Բ. Կ. ՂԱԶԱՐՅԱՆ

ՀԱՃԱԽԱԿԱՆՈՒԹՅԱՆ ԲԱՌԱՐԱՆԻ ՅՈՒՐԱՀԱՏՈՒԿ ԿՈՂՄԵՐԸ

Լեզվաբանության առջև ծառացած մի շարք պրոբլեմներ ներկայումս ուսումնասիրվում ու լուծվում են ճշգրիտ մեթոդների կիրառմամբ։

Ճշգրիտ մեթոդները՝ դրանք մաթեմատիկական մեթոդներ են, որոնց ստեղծումն ու կիրառումը լեզվաբանության մեջ ցույց է տալիս, որ լեզվաբանական գիտություններ (ինչպես նաև այլ գիտություններ) Մարքսի խոսքերով ասած՝ «հասնում է հասունացման»։

Լեզվաբանության մեջ ճշգրիտ մեթոդների մասին առանձին-առանձին խոսելն այստեղ մեր նպատակից դուրս է։ Մենք կնշենք միայն մի մեթոդի ու նրա կիրառման մասին։ Դա քանակական, իմաստային կամ վիճակագրական վերլուծության մեթոդն է, որի կիրառության առաջին պտուղները հանդիսացան աշխարհի տարբեր լեզուներով կազմված հաճախականության բառարանները։

Այստեղ էլ հենց կանգ կառնենք հաճախականության բառարանների մի քանի յուրահատուկ կողմերի վրա։

Հաճախականության բառարանի համար ելակետային հարց է հանդիսանում տեքստերի ընտրումը։ Լեզուն կամ լեզվի մի առանձին շրջանն ամբողջությամբ վերցրած հնարավոր չէ ենթարկել վիճակագրական վերլուծության։ Արդ, եթե հնարավոր չէ հետազոտել մեզ հետաքրքրող բոլոր տեքստերի ամբողջությունը, ապա մնում է ընտրանքային մեթոդի օգնությամբ հետազոտել ամբողջի մի մասը, որի ուսումնասիրությունը թույլ կտա որոշակի պատկերացում կազմել ամբողջի մասին։ Ավելին. մասի և ամբողջի դիալեկտիկական փոխակապակցության ճիշտ հաշվառումը կարևորագույն նշանակություն ունի ճանաչողության պրոցեսում։ Երբ մասի օգնությամբ քննում ենք ամբողջը, ուրեմն անհրաժեշտ է ի հայտ բերել, թե մասին վերաբերվող տվյալներն իրականում ինչ չափով են համապատասխանում ամբողջի մեջ տեղ գտած փաստերին։ Պարզ է, որ բացարձակ համապատասխանություն լինել չի կարող։ Համապատասխանության աստիճանը կարող է տատանվել, և ամբողջի մասին մեր դատողությունները, որ ստացել ենք մասի քննարկումից, կարող են լինել առավել հավաստի կամ նվազ հավաստի։

Մաթեմատիկական վիճակագրությունն ուսուցանում է, եթե անմիջական քննման համար մատչելի է միայն ամբողջի մի մասը, այսինքն՝ ընտրանքը, ապա մենք հնարավորությունից չենք զրկվում՝ ընտրանքի միջոցով ստանալ ճիշտ պատկերացում ամբողջի մասին։

Վիճակագրությունն ունի ստացված արդյունքների հավաստիությունը գնահատելու այնպիսի մեթոդներ, որոնց օգնությամբ կարողանում է որոշել, թե ինչպիսի ծավալ պետք է ունենա ընտրանքը, որպեսզի կարողանա ապահովել որոշակի աստիճանի հավաստիություն։

Որոշ հանգամանքներում ընտրանքը կարող է իր մեջ ընդգրկել ամբողջը՝ լրիվ վերցրած։ Այսպես. Պուշկինի լեզվի բառարանը կազմելիս, որտեղ յուրաքանչյուր բառի դիմաց նշանակվել է, թե այն ամբողջ ստեղծագործության մեջ քանի անգամ է գործածվել, ընտրանքի մեջ ընդգրկվել են հեղինակի բոլոր ստեղծագործությունները։ Դրությունն այլ է, երբ գործ ենք ունենում ոչ թե մի առանձին հեղինակի ստեղծագործությունների կամ մի որևէ առանձին աշխատության հետ, այլ ցանկանում ենք կազմել մի լեզվի որևէ մասի հաճախականության բառարան։ Այդ դեպքում խոսք լինել չի կարող այն մասին, թե ընտրանքն իր մեջ կընդգրկի ամբողջությունը՝ առանց մնացորդի։

Այդպես էլ՝ հայոց լեզվի (ինչպես նաև այլ լեզուների) հաճախականության բառարան կազմելիս չենք կարող լեզվի մի ամբողջ բնագավառ ենթարկել վիճակագրական վերլուծության, քանի որ դա կպահանջի մի անհամաստ ու, թերևս, անհաղթահարելի աշխատանք։ Մինչդեռ ընտրանքային մեթոդի կիրառմամբ լեզվի տվյալ բնագավառը կարող ենք ենթարկել վիճակագրական հետազոտության։ Այն կդիտենք որպես մի ամբողջություն։ Այդ ամբողջությունից կարող ենք ընտրել առանձին տեքստեր։ Հենց այդ տեքստերն էլ որպես ընտրանք, կենթարկենք վերլուծության։ Ստացված արդյունքները մեզ այս կամ այն չափով հավաստի տեղեկություններ կտան՝ ճիշտ դատողություններ անելու ամբողջի մասին։ Այլ խոսքով՝ լեզվից ջոկված տեքստերի վիճակագրական ուսումնասիրությունից պարզ կդառնա, թե տվյալ տեքստերում հանդես եկած բառերն ու քերականական ձևերն ինչպիսի գործածություն ունեն։ Ստացված տվյալների ճշտությունը առաջին հերթին կախված է ընտրանքի ծավալից։ Որքան մեծ լինի ընտրանքը, այսինքն՝ որքան շատ տեքստեր ուսումնասիրվեն, ստացված արդյունքներն էլ ավելի ճշգրտորեն կարտացոլեն լեզվի իրավիճակը։

Պետք է նկատի ունենալ, որ գիտության տարբեր բնագավառներից ընտրված մասնագիտական գրականության վերլուծության արդյունքները ճշգրիտ տեղեկություններ չեն կարող տալ գեղարվեստական գրականության մեջ գործածվող բառերի մասին։ Օրինակ՝ գիտատեխնիկական գրականության մեջ բարձր հաճախականություն ունեցող բառերը գեղարվեստական գրականության մեջ, բանավոր խոսքում հանդես կգան շատ սակավ։ Ուստի հաճախականության բառարան կազմելիս պետք է ելնել նրա նպատակից և ըստ այդմ էլ ընտրել համապատասխան բնագավառի գրականություն, որտեղ ավելի ճշգրտորեն կարտացոլվի տվյալ բնագավառի հաճախ ու հազվադեպ գործածվող բառերի ռեալ փոխհարաբերությունը։

. Ո՞ր բառերն են, որ հաճախականության բառարանում գրավում են առաջին տեղերը, այլ կերպ ասած՝ ո՞ր բառերն են, որ տվյալ լեզվում ավելի շատ են գործածվում։ Դրանք, առաջին հերթին, օժանդակ բառերն են՝ շաղկապները, կապերը, այնուհետև՝ դերանունները և այլն։

Վիճակագրական ուսումնասիրությունները ցույց են տվել, որ հայերեն «և» շաղկապը միայն «Աբու-լալա-Մահարի» պոեմում (ըստ Լ. Մոտալովայի «Ավ. Իսահակյանի «Աբու-լալա-Մահարի» պոեմի հաճախականության բառարանի»)[1] գործածվել է $\frac{161}{3328}$ անգամ, «ու»-ն՝ $\frac{73}{3328}$ անգամ, «դու» դերանունը՝

1 «Աբու-լալա-Մահարի» պոեմի ծավալը կազմում է 3328 բառ, տարբեր բառերը՝ 1242։

$\frac{57}{3328}$ անգամ և այլն։ Հասկանալի է, որ այդպիսի բարձր հաճախականությամբ հանդես եկող բառերը տվյալ բառարանում կգրավեն առաջին տեղերը։ Մինչդեռ պոեմում կան այնպիսի բառեր, ինչպիսիք են՝ ոսկեհոս, հրեշտակարև, հավերժափայլ, հոլանի, խոլական, խորամանկ, ժանտ և այլն, որոնց թիվը պոեմում կազմում է 769, օգտագործվել են միայն մեկ անգամ։ Իսկ 1242 տարբեր բառեր ունեցող նշված պոեմի հաճախականության բառարանը եղրափակվում է մեկ անգամ գործածվող բառերով։

Նույն օրինաչափությունը նկատվում է նաև այլ լեզուներում։ Օրինակ ռուսերեն «и» շաղկապը Պուշկինի բոլոր տեքստերում միջին հաշվով գործածվել է յուրաքանչյուր 20 բառից հետո ($^1/_{20}$), «я» դերանունը՝ 36 բառից հետո ($^1/_{36}$), իսկ горестно, горемыка, взвешивать և այլ բառեր գործածվել են մեկ անգամ[2]։

Հաճախականության բառարանը, որի կազմումը հետապնդում է մի շարք նպատակներ, սովորաբար ձգտում է այն բանին, որպեսզի իր մեջ ընդգրկված բառերի հաճախականությունների մասին տվյալներն իրենց տարածումն ունենան ոչ միայն վերլուծված տեքստերի վրա, որոնց հիման վրա կազմվել է տվյալ բառարանը, այլև այնպիսի տեքստերի, որոնք ընտրանքից դուրս են գտնվում։ Դրանով իսկ հաստատվում է այն միտքը, որ ընտրանքը ամբողջի օրգանական մասն է կազմում, և սրա արդյունքները կարող են մեզ հավաստի տեղեկություններ տալ այն տվյալների մասին, որոնք գոյություն ունեն ամբողջի մեջ։ Բնական է, որ հաճախականության բառարանը կունենա գործնական մեծ արժեք միայն այն դեպքում, եթե այն կարողանում է տալ բավական ճշգրիտ տեղեկություններ բառերի հաճախականությունների մասին։ Այդ տեսակետը «ղարգացնելու» միտումով չպետք է ենթադրել, թե բառերի հաճախականությունների մասին ընտրանքի տվյալներն ամբողջովին համապատասխանում են քննվող ճյուղի ամբողջության մեջ գտնվող բոլոր բառերի դործածություններին։ Եթե այդ ենթադրությունը համապատասխաներ իրականությանը, ապա ընտրանքը կդադարեր այդպիսին լինելուց։

Ընտրանքի մեջ գտնվող բառերի հաճախականությունը կտարբերվի ամբողջում եղած բառերի հաճախականությունից։ Սակայն այդ տարբերությունը չպետք է գերազանցի նախօրոք նշված հարաբերական սխալի մեծությունը։ Վերջինիս չափը կախված է այն բանից, թե ինչ նպատակի համար է կազմվում հաճախականության բառարանը։ Բայց անկախ բառարանի նպատակից ընտրանքի անհրաժեշտ ծավալի բացակայությունը հանգեցնում է սխալ արդյունքների։

Հայտնի մաթեմատիկոս Ա. Ա. Մարկովը դեռևս 1916 թ. քննադատելով Ն. Ա. Մորոզովի «լեզվաբանական սպեկտրներ» հոդվածը, որտեղ նա ընտրանքի ոչ ճիշտ ծավալի պատճառով ստացել էր այնպիսի տվյալներ, որոնք չէին համապատասխանում լեզվում եղած տվյալների գործածությանը, նշում է.

---

2 Տե՛ս «Словарь языка Пушкина», т. I, М., 1956 г. Պուշկինի բոլոր ստեղծագործությունների ընդհանուր ծավալը կազմում է 544777 բառ, տարբեր բառերի թիվը՝ 21197, մեկ անգամ գործածված բառերի թիվը՝ 6389։

«Միայն հետազոտական դաշտի նշանակալի ընդարձակումը (ոչ թե հինգ հազար, այլ հարյուր հազարավոր բառերի հաշվարկումը) կարող է տալ եզրահանգումներին որոշակի աստիճանի հիմնավորումներ»[3]։

Ուրեմն՝ հաճախականության բառարանի տվյալների ճշգրտությունն անմիջականորեն կապված է անհրաժեշտ ծավալի ընտրանքի հետ։ Վերջինիս օբյեկտիվ լուծումը հանդիսանում է տվյալ բնագավառի արմատական հարցի ճիշտ լմբռնում։

Նկարագրած պրոբլեմի լուծմանը իրավացի մոտեցում է հանդես բերում ՍՍՀՄ ԳԱ լեզվաբանության ինստիտուտի գիտական աշխատակից Ռ. Մ. Ֆրումկինան, որի կողմից մշակված մեթոդը[4] օգնում է ճիշտ կողմնորոշվելու հաճախականության բառարան կազմելու աշխատանքներում։

Միաժամանակ պետք է նշել, որ մինչև այժմ կազմված հաճախականության բառարանների գերակշռող մեծամասնության հեղինակներն իրենց առաջ հատուկ խնդիր չեն դրել սահմանելու ընտրանքի անհրաժեշտ ծավալ։ Նրանք ընտրանքի ծավալը բնորոշելիս առաջնորդվել են իրենց անձնական նախասիրություններով։

Եվ ահա, Ֆրանտիշկա Մալիրժի ղեկավարությամբ չեխ ընթերցողների համար կազմվել է ռուսաց լեզվի հաճախականության բառարան, որի ընտրանքը կատարվել է բացառապես պարբերական մամուլի նյութերի հիման վրա։ Պարզ է, որ այսպիսի միակողմանիության դեպքում արդյունքների հավաստիությունը այլ բնագավառի տեքստերի վրա տարածվել չի կարող։

Իսպաներեն լեզվի հաճախականության բառարանի կազմման համար ընտրանքի ծավալի բնորոշման յուրահատուկ մեթոդիկա է կիրառել Գարսիա Հոսը։ Սա ընտրած տեքստերը բաժանում է չորս խմբի (ըստ տարբեր բնագավառների գրականության), իսկ յուրաքանչյուր խմբի տեքստը բաժանում է այնպիսի մասերի, որոնք հավասարապես պարունակում են 10,000-ական բառ։ Այնուհետև հաշվում է, թե քանի տարբեր բառեր են հանդես եկել առաջին 10000 բառում և քանի տարբեր բառեր՝ երկրորդ 10000-ում, որոնք չեն հանդիպել առաջինում ու՝ այդպես շարունակ։ Հոսը գտավ, որ առաջին 10000-ում հանդես են եկել 1800 տարբեր բառեր (մնացածը կրկնվել է), երկրորդ 10000-ում՝ մոտ 1000 տարբեր բառեր, որոնք չեն հանդիպել առաջինում։ Իսկ երրորդ 10000-ում գործածվել են 500 տարբեր բառեր, որոնցից և ոչ մեկը հանդես չի եկել առաջին և երկրորդ մասերում։ Վերջապես տասնմեկերորդ 10000-ում հանդես է եկել 90 նոր բառ։ Այդ մինիմումով էլ հեղինակը բավարարվել է, որովհետև դրանից հետո նոր բառերը հանդես են եկել շատ դանդաղ։

Գարսիա Հոսը եկավ այն եզրակացության, որ յուրաքանչյուր խմբի հա-

---

3 А. А. Марков, Об одном применении статистического метода, «Известия Императорской Академии наук», серия VI, т. X, № 4, стр. 241.

4 Նշված մեթոդիկայի մասին տե՛ս. Р. М. Фрумкина, Некоторые вопросы методики составления частотных словарей, «Машинный перевод и прикладная лингвистика», М., 1959, № 2 [9]. Р. М. Фрумкина, Статистические методы изучения лексики, М., 1964. Р. М. Фрумкина, Некоторые практические рекомендации по составлению частотных словарей, «Русский язык в национальной школе», 1963, № 5.

մար բավական է քննել 100000-ական բառ (չորս խմբի համար ընդհանուր ծավալը կկազմի 400000 բառ), որպեսզի ստացված արդյունքները հավաստի տեղեկություններ տան նաև այլ տեքստերի մասին։ Նրա մոտեցումը ընտրանքի ծավալի որոշման հարցում կամայական է և ճիշտ չափանիշ լինել չի կարող, քանի որ այն ընդունակ չէ ապահովելու ճշգրիտ տեղեկություններ բառերի հաճախականության մասին։ Նա ծավալի չափանիշը համարում է նոր բառերի հանդես գալու արագությունը, չափանիշ, որը պարզապես չի կարող սահմանել ընտրանքի անհրաժեշտ ծավալ[5]։

Ընտրանքի անհրաժեշտ ծավալն ու դրա հետ կապված այլ հարցերը, ինչպես շոշափեցինք վերևում, որոշակի լուսաբանում են գտել Ռ. Մ. Ֆրումկինայի մոտ։

Բառի նվազ հաճախականության դեպքում Ֆրումկինան պաշտպանում է ընտրանքի ծավալը մեծացնելու տեսակետը, քանի որ փոքր ընտրանքի դեպքում հազվադեպ հանդիպող բառի մասին ստույգ տեղեկություն ստանալ հնարավոր չէ։ Միաժամանակ, նա խորհուրդ չի տալիս ընտրանքի ծավալը գերազանցել որոշակի մեծությունից։ «Ընտրանքը չպետք է գերազանցի որոշակի ծավալին, որը ենթակա է վերլուծության»[6],—գրում է Ֆրումկինան։ Սկզբունքորեն այդ տեսակետի հետ չի կարելի չհամաձայնվել։ Սակայն լեզվի մի քանի ճյուղերի տվյալներն ընդգրկող հաճախականության բառարանի ընտրանքի համար դա խիստ է ասված (անտեսված է ժամանակակից հաշվիչ տեխնիկայի միջամտությունը բառարանի կազմման աշխատանքներին)։

Ընտրանքի ծավալի, բառերի հաճախականության, հարաբերական սխալի ու այլ մեծությունների փոխադարձ կապի ու կախվածության բացահայտումը կարող ենք արտահայտել հետևյալ վիճակագրական բանաձևի միջոցով։

$$\delta = t \sqrt{\frac{q}{np}} \quad *$$

n-ն ընտրանքի ծավալն է,

δ-ն՝ հարաբերական սխալը,

p-ն հաճախականությունն է, որն ըստ տվյալ ընտրանքի պետք է բնորոշվի այնպիսի հնարավոր սխալով, որը չի գերազանցում δ-ին։

t-ն հաստատուն թիվ է,

q-ն p-ի հակառակ մեծությունն է։

Այս բանաձևի միջոցով կարող ենք լուծել մի քանի խնդիրներ, որոնք ծառանում են մեր առջև՝ հաճախականության բառարան կազմելիս։

Նախ որոշում ենք թույլատրելի հարաբերական սխալի (δ) մեծությունը և ապա սահմանում, թե ինչպիսին պետք է լինի ընտրանքը, որպեսզի համապատասխան բառերի հաճախականությունները հաշվի առնվեն այնպիսի սխալով,

---

5 Դարսիա Հոսի մեթոդի մասին մանրամասն տեղեկություններ ստանալու համար տե՛ս Р. М. Фрумкина, Статистические методы изучения лексики, М., 1964.

6 Р. М. Фрумкина, Статистические методы изучения лексики, М., 1964, стр. 16.

* Նշված մեծությունների փոխադարձ կապը բացահայտելու համար Ֆրումկինան օգտագործում է հետևյալ վիճակագրական բանաձևը (պարզեցված ձևով)՝

$$\delta = \frac{zp}{\sqrt{Np}}$$

Այս բանաձևում zp հաստատուն թիվ է, որը վերցվում է հատուկ տիպի աղյուսակներից։

որը չի գերազանցում հարաբերական սխալին։ Այսինքն, սահմանում ենք ընտրանքի այնպիսի ծավալ, որը կարողանա ապահովել այս կամ այն բառի համապատասխան հաճախականություններ՝ դուրս չգալով որոշված հարաբերական սխալի սահմաններից։

Մաթեմատիկական վիճակագրության դրույթների համաձայն հնարավոր են նաև այնպիսի դեպքեր, երբ սահմանված ծավալի ընտրանքի մեջ շեղումներն ավելի մեծ են լինում, քան ենթադրում ենք։ Իհարկե, այդպիսի շեղումներ հազվադեպ են պատահում։

Մեր կիրառած բանաձևում տեղ գտած հաստատունն (t) էլ բնորոշում է այդպիսի մեծ շեղման դեպքերի թիվը։ Եթե ընդունենք, որ t=2 (այսինքն, F(t)=0,95), ապա այդ նշանակում է, որ ընտրանքում հաճախականության 100 դեպքից 95-ը դուրս չի գա մեր նախանշած սահմաններից, իսկ մնացած 5-ը կարող է չհամապատասխանել պայմանին։

Հարաբերական սխալի մեծությունը սահմանելիս ելնում ենք գործնական ու տեսական պահանջներից։

Եթե հաճախականության բառարանը կազմված է ուսումնական նպատակների համար, ապա հարաբերական սխալին կարող ենք տալ այնպիսի մեծություն, որի համար շատ բարձր ճշտությունն այնքան էլ անհրաժեշտ չէ։ Իսկ եթե կազմում ենք մի այլ տիպի բառարան, որի նպատակն է՝ ծառայել հեռագրային հաղորդումների լեքսիկական կոդավորման սիստեմին, ապա այստեղ պահանջվում է հարաբերական սխալին տալ այնպիսի արժեք, որը բարձր ճշտությամբ արտացոլի ամբողջի տարրերի ու ընտրանքի միջոցով ստացված արդյունքների իրական փոխհարաբերությունը։ Սա նշանակում է, որ ընտրանքի միջոցով ստացված բառերի հաճախականությունը պետք է տարբերվի ամբողջի (լեզվի) մեջ գոյություն ունեցող բառերի հաճախականությունից՝ նվազագույն չափով՝ ինչպես ցույց է տալիս հարաբերական սխալի մեծությունը։

Եթե հարաբերական սխալին տալիս ենք այսպիսի մեծություն՝ $\sigma = 0{,}3$, ապա ընտրանքի տարրերի հաճախականությունը ամբողջի տվյալների հաճախականությունից կշեղվի մոտավորապես մեկ երրորդով $\left(\frac{1}{3}\right)$։ Եթե $\sigma = 0{,}1$, ապա ընտրանքի արդյունքը ամբողջի մեջ գոյություն ունեցող տարրերի հաճախականությունից կշեղվի մեկ տասներորդով $\left(\frac{1}{10}\right)$ և այլն։

Եթե կազմում ենք ուսումնական ու այլ նպատակների համար որևէ լեզվի հաճախականության բառարան, որը (հարաբերական սխալի $\sigma = 0{,}3$ դեպքում) կարողանա արտահայտել տեքստի ընդհանուր ծավալի մոտավորապես 80%, ապա անհրաժեշտ է առաջին հերթին սահմանել համապատասխան ծավալի ընտրանք։ Այստեղ կարևորն այն է, որ հմտությամբ բնորոշվի ամենահաճախ հանդես եկող բառերի առաջին խմբի վերջին բառի հաճախականությունը (p)։ Այդ բառի հաճախականությունը կարող ենք որոշել մեր վերոհիշյալ բանաձևի օգնությամբ։

Այդ բանաձևը կարելի է ենթարկել այսպիսի ձևափոխության՝

$$\delta = t\sqrt{\frac{q}{np}}\ ;\quad \sqrt{\frac{q}{np}} = \frac{\delta}{t}\ ;\quad \frac{q}{np} = \frac{\delta^2}{t^2}\ ;$$

$$\frac{1}{n} = \frac{\delta^2 p}{t^2 q}; \quad n = \frac{t^2 q}{\delta^2 p}$$

և վերջապես՝ $p = \frac{t^2 q}{\delta^2 n}$

ընդունենք՝ $n = 700000$*

$p = ?$

$q = 1 - p$

$\delta = 0{,}3$

$t = 2$

p-ն գտնելու համար տեղադրենք արժեքները.

$$p = \frac{2^2 (1 - p)}{0{,}3^2 \cdot 700000} = \frac{4 - 4p}{63000};$$

$p = \frac{4 - 4p}{63000}$, Այս հավասարության երկու կողմերն էլ բազմապատկում ենք միևնույն թվով.

$$63000\ p = \left(\frac{4 - 4p}{63000}\right) \cdot 63000$$

$$63000\ p = 4 - 4p$$

$$63000\ p + 4p = 4$$

$$63004p = 4$$

$$p = \frac{4}{63004} = 0{,}00007$$

Ստացանք, որ բարձր հաճախականությամբ հանդես եկած բառերի առաջին խմբի ամենավերջին բառը (ըստ նվազող հաճախականության) ուսումնասիրված տեքստերում օգտագործվել է 0,00007 հաճախականությամբ: Այսինքն, ուսումնասիրված տեքստերի յուրաքանչյուր հարյուր հազար բառում տվյալ բառը հանդիպել է յոթ անգամ: Քանի որ մենք հետազոտության ենք ենթարկել 700000 բառ, ուստի այդ ամբողջ ընտրանքում մեզ հետաքրքրող բառը հանդես կգա յոթ անգամ ավելի:

Ուրեմն՝ ամենահաճախ գործածվող բառերի առաջին խմբի վերջին բառը հանդես է եկել 700000 բառում 49 անգամ: Եթե կարողացել ենք հարաբերական սխալի հետ մեկտեղ ճիշտ որոշել այդ վերջին բառի հաճախականությունը, ապա նրանից առաջ դասավորված մյուս բառերը, որոնք ունեն ավելի բարձր հաճախականություն, ինքստինքյան կբնորոշվեն ավելի մեծ ճշգրտությամբ:

Ընտրանքի ծավալը սահմանելիս պետք է հիմք ընդունել առավել հաճախ գործածվող բառերի առաջին խմբի ամենավերջին բառի հաճախականությունը. որն իր նախորդների համեմատ ավելի նվազ գործածություն ունի: Նվազ ասելով, չպետք է հասկանալ, որ այդ բառը ընդհանրապես հազվադեպ է գործածվում, այլ պարզ է, որ այդ խմբի վերջին բառի հաճախականությունը պետք է ուննենա անհրաժեշտ չափի մեծություն, որպեսզի նրա բավարար աս-

* Այդ թիվը ներկա դեպքում կամայական է և ծառայում է լոկ որպես օրինակ:

տիճանի գործածության բնորոշումը հարաբերական սխալի հետ մեկտեղ չպահանջի ավելի մեծ ընտրանք, քան սահմանված անհրաժեշտ ծավալն է։

Իսկ ինչպե՞ս գտնել այդ վերջին բառի համարը։ Այլ կերպ ասած, մեր առանձնացրած առավել գործածված բառերի առաջին խումբը որքան բառ է պարունակում իր մեջ։ Այս հարցը պարզելու համար գիմննք այսպես կոչված Ցիպֆի օրենքին[7], որի մասին կխոսենք քիչ հետո։

Ըստ այդ օրենքի՝

$$p_r = kr^{-\gamma}, \quad \text{որտեղ}$$

p_r-ը բառի հաճախականությունն է,

r-ը բառի համարն է՝ ըստ նվազող հաճախականության,

k-ն և γ-ն որոշ սահմաններում հաստատուն թվեր են։

Այս բանաձևը կարող ենք ներկայել այսպիսի ձևափոխության.

$$p_r = k \cdot r^{-\gamma};$$

$$p_r = \frac{k}{r^{\gamma}}, \quad \text{որտեղից՝} \quad r^{\gamma} = \frac{k}{p_r}.$$

Ունենք

$$k = 0{,}1$$
$$\gamma = 1{,}01$$
$$p = 0{,}00007$$
$$r^{\gamma} = ?$$

Այժմ գտնենք, թե ինչի՞ է հավասար r^{γ}։

$$r^{\gamma} = \frac{0{,}1}{0{,}00007} = \frac{10000}{7} = 1428$$

$$r^{\gamma} = 1428$$

$$r^{1{,}01} = 1428^{1{,}01} = 1428$$

Պարզվեց, որ ընտրանքի վերլուծման հետևանքով ստացված բառերի առաջին խումբը, որն առանձնացվել էր մեր կողմից (ըստ նվազող հաճախականության), իր մեջ ընդգրկում է 1428 բառ։ Պարզվեց նաև, որ նշված խմբի վերջին բառն ունի 1428-րդ համարը։ Բայց դեռևս չլուծված է մնում այն հարցը, թե 1428 բառը որքանո՞վ կարող է արտահայտել որոշակի երկարության տեքստի ծավալը։

Այդ հարցը կարող ենք նախօրոք լուծել տեսականորեն, իսկ այնուհետև ստուգել նաև փորձնական ուղիով։ Սրա տեսական լուծման համար դարձյալ վկայակոչում ենք Ցիպֆի օրենքը։

Մինչ դրան անցնելը, հիշենք, որ հաճախականության բառարաններում բառերը դասավորված են ըստ նվազող հաճախականության։ Հարց է առաջանում՝ արդյոք բառերի այդպիսի նվազող հաճախականությունը ներկայո՞ւմ է որևէ օրինաչափության, թե ուղղակի տարերայնության հետևանք է։ Կամ՝ կարո՞ղ ենք հաճախականության բառարանի որևէ բառի համարի միջոցով գտնել, թե այդ բառը տեքստերում մոտավորապես ինչպիսի հաճախականությամբ է հանդես գալիս։ Այդ հարցերին է, որ Ցիպֆի օրենքը դրական պա-

7 Ցիպֆի օրենքի մասին մանրամասն տեղեկություններ ստանալու համար տե՛ս Р. М. Фрумкина, Статистические методы изучения лексики, М., 1964.

տասխան է տալիս (այդ բանում մենք որոշ չափով համոզվեցինք նաև վերևի օրինակում)։

Հետամուտ լինելով նշված օրինաչափությանը, տեսնենք, թե մեր առանձնացրած առավել հաճախ գործածվող բառերի առաջին խումբը (մոտ 1430 բառ) տեքստի ծավալի ո՞ր մասը կկազմի։

Եթե հաճախականության բառարանում բառի համարը նշանակենք r, իսկ տեքստերում հանդես եկող այդ նույն բառի հարաբերական հաճախականությանը՝ Pr, ապա բառի համարի ու նրա հարաբերական հաճախականության միջև նկատված օրինաչափությունը, որը ստուգվել է տարբեր լեզուների նյութերի հիման վրա, կարտահայտվի հետևյալ բանաձևով.

$$P_r = \frac{k}{r^{\gamma}};$$ (Այս բանաձևն արտահայտում է Ցիպֆի օրենքը)։

k և γ այս օրենքում հաստատուն մեծություններ են[8]։ Ըստ այդ օրենքի՝ մեր նշած 1430 ամենահաճախ գործածվող բառերը կկազմեն տեքստերի ծավալների մոտավորապես 80%, քանի որ՝

$$P_r = k \cdot r^{-\gamma}$$

Տեղադրելով արժեքները՝ կունենանք.

$$P_1 = 0{,}1 \cdot 1^{-1{,}01}$$
$$P_2 = 0{,}1 \cdot 2^{-1{,}01}$$
$$P_3 = 0{,}1 \cdot 3^{-1{,}01}$$

և վերջապես՝

$$P_r = 0{,}1 \cdot 1430^{-1{,}01}$$

Այստեղ հաշվի առնելով Ցիպֆի օրենքի սահմանափակությունը (ինչպես կտեսնենք ստորև), 1—50 հաշվում ենք առանձին, իսկ 51—1430՝ առանձին։

$$S_1 = \sum_{r=1}^{50} p_r = p_1 + p_2 + p_3 + \cdots + p_{50} = 0{,}1\,(1^{-1{,}01} + 2^{-1{,}01} + 3^{-1{,}01} + \cdots$$
$$\cdots + 50^{-1{,}01}) \approx 0{,}49$$

$$S_2 = \sum_{r=51}^{1430} p_r = p_{51} + p_{52} + p_{53} + \cdots + p_{1430} = 0{,}1\,(51^{-1{,}01} + 52^{-1{,}01} +$$
$$+ 53^{-1{,}01} + \cdots + 1430^{-1{,}01}) \approx 0{,}31$$

$$S = S_1 + S_2 = \sum_{r=1}^{50} p_r + \sum_{r=51}^{1430} p_r \approx 0{,}8 = 80\%$$

$$S \approx 80\%$$

Այսպիսով, տեսանք, որ մեզ հետաքրքրող 1430 բառը կարող է արտահայտել համապատասխան տեքստի ծավալի մոտավորապես 80%։ Միաժամանակ, չպետք է եզրակացնել, թե Ցիպֆի օրենքը կարող է տարածվել ամեն

---

8 Սրանց արժեքները (որոնք նշված են վերևում) տարբեր լեզուներում կրում են որոշ փոփոխություններ։ Այդ փոփոխությունները գոյություն ունեն նույնիսկ միևնույն լեզվի տարբեր հեղինակների մոտ։

մի երկարություն ունեցող տեքստի վրա։ Նրա կիրառումը սահմանափակվում է $50<r<1500$ շրջանակով։ Եթե r գերազանցում է 1500, այսինքն, եթե ընտրանքի միջոցով ստացված ամենահաճախ գործածվող բառերի առաջին խումբն իր մեջ պարունակում է 1500-ից ավելի բառ (ասենք՝ մոտ 2000 և այլն), ապա Ցիպֆի բանաձևում պարամետրը (γ) գագարում է հաստատուն թիվ լինելուց. այն փոփոխվում է r-ի մեծացմանը զուգընթաց։ γ-ի ոչ հաստատուն լինելը հատուկ է ոչ միայն տարբեր լեզուներին, այլև միևնույն լեզվի տարբեր տեքստերին։

Առանձին հետաքրքրություն կարող է առաջացնել պարամետրի (γ) ոչոշումը հայերենում, որը կապված է հաճախականության բառարանը կազմելու և վերջինիս արդյունավետության ստուգման հետ։

Ինչպես արդեն նկատել ենք՝ հաճախականության բառարաններ կազմելիս հարաբերական սխալին հատկացնում ենք տարբեր մեծություններ։ Իրոք, եթե մեզ չի բավարարում 0,3 մեծությունը, որը տալիս է համեմատաբար ավելի մեծ շեղում, ապա հարաբերական սխալին կարող ենք տալ այլ մեծություններ (օրինակ, 0,2, 0,1 և այլն), որոնց դեպքում շեղումը համապատասխանաբար ավելի կփոքրանա։ Այդ դեպքում էլ ընտրանքի ծավալը կհասնի այնպիսի չափերի, որի քննումը կպահանջի դժվարահաղթ աշխատանք։

Ընդունենք՝ $\varepsilon=0,3$, իսկ ընտրանքի ընդհանուր ծավալը՝ 750000 բառ, ապա ամենահաճախ գործածվող բառերի առաջին խմբի ամենանվազ գործածվող բառի հարաբերական հաճախականությունը (p) կկազմի 0,00006, այսինքն, ուսումնասիրված բոլոր տեքստերում այն կհանդիպի մոտ 45 անգամ ու հաճախականության բառարանում կգրավի 1666-րդ տեղը (կամ՝ այդ խումբն իր մեջ կպարունակի 1666 բառ)։ 1666 բառը կարող է արտահայտել ոչ միայն ուսումնասիրված, այլև ուրիշ տեքստերի ծավալների մոտ 84 տոկոսը։

Պատկերը փոխվում է երբ հարաբերական սխալին տալիս ենք այլ արժեք։ Եթե $\varepsilon=0,2$, ապա միևնույն ընտրանքի դեպքում (750000) ամենաբարձր հաճախականությունն ունեցող բառերի առաջին խմբի ամենավերջին բառի հարաբերական հաճախականությունը կկազմի 0,0001։ Ուսումնասիրված տեքստերում այս բառն արդեն հանդես կգա 75 անգամ։ Միաժամանակ, բառերի այդ խումբը կեղրափակվի 1000-րդ բառով։ Առանձնացված այդ 1000 բառը կարող է արտացոլել համապատասխան տեքստի ծավալի 70%։

Ինչպես տեսնում ենք՝ տարբերությունն ակնհայտ է։ Վերջինիս դեպքում վերլուծության հետևանքով բառերի հաճախականությունների մասին ստացված տվյալներն ավելի ստույգ են (հարաբերական սխալն այստեղ կազմում է երկու տասներորդ մասը, իսկ մյուսում՝ երեք տասներորդը)։

Շարադրանքի ամբողջ ընթացքում իրեն զգացնել է տալիս այն համոզմունքը, որ ինչպես բոլոր լեզուներում, այնպես էլ հայերենում ոչ բոլոր բառերն են, որոնք օժտված են մարդկային հասարակությանը միատեսակ ծառայելու ունակությամբ։ Նրանց մի մասը ցուցաբերում է մեծ ակտիվություն ու մարդկային հաղորդակցման բազմազան պրոցեսներում օգտագործվում է բարձր հաճախականությամբ։ Մյուս մասը զուրկ է այդպիսի ակտիվությունից[9]։

---

9 Առիթից օգտվելով, շեշտենք, որ նվազ գործածություն ունեցող բառերն իրենց մեջ պարունակում են ինֆորմացիայի ավելի մեծ բանակ, քան բարձր հաճախականություն ունեցող բառերը։

Օթե հարկ է լինում հազվադեպ գործածվող բառերն ընդգրկել հաճախականության բառարանի մեջ, ապա անհրաժեշտորեն պահանջվում է՝ հետազոտել անհամեմատ ավելի մեծ ծավալի ընտրանք, որպեսզի կարողանանք համոզված կերպով պնդել, թե հազվադեպ հանդես եկող բառերի մասին մեր ստացած տեղեկությունները հավաստի են նաև այլ տեքստերի համար, որոնք չեն ներգրավվել հետազոտված ընտրանքի մեջ։

Ուրեմն, բառը որքան հազվադեպ հանդես գա, ընտրանքն այնքան մեծ պետք է լինի և ընդհակառակը։ Օրինակ. վերցնենք հայերեն «գնալ» և «նշավակել» բառերը։ Զորյանի «Պապ թագավորը» վեպի առաջին գլխի 10000 բառում «գնալ» բառն օգտագործվել է 42 անգամ $\left(\frac{42}{10000}\right)$։ Այս բառի համար, հետևապես, հարկադրված չենք լինի ուսումնասիրել համեմատաբար մեծ ծավալի ընտրանք։ Մինչդեռ «նշավակել» բառն այդ նույն 10000-ում չի գործածվել և ոչ մի անգամ։ Ուրեմն, այդ բառի վերաբերյալ ստույգ տեղեկություն ստանալու համար, որը տարածվի նաև այլ տեքստերի վրա, ստիպված կլինենք հետազոտել այնպիսի ընտրանք, որի ծավալը կհասնի նույնիսկ ավելի քան տասը միլիոն բառի։ Պարզ է, որ այդպիսի ընտրանքը կպահանջի հսկայական աշխատանք ու ժամանակ և առանձին հետաքրքրություն էլ չի ներկայացնի։

Դրա համար էլ հաճախականության բառարանը կոչված չէ տեքստերի բոլոր բառերի մասին տեղեկություններ տալու։ Այն սահմանափակվում է միայն առավել հաճախ գործածվող բառերի բնութագրմամբ։ Դրանով իսկ նա ծառայում է այն նպատակին, ինչի համար կյանքի է կոչվել։

Նախքան ճյուղային բառարանների կազմումը, հայոց լեզվի համար այժմ նպատակահարմար է կազմել այնպիսի հաճախականության բառարան, որի ընտրանքի անհրաժեշտ ծավալը կազմի 1000000 բառ։ Մեր հանրապետությունում այդպիսի բառարանի բացակայությունն իր բացասական ազդեցությունն է թողնում տարբեր բնագավառների աշխատանքների վրա, ինչպես, օրինակ՝ մեքենայական թարգմանության, լեքսիկական կոդավորման սիստեմի (ինֆորմացիայի տեսության մեջ), լեզվի տեսության, լեզվի ուսուցման և այլ ճյուղերի հարցերում։ Ընտրանքի նշված ծավալի հետ գործ ունենալը կապված է մեծ դժվարությունների հետ։ Սակայն այդ դժվարությունների հաղթահարման աշխատանքում իր մասնակցությունը պետք է ցուցաբերի մեր հանրապետության գիտությունների ակադեմիայի և պետական համալսարանի հաշվողական կենտրոնը[10]։

Հաճախականության բառարանի համար, ինչպես տեսանք վերևում, իրոք հիմնական հարց է հանդիսանում անհրաժեշտ ծավալի տեքստերի ընտրումը։ Այո, հիմնական, բայց ոչ միակը։ Այդ բառարանի համար մեծ կարևորություն է ներկայացնում նաև այն հանգամանքը, թե ինչպիսի տեքստեր պետք է ընտ-

---

10 Հայաստանի գիտությունների ակադեմիայի հաշվողական կենտրոնի աշխատակիցներն իրենց ձեռքի տակ չունենալով համապատասխան բնագավառի հաճախականության բառարան, հետազոտել են ռուսերեն լեզվով որոշակի երկայնության մաթեմատիկական տեքստ ու առանձնացրել 5.000 առավել հաճախ գործածվող բառեր։ Այդ բառացանկը թարգմանել են հայերեն ու գործածել մեքենայական թարգմանության համար։ Հաճախականության բառարանի այս տեսակը կարելի է կոչել ճյուղային, ճյուղային բառարանների հիմնավորումն ու հետանկարայնությունը կտանք այլ առիթով։

րել: Ջպետք է կարծել, թե տեքստի ընտրումը հարցի ֆիզիկական կողմն է միայն: Բառարանի համար պետք է ընտրել այնպիսի տեքստեր, որոնց տվյալներն ավելի ճշգրտորեն ներկայացնեն տվյալ լեզվի համապատասխան բնագավառը: Միայն այդ գեպքում ընտրանքի մեջ ընդգրկված միավորը (տեքստը) կարող է ճիշտ արտացոլել ամբողջի համապատասխան հատկությունները:

Այդ տեսակետից խոցելի է Ջոսելսոնի ռուսաց լեզվի հաճախականության բառարանը[11]: Սրա ընտրանքի 50% կազմում է մինչհեղափոխական ռուսական գեղարվեստական գրականությունը: Այնպես է ստացվել, որ առավել հաճախ գործածվող բառերի շարքն են դասվել այնպիսի բառեր (барин, чиновник և այլն), որոնք ժամանակակից ռուս գրական լեզվում շատ հազվադեպ են գործածվում: Դա ցույց է տալիս, որ Ջոսելսոնը ընտրանքի հարցում ցուցաբերել է ոչ ճիշտ մոտեցում, և ռուս ժամանակակից գրական լեզուն, որը հեղափոխությունից հետո բավական փոփոխությունների է ենթարկվել, ըստ արժանվույն չի ներկայացվել ընտրանքի մեջ: Բացակայում են բանաստեղծական տեքստերը, շատ աղքատիկ է ներկայացված սովետական դրամատուրգիան: Վերջինս, առանձնապես, լավ տեղեկություններ կարող էր տալ ժամանակակից ռուս գրական-խոսակցական լեզվի տվյալների վերաբերյալ: Նշված թերությունների պատճառով էլ հեղինակին չի հաջողվել բառարանում ճիշտ արտացոլել ժամանակակից ռուս դրական լեզվի իսկական վիճակը:

Հաճախականության բառարան լեզվի գոյավիճակն ավելի լավ կներկայացնի միայն այն դեպքում, եթե հաշվի առնվի բառերի հաճախականությունը ոչ միայն գրավոր, այլև բանավոր խոսքում: Այդ երկուսի արդյունքների համեմատությունը մեծ հետաքրքրություն է ներկայացնում:

Մինչև այժմ կազմված հաճախականության բառարանները (բացառությամբ Գոգենհայմի բառարանի[12] հենվել են գրավոր խոսքի տվյալների վրա: Սրանց հեղինակներն անուշադրության են մատնել բանավոր խոսքի ուսումնասիրությունը:

Ճիշտ է, այդ նպատակով բանավոր խոսքի հետազոտությունը կապված է լուրջ դժվարությունների հետ, բայց դա չի նշանակում, որ պետք է անուշադրության մատնվի այդպիսի խոշոր կարևորություն ունեցող մի բնագավառ, ինչպիսին բանավոր խոսքն է:

Բանավոր խոսքի հետազոտության անհրաժեշտությունն ավելի ակնառու է դառնում նաև այն իմաստով, որ բանավոր ու գրավոր խոսքի ուսումնասիրության միասնական արդյունքները կարող են լիարժեք տեսք տալ հաճախականության բառարանին: Վերջինս, միաժամանակ, հանդիսանում է այն հուսալի հիմքը, որի վրա ուսումնական նպատակներով կազմվում է մինիմում-բառարանը:

Հաճախականության բառարանների վերջնական գնահատականը կարելի է տալ միայն այն ժամանակ, երբ նրանք համապատասխան բնագավառներում գործածության մեջ են դրվում՝ արդյունավետությունն ստուգելու համար:

---

[11] Josselson H., The Russian word count and frequency analysis of grammatical categories of standart literary Russian, Detroit, 1953.

[12] Gougenheim, L' élaboration du français elementaire, Paris, 1956.

Б. К. КАЗАРЯН

ХАРАКТЕРНЫЕ ОСОБЕННОСТИ ЧАСТОТНОГО СЛОВАРЯ

Р е з ю м е

Данная работа предетавляет собой первое на армянском языке изложение вопросов и проблем составления частотного словаря. Освещая эти вопросы, автор связывает их с задачами составления частотного словаря армянского языка.

Для составления частотного словаря мы употребляем следующую статистическую формулу — $\delta = t\sqrt{\frac{q}{np}}$, где отображены взаимосвязь зависимость между объемом выборки (n), частотностью слов (p), относительной ошибкой (δ) и другими величинами.

Исходя из требований науки и общественных отношений, осознаем необходимость составления нового типа частотных словарей. Их мы называем отраслевыми — профессиональными частотными словарями.