Slavko Stojanović
(Beograd)

JEZIK I INFORMACIJA

Poznato je da se jezička komunikacija ostvaruje kroz proces kombinovanja jezičkih jedinica koje se raspodjeljuju prema pravilima verovatnoće. To je osnovna karakteristika svakog prirodnog jezika. Rezultat takvog kombinovanja su, između ostalog, i reči od kojih neke imaju visoke (stabilne) frekvencije dok se druge retko pojavljuju tj. karakterišu ih niske frekvencije. Još je Pjer Giro (Pierre Guiraud) zapazio da je frekvencija jezičkih jedinica jedna od najvažnijih karakteristika jezika (1). Polazeći od ove činjenice, a na osnovu statističkih podataka, moguće je vršiti veoma različite kvalitativne analize jezika. Ovoga puta našu pažnju je zaokupio odnos jezika kao informacionog sistema i količine informacije koja se njime prenosi.

Međutim, pre toga je potrebno da se podsetimo da su stabilnost frekvencija jezičkih jedinica i njihova distribucija dva osnovna aksioma od kojih polazi teorija informacija i komunikacije (2,162–164). Teorija informacija i komunikacija se bavi problemima prenosa informacija, ali je ona isto tako našla primenu i u lingvistici.

Šta je zapravo informacija? U svakodnevnom životu za nas je informacija spoznaja o nekoj stvari, događaju, pojmu, pojavi itd. Ako je poruka koju primamo nova, ako nosi veći stepen iznenađenja, onda je i informacija veća ili kako Giro kaže: »Veličina informacije ili količina informacije sadržana u jednoj poruci zavisi od našeg poznavanja koje imamo o pojavi na koju se poruka odnosi« (1,72).

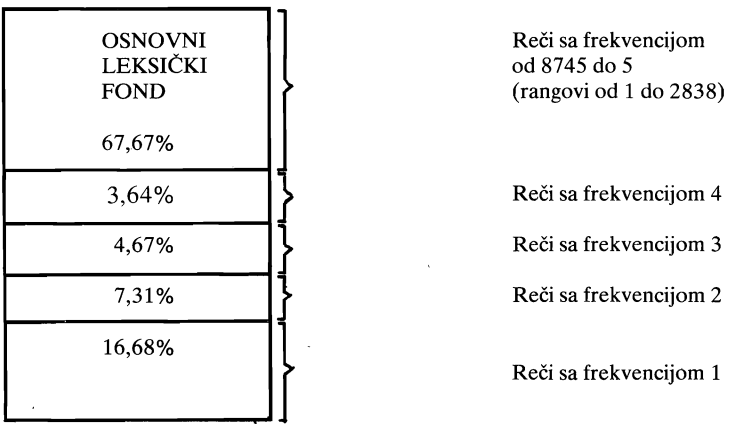
Koristeći se nekim osnovnim statističkim podacima frekventne liste arapskog jezika dnevne štampe dobijene pomoću kompjutera i sređene prema opadajućim frekvencijama, pokazaćemo u kakvom su odnosu jezik i informacija. Na slici broj 1 je prikazana raspodela frekvencija reči frekventne liste dobijene na uzorku nešto većem od 100.000 reči. Uzećemo prvih 50 reči sa frekventne liste. Odgovarajuće podatke ćemo uneti u koordinatni sistem tako što ćemo brojčane vrednosti rangova (redni broj reči na frekvntnoj listi) naneti na apscisu, a frekvenciju reči na ordinatu.

Prva reč sa rangom 1 ima frekvenciju 8745, sledeća sa rangom 2 ima frekvenciju 3317 i tako redom. Pošto se radi o frekventnoj listi sređenoj prema opadajućoj frekvenciji, sa porastom ranga, frekvencija reči opada, te prema tome u našem slučaju negde oko 10.000 ranga frekvencija reči ima vrednost 1.

To se jasno vidi na slici 2, gde se sa opadanjem frekvencije reči kriva postepeno, u početku brže a zatim sporije, približava apscisi da bi kod reči sa frekvencijom 1 dostigla najnižu vrednost. Shodno onome što smo rekli o informaciji moglo bi se zaključiti da reči koje se u jeziku često pojavljuju, tj. reči koje imaju visoke i stabilne frekvencije nose malu količinu informacije, dok, suprotno tome, retke reči nose veću količinu informacije.

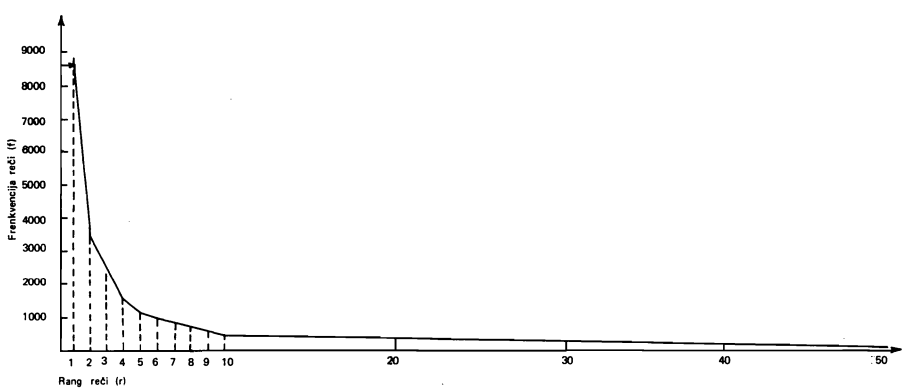
Prema tome, od frekvencije reči (njene verovatnoće) zavisi i količina informacije. Znači, reči sa visokim frekvencijama u jeziku nose malu količinu informacije, dok manje prisutne reči sa niskim frekvencijama nose veću količinu informacije. Zaključak koji u tom smislu važi za sve prirodne jezike je da su frekvencija reči (njena verovat-

RASPODELA FREKVENCIJA REČI FREKVENTNE LISTE



Slika 1

noća) i količina informacije dve obrnuto proporcionalne veličine. Ovde nije potrebno naglašavati da statistiku, matematiku i brojne matematičke modele (koji se danas sve više primenjuju u lingvistici) kao i rezultate koji se na taj način dobijaju o činjenicama jezika treba, s obzirom na prirodu jezika, posmatrati sa izvesnim stepenom fleksibilnosti.



Slika 2

Ono što je prikazano na slici 2 prikazaćemo na drugi način, tj. izračunaćemo količinu informacije za nekoliko reči prema formuli $I = -\log_2 p_i$ koju koristi Glison (Gleason) (3). Uzeli smo nekoliko reči sa visokim frekvencijama i jednu sa frekvencijom 1. To je prikazano na slici broj 3 gde se može videti da prva reč sa frekventne liste koja ima frekvenciju 8745 nosi svega 3,52 bita informacije da bi reč na 10.000 rangu nosila neuporedivo veću količinu informacije od 102,040 bita.

Kada frekventnu listu posmatramo sa aspekta količine informacije koju nosi svaka reč, neminovno se nameće pitanje korisnosti frekventne liste za učenje jezika. Naime,

da li uopšte ima smisla praviti frekventne liste i iz njih izdvajati osnovni leksički fond koji sadrži najfrekventnije reči, kada su one najmanje informativne.

rang reči	reč	aps. fr.	p_i (rel. fr.)	I (informacija)
1	WA	8745	0,086	3,52 bita
2	FĪ	3317	0,033	4,91 bita
3	MIN	2260	0,022	5,51 bita
4	ʿALĀ	1396	0,014	6,19 bita
5	ʾILA	1076	0,011	6,50 bita
10	HĀḌA	498	0,005	7,64 bita
50	ʿARABIYYU	118	0,0012	9,70 bita
10.000	ʾANHARI	1	0,0000098	102,040 bita

Slika 3

Ovo pitanje bi moglo biti predmet zasebnog rada, jer su frekventne liste od značaja za različita lingvistička istraživanja, teoriju informacija i komunikacije, medicinu, pedagogiju, psihologiju, fiziku itd. i naravno za učenje stranih jezika. Prema tome, frekventne liste imaju određeni značaj, jer i u učenju jezika kao i u svakom drugom poslu, treba početi od osnovnih stvari koje će poslužiti kao baza za učenje i nadgradnju.

Posmatrajući frekventnu listu kao izvor podataka za različita lingvistička i vanlingvistička istraživanja, Giro kaže: »Frekventna reč je najkorisnija u celini, dok je retka reč najkorisnija u svakom pojedinačnom slučaju« (1,95).

Posmatraćemo 1. i 10.000. reč sa frekventne liste. Uzećemo jednu meru koja se naziva koeficijentom korisnosti C_k, a koja predstavlja odnos apsolutne frekvencije date reči i ukupnog broja upotrebljenih reči u uzorku. Prema tome, za prvu reč sa frekvencijom 8745 imamo 8745/101.192 = 0,0861, dok za 10.000. reč sa frekvencijom 1 koeficijent korisnosti je 1/101.192 = 0,0000098.

Za našu frekventnu listu koeficijent korisnosti se kreće između ove dve vrednosti (od 0,0861 do 0,0000098), što znači da sa padom frekvencije reči opada i vrednost ovoga koeficijenta.

Valja skrenuti pažnju da smo navedeni koeficijent primenili na arapski jezik da bismo i ovde pokazali da reči sa visokim frekvencijama imaju određeni značaj za učenje jezika, jer kada usvojimo jednu reč sa visokom frekvencijom, na primer prvu WA, mi smo iz uzorka od 101.192 reči uslovno rečeno »savladali« 8745 reči, jer je frekvencija prve reči tolika. Međutim, s obzirom na to da ovde govorimo o koeficijentu korisnosti reči neophodno je dati jedno objašnjenje. Naime, u obradi ove teme korišćeni su čisto kompjuterski podaci, odnosno frekvencije reči – gramatičkih oblika, bez prethodne analize semantike svake reči. Dakle, u ovom slučaju frekvencija reči je odraz njene grafičke identičnosti u ispitivanom uzorku, odnosno u pitanju je jedno formalno posmatranje reči. Izvesno je da bi prethodna ručna priprema celokupnog uzorka u semantičkom smislu, posle koje bi usledila kompjuterska obrada, dala više materijala za različite lingvističke analize. No, to bi bio jedan veoma složen i opsežan posao. Ipak, čini se da se i kroz jedan ovakav pristup jezičkom materijalu mogu prikazati opšti stavovi o odnosu jezika i informacije. Međutim, kada se govori o koeficijentu korisnosti reči izračunatoj na bazi njene grafičke identičnosti (bez ulaženja u njenu semantiku) mora se imati izvesna rezerva. Razlog ovome je što mnoge reči u arapskom jeziku, prevashodno one sa visokim frekvencijama, mogu imati različita pa čak i oprečna značenja. Dakle, vrednost koeficijenta korisnosti reči, računata prema njenoj grafičkoj identičnosti u tekstu, bila bi donekle različita od vrednosti ovog

koeficijenta dobijenog na osnovu frekvencije reči i njene semantike. Na primer, uzmimo reč X sa frekvencijom f_i koja ima pet različitih značenja. Ako u ovom slučaju zapostavimo broj značenja reči X, odnosno njenu semantiku i uzmemo u obzir samo njenu frekvenciju, dobićemo određeni koeficijent korisnosti date reči. Međutim, vrednost ovako dobijenog koeficijenta reči će se unekoliko razlikovati od vrednosti koeficijenta za istu reč u situaciji kada se uzima u obzir njena semantika. Do razlike u vrednostima koeficijenta korisnosti ne dolazi u slučajevima kada reč ima samo jedno značenje.

I na kraju valja reći da nam nije bila namera da se detaljnije bavimo odnosom jezika i informacije, već da, neke osnovne statističke podatke frekventne liste koja je rezultat jednog našeg ranijeg istraživanja, primenimo na arapski jezik i potkrepimo jednu opštu relaciju između jezika i informacije.

Rezime

JEZIK I INFORMACIJA

Arapski, kao i svaki drugi prirodni jezik, odlikuje se različitim frekvencijama jezičnih elemenata, te prema tome i različitim frekvencijama reči. Svaka reč u jeziku ima različitu verovatnoću upotrebe, odnosno pojavljivanja. Frekvencija reči određuje količinu informacije koju ona nosi, odnosno ukoliko je reč frekventnija njena informativnost je manja i obrnuto, što je potvrđeno na većini prirodnih jezika.

Resume

LA LANGUE ET L'INFORMATION

Dans cet article on a montré certaines rélations entre la langue arabe et l'information. Sur la base des études déjà connues dans la statistique linguistique et la théorie de l'information on a construit une liste des fréquences des mots arabes pris des quotidiens arabes. Une liste tellement formée nous donne la possibilité de dire et de concluire que la fréquence du mot (ou sa probabilité d'occurrence) et la quantité de l'information qu'il porte sont les deux valeurs inversement proportionnelles, c'est-à-dire, si le mot est plus fréquent la quantité de l'information qu'il porte est plus petite et vice-versa.

Aussi, on a montré que le mot le plus fréquent est en général le plus utile, tandis que le mot rare est le plus utile dans chaque cas séparé. Cela peut-être illustré par le coefficient de valeur du mot. Mais, dans notre échantillon les mots se trouvent dans sa forme gramaticale, c'est-a-dire, qu'il s'agit de sa présentation graphique dans le texte et que le coefficient de valeur du mot diffère de celui quand la sémantique du mot est incule dans l'étude. Bien sûr, cela est valable pour les mots qui ont plus d'un sens.

Literatura:

1. Guiraud Pierre, *Problèms et methodes de la statistique linguistique*, PUF, Paris 1960.
2. Herdan G., *Language as Choice and Chance*, Groningen, 1956.
3. Martinet André, *Elements of General lingvistics*, The University of Chicago, Press, Chicago, 1966.