Preliminarna studija definiranja neizrazite funkcije članstva i neprecizno određivanje kvantifikacije na hrvatskom jeziku temeljene na istraživanju

doc.dr.sc. Leo Mršić
Visoko učilište Algebra
Zagreb, Hrvatska
leo.mrsic@algebra.hr

doc.dr.sc. Sandro Skansi
Sveučilište u Zagrebu
Zagreb, Hrvatska
skansi.sandro@gmail.com

doc.dr.sc. Robert Kopal
Visoko učilište Algebra
Zagreb, Hrvatska
robert.kopal@algebra.hr

***Sažetak**. U ovoj preliminarnoj studiji predlažemo metode za definiranje nepreciznog kvantifikacije za hrvatski jezik utemeljene na istraživanju. Koristeći rezultate ankete provedene među studentima, rad se bavi definiranjem neizrazita funkcija članstva za svaki od preciznih i nepreciznih kvantifikacijskih uvjeta, s mogućim proširenjima neizrazite (engl. „fuzzy") funkcije članstva tipa 2. Ranija verzija ovog rada bila je poslana i naknadno povučena s konferencije ACE-X 2017.*

Ključne riječi. Neizrazita funkcija članstva, neizraziti kvantifikatori, lingvističke varijable, kvantifikatori za hrvatski

1 Uvod

Kvantifikatori jezičnih i lingvističkih varijabli u neizrazitoj logici dijele zajedničke početne pretpostavke. Proučavanje kvantifikacije potječe od Aristotelovog Organona, a u modernom je razdoblju oživljeno kroz razne autore poput G. Frega (Frege 1879) odnosno usavršavano kroz radove C. S. Peirca (Peirce 1885). Kvantifikatori se danas bitna značajka svih glavnih primijenjenih logičkih sustava s nekoliko primjetnih iznimaka (SAT bazirana logika (Marek 2009), propozicionalna modalna logika (Blackburn, de Rijke, Venema 2002)). Kvantifikatori su, također, jezična konstrukcija koji omogućuje referencu prema više pojmova (Peters i Westerstahl 2006). Neprecizni kvantifikatori bili su motivacijski čimbenik razvoja matematičke neizrazite (engl. „fuzzy") logike (Hajek 1998), a prirodni pristup njihovom značenju je putem neizrazitih funkcija članstva. Jezične varijable u početku su istraživane u kontekstu teorije procesa (D'Ambrosio 1989), no kako neizrazita (engl. „fuzzy") logika danas postaje poznatija, oni se raspravljaju u kontekstu neizrazite logike (Ross 2010) (Mrsic 2017).

Druga zanimljiva pojava je da su nekonvencionalni kvantifikatori (Torza 2015) (Peters i Westerstahl 2006) relativni prema određenom jeziku, a različiti jezici posjeduju prirodne kvantificirane pojmove kako za precizne tako i za neprecizne brojeve. Precizni kvantifikatori mogu se smatrati terminima brojeva, koji jedinstveno označavaju preciznu količinu. Primjer takvog izraza može biti "deset", ali njihova vidljiva preciznost ne uzima u obzir njihovu sposobnost referiranja entiteta (Donnellan 1966) (Donnellan 1972). Izjavu "Dohvatite mi tu posudu s deset vijaka" može biti uspješna ako se odnosi na ispravnu posudu (koja sadrži npr. 12 vijaka). Primjer možemo prisnažiti tako što će odrediti prisustvo još jedne, druge, posude koja sadrži npr. 143 vijka a nalazi se neposredno pored prve. To se može činiti kao manje značajna odrednica, no ukazuje na inherentnu nepreciznost čak i uz precizne izraze numeriranja pri razmatranju svakodnevnih komunikacijskih aspekata jezičnih kvantfikatora.

(Prva verzija ovog istraživanja pod naslovom "Učenje neizrazite funkcije članstva za odrednice kvantificiranja slavenskih jezika" namijenjeno je objavljivanja na konferenciji ACE-X 2017, međutim smo nakon razmatranja zaključili da rad iz tog doba zahtjeva značajnu reviziju i proširenje te dodano usmjeravanje te smo ga slijedom toga povukli prije konferencije. Ovaj je rad revidirana i dodatno usmjerena verzija ranijeg, neobjavljenog rada.)

2 Osnovna kvantifikacija i korištenje slovnih naziva numerika

Osnovni kvantifikatori u logici prvog reda su "Svi" i "Postoje", a definirani su kao istiniti u kombinaciji s značenjem koje nose. Primjer toga može biti "Za sve x, P (x)" („za sve x vjerojatnost P od x"). Naivna skupna teorija, prije Russellovog paradoksa (van Heijenoort 1967) tvrdi da svaka svojstva P (*) definiraju skup objekata x koji zadovoljavaju P (*). Klasična logika, zajedno s naivnom skupom teorije, dokazala se nedosljednom Russellovom paradoksu. Otvoreno je pitanje je li naivna postavljena teorija i neizrazita (engl. „fuzzy") logika nedosljedna (Behounek i Hanikova 2014), no pretpostavit ćemo tako u kontekstu opsega ovog rada u svrhu pojednostavljenja samog izlaganja.

Kvantifikator "Postoji" (engl. „Exists“) vrijedi ako se to značenje odnosi na barem jedan objekt. Moglo bi se reći da kvantifikatori poput "Deset" mogu biti lako definirani produženjem ovog načela, međutim time ulazimo u područje referentnog problema. Naime, ako definiramo da "Deset" znači točno 10, referenca bi trebala imati neuspjeh za staklenku s deset vijaka.

Činjenica je da je referenca u takvim slučajevima uspješna, a rješenje je modeliranje željenih kvantificiranih pojmova s neizrazitim članstvom. Dajmo primjer. Kao što smo ranije izjavili, "Postoji x takav da je P (x)" istinit, pod uvjetom da postoji objekt s imenom P (*). Možemo napraviti istu analogiju za "Deset", što nas vodi prema tome da u tom slučaju moramo pronaći istinsku vrijednost za "Postoji deset takvih da je P (*)". Mogli bismo reći da je to istina kada ima deset predmeta (a inače lažno), ali možemo ovo stanje učiniti opuštenijim i prihvatiti neke granične slučajeve s nekom razinom istine, npr. s vrijednosti "istine" recimo 0.8 za 9 i 11 stavki odnosno predmeta.

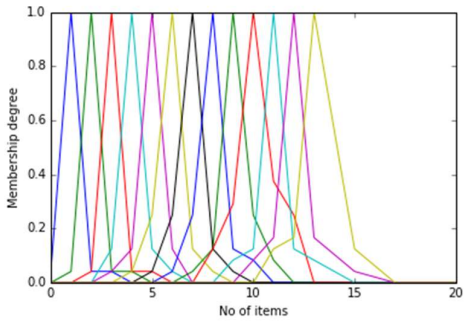
Slijedom navedenog, izraze koji su analizirani kroz ovo istraživanje možemo grupirati kao (i) precizne uvjete i (ii) neprecizne uvjete. Precizni termini analizirani su kao "Jedan", "Dva", "Tri", "Četiri", "Pet", "Šest", "Sedam", "Osam", "Devet", "Deset", "Jedanaest" „Dvanaest“, “Trinaest“ (brojevi od jedan do trinaest). Kao neprecizni termini korišteni su "Jedva išta", "Par", „Nekolicina“, „Nekoliko“, „Brojni“, „Dosta“, „Mnogo“, „Puno“, „Malo“, „Nešto“, „Osjetno“, „Više“. Kao što se može uočiti, uključili smo nekoliko različitih riječi, ali s vrlo sličnim značenjem, pa se razlika između njihovih funkcija članstva može smatrati doprinosom razumijevanju njihove semantike. Ovaj rad istovremeno je obavljen i na engleskom jeziku te je njihov prijevod na engleski jezik također izveden na najprikladniji način, obzirom se mnogi smatraju sinonimima.

Za pronalaženje dobre reprezentacije jezičnih kvantifikatora provodili smo anketu među 43 studenta Visokog učilišta Algebra tražeći od njih da popunjavaju grafikon s rezultatima od 1 do 5 koji predstavljaju mjeru koliko dobro određeni pojam opisuje količinu. Na primjer, ocjena 4 za "Mnogo" ispod stupca "Količina: 20" značila je da je "Mnogo" bio 80% prikladan pojam za opisivanje količine od 20 jedinica. Rezultati 1-5 su naknadno normalizirani na skali 0-1.

Za daljnja istraživanja planira se sveobuhvatnija anketa, kao i tumačenje rezultata korištenjem „type-2 fuzzy“ seta kroz dalju razradu. Uz navedeno, dodatni smjer za daljnja istraživanja bazirana na ovom radu jest uporaba opisanog pristupa kako bi se olakšala anafora rezolucija južnoslavenskih jezika.

2.1 Precizni uvjeti numeričkih izraza

Naše je istraživanje pokazalo niže razine „jasnoće“ čak i za precizne brojeve, stoga smo interpolirali vrijednosti u rezultatima u svrhu pronalaženja funkcije za opisivanje odnosno korištenje u grafičkim prikazima. Razmatrani su najčešći termini na hrvatskom jeziku, a većina ih je pokazala relativno visoku preciznost, uz malo nejasnoće na graničnim slučajevima, koja su se povećavala dok su se brojke povećavale. Značajna iznimka bila je "Deset" koji pokazuje znatno više zamućenosti nego "Jedanaest".

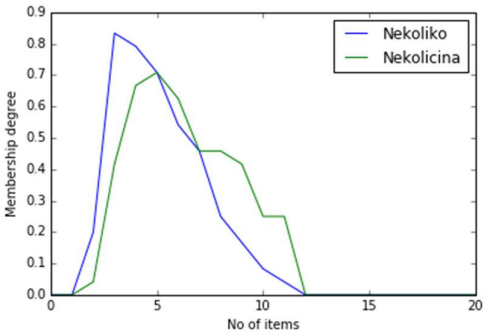


Slika 1. Precizni uvjeti numeričkih izraza

Sve te funkcije mogu biti modelirane minimalnom prilagodbom Gaussovoj funkciji, ili, prema našim potrebama, s jednostavnom funkcijom vrha. Mogući nedostaci u vrijednostima su beznačajni, sve dok je funkcija definirana za sve argumente. Isto tako, funkciju se treba koristiti za procjenu cjelobrojnih vrijednosti.

2.2 Neprecizni uvjeti numeričkih izraza

Prvi precizni pojam kvantificiranja koji je analiziran bio je "Nekolicina", koji je pokazivao zvonoliku formu i lako se može aproksimirati Gaussovom funkcijom. Vidljive jabučice na desnoj strani nisu važne za modeliranje, međutim granice jesu, tako da funkcija vraća vrijednost 0 na 0, 0.7 na 5, te ponovno ide na 0 na 12. Grafikon je prikazan na donjoj slici.

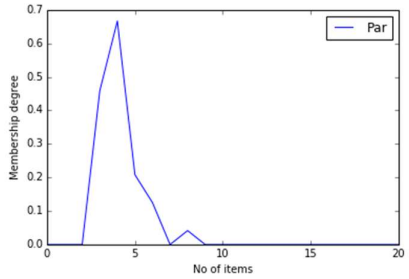


Slika 2. Neprecizni uvjeti numeričkih izraza

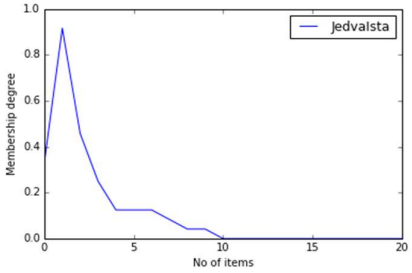
Drugi neprecizni numerički izraz koji smo modelirali bio je "Nekoliko". Prvi dio funkcije djeluje

slično kao i prethodni, ali prelazi na 0.8 kada je dosegnuta vrijednost 3 te silazi strmo na 12. Primijetite kako se vrijednosti u 10 znatno razlikuju. Najvažnija razlika je argument za koji funkcije postižu maksimalnu vrijednost (u prethodnom slučaju 5, a ovdje 3), što upućuje na semantičke razlike u ova dva pojma (u engleskoj verziji, obje adekvatno prevedene na engleski sa terminom "few").

Ovo upućuje na činjenicu (koju ćemo kasnije pokazati) da za većinu kvantificiranih pojmova funkcije članstva arg(max), arg(min), maksimalne i minimalne vrijednosti daju vrlo preciznu definiciju za neizrazitu funkciju članstva. Mogli bismo pristupiti ovom problemu korištenjem Gaussove funkcije kako bismo ih modelirali, ali umjesto toga koristit ćemo djelomične linearne funkcije (obzirom ih je lakše moguće računalno izvesti).



Slika 3. Grafikon termina “Par”



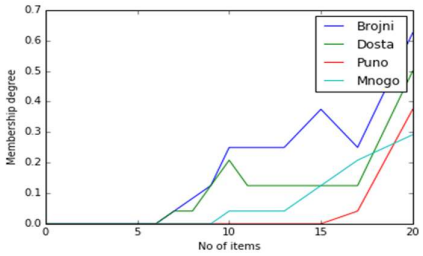
Slika 4. Grafikon termina “Jedva išta”

Termin "Par" doseže svoj maksimum na 4, s vrlo strmim nagibom (završava na 6), ali u praksi je isti kao "Nekoliko", sa samo maksimumom prevedenim na 4 i strmom krivuljom. "Jedva ništa" prikazuje sličan uzorak s visokim maksimumom na 0.95 i arg(max) na 1.

Sljedeći tip funkcija članstva je ReLU tip funkcije, koji je sličan izgledu ReLU funkcije (f (x) = max (0, x)). Radi se o veliki kvantificiranim pojmovima koji ovise o skali: ako se ljestvica popne do 1.000, tada na 1.000 dostignu vrijednost koja je najbliža broju 1. Ako su s druge strane skalirane do 10.000, onda će funkcije članstva biti bliže vrijednosti 1.

Pojam "Dosta" pokazao se najosebujniji, s početkom u 5, s lokalnim vrhom na 10 te vrhuncem između 10-17, nakon toga odlazi. To ukazuje na problem s ograničenjem ljestvice, a može se prepraviti kako bi se utvrdilo da "Dosta" ima šiljak srednjeg raspona, maksimum na kraju raspona i visoravan između.

"Puno" i "Mnogo" ponekad se smatraju sinonimima na južnoslavenskim jezicima, ali se čini da su različiti u semantici i prikazani su na usporednom grafikon ispod, budući da je "Mnogo" ocijenjen realnijim od manjeg broja stavki u odnosu na "Puno".

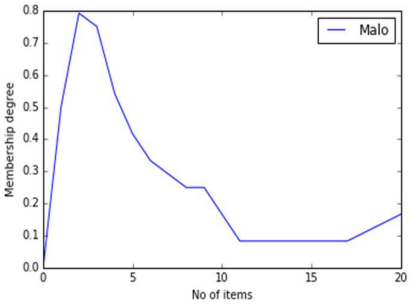


Slika 5. Usporedni grafikon

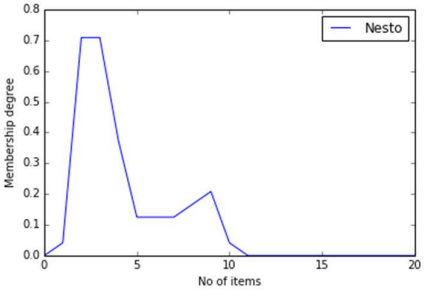
Koraci u "Mnogo" su zanimljiva pojava, međutim nema veću značajnost. Semantička razlika zapravo je u ovom slučaju stvar detalja, evidentno ne postoji praktična primjenjivost, ali kao pojava je ipak vrlo zanimljiva..

Pojam "Brojni“ ima neuobičajeniju funkciju članstva prikazanu na donjoj slici, ali se ipak može aproksimirati funkcijom sličnom ReLU-u. Detaljno izlaganje kako su relevantni ReLU-ovi definirani dan je u sljedećem odjeljku.

Pojmovi "Malo" i "Nešto" imaju šiljak na 2 i 3, da bi kasnije opadali. Oni su slični u ponašanju prema terminima brojeva, samo s produženim repom dok vrijednosti rastu.

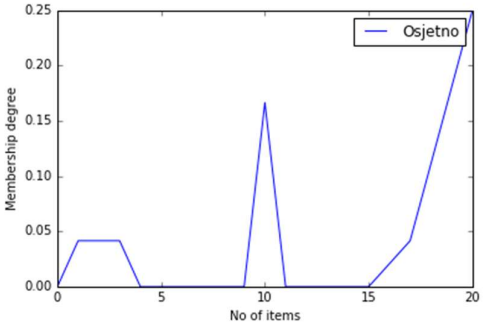


Slika 6. Grafikon termina “Malo”

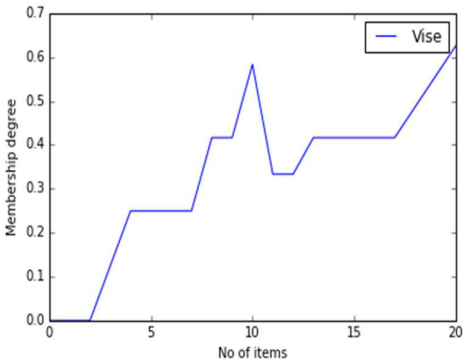


Slika 7. Grafikon termina “Nešto”

Posljednja dva termina su "Osjetno" i "Više", koji dijele srednji šiljak na 10, a isto tako i kasniji pad, nakon čega slijedi porast do broja 1 dok se broj izbrojenih pojmova povećava. To je na neki način iznenađujuće, budući da se njihova semantika tradicionalno ne smatra bliskom, odnosno dok se pojam "Više" smatra gotovo sinonimom "Puno" i "Mnogo". U svrhu vađenja članstva, međutim, smatrat ćemo "Više" kao funkciju sličnom ReLU funkciji.



Slika 8. Grafikon termina “Osjetno”



Slika 9. Grafikon termina “Više”

3 Ekstrapolirane funkcije

U nastavku je predstavljena tablice ekstrapoliranih funkcija:

Tablica 1. Ekstrapolirane funkcije: translacije, Non-Zero vrijednosti

Termin	Translacija	Non-Zero vrijednost
Nekoliko	Few	1-12
Nekolicina	Few	1-12
Par	A couple	2-6
Jedva išta	Barely Anything	0-10
Brojni	Numerous	6-Inf(20)
Dosta	Plenty	6-Inf(20)
Puno	A lot	15-Inf(20)
Mnogo	A lot	9-Inf(20)
Više	A number of	2-Inf(20)
Malo	Few	0-20
Nešto	Few	1-11
Osjetno	Considerably	0-4, 9-11, 15-Inf(20)
[Pojavnost N]	Precise number N	(N-1)-(N+2)

Table 2. Ekstrapolirane funkcije: Maksimalna vrijednost At, MF tip, ReLU kickoff

Termin	Maksimalna vrijednost (Vrijednost argumenta)	MF tipe	ReLU kickoff
Nekoliko	(3, 0.85)	Spike	--
Nekolicina	(5, 0.7)	Spike	--
Par	(4, 0.65)	Spike	--
Jedva išta	(1, 0.9)	Spike	--
Brojni	(20, 0.65)	ReLU	0.33
Dosta	(20, 0.5)	ReLU	0.33
Puno	(20, 0.38)	ReLU	0.75
Mnogo	(20, 0.3)	ReLU	0.45
Više	(20, 0.6)	ReLU	0.1
Malo	(2, 0.8)	Spike	--
Nešto	(3, 0.7)	Spike	--
Osjetno	(2, 0.05), (10, 0.16), (20, 0.25)	Other	--
[Pojavnost N]	(N, 1.0)	Spike	--

Da bismo kreirali što precizniju verziju funkcije članstva koja je potrebna iz gornje tablice, u prvom koraku moramo koristiti stup funkcije članstva. U slučaju ReLU funkcija koristimo sljedeći opći obrazac:

$$ReLU(x) = max(0, f(x)) \quad (1)$$

Gdje je f (x) linearna funkcija izračunata kroz dvije točke nakon početka nulte linije. Ovo je trivijalan zadatak, ali ponavljamo postupak za praktičnost čitatelja (za pojedinosti pogledajte Bronshtein et al., 2007). Prvo je nagib izračunat kroz dvije točke Non-Zero vrijednosti s nagibom (f) = (f (x2) -f (x1)) / (x2-x1), gdje se provode uobičajene odredbe nazivnika (ali čak i ako ne, to nije posljedica jer se odabire drugi par

točaka i izračunava nagib). Nakon izračunavanja nagiba, pomoću y-y1 = nagib (f) (x2-x1) dobiva se eksplicitna reprezentacija. Uputno je napomenuti: prilikom odabira točaka za izračun, zbog pogrešaka uzrokovanih približavanjem najbolje je odabrati točke na kojima se funkcija djelomično povezuje s ostalim dijelovima. To znači da pri izračunavanju nagiba, posljednja točka za koju f (x) = 0 treba biti jedna od točaka koja se koristi, a drugi kraj trebao bi biti kraj točke raspona (što također ima maksimalnu vrijednost u funkcijama sličnim ReLU). To je još važnije u funkcijama šiljaka, koje imaju opći oblik:

$$Spike(x) = max(0, (up(x), down(x))) \quad (2)$$

Gdje je (gore, infleksija, dolje) kratica za dvodijelnu funkciju s (globalnim) maksimumom u sredini (ovo je točka infleksije). Za dalju analizu potrebno je primijetiti dva nagiba. Prvi nagib je "gore" dio. Njegov izračun sličan je ReLU-ima. Koriste se dvije točke, a lijeva je posljednja točka za koju f (x) = 0, što je prva točka u rasponu koji nije nula od gornje tablice. Za drugu točku potrebno je konzultirati par u stupcu „Maksimalna vrijednost“ iz gornje tablice. Za "dolje" dio, prva točka za nagib treba biti par u stupcu „Maksimalna vrijednost“ iz gornje tablice, a druga točka prva točka za koju je f (x) = 0 nakon točke infleksije.

4 Zaključak

U našem istraživanju usredotočili smo se na manji broj pojmova obzirom na činjenicu da su oni učinkovito ograničeni s 0. Ovi su pojmovi funkcije tipa šiljaka. Zbog cjelovitosti, uključili smo i velike kvantificirane pojmove koje predstavljaju funkcije članstva poput ReLU-a, ali postoji inherentan problem s tim pojmovima tj. da ih se interpretira u odnosu na ponuđene skale. U nekim slučajevima 20 može biti „Mnogo“, dok se u nekim slučajevima i 100 možda neće dobro uklopiti u taj pojam. Tako smo koristili raspon od 0 do 20, da bismo osigurali adekvatan pristup u odnosu na ovaj problem, u tablici smo dodali parametar "ReLU kickoff" koji definira nakon kojeg postotka raspona funkcija prestaje biti nula i zaustavi se. To je daleko bolje od pristupa s logaritamskim ljestvicama, budući da bi ljestvica zapisivanja zahtijevala nelinearnu funkciju u ReLU-u, a ipak bi samo smanjili problem i ne bismo ga potpuno uklonili, budući da bi još uvijek trebalo predvidjeti drugačiji pristup na desnoj strani.

Funkcija članstva za "Osjetno" je izostavljena, jer nije dala jasnu diferencijaciju (ima maksimum na samo 25%), a i prilično ju je teško opisati. Smatramo da bi za ovaj pojam trebalo koristiti prikladnu zastupljenost za "Osjetno", veći skup podataka i upotrijebiti dublji polinomični algoritam. Na taj način može se pojaviti i više pravilnosti, pa to ostavljamo kao otvoreni problem za daljnja istraživanja. Naravno, postoji i niz drugih pojmova jednako složenih kao i "Osjetno", što svakako treba uzeti u obzir odnosno pokušati adresirati.

Vjerujemo da bi naša istraživanja mogla biti od velike koristi za računalnu semantiku i anafora rezoluciju na južnoslavenskim jezicima. Prvo, zbog sličnosti vjerujemo da bi za funkcionalne inženjerske aplikacije trenutni prikazi kvantificiranih izraza na hrvatskom jeziku mogli koristiti za sve južnoslavenske jezike (bosanski, bugarski, hrvatski, makedonski, crnogorski, srpski, slovenski). Primjena na računalnu semantiku je sasvim jasna: dva glavna segmenta računalne semantike su odnosi (ekstrahirani obično strojnim učenjem) i kvantificiranje, koje smo evidentno dobro adresirali i djelomično riješili.

Dodatna primjena očituje se u sposobnosti procjene anaforičkog ponašanja kvantificiranih pojmova tj. pristupa u kojem se pristupa pitanju „da li se trenutna neodređena količina odnosi na prethodnu neodređenu količinu“ i, ako da, „na koju se od njih odnosi u većoj mjeri“. To se može učiniti učenjem odnosno usporedbom vrijednosti hipotetske količine za dvije slične funkcije članstva. Odgovor na ovo pitanje ne može se tražiti jednostavnim linearnim prilagodbom, a što ostavljamo kao predmet za daljnja istraživanja.

Reference

Behounek, L. and Hanikova, Z. (2014). Set Theory and Arithmetic in Fuzzy Logic. In Petr Hajek on Mathematical Fuzzy Logic, ed. F. Montagna, str. 63-89.

Blackburn, P., de Rijke, M., Venema, Y. (2002). Modal Logic (Cambridge Tracts in Theoretical Computer Science). Cambridge: Cambridge University Press.

Bronshtein, I. N., Semendyayev, K. A., Musiol, G. and Muehlig, H. (2007). Handbook of Mathematics (Fifth Edition). Berlin: Springer.

D'Ambrosio, B. (1989). Qualitative Process Theory Using Linguistic Variables. Berlin: Springer.

Donnellan, K. S. (1966). Reference and Definite Descriptions. The Philosophical Review, vol. 75 br. 3, str. 281–304.

Donnellan, K. S. (1972). Proper Names and Identifying Descriptions. U D. Davidson i G. Harman (ur.). Semantics of Natural Language.

Frege, G. (1879). Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reinen Denkens. Halle.

Hajek, P. (1998). Metamathematics of Fuzzy Logic. Amsterdam: Kluwer Academic Press.

van Heijenoort, J. (1967). From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931, str. 124-125. Cambridge: Harvard University Press.

Marek, V. W. (2009). Introduction to Mathematics of Satisfiability. Boca Raton, FL: Chapman & Hall/CRC Studies in Informatics Series.

Mrsic, L. and Klepac, G and Kopal R (2017). A New Paradigm in Fraud Detection Modeling Using Predictive Models, Fuzzy Expert Systems, Social Network Analysis, and Unstructured Data, Computational Intelligence Applications in Business Intelligence and Big Data Analytics, Auerbach Publications, pp. 157-194

Peirce, C. S. (1885). "On the Algebra of Logic: A Contribution to the Philosophy of Notation, American Journal of Mathematics, vol. 7, str. 180–202.

Peters, S. and Westerstahl, D. (2006). Quantifiers in Language and Logic. Oxford: Oxford University Press

Ross, T. (2010). Fuzzy Logic with Engineering Applications. New York: Wiley Press.

Torza. A. (2015). Quantifiers, Quantifiers, Quantifiers: Themes in Logic, Mataphysics and Language. New York: Springer.