**

Babić, S. i sur. (1991) *Povijesni pregled, glasovi i oblici hrvatskoga književnoga jezika: Nacrti za gramatiku*. Zagreb: HAZU, Nakladni zavod Globus.

Barić, E. i sur. (1979) *Priručna gramatika hrvatskoga književnog jezika*. Zagreb: Školska knjiga.

Bekavac, B. (2002) Strojno obilježavanje hrvatskih tekstova – stanje i perspektive. *Suvremena lingvistika*. 27 (53-54 (1-2)), str. 173-182.

Hrvatski jezični portal. https://hjp.znanje.hr [pristup: 14.09.2021.]

Hrvatski mrežni korpus: hrWaC. http://nlp.ffzg.hr/resources/corpora/hrwac/ [pristup: 10.09.2021.]

Hudeček, L. i Mihaljević, M. (2009) Homonimija kao leksikografski problem. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 35, str. 159-186.

Josić, Lj. (2011) Obradba čestica u suvremenim gramatikama hrvatskoga jezika. *Jezik: časopis za kulturu hrvatskoga književnog jezika*, 58, str. 7-16.

Karlić, V. i Tušek, J. (2013) Čestice u nastavi južnoslavenskih jezika. *Opera slavica*, str. 208-214.

Ma, J. i sur. (2011) POS Tagging of English Particles for Machine Translation. *Proceedings of Machine Translation Summit XIII: Papers*, str. 57-63.

Pranjković, I. (2008) Zamjenica, prilog, čestica i veznik „što“. U: Badurina, L. i Bačić-Karković, D. (ur.) *Riječki filološki dani 8*, str. 238-246.

Raguž, D. (1997) *Praktična hrvatska gramatika*. Zagreb: Medicinska naklada.

Silić, J. i Pranjković, I. (2005) *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*. Zagreb: Školska knjiga.