Analiza i usporedba entropija službenih jezika Europske unije

Prpić, Luka

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2022./2023.

Luka Prpić

Analiza i usporedba entropija službenih jezika Europske unije

Završni rad

Mentor: izv. prof. dr. sc. Petra Bago

Zagreb, rujan 2023.

Izjava o akademskoj čestitosti

Izjavljujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

(potpis)

Želim izraziti iskrenu zahvalnost profesorici Bago na velikodušnoj podršci, stručnosti i vodstvu koje mi je pružila tijekom procesa izrade mog završnog rada. Njezina posvećenost i mentorstvo bili su od neprocjenjive važnosti za moj akademski razvoj, te su mi omogućili dublje razumijevanje i uspješno završavanje rada.

Također, želim zahvaliti doktorandu na poslijediplomskom doktorskom studiju „Lingvistika" Diegu Váliu Antunesu Alvesu na dragocjenoj pomoći oko proučavanja europskih jezika i strpljenju koje je pokazao tokom toga.

Bez njihove podrške, ova akademska avantura ne bi bila moguća. Duboko sam zahvalan na njihovom nesebičnom angažmanu i žrtvovanju vremena kako bi mi pomogli ostvariti ciljeve.

Sadržaj

1. Uvod

Jezik, kao jedinstveno obilježje ljudske spoznaje i komunikacije, stoljećima je fascinirao lingviste, psihologe, informacijske stručnjake i istraživače u različitim disciplinama. Proučavanje jezika nadilazi puku jezičnu analizu i uključuje dublje u njegove temeljne strukture, evoluciju, usvajanje i obradu. Jedan kritični aspekt koji je privukao sve veću pozornost u polju lingvistike je koncept entropije jezika.

Claude Shannon definirao je entropiju - ili prosječni sadržaj informacija kao mjeru nesigurnosti ili „izbora" svojstvenu nizovima simbola (Shannon, 1948).

Jezična entropija služi kao mjerljivo svojstvo jezične složenosti i komunikacijske učinkovitosti unutar danog jezičnog sustava. Uvedena kroz Shannonovu informacijsku teoriju, entropija pruža sredstvo za kvantificiranje nesigurnosti i predvidljivosti jezičnih jedinica, kao što su fonemi, morfemi, riječi i sintaktičke strukture. Razumijevanjem entropije različitih jezika, istraživači mogu dobiti dragocjene uvide u to kako se informacije kodiraju, prenose i dekodiraju u različitim jezičnim sustavima.

Motivacija iza ovog istraživanja je stjecanje dubljeg razumijevanja jezične entropije koje može poboljšati naše znanje o temeljnim principima koji upravljaju ljudskim jezikom. Istraživanje jezične entropije u raznim znanstvenim radovima ima praktične implikacije za različita područja, uključujući obradu prirodnog jezika (eng. *Natural language processing, NLP*)[1], umjetnu inteligenciju (eng. *Artificial intelligence, AI*)[2], jezično obrazovanje[3] i sociolingvistiku[4]. Analizirajući obrasce jezične entropije, istraživači mogu razviti razne sofisticiranije jezične modele, sustave strojnog prevođenja i alate za analizu sentimenata. Štoviše, razumijevanje načina na koji entropija utječe na usvajanje jezika i evoluciju može informirati pristupe jezičnom obrazovanju i pridonijeti očuvanju ugroženih jezika.

Prepoznavanje jezika na temelju raznih jezičnih karakteristika i parametara jedan je od temeljnih koncepata strojnog učenja i obrade prirodnog jezika, a posebno strojnog prevođenja. U računalnoj lingvistici, jezična entropija i srodne mjere bile su naširoko primjenjive za rješavanje problema povezanih sa strojnim i strojno potpomognutim

[1] Benz, C. Alikaniotis, D. Cysouw, M. The Entropy of Words: Learnability and Expressivity across More than 1000 Languages, 2017.
[2] Poir, M. Entropy in Machine Learning. Java T Point, 2011.
[3] Ahmad et al. Entropy in Education System: Transformation of an Individual Through Meaningful Interactions in a Community of Inquiry, 2022.
[4] Razumovskaya, V. Information Entropy of Literary Text and its Surmounting in Understanding and Translation, 2010.

prevođenjem[5], pronalaženjem informacija[6] i distribucijskom semantikom[7]. Sva ova područja presudno ovise o procjeni vjerojatnosti i neizvjesnosti povezanih s riječima u danom tekstu i jeziku. Dva su središnja istraživačka pitanja koja se pojavljuju u ovome radu:

1) Može li se uspostaviti odnos između jezika i vrijednosti pripadajućih entropija?

2) Može li se predvidjeti prema vrijednosti entropije o kojem jeziku ili jezičnoj skupini se radi?

Prvo i drugo istraživačko pitanje odnose se na problem neizvjesnosti fenomena i primjene obrade prirodnog jezika na različitim jezicima. U radu se uspoređuju entropije službenih jezika Europske unije.

Razlog odabira ove teme je i postojanje široko rasprostranjenog pogrešnog shvaćanja da je identifikacija jezika „riješena stvar“. Potrebno je skrenuti pozornost na činjenicu da smo još daleko od stvaranja savršene jezične identifikacije *web* dokumenata i tekstova.

Slijedom navedenoga, sam rad podijeljen je u četiri dijela, od kojih se u prvom piše o entropiji i kondicionalnoj entropiji kao fenomenu područja teorije informacije, entropiji jezika, a potom o dosadašnjim istraživanjima o entropiji. U drugom dijelu rada opisana je procedura prikupljanja podataka i stvaranja korpusa. Treći dio rada olakšava razumijevanje programa korištenog u ovome radu zbog lakše interpretacije dobivenih rezultata. Zadnji, četvrti dio rada piše o predviđanju rezultata te potom o analizi i interpretaciji dobivenih entropija.

Na kraju rada dan je kritički osvrt u obliku zaključka na temu usporedbe entropija službenih jezika Europske unije te se sumira i rekapitulira sve što se obradilo u ovome radu.

[5] Wei Y. Entropy as a Measurement of Cognitive Load in Translation, 2022.
[6] Henning, P. Schuler, C. J. Entropy Search for Information-Efficient Global Optimization, 2012.
[7] Melamed I. Measuring Semantic Entropy, 1994.

2. Razrada

U razradi su pojašnjeni teorijski temelji jezične entropije i njezine veze s teorijom informacija. Objašnjen je koncept entropije kako ga je uveo Claude Shannon i njegova prilagodba lingvistici. Razumijevanje teorijskih temelja entropije i kondicionalne entropije ključno je za razumijevanje načina na koji se kvantificira jezična složenost i kako se informacije kodiraju i prenose unutar jezičnih sustava. Pojašnjen je i pojam jezične entropije te su navedene razlike među visokoentropijskim i niskoentropijskim jezicima. Na kraju poglavlja sumirani su prijašnji radovi na sličnu temu te su ukratko pojašnjeni.

2.1. Pojam entropije u teoriji informacije

Riječ entropija prvi put je uporabio Rudolph Clausius 1867. godine u svome radu „O kretanju topline“. Naziv je nastao od grčkih riječi érgon (grč. ἔργον, raditi) i tropḗ (grč. τροπή, mijenjanje) što označava „djelovanje promjene“. Iako je primarno entropiju koristio u termodinamici, nagovijestio je kako će entropija imati široku primjenu u raznim znanstvenim disciplinama.

Teorija informacija, revolucionarna disciplina koju je postavio Claude Shannon sredinom 20. stoljeća, pruža matematički i teorijski okvir za proučavanje prijenosa informacija u komunikacijskim sustavima. Shannon je informaciju definirao kao smanjenje neizvjesnosti i iznenađenja koje se događa kada primimo određenu poruku (Shannon, 1948). Ovaj koncept omogućio je kvantificiranje informacija pomoću bitova, osnovnih jedinica informacija, gdje jedan bit predstavlja količinu informacija potrebnu za izbor između dvije jednako vjerojatne alternative.

Entropija je ključni pojam u teoriji informacija koji označava mjeru nepredvidivosti ili nesigurnosti u slučajnoj varijabli. Preciznije, entropija mjeri prosječnu količinu informacija koja je potrebna kako bismo opisali ili predstavili slučajnu varijablu. U slučaju kada je vjerojatnost pojedinih događaja ravnomjerno raspoređena, tada je entropija najveća, jer su događaji najmanje predvidljivi. S druge strane, ako je vjerojatnost koncentrirana oko određenih vrijednosti, entropija je manja jer su događaji predvidljiviji.

Shannon je prvi definirao pojam entropije informacije 1948. godine u svom radu „A Mathematical Theory of Communication“. Ovaj pojam omogućio je precizno mjerenje koliko je informacija sadržano u određenoj poruci ili varijabli.

Teorija informacija ima široku primjenu u gotovo svim područjima gdje se informacije prenose, pohranjuju i obrađuju. Shannon je postavio temelje za modernu ergodičku teoriju i dublje razumijevanje komunikacije i informacija u digitalnom svijetu.

2.2. Entropija diskretne slučajne varijable

Diskretna slučajna varijabla ima prebrojiv broj mogućih vrijednosti. Zbroj vjerojatnosti svih vrijednosti jednak je 1. Očekivana vrijednost varijable X je prosječna vrijednost koju bismo mogli očekivati ako ponovimo eksperiment mnogo puta (Petersen, 2018). Računa se sa formulom 1.

Formula 1: $\mu = E(X) = \sum x \cdot P(X = x)$.

μ je simbol koji označava očekivanu vrijednost slučajne varijable što ujedno označava i E(X), P(X=x) označava vjerojatnost da slučajna varijabla X poprimi vrijednost x. U kontekstu ove formule, to je vjerojatnost da se događaj X=x ostvari.

Zbog više vrijednosti, dolazi se do problema neizvjesnosti (eng. *uncertainty*), gdje je nepoznato koji ishod će biti odabran. Neizvjesnost je najveća kada je vjerojatnost svih vrijednosti jednaka, a najmanja kada je vjerojatnost jednog događaja 1, a svih ostalih događaja 0. Entropija je mjera koja nam pomaže da izračunamo koliko je dobro neizvjesnost raspoređena (Shannon, 1948).

Entropija diskretne slučajne varijable označava prosječnu informaciju sadržanu u varijabli. Ona ne ovisi o stvarnim vrijednostima koje uzima slučajna varijabla, već samo o vjerojatnostima (Douglas, 1987.). Entropija diskretne slučajne varijable X definirane na prostoru vjerojatnosti definirana je pomoću formule 2.

Formula 2: $H(X) = E \log\frac{1}{p(x)} = \sum_{x \in X} p(x) \log\frac{1}{p(x)}$

gdje $H(X)$ predstavlja entropiju, a p(x) vjerojatnost da će se ostvariti događaj x.

Izbor baze logaritma varira za različite primjene. Baza 2 daje jedinicu bitova (ili „shannons“), baza *e* daje prirodnu jedinicu „nat“, a baza 10 daje jedinice „dits“, „bans“ i „hartleys“ (Cook, 2021.). Prirodni logaritmi sa bazom e obično se koriste u matematici i fizici, dok se logaritmi sa bazom 2 koriste najčešće u teoriji informacije.

2.3. Kondicionalna entropija

U teoriji informacije, uvjetna ili kondicionalna entropija (eng. *conditional entropy*) kvantificira preostalu entropiju slučajne varijable Y s obzirom da je poznata vrijednost druge

slučajne varijable X (Orlinsky, 2016). Matematički se zapisuje kao $H(X|Y)$. Kao i druge entropije, mjeri se u bitovima, natovima, ditsovima… Formule 3 i 4 koristi se za računanje kondicionalne entropije. Obje formule entropije pomažu u razumijevanju količine neizvjesnosti ili informacija povezanih s jednom slučajnom varijablom kada imate djelomične informacije o drugoj slučajnoj varijabli.

Formula 3: $H(X|Y = yj) = -K \sum_i pi|j \; log_d \;\; pi|j = 1,2, \dots, n,$

Formula 4: $H(Y|X = xi) = -K \sum_i p'j|i \; log_d \;\; p'j|i = 1,2, \dots, m.$

$H(X|Y = yj)$ i $H(Y|X = xi)$ predstavljaju kondicionalnu entropiju, K koeficijent entropije, a $pi|j$ vjerojatnost događaja i uz uvjet da se dogodio događaj j.

Vrijednost kondicionalne entropije biti će 0 ako i samo ako u entropiji $H(X|Y)$ vrijednost Y je potpuno određena vrijednošću X. Najveća moguća vrijednost kondicionalne entropije postići će se jedino u slučaju kada su X i Y nezavisne slučajne varijable, pa je $H(X|Y) = H(Y)$.

Kondicionalna entropija u ovome radu proučavati će odnose bigrama, to jest sekvencu koja se sastoji od dviju riječi koje se nalaze unutar teksta korpusa. Uvjet kondicionalne entropije predstavljat će riječ prije proučavane riječi. Važnost mjerenja kondicionalne entropije u ovome radu je upravo zbog ograničenja broja mogućih riječi koje slijede nakon prethodnih riječi. Na primjer, u engleskom jeziku, nakon subjekta nećemo očekivati pojavu imenice, pridržavajući se reda riječi u jeziku.

2.4. Jezična entropija

Prirodni jezik daje nam izvanredan primjer sustava koji se koristi za generiranje duge sekvence simbola koji posjeduju neka izuzetna svojstva. Tekst koji se razvija u vremenu i u prostoru je „slučajan“, tj. nije potpuno predvidljiv, stoga se dolazi do zaključka da je tekst pisan u nekom prirodnom jeziku realizacija slučajnog procesa te kao nepredvidivi niz simbola ima svoju vrijednost entropija. (Shannon, 1950).

U izračunavanju jezične entropije, izbor jezičnih jedinica je proizvoljan. Fonemi, najmanji različiti glasovi govora, služe kao jedna od temeljnih jedinica za izračunavanje entropije u fonologiji. Morfemi, najmanje jedinice značenja, relevantni su za morfološku entropiju. Za leksičku entropiju riječi se smatraju jezičnim jedinicama, dok se u sintaksi entropija izračunava na temelju rasporeda riječi i fraza. Osnovna informacijsko-teorijska karakteristika jezika je entropija po simbolu teksta, no u ovom će se radu obrađivati entropija na razini

riječi. Jezična entropija povezana s riječima je koncept informacijske teorije u srcu kvantitativne i računalne lingvistike. Entropija je uspostavljena kao mjera ove prosječne nesigurnosti, koja se naziva i prosječni sadržaj informacija (Hausser, 2014). Jedno od temeljnih svojstava riječi je njihova učestalost ponavljanja, ali također i učestalost s obzirom na okolinu u kojoj se nalaze riječi (Baayen, 1997). Entropija nad raspodjelom vjerojatnosti riječi mjeri upravo „izbor“ riječi. Analiza entropije jezika otkriva zanimljive obrasce u jezičnoj tipologiji. Neki jezici, poput izolacijskih jezika s jednostavnim strukturama riječi, imaju tendenciju veće entropije, dok drugi, poput polisintetičkih jezika sa složenim morfološkim strukturama, imaju tendenciju niže entropije (Lo, 2018). Analiza jezične entropije kroz tipološke kategorije daje uvid u odnos između jezične raznolikosti i prijenosa informacija.

2.4.1. Kondicionalna jezična entropija

Kondicionalna entropija u ovome radu proučavati će odnose bigrama, to jest sekvencu koja se sastoji od dviju riječi koje se nalaze unutar teksta istog korpusa. Uvjet kondicionalne entropije predstavljat će riječ prije proučavane riječi. Kondicionalna entropija procjenjuje se na temelju učestalosti pojavljivanja bigrama, to jest koja pojavnica je najvjerojatnija da slijedi određenoj prethodnoj pojavnici. Važnost mjerenja kondicionalne entropije u ovome radu je upravo zbog ograničenja broja mogućih riječi koje slijede nakon prethodnih riječi. Ova ideja ima duboke veze s kognitivnim procesima. Kada obrađujemo jezik, posebno tijekom razumijevanja i generiranja rečenica, naš mozak koristi kontekstualne informacije kako bi predvidio najvjerojatniji sljedeći jezični element (Gabora, 2016). Ovaj proces temelji se na principu kondicionalne entropije. Ako je kontekst jasan i konzistentan, tada će i kondicionalna entropija biti visoka, što olakšava jezičnu obradu. Međutim, kada se susretnemo s nejasnim ili ambivalentnim kontekstom, kondicionalna entropija će biti niska, što otežava proces kognitivne obrade jezika (Wei, 2022). Visokoentropijski jezici imaju manje vrijednosti kondicionalne entropije zbog svoje nepredvidivosti, dok niskoentropijski jezici imaju veće radi predvidljivijeg rasporeda jezičnih elemenata. Na primjer, u engleskom jeziku, nakon subjekta nećemo očekivati pojavu imenice, pridržavajući se reda riječi u jeziku, dok u njemačkom jeziku glagol dolazi nakon subjekta.

2.4.2. Visokoentropijski jezici

Jezici s visokom entropijom pokazuju veću nepredvidljivost u svojim jezičnim jedinicama. U takvim jezicima postoji šira distribucija jezičnih jedinica, što dovodi do većeg stupnja

nesigurnosti u predviđanju sljedećeg elementa u nizu (Vladisavljević, 2021). Kao posljedica toga, jezici visoke entropije prenose informacije na gušći način, zahtijevajući od slušatelja da obrade opsežniji raspon mogućnosti prilikom tumačenja poruke. Kako bi kompenzirali povećanu nesigurnost, jezici visoke entropije često pokazuju veću redundantnost informacija. Redundancija uključuje prisutnost više signala koji prenose istu poruku, povećavajući vjerojatnost uspješne komunikacije čak i uz prisutnost šuma ili pogrešaka (Darian, 1979). Redundancija u jezicima visoke entropije može olakšati oporavak informacija i poboljšati otpornost komunikacije. U takvim jezicima kontekst igra ključnu ulogu u razjašnjavanju značenja. Zbog veće nepredvidivosti jezičnih jedinica, slušatelji se uvelike oslanjaju na kontekstualne znakove kako bi točno protumačili željenu poruku. Ovo oslanjanje na kontekst može dovesti do učinkovite komunikacije unutar specifičnih konteksta diskursa, ali može predstavljati izazov kada se radi o dvosmislenim ili nepoznatim situacijama (Finn, 2012).

U visokoentropijskim jezicima, veća nepredvidljivost jezičnih jedinica može dovesti do povećanog kognitivnog opterećenja tijekom razumijevanja jezika. Slušatelji moraju obraditi širi raspon mogućnosti, aktivirajući dodatne kognitivne resurse kako bi razjasnili značenje i napravili točna tumačenja. Visokoentropijski jezici mogu doživjeti češću promjenu jezika jer govornici eksperimentiraju s novim jezičnim elementima kako bi prenijeli evoluirajuća značenja (Montemurro, 2011). Jezici s visokom entropijom pružaju govornicima raznolik skup jezičnih jedinica, potičući jezične inovacije i kreativnost. Nove riječi, izrazi i gramatičke strukture mogu se lako pojaviti u lingvističkim sustavima visoke entropije.

2.4.3. Niskoentropijski jezici

Jezici s niskom entropijom imaju predvidljivije jezične jedinice, a određeni elementi pojavljuju se s višim frekvencijama. Posljedično, ovi jezici obično nose manje informacija po jedinici, što rezultira manjim opterećenjem informacijama po poruci (Shannon, 1950). Zbog veće predvidljivosti, obrada jezika s niskom entropijom može zahtijevati manje kognitivnog napora, budući da slušatelji mogu predvidjeti jezične jedinice s većom točnošću. Ova povećana predvidljivost može dovesti do brže i učinkovitije komunikacije. U jezicima s niskom entropijom postoji tendencija ekonomiziranja izraza korištenjem kraćih jezičnih jedinica za prenošenje specifičnih značenja. Riječi ili morfemi mogu nositi više informacija, što rezultira sažetim i učinkovitim kodiranjem informacija. Niskoentropijski jezici, s predvidljivijim jezičnim jedinicama, omogućuju slušateljima da predvide nadolazeće elemente i lakše integriraju kontekstualne informacije. Smanjeno kognitivno opterećenje pri obradi jezika s niskom entropijom omogućuje brže i učinkovitije učenje, pamćenje i

razumijevanje tog jezika (Klein, 2021). Niskoentropijski jezici mogu zadržati arhaične oblike i gramatičke značajke tijekom duljeg razdoblja. Također se mogu oduprijeti vanjskim utjecajima i posuđivanjima iz drugih jezika zbog svoje konzervativne prirode (Montemurro, 2011). Taj otpor doprinosi očuvanju jezičnih tradicija i identiteta.

2.5. Dosadašnja istraživanja

Entropijom engleskog jezika bavio se Shannon još davne 1950. godine u svome radu „Prediction and Entropy of Printed English“. Razmatrajući vjerojatnost pojavljivanja slova, uspostavio je da je prosječna entropija slova u engleskom jeziku 4,11 bita. Govorio je i o entropiji riječi u engleskom jeziku. Engleske riječi poput „the“ imaju manju entropiju jer su uobičajene i obično prethode imenici, dok neuobičajene riječi poput „agriculturist“ imaju veću entropiju jer su manje predvidljive u danom kontekstu.

Christian Benz je računao entropiju značenja riječi u svome radu „The Entropy of Words - Learnability and Expressivity across More than 1000 Languages“. Uspostavilo se da neodređenost treba mjeriti u dva oblika, vrsti značenja i njenom sveukupnom obuhvatu. Kako bi se poboljšala učinkovitost korištenja vokabulara, nekim se riječima pripisuje više značenja, ali to također znači da imamo nekoliko mogućih tumačenja značenja riječi. Niža entropija tumačenja vokabulara znači da je prostor za dvosmislenost manji i obratno. Na ovaj način, nejasnoća riječi je vrijednost entropija. U radu se zaključuje da bi se prenijelo točnije značenje, potrebno je više riječi koje predstavljaju isto značenje u različitim slučajevima. Na taj način se povećanjem broja riječi smanjuje nejasnoća značenja riječi.

Drugačiji pristup jezičnoj entropiji istražio je Lev B. Levitin u radu „Entropy of Natural Languages: Theory and Experiment“ gdje je uspoređivao entropiju po simbolu. U radu se mjerila nesigurnost simbola koja slijedi nakon danog niza simbola. U radu su mjerene 4 nesigurnosti: slovo nakon prefiksa, slovo u sredini riječi, završno slovo i razmak. Došlo se do zaključka da počeci riječi predstavljaju najteže situacije za pogađače riječi na većini jezika te da nesigurnost simbola uvelike ovisi o njegovom položaju u riječi. U tablici 1 prikazane su gornje i donje granice vrijednosti entropije za pojedini simbol (slovo).

Tablica 1: Entropija po simbolu za razne vrste jezičnih situacija

	After prefix	Middle letter	End letter	Space
Upper bound	3.572	1.48	1.01	0.176
Lower bound	2.731	0.87	0.475	0.0816

Izvor: https://www.sciencedirect.com/science/article/pii/0960077994900795/

3. Prikupljeni podaci

Poglavlje opisuje postupke i metode prikupljanja podataka korištenih za računanje entropija službenih jezika Europske unije.

3.1. Opis jezika i razvoj korpusa

U pozadini velike raznolikosti, europski jezici imaju mnogo dodirnih točaka zbog kontakta i miješanja stanovništva tisućama godina (Europa u pokretu, 2020.). Jezici dijele dosta riječi koje imaju isti korijen ili značenje.

Postavlja se pitanje možemo li na temelju različitih iznosa entropija za pojedine jezike znati o kojem se jeziku radi ili kojoj jezičnoj skupini pripada. Za potrebe rada prikupljeno je 24 jednojezična korpusa za 24 službena jezika Europske unije (bugarski, hrvatski, češki, danski, engleski, estonski, finski, francuski, njemački, grčki, mađarski, irski, talijanski, latvijski, litavski, malteški, nizozemski, poljski, portugalski, rumunjski, slovački, slovenski, španjolski i švedski). Većina europskih jezika pripada indoeuropskoj jezičnoj skupini. Germanski jezici dijele se na sjevernogermanske jezike i zapadnogermanske jezike. Od navedena 24 jezika, pod sjevernogermanske spada danski i švedski, a pod zapadnogermanske engleski, njemački i nizozemski. Podjela romanskih jezika na skupine je malo teža, a tu ubrajamo: francuski, talijanski, portugalski, rumunjski i španjolski. Zapadnoslavenskim jezici pripada: poljski, češki i slovački, a južnoslavenskim hrvatski, bugarski i slovenski. U baltičke jezike ubrajamo litavski i latvijski. U indoeuropske jezike spadaju još irski i grčki. Irski, zajedno sa škotskim, manskim i hiberno-škotskim gaelskim čini goidelsku ili gaelsku podskupinu jezika (Jalde, 2018). Grčki pripada helenskoj grani indoeuropske jezične porodice i najranije je zabilježeni indoeuropski jezik koji se govori još od vremena antike.

Druga najbrojnija jezična skupina u europi su ugro-finski jezici, a to su finski, mađarski i estonski.

Malteški je semitski jezik iz afrazijske porodice jezika. Najsličniji je arapskom jeziku, i to tuniškoj verziji, no zbog geografskog položaja i povijesnih okolnosti ima osobine talijanskog i engleskog (Brincat, 2022.).

Na slici 1 prikazan je geografski raspored europskih jezičnih skupina koje su označene različitim bojama. Crvenom bojom su prikazane države gdje se govore romanski jezici, smeđom germanski, plavom slavenski, ljubičastom ugro-finski i tamnoplavom baltički.

Tamnozelenom bojom prikazano je govorno područje irskog, žutom grčkog, sivom albanskog i zelenom turskog jezika.

Slika 1: Jezici u Europi

Izvor: https://jakubmarian.com/map-of-languages-and-language-families-of-europe/

Radi dosljednosti rezultata, odabrani su korpusi koji sadrže istu temu, a to su parlamentarne rasprave i debate, preuzeti sa stranica Clarin[8] i Sketch Engine[9]. Pristup Sketch Enginu studentima Filozofskog fakulteta Sveučilišta u Zagrebu osiguran je preko AAI identiteta, a korpusi preuzeti s Clarina su javno dostupni. Iz preuzetih korpusa europskih jezika maknuti su nepotrebni dijelovi koji ne doprinose u obradi i istraživanju. Sav tekst u korpusu prebačen je isključivo u mala slova abecede, uklonjeni su interpunkcijski znakovi te je tekst raščlanjen samo na riječi. Takvim čišćenjem pripremljeni su korpusi koji će se naknadno analizirati.

3.2. Opis korpusa

Korpusi na koje se programski algoritam računanja entropija oslanja predstavljaju donju granicu minimalne veličine teksta pri kojoj entropije riječi još mogu biti pouzdano procijenjene (Benz, 2017). Preuzetim korpusima sa stranica Clarin i Sketch Engine znatno je reduciran broj pojavnica, jer je preko pilot projekta utvrđeno da su korpusi s 30 000 pojavnica

[8] Poveznica na stranicu Clarin: https://www.clarin.eu
[9] Poveznica na stranicu Sketch Engine: https://www.sketchengine.eu

dovoljni za pouzdano računanje vrijednosti entropija jezika. Korpusi su pohranjeni u .txt datoteke. Svakom jeziku pripada jedan korpus (ukupno 24 korpusa).

4. Metodologija

4.1. Procjena vjerojatnosti

Prvi korak u izračunavanju jezične entropije je procjena distribucije vjerojatnosti jezičnih jedinica unutar danog korpusa ili skupa podataka. Postoji nekoliko pristupa za procjenu vjerojatnosti, u rasponu od jednostavnih metoda temeljenih na učestalosti do naprednijih statističkih tehnika. Tehnike procjene vjerojatnosti koje će se koristiti u ovome radu su:

a) Vjerojatnosti temeljene na frekvenciji.

U ovom osnovnom pristupu, vjerojatnost jezične jedinice procjenjuje se dijeljenjem njezine učestalosti pojavljivanja u skupu podataka s ukupnim brojem jezičnih jedinica. Iako je jednostavna, ova metoda možda nije prikladna za rijetke ili niskofrekventne jezične jedinice, što dovodi do netočnih izračuna entropije.

b) Procjena maksimalne vjerojatnosti (eng. *Maximum Likelihood Estimation*)

Procjena maksimalne vjerojatnosti je široko korišten statistički pristup. Izračunava vjerojatnost jezične jedinice na temelju njezine opažene učestalosti u skupu podataka, uz pretpostavku da promatrani podaci predstavljaju pravu temeljnu distribuciju (Brooks-Bartlett, 2018). Ovaj pristup može biti robusniji od jednostavnih metoda temeljenih na frekvenciji, osobito za veće skupove podataka.

4.2. Program za računanje entropija

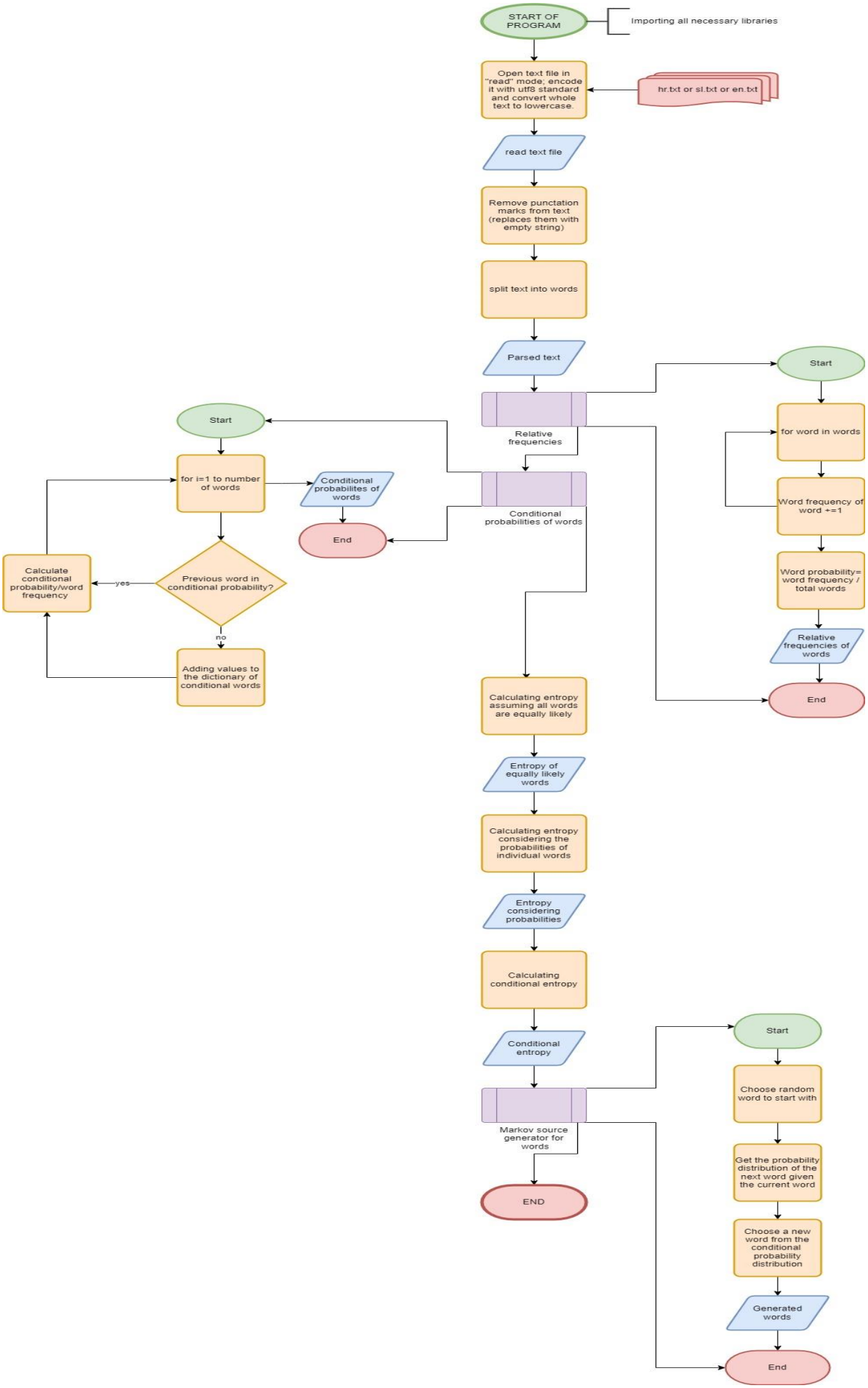
Program korišten u svrhe ovoga rada ima jasnu definiciju; jednostavan je i lako razumljiv. Napisan je u programskom jeziku Pythonu. Program analizira riječi iz teksta i pohranjuje vrijednosti relativnih frekvencija pojavnica i uvjetnih vjerojatnosti. Program računa:

1) entropiju pod pretpostavkom da su sve riječi jednako vjerojatne
2) entropiju uzimajući u obzir vjerojatnosti pojedinačnih riječi
3) kondicionalnu entropiju, uzimajući u obzir dvije susjedne pojavnice.

Program nema semantičkih pogrešaka i postiže očekivane rezultate. Napisan je u 95 linija koda. Grafički prikaz algoritma prikazan je pomoću dijagrama toka (eng. *flowchart*) na slici 2 kako bi olakšao razumijevanje složene logike unutar programa. Dijagram toka napravljen je u javno dostupnom i studentima besplatnom alatu SmartDrawFlowchart[10].

[10] Poveznica na stranicu SmartDrawFlowchart: https://www.smartdraw.com/flowchart/flowchart-maker.htm

Slika 2: Dijagram toka programa



5. Rezultati i analiza

U ovom dijelu rada testirano je predviđanje rezultata na temelju matematičkih, morfoloških i gramatičkih spoznaja te se nakon toga tablično prikazuju rezultati različitih vrsta entropija svih službenih jezika Europske unije. Nakon tabličnog prikaza rezultata detaljno se analiziraju i uspoređuju vrijednosti entropija.

5.1. Predviđanje rezultata pomoću spoznaja teorije informacije

U radu su odabrana 24 službena jezika Europske unije kako bi se vidjeli odnosi između njihovih entropija. Svakom jeziku pripada jedan zasebni korpus.

Prva entropija koju program računa je entropija pod pretpostavkom da su sve riječi jednako vjerojatne. Ova entropija bi uvijek trebala biti veća od drugih, jer se savršenom nesigurnošću smatra kada je od *n* elemenata vjerojatnost pojavljivanja elementa 1/*n*. Vrijednost navedene entropije raste prema broju različnica u tekstu.

Druga entropija uzima u obzir vjerojatnosti pojavljivanja riječi u tekstu. Za računanje ove entropije koristimo relativne frekvencije. Gornja granica ove entropije je prethodna entropija, ali samo u slučaju kada su sve pojavnice ujedno i različnice (svaka riječ se pojavljuje samo jednom). U tom konkretnom slučaju količina neizvjesnosti je najveća. U velikim tekstovima s preko 30 000 riječi bi vrijednost druge entropije uvijek trebala biti manja od prve entropije.

Posljednja entropija koju program računa je kondicionalna entropija, entropija slučajne varijable s obzirom na vrijednost druge slučajne varijable. Smanjenje nesigurnosti zbog druge slučajne varijable naziva se uzajamna informacija (eng. *mutual information*). Gornja granica uvjetne entropije je prethodna entropija s razmatranim vjerojatnostima riječi. U velikim tekstovima, ova entropija bi uvijek trebala biti najmanja zbog lančanog pravila prikazanog u formuli 5:

Formula 5: $H(X|Y) = H(X,Y) - H(X)$

5.2. Predviđanje rezultata pomoću morfoloških obilježja jezika

Postoje otprilike četiri vrste morfologije koje jezici koriste: analitička, flekcijska, aglutinativna i polisintetička. Morfološke klasifikacije rade se na temelju toga kako funkcionira morfologija jezika, odnosno kako se riječi tvore, spajaju i sklanjaju (Kantaet, 2018).

Analitički jezici se ne sklanjaju, to jest, imenice i pridjevi se ne dekliniraju, već imaju stalan oblik. U analitičkim jezicima redoslijed riječi određuje gramatičke odnose. Budući da ovim jezicima ono što nedostaje u fleksiji (ili morfologiji) mora biti nadoknađeno redoslijedom riječi ili sintaksom, oni trebaju imati vrlo precizan redoslijed riječi u rečenici, kako bi pokazali gramatičke odnose, a samim time i nešto veću kondicionalnu entropiju. Na primjer, subjekt mora biti u određenom položaju u odnosu na glagol (u engleskom mora biti prije glagola). Od promatranih jezika u radu, analitičkim jezicima pripadaju: engleski, danski, švedski te djelomično francuski, nizozemski i bugarski. Bugarski je uz makedonski jedini djelomično analitički slavenski jezik.

Flektivni jezici razlikuju se od analitičkih jezika, jer oni flektiraju, to jest mijenjaju oblik.

Flektivni jezici bi u prosjeku trebali imati veću entropiju od analitičkih jezika. Kada riječi u rečenici flektiraju kako bi se pokazalo slaganje sa svim njihovim subjektima, objektima i drugim argumentima, tada je red riječi vrlo fluidan i fleksibilan (Kantaet, 2018). Svi slavenski i baltički jezici osim bugarskog pripadaju flektivnima, ujedno s njemačkim, portugalskim, rumunjskim, španjolskim, talijanskim, grčkim, irskim i malteškim.

Aglutinativni jezici su oni u kojima se riječi mogu lako kombinirati. U tim jezicima, nove riječi nastaju spajanjem starih, tako da riječi mogu postati izuzetno duge i sadržavati puno informacija. Ova vrsta morfologije naziva se produktivnom, jer se nove riječi tvore na predvidiv način (Jäger, 2019). Aglutinativni jezici mogu imati puno slaganja i fleksija te imati slobodan redoslijed riječi. Na primjer, u aglutinirajućim jezicima nema prijedloga ispred imenica. Umjesto toga, odgovarajuće se riječi mijenjaju prema padežima. Ostale riječi, poput glagola, modificiraju se uz pomoć afiksa, što znači da se riječi mogu dodati dodatna slova ovisno o subjektu, vremenu i padežu (Jäger, 2019). Aglutinativni jezici bi zasigurno trebali imati najveće iznose entropija, a male iznose kondicionalnih entropija. Finski, mađarski, estonski i djelomično njemački su europski jezici koji se smatraju aglutinirajućima. Drugi jezici, kao što su japanski, korejski, turski, malezijski, svahili i baskijski, na primjer, nemaju nikakve veze s finskim ili mađarskim, ali ipak pripadaju aglutinirajućim jezicima.

Posljednja vrsta morfološke jezične kategorije su polisintetički jezici. Lingvisti ih najmanje razumiju, jer nijedan službeni glavni jezik u svijetu nije polisintetički Najpoznatiji polisintetički jezici su inuitski, nahuatlanski, navajo i cree jezik. Nijedan od službenih jezika Europske unije nema polisintetička obilježja, stoga se neće proučavati i analizirati u ovome radu.

Morfološki bogati jezici često su teški za obradu u mnogim zadacima. Za razliku od analitičkih jezika poput engleskog, morfološki bogati jezici kodiraju raznolike skupove gramatičkih informacija unutar svake riječi korištenjem fleksije, koje prenose karakteristike kao što su padež, rod i vrijeme. Dodatak nekoliko flektivnih inačica preko riječi dramatično povećava veličinu vokabulara, što rezultira nedostatkom podataka i problemima izvan rječnika (Balog, 2018). Zbog ovog problema, strojno prevođenje morfološki bogatih jezika je često otežano.

Sveukupno uzevši, morfološki bogatiji jezici poput slavenskih, baltičkih i ugro-finskih jezika zasigurno bi u prosjeku trebali imati veće vrijednosti entropije gdje su sve pojavnice jednako vjerojatne i entropije koja uzima u obzir vjerojatnost pojavljivanja riječi od germanskih i romanskih jezika. Ugro-finski jezici bi trebali imati najveće vrijednosti tih dviju entropija, zbog pripadnosti aglutinirajućim jezicima.

5.3. Predviđanje rezultata pomoću reda riječi u rečenici

U lingvistici, poredak riječi je poredak sintaktičkih sastavnih jedinica. Red riječi podrazumijeva poredak subjekta, predikata i objekta. Neki jezici koriste relativno fiksni red riječi, često se oslanjajući na redoslijed sastavnih dijelova za prenošenje gramatičkih informacija, poput engleskog. Drugi jezici koji informacije o gramatici prenose putem fleksije dopuštaju fleksibilniji redoslijed riječi, što rezultira manjim iznosom kondicionalne entropije.

Više od polovine jezika na svijetu ima redoslijed subjekt – objekt – predikat (SOP), no u europskim jezicima je najzastupljeniji redoslijed subjekt – predikat – objekt (SPO). Gotovo svi jezici sa SPO strukturom imaju razmjerno slobodan red riječi s gramatičkog gledišta, stoga bi trebali imati manje vrijednosti kondicionalne entropije od jezika sa SOP ili PSO strukturom.

Predikat – subjekt – objekt (PSO) je treći najčešći red riječi u rečenici. Jezici s PSO strukturom imaju izrazito fiksiran redoslijed riječi te se treba strogo pridržavati pravila kako bi rečenica imala smisla. Irski je jedini službeni jezik Europske unije s PSO strukturom te bi trebao imati najveću kondicionalnu entropiju.

5.4. Rezultati i analiza

Entropije 24 službena jezika Europske unije prikazane su u tablici 2. Prvi stupac prikazuje jezik koji se govori u promatranoj državi, drugi stupac prikazuje vrijednosti entropije gdje se nisu uzimale u obzir vjerojatnosti pojavljivanja riječi te se smatra da su sve riječi jednako

vjerojatne, treći stupac prikazuje vrijednosti entropije koja uzima u obzir vjerojatnost pojavljivanja riječi, a četvrti stupac prikazuje vrijednosti kondicionalne entropije za promatrani jezik. Jezici su prikazani različitim bojama po jezičnim skupinama, te se nakon svake jezične skupine nalazi redak „PROSJEK“ gdje su izračunate prosječne vrijednosti entropija za jezičnu skupinu. Zadnji redak prikazuje ukupni prosjek svih entropija u jednom stupcu. Ispod tablice 2 nalazi se kazalo gdje je pojašnjeno koja boja prikazuje odgovarajuću jezičnu skupinu.

Tablica 2: Vrijednosti entropija službenih jezika Europske unije

Jezik	Entropija-jednake vjerojatnosti	Entropija	Kondicionalna entropija
danski	12,328394438351941 bit	10,0459155992856 bit	5,29618422593958 bit
švedski	12,453515216150654 bit	10,5021327589851 bit	4,7855447579756545 bit
engleski	12,514795455858846 bit	9,39097322212765 bit	5,217754562842618 bit
njemački	12,710451351505047 bit	10,050096654866601 bit	5,217005840509424 bit
nizozemski	12,388555503981635 bit	10,2571371226658 bit	5,126106043609722 bit
PROSJEK	**12,45914239 bit**	**10,15725107 bit**	**5,128519086 bit**
francuski	12,578136783415047 bit	10,10809208507627 bit	5,065930458266714 bit
portugalski	12,699355555588197 bit	10,24050210413784 bit	4,911977242887317 bit
rumunjski	12,868243817706439 bit	10,63176663095894 bit	4,563606344728425 bit
španjolski	12,627533884471145 bit	10,03838469089112 bit	5,006280841282505 bit
talijanski	12,370959793299013 bit	10,43196999636297 bit	5,212729309174279 bit
PROSJEK	**12,62884597 bit**	**10,1963885 bit**	**4,952104839 bit**
bugarski	13,279320387940494 bit	11,12212802569278 bit	4,057935824415646 bit
češki	12,890454292833704 bit	11,02261051545075 bit	4,165849880593503 bit
poljski	13,295912393378881 bit	11,01777675452583 bit	3,851117173898693 bit
slovački	12,804728239244283 bit	11,12191828320113 bit	4,172747383783744 bit
hrvatski	13,140030892055576 bit	11,03007894881778 bit	4,150594445729671 bit
slovenski	13,194430852555624 bit	11,11077494628379 bit	4,095749382734822 bit
PROSJEK	**13,10081284 bit**	**11,07088125 bit**	**4,082332349 bit**
latvijski	12,656871725614936 bit	11,102414166886124 bit	3,768041842668554 bit
litavski	12,259847483939472 bit	11,198748319365494 bit	3,884619473617234 bit
PROSJEK	**12,4583596 bit**	**11,15058124 bit**	**3,826330658 bit**
grčki	12,952195219912486 bit	10,258846642700288 bit	4,654276987097419 bit
irski	**12,209453365630207 bit**	**9,68316836707176 bit**	**5,377616869416009 bit**
estonski	13,706147528020653 bit	11,578674570287975 bit	**3,387247713308328 bit**
finski	**13,836266686773508 bit**	**11,870843593331233 bit**	3,659460953701580 bit
mađarski	13,535562522934518 bit	11,311736342464039 bit	3,905567597106103 bit
PROSJEK	**13,69265891 bit**	**11,58708484 bit**	**3,650758755 bit**
malteški	12,577361917374636 bit	10,965819231315312 bit	4,974612645162664 bit
UKUPNI PROSJEK	**12,81160535 bit**	**10,67348902 bit**	**4,533314908 bit**

Iz tablice vidljivo je da najveći ukupni prosjek ima entropija pod pretpostavkom da su sve riječi jednako vjerojatne (12,81160535 bitova), potom entropija koja uzima u obzir vjerojatnosti pojavljivanja riječi (10,67348902 bitova), a najmanji iznos ima kondicionalna entropija, kao što je i bilo pretpostavljeno zbog lančanog pravila.

Najveću entropiju jednakih vjerojatnosti ima finski jezik, koja iznosi 13,836266686773508 bitova. Zajedno s finskim, jedini jezici koji imaju prvu entropiju preko 13,5 su estonski i mađarski, a sva tri jezika pripadaju istoj jezičnoj skupini, ugrofinskoj. To je zato što su to aglutinativni jezici, što dolazi od latinske riječi *agglutinare*, „lijepiti zajedno“. U ovoj obiteljskoj skupini dijelovi jezika nanizani su zajedno kako bi tvorili složenije riječi, stoga imaju veći broj različnica, a ujedno i veće vrijednosti entropije s jednakom vjerojatnošću riječi i entropije koja uzima u obzir relativne frekvencije riječi. Ugro-finski jezici također imaju i drugu entropiju koja uzima u obzir vjerojatnost pojavljivanja riječi poprilično veću od ostalih jezika. Zbog velikog broja različnica, ugro-finskim jezicima slabo je ograničen broj riječi koje mogu doći nakon sljedeće riječi, stoga su im vrijednosti kondicionalnih entropija male. Estonski jezik ima najmanju kondicionalnu entropiju od svih promatranih jezika.

Germanski i romanski jezici imaju vrlo slične vrijednosti svih triju entropija, stoga je malo vjerojatno ili gotovo nemoguće da se po iznosu entropija prepozna jezik ili pripadnost jezika jezičnoj skupini. Engleski je jedini u ovim jezičnim skupinama koji ima drugu entropiju manju od 10 bitova, vjerojatno zbog učestalog pojavljivanja članova „the“, „a“ i „an“ koje predstavljaju vrstu odrednice te idu ispred imenice. Neodređeni članovi „a“ i „an“ koriste se prije imenice kako bi se označilo da je identitet imenice nepoznat čitatelju, dok određeni član „the“ ukazuje da je identitet imenice poznat čitatelju. Germanski i romanski jezici imaju u prosjeku visoke iznose kondicionalnih entropija zbog znatno fiksnijeg redoslijeda riječi u rečenici s obzirom na ostale službene jezike Europske unije, izuzev irskog koji ima najveću kondicionalnu entropiju.

Slavenski i baltički jezici imaju slične iznose entropija te se smatraju bliskim jezicima. Većina jezičnih stručnjaka svrstava ih u baltoslavensku jezičnu obitelj te smatraju da se ova obitelj razvila od prabaltoslavenskog jezika. Imaju u prosjeku veće entropije od germanskih i

romanskih jezika jer pripadaju u flektivnije jezike, s dominacijom vanjske fleksije, tj. dodavanjem sufiksa ili prefiksa. U prosjeku imaju manju kondicionalnu entropiju nego germanski i romanski jezici jer ih karakterizira bogata morfologija i fleksibilniji redoslijed riječi u rečenici.

Irski, koji spada u keltsku jezičnu obitelj ima najmanje iznose prve i druge entropije (12,209453365630207 bitova i 9,68316836707176 bitova) od svih promatranih jezika, ali ima najveću kondicionalnu entropiju (5,377616869416009 bitova). Kao u većini keltskih jezika, poredak riječi je vrlo fiksan (predikat – subjekt – objekt), stoga je iznos kondicionalne entropije izuzetno visok.

6. Zaključak

U ovom radu istražuje se povezanost iznosa jezičnih entropija svih službenih jezika Europske unije s njihovom jezičnom skupinom kako bi se poboljšali sustavi prepoznavanja jezika. Kroz ovo sveobuhvatno istraživanje jezične entropije, razotkrili smo višestruku prirodu jezične složenosti i njezin utjecaj na ljudsku komunikaciju. Teorijski temelji jezične entropije, ukorijenjeni u informacijskoj teoriji Claudea Shannona, pružili su osnovu za kvantificiranje nesigurnosti i gustoće informacija unutar jezičnih sustava.

Pomoću računalne metode za mjerenje entropija jezika u ovome radu se analiziraju svi službeni jezici Europske unije. Ispitali smo međusobni odnos entropija, odnos između entropije jezika i morfoloških jezičnih obilježja te odnos entropije i redoslijeda riječi u rečenici, ističući kako jezici s visokom entropijom nude prilagodljivost i inovativnost, dok jezici s niskom entropijom daju prednost stabilnosti i učinkovitosti.

Međujezična entropijska analiza pružila je uvid u jezičnu raznolikost. Promatramo kako različite jezične obitelji i tipološke kategorije pokazuju različite razine entropije, odražavajući njihovu komunikacijsku učinkovitost i preferencije složenosti. Takve analize pridonose našem razumijevanju lingvističke tipologije između jezičnog kontakta i prijenosa entropije.

Rezultati rada potvrđuju da se mnogo informacija o jezicima može naučiti iz vrijednosti entropija, poput fiksnog ili fleksibilnog redoslijeda riječi, pripada li jezik morfološko bogatim ili siromašnim jezicima i slično. Sveukupno uzevši, sami rezultati entropija nisu dovoljni da bi se odredilo kojem jeziku pripadaju, eventualno bi se moglo odrediti kojoj jezičnoj skupini entropija promatranog jezika pripada.

Čini se da bi ovakav pristup proučavanja jezika mogao pružiti nove uvide jezičnim stručnjacima te ih potaknuti na neka nova istraživanja. Kako se polje jezične entropije nastavlja razvijati, postoje različiti putevi za buduća istraživanja. To uključuje inovacije novih dodatnih tehnika mjerenja entropije, promatranje utjecaja jezične entropije na višejezičnu kogniciju i ispitivanje dinamičke interakcije između entropije i jezične promjene tijekom duljih razdoblja.

Zaključno, ovo sveobuhvatno istraživanje usporedba jezičnih entropija 24 službena jezika Europske unije rasvijetlilo je temeljne aspekte jezične složenosti i njezinu višestruku ulogu u ljudskoj komunikaciji. Analizirajući entropiju jezika iz teorijske, računalne i praktične perspektive, dobili smo dragocjene uvide između entropije i jezične evolucije. Proučavanje

jezične entropije ima duboke implikacije na jezičnu raznolikost, međukulturnu komunikaciju, učenje jezika i razvoj naprednih jezičnih tehnologija.

Dok nastavljamo ulaziti u složenost jezične entropije, utiremo put dubljem razumijevanju ljudske sposobnosti za jezik i njegove izvanredne prilagodljivosti u različitim jezičnim kontekstima. Prihvaćanjem izazova i prilika koje predstavlja istraživanje jezične entropije, možemo obogatiti svoje razumijevanje prirodnog jezika, potičući inkluzivne jezične prakse i promičući smislene interakcije u našem sve više međusobno povezanom svijetu.

7. Literatura

1. Ahmad, M.I., Khan, I.A., Ahmad, M. et al. Entropy in Education System: Transformation of an Individual Through Meaningful Interactions in a Community of Inquiry. Syst Pract Action Res 35, 591–606, 2022. https://doi.org/10.1007/s11213-021-09585-6b
2. Baayen H.; Lieber R. Word Frequency Distributions and Lexical Semantics. Computers and the Humanities. Vol.30, pp. 281–297, Netherlands, Amsterdam: mo Kluwer Academic Knowledge Publishers, 1997. Dostupno na: https://pure.mpg.de/rest/items/item_2629773_3/component/file_2630611/content
3. Balog, N.. Morfološke pogreške kod govornika s afazijom, Diplomski rad, Sveučilište u Zagrebu, Edukacijsko-rehabilitacijski fakultet, 2018. Dostupno na: https://urn.nsk.hr/urn:nbn:hr:158:061303
4. Benz, C. Alikaniotis, D. Cysouw, M. The Entropy of Words: Learnability and Expressivity across More than 1000 Languages. Non-Equilibrium Dynamics and Self-Organisation. 2017. Dostupno na: https://www.mdpi.com/1099-4300/19/6/275
5. Brincat, J. M. Maltese, a Language so Unique in Europe. Orient XXI, 2022. Dostupno na: https://orientxxi.info/magazine/maltese-a-language-so-unique-in-europe,5590
6. Brooks-Bartlett, J. Probability concepts explained: Maximum likelihood estimation. Towards Dana Science, U.S.A., 2018. Dostupno na: https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1
7. Cook, J. D. Cologarithms and Entropy, D. Consulting, 2021. Dostupno na: https://www.johndcook.com/blog/2021/04/30/cologarithm/
8. Darian, S. The role of redundancy in language and language teaching. System. Vol. 7, 1, pp. 47–59. U.S.A, New Jersey: Rutgers University, 1979. Dostupno na: https://www.sciencedirect.com/science/article/abs/pii/0346251X79900228
9. Douglas, F. E. Discrete Random Variable: Transforms and Transform Properties.Handbook of Digital Signal Processing. USA: Chicago, 1987. Dostupno na: https://www.sciencedirect.com/topics/engineering/discrete-random-variable .

10. Europa u pokretu: Mnogo jezika, jedna obitelj. 2023. Dostupno na: http://www.europe.hr/documents/knjiga%20%20mnogo%20jezika,%20jedna%20obitelj.pdf
11. Finn, E. The advantage of ambiguity. MIT News Office, 2012. Dostupno na: https://news.mit.edu/2012/ambiguity-in-language-0119
12. Gabora, L. A possible role for entropy in creative cognition. In Proceedings of the 3rd International Electronic Conference on Entropy and its Applications. Sciforum Electronic Conference Series, Vol.3, E001, 2016. Dostupno na: https://sciforum.net/conference/84/paper/3652
13. Hausser, J.; Strimmer, K. Entropy: Estimation of Entropy, Mutual Information and Related Quantities; Package Version 1.2.1, 2014. Dostupno na: https://CRAN.R-project.org/package=entropy
14. Henning, P. Schuler, C. J. Entropy Search for Information: Efficient Global Optimization. Journal of Machine Learning Research Vol.13, 1809-1837, 2012. Dostupno na: https://jmlr.csail.mit.edu/papers/volume13/hennig12a/hennig12a.pdf
15. Jalde, M. Celtic Languages, 2018. Dostupno na: https://hr.wikipedia.org/wiki/Keltski_jezici
16. Jäger G. Languages of the World: Morphological language classification. Universität Tübingen, Njemačka, 2019. Dostupno na: http://www.sfs.uni-tuebingen.de/~gjaeger/lehre/ws0910/languagesOfTheWorld/morphologicalTypology.pdf
17. Kantaet, K. Morphological Typology. Lingonet, 2018. Dostupno na: https://www.linguisticsnetwork.com/morphological-typology/
18. Klein, A. Intuitive Explanation of the Concept of Entropy in Information Theory? Stanford, 2021. Dostupno na: https://www.quora.com/What-natural-languages-have-the-highest-information-entropy-on-the-character-level
19. Levitin L. B.; Reingold Z. Entropy of natural languages: Theory and experiment. Chaos, Solitons & Fractals Vol.4, Issue 5, pp. 709-743, 1994. Dostupno na: https://www.sciencedirect.com/science/article/pii/0960077994900795/
20. Lo, J. The Entropy of Language. Language Insights, 2018. Dostupno na: https://medium.com/language-insights/the-entropy-of-system-life-and-language-43d89c0d185b

21. Montemurro, M. A. Universal Entropy of Word Ordering Across Linguistic Families. PLoS One.; v6(5): e19875, 2011. Dostupno na: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3094390/
22. Orlinsky, A. Information Theory: Conditional Entropy. Journal of Network and Computer Applications. Vol.60 pp. 19–31, 2016. Dostupno na: https://www.sciencedirect.com/science/article/abs/pii/S1084804515002891
23. Petersen, K. Information and Coding. AMS 2018. London: Chelsea Publishing, 2018.
24. Poir, M. Entropy in Machine Learning. Java T Point, 2011. Dostupno na: https://www.javatpoint.com/entropy-in-machine-learning
25. Razumovskaya, V. Information Entropy of Literary Text and its Surmounting in Understanding and Translation. Journal of Siberian Federal University. Humanities & Social Sciences Vol.2, 259-267, 2010. Dostupno na: https://core.ac.uk/download/pdf/38633204.pdf
26. Shannon, C. E. A Mathematical Theory of Communication. The Bell System Technical Journal. Vol.27, pp. 379–423, 623–656, 1948. Dostupno na: https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf
27. Shannnon, C. E. Prediction and Entropy of Printed English. New Jersey: Princeton University, 1950. Dostupno na: https://www.princeton.edu/~wbi /refs/shannon_51.pdf
28. Vladisavljević, M.; Bogetić, K. Kognitivna obrada sintaktički dvosmislenih konstrukcija:Uvidi iz obrade konstrukcija s globalno dvosmislenim rečenicama. Psihologijske teme, 30, Pregledni rad 3, 421-446, 2021. Dostupno na : https://doi.org/10.31820/pt.30.3.2
29. Wei, Y. Entropy as a measurement of cognitive load in translation. Centre for Translation and Textual Studies, Dublin City University, Ireland. 2022. Dostupno na: https://aclanthology.org/2022.amta-wetpr.8.pdf

8. Popis slika i tablica

Analiza i usporedba entropija službenih jezika Europske unije

Sažetak

Entropija je znanstveni koncept popularan među područjima znanosti koje se bave kvantificiranjem raznolikosti i nesigurnosti fenomena. Jezična entropija je statistički parametar koji mjeri količinu informacije sadržane u promatranom tekstu. Ovo područje objedinjuje lingvistiku, teoriju informacije i računalnu tehnologiju te zahtijeva interdisciplinarni pristup. U radu se uspoređuju entropije službenih jezika Europske unije. Jezik nosi značenje u riječima koje biramo, ali isto tako i u redoslijedu kojim ih postavljamo, stoga će se u radu uspoređivati i kondicionalne entropije. Mogu li se uspostaviti odnosi između jezika i vrijednosti pripadajućih entropija te može li se predvidjeti prema iznosu entropija o kojem je europskom jeziku riječ, glavna su pitanja kojima se bavi ovaj rad. Identificirana je minimalna veličina korpusa potrebna za ovu temu. Radi dosljednosti rezultata, odabrani su korpusi koji sadrže istu temu, a to su parlamentarne rasprave i debate, preuzeti sa stranica Clarin i Sketch Engine.

Ključne riječi: entropija, kondicionalna entropija, jezik, korpus

Analysis and comparison of the entropy of European Union languages

Summary

Entropy is a scientific concept popular among fields of science that deal with quantifying the diversity and uncertainty of phenomena. Linguistic entropy is a statistical parameter that measures the amount of information contained in the observed text. This field combines linguistics, information theory, computer technology and requires an interdisciplinary approach. The paper investigates the entropy comparisons of the languages of the European Union. Language carries meaning in the words we choose, but also in the order in which we place them, therefore the paper will also compare conditional entropies. Whether it is possible to establish relationships between languages and the value of the corresponding entropies and whether it can be predicted according to the amount of entropy of the European language are the main questions that this paper deals with. The minimum corpus size required for this topic is identified. The amounts of entropy depend on the jargon and the amount of vocabulary, therefore, for the sake of consistency of the results, parliamentary discussions and debates are the topics of all corpora.

Key words: entropy, conditional entropy, language, corpus.