NASLOV: **Računalno jezikoslovlje**

PREDAVAČI:
Prof.dr.sc. Mario Essert, *Fakultet strojarstva i brodogradnje*, Zagreb (messert@fsb.hr)
Dr.sc. Kristina Štrkalj Despot, *Institut za hrvatski jezik i jezikoslovlje*, Zagreb (kdespot@ihjj.hr)

Sažetak

Prema Rolandu Hausseru[2] tri su osnovna pristupa prirodnomu jeziku:

(i) *tradicionalna gramatika* - koristi se metodom neformalne klasifikacije i opisa koji se temelje na tradiciji i iskustvu,
(ii) *teorijska lingvistika* – koristi se metodama matematičke logike za opisivanje prirodnih jezika uz pomoć sustava formalnih pravila namijenjenog za dobivanje svih, ali samo gramatički ovjerenih, jezikoslovnih izraza,
(iii) *računalno jezikoslovlje* – kombinira metode tradicijske gramatike i teorijske lingvistike s metodom učinkovite provjere eksplicitnih hipoteza, izvedbom formalnih gramatika kao učinkovitih računalnih programa i njihovim automatskim testiranjem na realnoj količini stvarnih podataka.

Unatoč različitim metodama, ciljevima i primjenama, sve tri varijante jezikoslovne znanosti dijele područje na jednake razine: na *fonologiju*, *morfologiju* (oblikoslovlje), *leksikon* (rječnik), *sintaksu* (skladnju), *semantiku* (značenje) i *pragmatiku* (područje primjene). Formalna teorija jezika (točke ii i iii) radi s matematičkim metodama koje obrađuju empirijski sadržaj gramatičke analize i funkcioniranje komunikacije što je moguće neutralnije – jezik postaje skup konačnih slijedova riječi (slobodni monoid).

S druge pak strane, budući da obradba prirodnog jezika (NLP) pripada području umjetne inteligencije (AI) i tradicionalne lingvistike, ona se bavi strukturom teksta i algoritmima koji izvlače smislenu informaciju iz teksta. Dobro poznata i učinkovita tehnika je model vektorskog prostora, koji predočava dokumente kao matricu n × m dimenzija (*Salton, Wong & Yang*, 1975). Metrička udaljenost može se tako upotrijebiti kao funkcija matrice za izračunavanje sličnosti između dokumenata. Ova vrsta algoritama strojnog učenja potiče statistički pristup jeziku i povezana je s istraživačkim poljima poput rudarenja teksta, kategorizacije teksta i dohvaćanja informacije. Model vektorskoga prostora temelj je mnogim zadaćama u obradbi prirodnoga jezika i strojnoga učenja, od pretraživačkih upita do razvrstavanja tekstova i njihova grupiranja (klasterizacije).

Ovo predavanje predstavit će teorijske i praktične korake u procesu izvlačenja informacije iz sirovih podataka (npr. pronalaženje riječi i rečenica u nizu znakova), prepoznavanja i stvaranja vrste riječi (imenica, glagola, pridjeva itd.), prepoznavanja logičkih zakona u rečenicama te pronalaženja značenja iz odnosa među riječima. Na koncu, pokazat će se i neki programi s modelom vektorskog prostora u jezikoslovlju.

[2] Foundations of Computational Linguistics, Human-Computer Communication in Natural Language, Third Edition, Springer-Verlag, 2014.