

Ռ. Ս. Ղազարյան

ՏԱՌԵՐԻ ՀԱՃԱԽՈՒԹՅՈՒՆԸ ՀԱՅԵՐԵՆ ՏԵՔՍՏՈՒՄ

Հայերեն տեքստի էնտրոպիայի գնահատման հետ կապված աշխատանքների համար անհրաժեշտ եղավ հաշվել հալոց տառերի հաճախությունը իմաստավորված տեքստում: Նպատակահարմար ենք գտնում հրապարակել այդ աշխատանքը, քանի որ նրանք կարող են հետաքրքրություններ կալացնել ինքնուրույն:

Հաշվիչ մեքենաների օգտագործման հնարավորությունը բացառություններ պայմանավորեց վիճակագրական աշխատանքները կրճատելու բնական տեղեկները: Այդ նպատակով կիրառվել է հետևյալ մեթոդիկան՝ տեքստի համեմատարար փոքր ժամալի քաղվածքից (3—4 հազար նիշ) հաշվել է տառերի հաճախությունը: Այնուհետև, տառերի հաճախությունների այդ կոպիտ արժեքները օգտագործվել են (իրեն հավանականություններ) այն անհրաժեշտ քաղվածքի ժամալը որոշելու համար, որի դեպքում ապահովվում է հաճախության տրված ճշտությամբ մոտեցումը իր հավանականությունները: Վերջինն ստացվել է Մուավրի-Լապլասի բանաձևից, այն ենթադրությամբ (ինչպես այդ ընդունված է զործնական վիճակագրությունում), որ 0.9973 հավանականություն ունեցող դեպքը կտրելի է համարել ստույգ: Այն ժամանակ Մուավրի-Լապլասի բանաձևն ընդունում է հետևյալ պարզ տեսքը՝

$$|f_N - P| < 3 \sqrt{\frac{P(1-P)}{N}},$$

այսինքն գործնականորեն ստույգ է, որ N ժամալի քաղվածքից հաշված f_N հաճախության շեղումը իր P հավանականությունից չի գերազանցի

$$3 \sqrt{\frac{P(1-P)}{N}}$$

արժեքը:

Քաղվածքի ժամալի (N) վերջնական կոտորված կառուցվել է տառերը հաշվելու պրոցեսի ընթացքում:

Բերված աղյուսակը կաղծելիս, ընդունվել է, որ հավանականություն որոշման մաքրիմալ հարաբերական սխալանքը հավասար է 10^{-3} -ի:

Հետադրուկվել է ժամանակակից խառը գեղարվեստական արձակ տեքստը՝ Ինտերվալը հաշվված է ինչպես տաս:

* Օգտագործվել են «Սովետական զրահանություն» ամսագրի 1959 թ. 1—7 համարները:

Ավելորդ չէ նշել, որ անհրաժեշտ քաղվածքի ծավալը զգալիորեն տարբերվում է առանձին տառերի համար (մի քանի հազար նիշից, ամենահաճախակի տառերի համար, մինչև մի քանի հարյուր հազար և միլիոն նիշ հազվագյուտ տառերի համար):

Առաջին տողի տվյալները հնարավորություն են տալիս հաշվել ժամանակակից հայերեն բառի l_s միջին երկարությունը (տառերով) գեղարվեստական տեքստում՝

$$l_s = \frac{1}{0,177} - 1 \simeq 4,7 \text{ տառ:}$$

Համեմատության համար ասենք, որ անդրերեն բառի միջին երկարությունը 4.5 տառ է, սուսերենինը՝ 5,1 տառ: Պետք է նշել, որ այդ մեծությունը զգալի տատանումների է ենթարկվում տարբեր ախպի տեքստերում և տարբեր ժամանակաշրջանների համար:

Մի առանձնահատկություն ևս, կթև դիտենք այբուբենը էֆեկտիվ կոդերի տեսանկյունից (ամենահաճախակի տառին՝ ամենակարճ և հասարակ նիշը), ապա պիտի ընդունենք, որ «ա» տառի նշանը հաջող չէ: Իրոք, կթև միայն «ա» տառի նշանակումը լինել կրկնակի ավելի կարճ (ասենք, «ո»-ի կամ «օ»-ի նման), ապա տեքստի միջին երկարությունը կնվազել մոտավորապես $5-6\%$ -ով:

№	Տառ	Հաճախությունը	№	Տառ	Հաճախությունը	№	Տառ	Հաճախությունը	№	Տառ	Հաճախությունը
1	Ինտերվալ	0,177	13	ժ	0,017	25	բ	0,0088	37	փ	0,0021
2	ա	0,114	14	ց	0,016	26	պ	0,0082	38	ճ	0,0016
3	ն	0,083	15	լ	0,016	27	չ	0,0076	39	օ	0,00072
4	բ	0,075	16	դ	0,015	28	ծ	0,0073	40	Ֆ	0,00025
5	ե	0,054	17	հ	0,014	29	չ	0,0070	41	(.)	0,014
6	ի	0,048	18	ք	0,014	30	դ	0,0051	42	(:)	0,010
7	ո	0,033	19	զ	0,013	31	խ	0,0044	43	(')	0,0014
8	ու	0,030	20	թ	0,012	32	ե	0,0043	44	(.)	0,0012
9	մ	0,025	21	ը	0,012	33	ջ	0,0042	45	(...)	0,0011
10	կ	0,024	22	է	0,011	34	դ	0,0041	46	(")	0,0011
11	տ	0,024	23	դ	0,0094	35	ձ	0,0035	47	(')	0,00046
12	ս	0,020	24	ս	0,0090	36	յ	0,0033	48	(')	0,00040

Վիճակագրական աշխատանքը պրոֆ. Գ. Սեակի սիրալիի համաձայնությամբ կատարել են համալսարանի հայկական բանասիրական ֆակուլտետի IV կուրսի ուսանողները, որոնց և հեղինակը հալտնում է իր խորին շնորհակալությունը:

Р. А. Казарян

Частота букв в армянском тексте

Р е з ю м е

Подсчитаны частоты букв в современном армянском смешанном художественном тексте. В целях сокращения статистической работы, из выборки малого объема (3—4 тыс. знаков) подсчитаны грубые частоты, которые затем использованы как вероятности в формуле Муавра-Лапласа (по правилу трех сигма) для нахождения выборки, обеспечивающей заданное приближение частоты к вероятности.

Окончательная коррекция объемов выборок проводилась в процессе подсчетов. При составлении таблицы принято, что максимально возможное отклонение частоты от вероятности равно 10%.