Ռ. Լ. ՈՒՌՈՒՏՅԱՆ, Պ. Կ. ՂԱԶԱՆՉՅԱՆ, Ե. Ս. ՍՏԵՓԱՆՅԱՆ, Գ. Օ. ՍՈՒՔԻԱՍՅԱՆ

ՀԱՇՎՈՂԱԿԱՆ ՏԵԽՆԻԿԱՅԻ ԿԻՐԱՌՈՒՄԸ ՏԵՔՍՏԻ ՈՐՈՇԱԿԻԱՑՄԱՆ ՀԱՐՑՈՒՄ

Ինֆորմացիայի տեսության մեջ կապակցված տեքստը դիտվում է ինֆորմացիայի էրգոդիկ աղբյուր։ Իսկ վերջինիս հատուկ են վիճակագրական որոշակի օրինաչափություններ, որոնք ենթարկվում են բաշխվածության օրենքի և հաղորդագրության երկարության ավելացման միջոցով կարելի է հասնել կայուն ու ճշգրիտ գնահատականների։ Այդ նշանակում է, որ որևէ լեզվով գրված ամեն մի տեքստի համար կարող ենք ստանալ բնութագրիչների համախումբ, որի օգնությամբ մի տեքստը տարբերվում է մեկ այլ տեքստից։

Դժվար է պնդել, թե ամեն մի տեքստ պատկանում է ինֆորմացիայի իդեալական էրգոդիկ աղբյուրների թվին, մանավանդ, երբ խոսքը վերաբերում է գրական հուշարձաններին։ Ցանկացած տեքստում առկա են լեզվական ընդհանուր ու հեղինակային անհատականությունը, դարաշրջանային համաժամանակությունն ու տարժամանակությունը, հնարանությունն ու նորարանությունը, և եթե այս ամենին ավելացնենք այն, որ գրական հուշարձանները մեզ հասել են ոչ թե բնագրային տեսքով, այլ տարբեր ժամանակներում և տարբեր կրթօջախներում արված կրկնօրինակների, այսպես կոչված ցուցակների (կամ ընդօրինակությունների) տեսքով, որոնք իրենց վրա կրելով դարաշրջանային, տեղային աղդեցությունները, պարունակում են օրինաչափ ու պատահական, սուբյեկտիվ ու օբյեկտիվ բնույթի աղավաղումներ։ Վերջիններս կարող են հասնել այն աստիճանի, որ կասկածի տակ առնվի տվյալ ընդօրինակությունը տվյալ հեղինակին պատկանելու հարցը։ Այդ աղավաղումների պատճառով առանձնապես դժվարանում է տեքստային բնութագրիչների որոշումը և դրանց գնահատումը։

Ընդհանրապես, տեքստաչափական աշխատանքներին հատուկ են տեքստային լայն զանգվածների ներառմամբ գործողությունների միօրինակություն, երբ անհրաժեշտ է հաշվի առնել հարյուրավոր փաստեր, որը հաճախ գերազանցում է մարդու հնարավորությունները։ Մյուս կողմից, ժամանակակից հաշվողական տեխնիկան թույլ է տալիս ընդգրկել ավելի

լայն տեքստային զանգվածներ, քննության առնել ամեն մի փաստ, առաջ քաշել խնդիրներ, որոց վճռումը ավանդական եղանակով գործնականում հնարավոր չէ։

Այս առումով ներկա հոդվածում դրվում է երկու խնդիր կապված Մ. Խորենացու «Հայոց պատմության» ստեղծագործության քննության հետ և նշվում այդ խնդիրների լուծման եղանակները ժամանակակից էլեկտրոնային հաշվողական մեքենաների օգնությամբ։

Ժամանակակից բանասիրության մեջ գրական ստեղծագործությունների որոշակիացման խնդիրը քննվում է երեք ուղղություններով՝ փաստաբանավերագրական գաղափարաթեմատիկ և լեզվաոճական։ Այս կապակցությամբ Վ. Վ. Վինոգրադովը գրում է. «Ոճական վերլուծության մարքսիստական մեթոդաբանությունը, մեր կարծիքով, պետք է բխի գրական ստեղծագործության իմացությունից՝ ինչպես գաղափարական, թեմատիկ, կոմպոզիցիոն և լեզվական (ընդգծումը մերն է) դիալեկտիկորեն կապված տարրերի կառուցվածքային միասնությունից» (Վ. Վ. Վինոգրադով, էջ 197)։

Հայտնի է, որ ինչպես տեքստի, այնպես էլ նրա կազմի մեջ մտնող ամեն մի նախադասության տարրերը, ոչ թե մեկուսացված, տարանջատ այլ համակարգայնության հարաբերություններով կապված լեզվական միավորներ են, որոնք հանդես են գալիս որոշակի օրենքների ու կանոնների համապատասխան։ «Ամեն մի գործողություն,— գրում է Ջ. Շմիդտը,— լինի դա լեզվական, թե ոչ լեզվական, նկատի ունի որոշակի նորմա և տեխնիկա, որոնք առանց էական շեղումների ընդունվում ու գործածվում են հասարակության բոլոր անդամների կողմից» (Շմիդտ, էջ 100)։

Լայն ըմբռնմամբ նորմայի ու տեխնիկայի համախումբը կարելի է անվանել քերականություն, որը, փաստորեն, տրված էլեկտրային միավորների հետ նախօրոք սահմանված գործողությունների հաջորդակարգ է։ Եթե ոչ լեզվական իրողություններում օբյեկտների միջև գոյություն ունեցող հարաբերությունները կարելի է նկարագրել այսպես կոչված սոցիալական քերականության օգնությամբ, ապա լեզվական իրողություններում լեզվական միավորների միջև հարաբերությունները նկարագրվում են լեզվի քերականության՝ շարահյուսության և ձևաբանության օգնությամբ։

Այսպիսով, տեքստի բառակազմի ընտրությունը և նրանում բառերի հաջորդման կարգը հիմնականում որոշվում է լեզվական նորմայով։ Սակայն լեզվական նորման քարացած հրահանգների հավաքածու չէ։ Լեզվական նորմայի կիրառության վրա էապես ազդում է հեղինակային անհատականությունը, որով և բնորոշվում է ոճը։ Հետևաբար, տեքստի հե-

ղինակային ոճի որոշումը հանգում է լեզվական նորմայի կիրառման առանձնահատկությունների բացահայտմանը։

Տեքստի վիճակագրությունը ցույց է տալիս, որ հեղինակային ոճի որոշման հարցում տեքստի առանձին միավորների արձանագրումը խիստ կարևոր է, սակայն հուսալի արդյունքներ չի ապահովում։ Արդյունքներն ավելի հավաստի ու ծանրակշիռ են դառնում, երբ իբրև հետազոտության օբյեկտ ընտրվում են ոչ միայն տեքստի առանձին միավորները, այլև նրանց տարբեր զուգորդումները։ Հարկ է նշել, որ քննության համար տեքստից կարող են ընտրվել ինչպես բառային միավորները, այնպես էլ դրանք բնութագրող քերականական կարգերը։

Սահմանում։ Տեքստում լեզվական A միավորի (կամ նրա բնութագրող քերականական կարգի) հանդես գալը անվանենք լեզվական միավորի (կամ քերականական կարգի) հանդիպում և նշանակենք A(n)-ով, որտեղ n-ը A միավորի հանդես գալու թիվն է։

Լեզվական A և B միավորների (կամ նրանց քերականական կարգերի) հարադիր ձևով տեքստում հանդես գալը անվանենք A և B միավորների զույգ հանդիպում և նշանակենք A n B, որտեղ n-ը այդպիսի հարադրության հանդիպումների թիվն է, երբ A-ն գրավում է ձախադաս դիրք։

Տեքստում լեզվական միավորների հանդիպման քննությունը հանգեցնում է միաչափ աղյուսակի, որտեղ տրված են լեզվական միավորները և հանդիպման n-հաճախությունները. այդ միավորների զույգ հանդիպումների քննությունը՝ երկչափ աղյուսակի՝ քառակուսի մատրիցի, որի տողերն ու սյուները փոխհամապատասխանության մեջ են դրվում A և B լեզվական միավորների հետ, իսկ մատրիցի յուրաքանչյուր տարր ներկայացնում է զույգ հանդիպման թիվը կամ հաճախությունը։

Տեքստի կարևոր բնութագրիչներից է բառաձևերը բնորոշող քերականական կարգերի կամ նրանց համախմբերի փոխդասավորությունը։ Այն որոշելու համար տեքստի յուրաքանչյուր բառը փոխարինվում է նրան համադրվող քերականական կարգերի համախմբով, գոյական՝ անուն, հոլով, թիվ, հոլովում, բայ՝ եղանակ, ժամանակ, թիվ, դեմք և այլն։

Մ. Խորենացու «Հայոց պատմության» քննության համար ընտրել ենք քերականական կարգերի 360 համախումբ։

Գոյական գն<անուն, հոլով, թիվ, հոլովում>,
ածուն <հատուկ (հտ), հասարակ հս>,
հոլով <ուղղական (ուղղ), սեռական (սռ), արական (ար), հայցական (հյ), բացառական (բց), գործիական (գծ)>,
թիվ <եզակի (եթ), հոգնակի (հթ)>,
հոլովում<ա, ե, ի, ո, ու, ի-ա (իա), ո-ա, (ոա)>,

ածական *այն <ածականի տեսակ, հոլով>*,
ածականի տեսակ *<դրական (դր), համեմատական (հմ), գերադրական (գր), հարաբերական (հր)>*,
թվական *թն<թվականի տեսակ, հոլով, թիվ>*,
թվականի տեսակ *<քանակական (քն), դասական(դս), անձնական (ած)>*,
դերանուն *դն<դերանվան տեսակ, հոլով>*,
դերանվան տեսակ *<անձնական (ան), փոխադարձ (փձ), ցուցական (ցց), ստացական (սց), հարցական (հց) հարաբերական (հր), անորոշ (աշ), որոշյալ(ոշ)>*,
բայ *բյ <եղանակ, ժամանակ, թիվ, դեմք>*,
եղանակ *<սահմանական (սհ), ստորադասական (սդ), հրամայական (հր), սահմանական եղանակի ժամանակ ներկա (նկ), անցյալ կատարյալ (ակ)>*,
ստորադասական եղանակի ժամանակ *<առաջին ապառնի (առ), երկրորդ, ապառնի (եր)>*,
հրամայական եղանակի ժամանակ *<բուն (բն), հորդորական (հն), արդելական (ար)>*,
դեմք *<առաջին (ադ), երկրորդ (բդ), երրորդ (դդ),*
դերբայ *դբ <տեսակ, հոլով>*,
դերբայի տեսակ *<անորոշ (աշ), անցյալ (ացյ), ապառնի (ապ) ենթադրական (են)>*,
շաղկապ *շղ <համադասական (հդ), ստորադասական (սդ)>*,
մակբայ *մբ <տեղի (տղ,) ժամանակի (ժմ,) ձևի (ձվ), չափի (չփ), աստիճանի (աս)>*,
ժխտական մասնիկ մս,
նախդիր նդ <հոլով>,
նախադրություն նխ <հոլով>,
կետադրություն <միջակետ (մտ), ստորակետ (ստ), բութ (բթ), վերջակետ (վջ)>։

Քերականական կարգերի ընտրված համակարգում գոյականին համադրվում է 168 համախումբ (1-168), ածականին՝ 24 (168-12), թվականին՝ 36 (298-228), դերանվանը՝ 48 (229-276), բային՝ 36 (277-312), դերբային՝ 24 (213-236), շաղկապին՝ 2 (237-238), մակբային՝ 5 (338-343), ժխտական մասնիկին՝ 1 (348), նախդիրներին՝ 6 (345-350), նախադասություններին՝ 6 (351-356) և կետադրական նշաններին՝ 4 (357-360)։

Հարկ է նշել, որ ընտրված քերականական կարգերի համախմբերի թիվը գերազանցում է մինչև այժմ կատարված փորձերում ընդգրկված

համախմբերի թվից։ Մասնագիտական գրականության մեջ հայտնի է 120-150-ական կարգերի համախմբերի ընտրման դեպք։ Մեր փորձի համար, ինչպես նշվեց, ընտրել ենք քերականական կարգերի թվով 360 համախումբ։ Դա ունի և՛ դրական, և՛ բացասական կողմեր։

Դրականն այն է, որ տեքստի ամեն մի բառ նկարագրվում է ավելի մանրամասն, որը թույլ է տալիս հաշվի առնելու հնարավորին չափ ավելի շատ փաստեր։ Այսպես, առանձին ուշադրության է արժանացել կետադրության հարցը, որը այլ փորձերում բաց է թողնվել (Բորոդկին)։ Մեր կարծիքով, տեքստում առկա կետադրությունը կարևոր ինֆորմացիա է կրում նրա կառուցվածքի վերաբերյալ։ Տեքստի կետադրությունը կարելի է օգտագործել նախադասությունների դասակարգման, նրանց սկզբնական և վերջնական հատվածների քննության, բառային կապակցությունների վերլուծության ժամանակ և այլն։ Իհարկե, այն տեքստերում, որտեղ կետադրությունը բացակայում է, քերականական կարգերի ցանկից կարելի է հանել նաև կետադրությանը համադրվող համախմբերը։

Քերականական կարգերի համախմբերի մեծ ցանկի ընտրության բացասական կողմն այն է, որ ստեղծվում են լրացուցիչ դժվարություններ զույգ հանդիպումների մատրիցի կազմման ու հետագա մշակման հարցում։

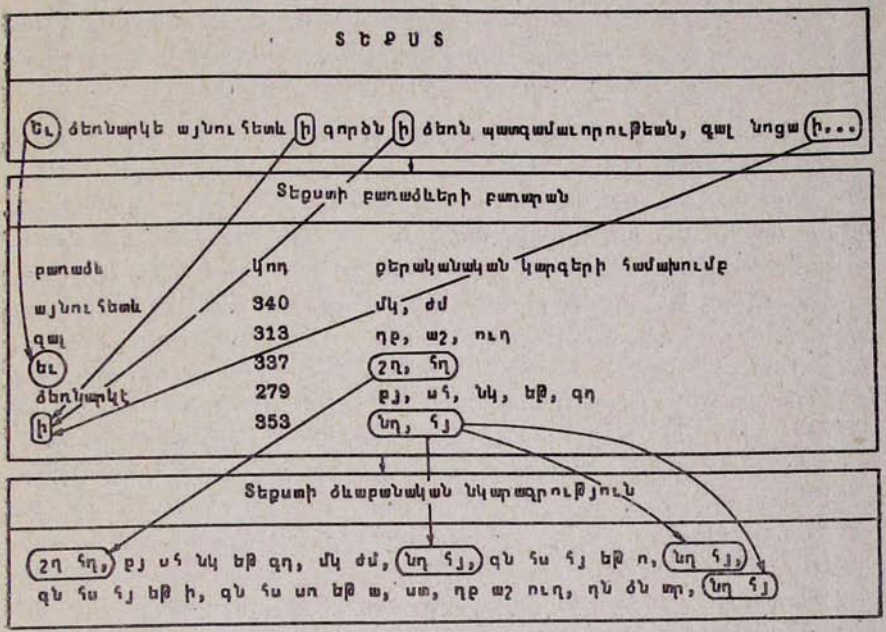
Որո՞նք են այդ դժվարությունների հաղթահարման միջոցները։ Իհարկե, քերականական կարգերի 360 համախմբերի օգնությամբ տեքստի ձևաբանական վերլուծությունը և քերականական կարգերի զույգ հանդիպումների մատրիցի կառուցումը և մշակումը ավանդական եղանակներով չափազանց դժվար է, գուցե և անհնար։ Սակայն քերականական կարգերի այդպիսի քանակությունը ընտրվել է նկատի ունենալով աշխատանքների ավտոմատացումը EC-1045 հզոր հաշվողական մեքենայի օգնությամբ։ Մյուս կողմից ծրագրային ապահովման համակարգը նախագծվել է այնպես, որ անհրաժեշտության դեպքում կարելի է նվազեցնել ելակետային կարգերի թիվը։ Այսպես, եթե գոյականական կարգերի քննության ժամանակ հաշվի չառնվի հոլովումը, ապա 7 անդամ կպակասի գոյականին համադրվող քերականական կարգերի համախմբերի թիվը, և այն 168-ից կդառնա 24։ Նման փորձեր կարելի է կատարել նաև մյուս խոսքի մասերի հետ։ Պարզագույն դեպքում կարելի է փորձը դնել միայն խոսքամասային մակարդակով։

Այսպիսով, քերականական կարգերի զույգ հանդիպումների մատրիցի կազմման առաջին փուլը տեքստի ձևաբանական նկարագրությունն է, որը տեքստաբանության հիմնական, աշխատատար փուլերից է։ Այդ նկատառումով տեքստի վերլուծության ժամանակ առաջին հերթին դրվում է ձևաբանական վերլուծության ավտոմատացման խնդիրը։ Ժամանակա-

կից լեզուների համար արդեն գոյություն ունեն ձևաբանական վերլուծության ավտոմատ համակարգեր։

Հր. Աճառյանի անվան լեզվի ինստիտուտում մշակված է ժամանակակից հայերենի ձևաբանական վերլուծության ալգորիթմ (Ռ. Ուռուտյան)։ Ինչ վերաբերում է հին հայերենին, ապա համապատասխան ալգորիթմներ դեռ մշակված չենք։ Այստեղ, եթե դիմենք ձևաբանական վերլուծության ավանդական եղանակներին, այսինքն՝ կատարենք ձեռքով, ապա խիստ նվազում է ընտրանքը, որը բացասաբար կանդրադառնա արդյունքների հավաստիության և ճշգրտության վրա։

Այդ պատճառով մենք ընտրել ենք տեքստի ձևաբանական նկարագրության ավտոմատացման եղանակ, որն այսօր էլ լայն կիրառում ունի տեքստերի ավտոմատ մշակման գործում (Լենդերս)։ Դա ձևաբանական վերլուծության բառարանային եղանակն է։



Եթե կազմենք տեքստի բառաձևերի բառարան, որտեղ ամեն մի բառաձև համադրվում է քերականական կարգերի համախմբով, ապա այն կարող ենք օգտագործել ձևաբանական նպատակով։ Դրա համար անհրաժեշտ է տեքստից ընտրած յուրաքանչյուր բառի համար բառարանում գտնել համապատասխան գլխաբառը և ստանալ նրան համադրվող քերականական կարգերի համախումբը։

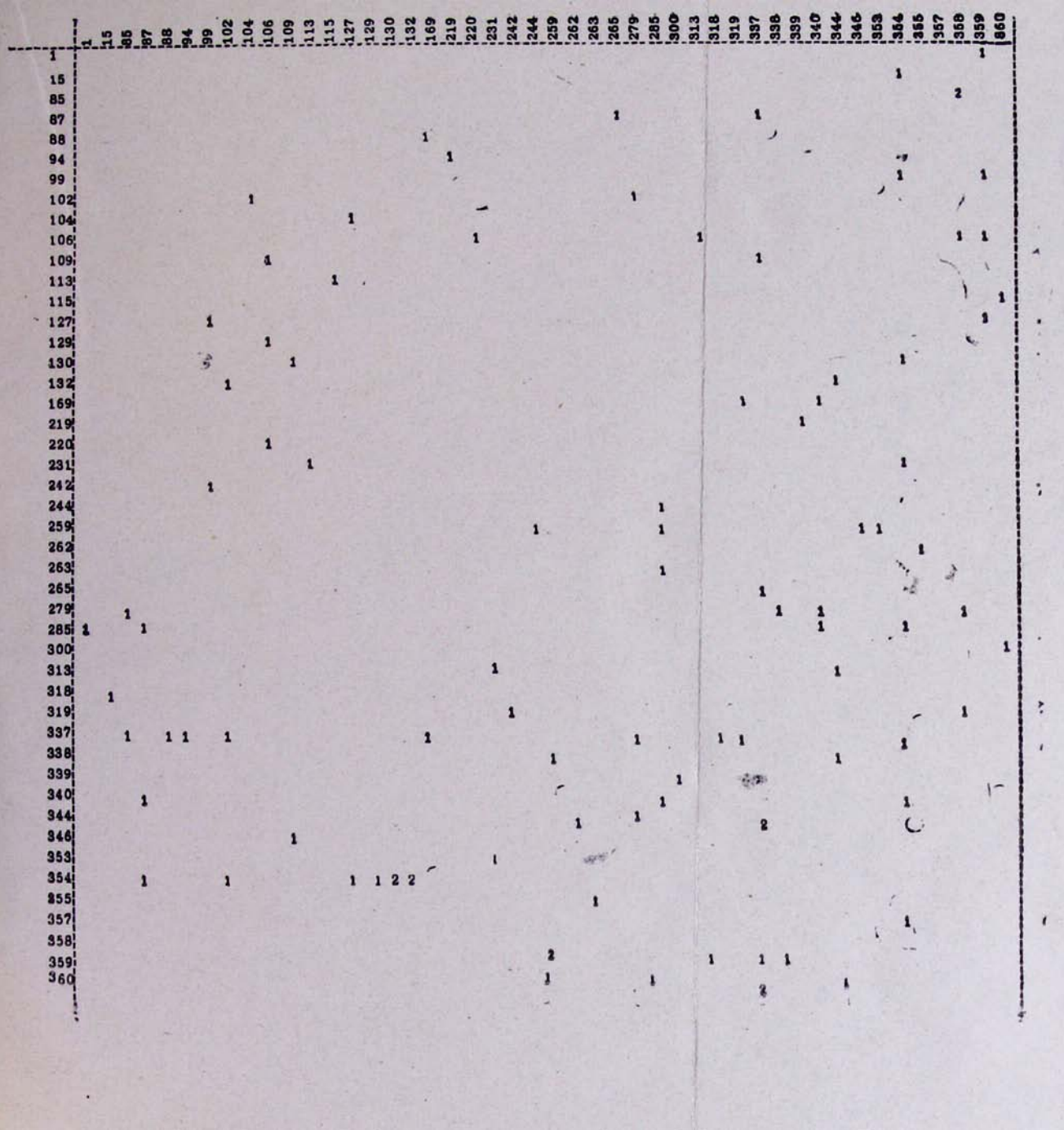
Տեքստի ձևաբանական վերլուծության առաջարկվող եղանակը կարելի է ներկայացնել հետևյալ կերպ.

Որպես օրինակ բերենք Մ. Խորենացու «Հայոց պատմություն» ստեղծագործությունից (Մ. Խորենացի, Հայոց պատմություն, Երևան, 1981) 96 էջի առաջին պարբերության ձևաբանական նկարագրությունը

Տեքստ Եւ ձեռնարկէ այնուհետեւ ի գործն ի ձեռն պատգամաւորութեան, գալ նոցա ի տեսութիւն միաբանութեան ի տեղի միջոց սահմանաց երկոցունց թագավորութեանցն. իբր բան ինչ և գործ հարկաւոր հասեալ, որ ի ձեռն գրոյ և հրեշտակութեան կատարել ոչ է կարողութիւն, եթե ոչ և դէմք երկոցունց հանդէպ լինիցին։ Այլ գիտելով Տիգրանայ զառաքելոյ իրին կատարումն՝ ոչ ինչ յորոց խորհէրն Աժդահակ՝ ծածկէ, այլ ի ձեռն գրոյ յայտնէ որ ինչ ի նորայն խորութեան սրտի։ Եւ յայտնեալ այսպիսւոյ չարութեան՝ ոչ ինչ էր այնուհետեւ բան և խորամանկութիւն, որ զայսպիսի առագաստէր զչարութիւն. այլ յայտնի այնուհետեւ գրգռէր խաղմն։

Հատվածի ձևաբանական նկարագրություն. /շղ, հդ/ բյ, սհ, նկ, եթ,դդ /մկ, ժմ/ նդ, հյ/ գն, ես, հյ, եբ, ռ/ նդ, հյ /գն, հս, հյ, եթ, ի/ գն, հս, սռ, եթ, ա /ստ/ դբ, աշ, ուղ, /ղն, ձն, տր/ նդ, հյ /գն, հս, հյ, եթ, ա/ գն, հս, սռ, եթ, ա /նդ հյ/ գն, ես, հյ, եբ, ռ/գն, հս, սռ, եթ, ո/գն, հս, սռ, հթ, ա/ թն, ած, սռ, հթ /գն, հս, սռ, հթ,ա/ մտ /նխ, հյ/ գն, հս, ուղ, եթ,ի /ղն, աշ, ուղ /շղ, հդ/ գն, հս, ուղ, եթ, ո /ան, դբ, ուղ/ դբ, աց, ուղ /ստ/ դն, հբ, ուղ /նխ, սռ/ գն, հս, սռ, եթ, ո /շղ, հդ/ գն, հս, սռ, եթ, ա /դբ, աշ, ուղ,/ մս /բյ, սհ, նկ, եթ, գդ/ գն, հս, ուղ, եթ, ա /ստ/ շղ, հդ /մս/ գն, հս, ուղ, հթ,ի /թն, ած, սռ, եթ/ մկ, տղ /բյ, սդ, առ, հթ, գդ/ վջ /շղ հդ /դբ, աշ, գծ/ գն, հտ, սռ, եթ, ա /նդ, հյ/ գն, հս, սռ, եթ, ռ/ գն, հս, սռ, եթ, ա /գն, հս, հյ, եթ, ա/ բթ /մս/ դն, հբ, հյ /նդ, բց/ դն, հբ, բց /բյ, սհ, աա, եթ, գդ/ գն,հտ, ուղ, եթ, ա /բթ/ բյ, սհ, նկ, եթ, գդ /ստ/ շղ, հդ /նդ, հյ/ գն, հս, հյ, եթ, իա /գն, հս, սռ, եթ,ո/ բյ, սհ, նկ, եթ, գդ /շղ, սդ/ դն, հբ, ուղ /նդ, տր/ դն, ձն, տր /գն, հս, տր, եթ, ա/ գն, հս, տր,եթ, ի/ վջ /շղ, հդ/ դբ, աց, ուղ /դն, ցց, սռ/ գն, հս, սռ, եթ, ա /բթ/ մս /դն, հբ, ուղ/ բյ, սհ, աա, եթ, գդ /մկ, ժմ/ գն, հս, ուղ, եթ, ի /շղ, հդ/ գն, հս, ուղ, եթ, ա /ստ/ դն, հբ, ուղ /դն, ցց, հյ/ բյ, սհ, աա, եթ, գդ /նդ, հյ/ գն, հս, հյ, եթ, իա /մտ/ շղ, հդ /ան, դբ, ուղ/ մկ, ժմ /բյ, սհ, աա, եթ, գդ/ գն, հս, ուղ, եթ, ի /վջ/

Եթե տեքստի ձևաբանական նկարագրությունը պատրաստ է, արդեն կարելի է անցնել քերականական կարգերի զույգ հանդիպումների մատրիցների կազմմանը։ Մեր կողմից բերած պարբերության մեջ ներառված է 100 բառ, հաշված նաև կետադրական նշանները։

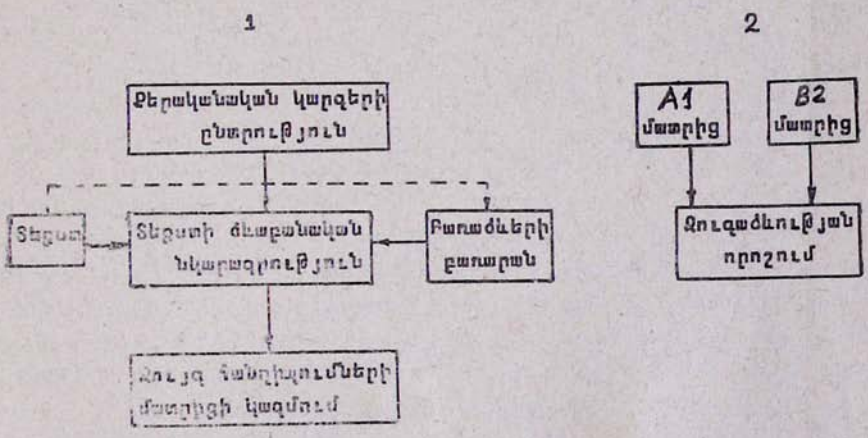


Սովորաբար փորձարկվում են 1000-1500 լիիմաստ բառերով հատվածներ։ 100 բառը նկարագրվում են քերականական կարգերի 46 համախմբով (քերականական կարգերի համախմբերի ցանկը տրված է հավելվածում)։ Հատվածի բառերի և քերականական համախմբերի հարաբերությունը 2:1 է։ Սակայն բառերի հետագա ավելացումը չի բերի քերականական կարգերի համեմատական աճի, քանի որ համախմբերի թիվը 360 է։ 1500 բառի դեպքում այդ հարաբերությունը կլինի մոտավորապես 5:1։

Բերված 100 բառն արդեն ցույց է տալիս քերականական կարգերի և նրանց զուգորդումների որոշ բաշխվածություն։ Այն ավելի ակնառու դարձնելու համար զույգ հանդիպումների մատրիցի հիման վրա կառուցվում է զույգ հանդիպումների գրաֆ։ Դա մի ուղղորդված հարթ գրաֆ է, որի գագաթները փոխհամապատասխանության մեջ են դրվում քերականական կարգերի համախմբերի, իսկ աղեղները, որոնք ուղղորդված են ըստ տեքստի ձախից աջ, ցույց են տալիս զույգ հանդիպումների թիվը։

Փորձնական եղանակով հաստատված է, որ նույն ոճի տեքստերին համապատասխանող գրաֆների զուգաձևության (изоморфность) աստիճանը բարձր է։ Զուգաձևության արդեն 25—30 տոկոսի դեպքում կարելի է խոսել երկու տեքստերի միևնույն ոճի մասին։

Տեքստում քերականական կարգերի զույգ հանդիպման բաշխվածության խնդրի մաթեմատիկական ապահովման համակարգի սկզբունքային սխեման բերվում է ստորև (նկ. 1)։



Համակարգը բաղկացած է երկու ենթահամակարգերից՝ 1) զույգ հանդիպումների մատրիցի կազմման և 2) զույգ հանդիպումների մատրիցների համեմատության։

Առաջին ենթահամակարգի ելակետային տվյալները տեքստն է, աշխատանքի արդյունքը՝ քերականական կապերի զույգ հանդիպումների մատրիցը: Աշխատանքի սկզբունքը հետևյալն է. նախ բառաձևերի բառարանի օգնությամբ ստացվում է տեքստի ձևաբանական նկարագրությունը, որը, այնուհետև վերածվում է քերականական կարգերի համախմբերի կոդերի հաջորդականության: Այսպես, եթե վերևում բերած պարբերության ձևաբանական նկարագրության մեջ (շղ, հդ) քերականական կարգը փոխարինենք համապատասխան կոդով՝ 337 (տես ծանոթությունը), բյ, սև, նկ, եր, գդ, կարգը՝ 297-ով և այլն, ապա կստանանք կոդերի հետևյալ հաջորդականությունը 337, 279, 340, 354, 130, 354, 129, 99, 359, 313, 231, 54 ...

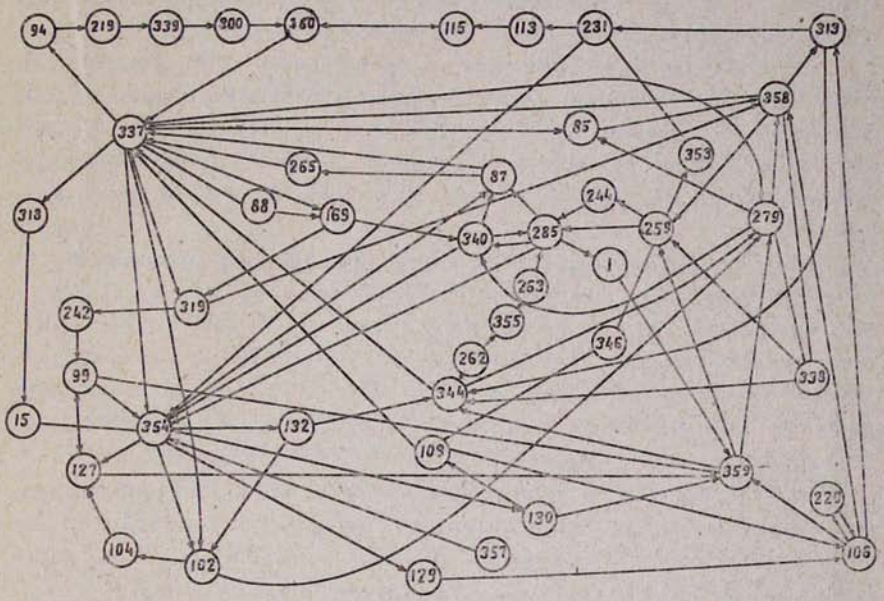
Վերցնենք հաջորդականության առաջին զույգը՝ 337,279:
Այն ցույց է տալիս, որ 337 և 279 կոդեր ունեցող քերականական կարգերով բնորոշվող բառերը տեքստում հարադիր են, որտեղ 337 կոդով բնորոշվող բառը զբաղեցնում է ձախադաս դիրք: Ըստ զույգ հանդիպումների մատրիցի սահմանման 337, 269, կոդերի զույգին համապատասխանության մեջ է դրվում qij=337, 279 տարրը, որովհետև, ինչպես ընդունել ենք, տեքստում մատրիցի տողերին համապատասխանող կարգերին հաջորդում են սյուներին համապատասխանող կարգերը: Հետևաբար, եթե վերևում ստացված հաջորդականության անդամները զույգ առ զույգ խմբավորենք այնպես, որ միևնույն կոդը առաջին խմբում ունենա ձախադաս դիրք (i-ի) տեղում, ապա կստանանք զույգ հանդիպումների մատրիցի տարրերը:
(337, 279), (279, 340), (340, 354), (354, 130), (130, 354), (354, 129), (129, 99), (99, 359), (395, 313), (313, 231), (231, 354)...

Ակնհայտ է, որ նման խմբավորման դեպքերում կլինեն խմբեր, որոնց բովանդակությունը կհամընկնեն: Դա ստացվում է այն դեպքում, երբ երկու հանդիպումներում համապատասխանաբար ձախադաս և աջադաս բառերի քերականական կարգերը համընկնում են: Մեր կողմից բերված օրինակներում (նդ, հյ) և (գն, հս, հյ, եթ, ի) կարգերը հանդիպում են երկու անգամ: Եվ քանի որ զույգ հանդիպումների մատրիցի տարրերը ցույց են տալիս տողերին ու սյուներին համապատասխանող կոդերի (քերականական կարգերի) հանդիպումների թիվը, ապա հաջորդ փուլում ստացված հաջորդակցությունից պետք է հեռացնել կրկնվող խմբերը, նախապես գրանցելով նրանց հանդիպման թիվը.
(337, 279), (279, 340), (340, 354), (354, 130), (130, 354), (354, 129), (129, 99), (99, 359), (359, 313), (313, 321), (231, 354) ...

Այժմ ստացված հաջորդականության հիման վրա զույգ հանդիպումների մատրիցի կազմումը առանձին դժվարություն չի ներկայացնում:

Ջույգ հանդիպումների մատրիցի կազման հարցում կարևոր խնդիր է տեքստի ձևային նկարագրության ստացումը։ Ինչպես նշվեց վերևում, այստեղ այդ խնդիրը վճռվում է տեքստի լեզվի բառաձևերի բառարանի օգնությամբ։ Դա մի ցուցակ է, որի մուտքը բառաձևեր են, իսկ ելքը՝ դրանք բնորոշող քերականական կարգերի համախմբեր։ Առանձին խնդիր է այդ ցուցակի կազմումը։ Այն կարելի է վճռել երկու եղանակով.

ա) նախապես կազմվում է լեզվի ամբողջ բառապաշարի բառաձևերի բառարանը բ) կազմվում է հաշվողական մեքենայի մեջ գրանցված տեքստի բառաձևերի բառարանը։ Մենք ընտրել ենք վերջինը։



Հարկ է նշել, որ եթե բառաձևերի բառարանը դիտենք իբրև առանձին հանգույց, որը կարելի օգտագործել տարբեր խնդիրների վճռման համար (տես ներքևում), ապա, հատուկ մաթեմատիկական ապահովման դեպքում, երկրորդ եղանակով կարող ենք հասնել ամբողջ բառապաշարի բառաձևերի բառարանի կազմմանը։ Դրա համար անհրաժեշտ է նորանոր տեքստերի մշակման ժամանակ զուգահեռաբար լրացնել բառաձևերի բառարանը, մինչև ստացվի այնպիսի վիճակ, երբ այն այլևս լրացման կարիք չունենա։

Համակարգի մաթեմատիկական ապահովման առումով կարևոր հարց

է քերականական կարգերի ընդհանուր ցանկից տվյալ փորձի համար անհրաժեշտ համախմբերի ընտրությունը։

Քերականական կարգերի համախմբերի ընդհանուր թիվը 360 է։ Սակայն քերականական կարգերի այդպիսի մեծ քանակ բոլորովին էլ պետք չէ ամեն մի փորձի համար, և փորձի կառավարումն ավելի ճկուն դարձնելու նպատակով մենք պետք է հնարավորություն ունենանք ընտրելու այն քերականական կարգերը, որոնք անհրաժեշտ են տվյալ տեքստի նկարագրության համար։ Այլ կերպ ասած՝ համակարգը մեզ պետք է հնարավորություն տա թվարկել այն քերականական կարգերը, որոնց օգնությամբ կազմելու է տեքստի ձևաբանական նկարագրությունը։ Այդ խնդիրը լուծվում է հատուկ երկխոսական ծրագրի օգնությամբ։

Եթե երկխոսության ընթացքում համակարգին որևէ լրացուցիչ ինֆորմացիա չի հաղորդվում, ապա տեքստի ձևաբանական վերլուծությունը կատարվում է այսպես կոչված խոսքիմասային մակարդակով՝ գոյական, ածական, թվական, դերանուն, բայ, դերբայ, շաղկապ, մակբայ, նախդիր, նախադրություն, ժխտական մասնիկ, միջակետ, ստորակետ, բութ, վերջակետ։

Հետևաբար, զույգ հանդիպումների մատրիցի նվազագույն չափը 16×16 է, այն դեպքում, երբ առավելագույնը՝ 360×360 է։

Այսպիսով, համակարգը մեզ թույլ է տալիս, ըստ անհրաժեշտության, տեքստի ձևաբանական վերլուծության խորության (մանրամասնության) աստիճանն ավելացնել ավելի քան 500 անգամ (եթե 16×16 մատրիցի դեպքում առավելագույնը կունենանք 256 հանդիպում ապա 360×360 մատրիցի դեպքում՝ 129600 հանդիպում)։

Երկրորդ ենթահամակարգը նախատեսված է երկու մատրիցների զուգաձևության աստիճանի որոշման համար։ Ստացված մատրիցները վերածելով երկուորդական մատրիցների՝ դիտելով յուրաքանչյուր հանդիպումը իրեն կապ տողերին և սյուներին համապատասխանող տարրերի միջև, զուգահեռաբար կարող ենք հաշվել այդ կապերի որոշ բնութագրիչներ՝ մեկ, երկու և ավելի կապերով հեռացված տարրերի մատրիցներ։

Անդրադառնանք երկրորդ խնդրին։

Ամեն մի տեքստի առանձնահատուկ է տեքստաչափական այնպիսի մի բնութագրիչ, ինչպիսին բառարանի ծավալն է։

Եթե համեմատենք լեզվի բառարանային միավորները տեքստային միավորների հետ, ապա նրանք տեքստում հանդես են գալիս անփոփոխ կամ փոփոխված՝ հարուցուցային (հոլովում, խոնարհում) շարքի որևէ անդամի տեսքով։

Հայտնի է, որ ամեն մի հարուցույց ունի իր էլակետային, կամ ինչպես ասվում է չնշույթավորված, միավորը, որին հակադրվում են շարքի մյուս՝

նշույթավորված միավորները։ Տեքստի բառերի՝ որպես նշույթավորված միավորների, ելակետային ձևերի ամբողջությունը կազմում է տեքստի բառարանը։

Վերջինի ընտրությունը որոշվում է հեղինակի անհատականությամբ. մի հեղինակ հակում ունի լավագույն կերպով կիրառել լեզվի բառափոփոխման (հոլովման, խոնարհման) համակարգը և փոքր քանակի բառերով կերտել տեքստ, իսկ մեկ այլ հեղինակ նույն երկարության տեքստ կերտում է համեմատաբար մեծ բառարանի օգնությամբ։

Տեքստաչափության մեջ ընդունված է, այսպես կոչված, զանազանության գործակիցը, որը ցույց է տալիս, թե հեղինակն ինչ չափով է օգտվում լեզվի բառափոփոխման համակարգից։ Դա տեքստի բառերի քանակի և նրա բառարանի ծավալի հարաբերությունն է։

Զանազանության գործակիցը ստատիկ պարամետր է, որը միայն ցույց է տալիս տեքստը որքան անգամ է մեծ բառարանից։

Տեքստաչափության մեջ կիրառվում է մեկ այլ բնութագրիչ ևս, որը ցույց է տալիս նոր բառերի կախումը բառակիրառությունների թվից։ Ակնհայտ է, որ այն գծային չէ։

Փորձնական եղանակով որոշված է, որ տեքստում նոր բառերի և բառակիրառությունների թիվը կապված են հետևյալ առնչությամբ՝

$$V = RN^a$$

որտեղ

V-ն նոր բառերի թիվն է։

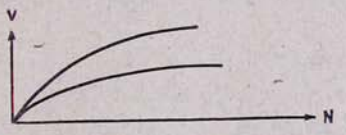
N-ը բառակիրառությունների թիվն է.

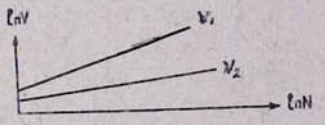
R-ը և a-ն տվյալ տեքստի համար բնորոշ հաստատուններ են

$$(o < a < 1)։$$

Պարզ է, որ

R-ի և a-ի տարբեր արժեքների դեպքում կստանանք տարբեր կորերի բազմություն։





Կորերի տեսքից պարզ երևում է, որ տեքստի սկզբնական մասում նոր բառերի թիվն արագ աճում է, իսկ այնուհետև հասնում մի սահմանի, որից աճը դադարում է։

Այդ դեպքում, երբ ընտրանքը մեծ է, հարմար է օգտվել լոգարիթմական կոորդինատային համակարգից։ Լոգարիթմելով վերևում տրված առնչությունը կստանանք.

$$\lg V=\lg R+a\lg N$$

լոգարիթմական կոորդինատային համակարգում սա ուղիղ գծի հավասարում է։

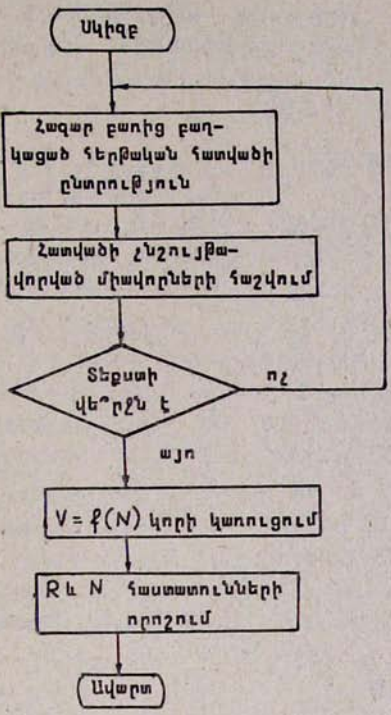
Այսպիսով, Մ. Խորենացու «Հայոց պատմություն» ստեղծագործության տեքստի համար զանազանության գործակցի և բառակիրառությունններից կախված բառարանի աճի որոշման համար անհրաժեշտ է տեքստի ամեն մի բառը ձևափոխել չհնչյունափոխված տեսքի։ Դրա համար իբրև հիմք կարող է ծառայել բառաձևերի բառարանը, որ կազմել ենք նախորդ խնդրի վճռման համար։ Այդ բառարանի յուրաքանչյուր գլխաբառը լրացվում է նշույթավորված ձևով։

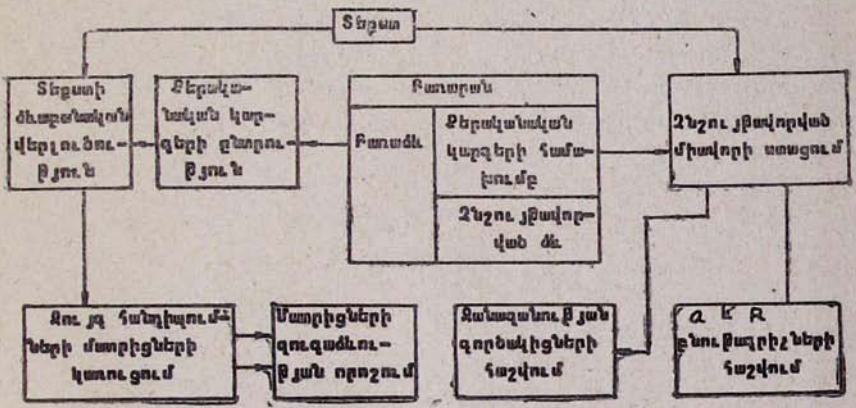
Ստորև բերում ենք բառաձևերի բառարանից մի հատված, որտեղ գլխաբառերը լրացված են նշույթավորված ձևերով։

Բառաձևերի բառարան	
Նշույթավորված ձևը	Նշույթավորված ձևը
այնուհետև	այնուհետև
գալ	գալ
գորշն	գորշ
ձևոն	ձեռն
ձեռնարկէ	ձեռնարկել
նոցա	նա
պատգամաւորութեան	պատգամաւորութիւն

Ունենալով այսպիսի բառարան՝ ալգորիթմական եղանակով կարող ենք կազմել տեքստի նշույթավորված ձևերի բառարանը։ Որպեսզի ստանանք $V=RN^a$ բանաձևերի R և a հաստատունները, անհրաժեշտ է տեքստի ամեն մի հազարական հատվածի համար հաշվել նոր գոյացող նշույթավորված ձևերի թիվը։

$V=RN^a$ կախման R և a հաստատունների որոշման ալգորիթմը կարելի է ներկայացնել հետևյալ կերպ (տե՛ս ալգորիթմը՝ էջ 149)։ 3-րդ նկարում տրվում է տեքստում քերականական կարգերի զույգ հանդիպումների մատրիցի կազմման ու վերլուծության, բառերի զանազանության գործակիցների հաշվման համակարգի սկզբունքային սխեման։


Սկիզբ
Հազար բառից բաղ-
կացած հերթական հատվածի
ընտրություն
Հատվածի չնշություն-
վորված միավորների հաշվում
Տեքստի
վե՞րջն է
ոչ
այո
V = f(N) կորի կառուցում
R և N հաստատունների
որոշում
Ավարտ


Տեքստ
Տեքստի
ձևաբանական
վերլուծու-
թյուն
Քերակա-
նական կար-
գերի ընտրու-
թյուն
Բառարան
Բառաձև
Քերականական
կարգերի համա-
խումբ
Չնշություն-
ված ձև
Չնշություն-
վող միավորի առացում
Ջուր չկ հանդիպումնե-
րի մատրիցների
կառուցում
Մատրիցների
զուգաձուգ-
թյան որոշում
Ջանազանություն-
գործակիցների
հաշվում
a և R
գնում թափրիչների
հաշվում

Ցանկություն. *Տեքստում բերված հատվածի քերականական կարգերի զույգ հանդիպումների մատրիցի կազմման համար քերականական կարգերի ընդհանուր ցուցակից ընտրված համախմբերը համապատասխան կոդերով.*

գն, հտ, ուղ, եթ, ա—1
գն, հտ, սո, ա—15
գն, հս, ուղ, եթ, ա—85
գն, հս, ուղ, եթ, ի—87
գն, հս, ուղ, եթ, ո—88
գն, հս, ուղ, հթ, ի—94
գն, հս, սո, սո, եթ, ա—99
գն, հս, սո, եթ, ո—102
գն, հս, սո, եթ, իա—104
գն, հս, սո, հթ, ա—106
գն, հս, սո, հթ, ո—109
գն, հս, տր, եթ, ա—113
գն, հս, տր, եթ, ի—115
գն, հս, հյ, եթ, ա—127
գն, հս, հյ, եթ, ի—129
գն, հս, հյ, եթ, ո—130
դն, հր, հյ—262
դն, հբ, բց—263
դն, աշ, ուղ—265
բյ, սհ, նկ, եթ, գդ—279
բյ, սհ, աա, եթ, գդ—295
բյ, սդ, առ, հթ, գդ—300
դբ, աշ, ուղ—313
դբ, աշ գծ—318
դբ, ացֆ, ուղ—319
ղղ, հդ—337
ղղ, սդ—338
մկ, տղ—339
մկ, ժմ—340
մս—344
նխ, սո—346
նխ, հյ—348
գն, հս, հյ, եթ, իս—132
ան, գր, ուղ—169
քն, աձ, սռ, եթ—219
քն, աձ, սո, հթ—220
դն, ձն, տր—231
դն, ցց, սո—242
դն, ցց, հյ—244
դն, հր, ուղ—259
նդ, տր—353
նդ, հյ—354
նդ, բց—355
մտ—357
ստ—359
վջ—360

ՕԳՏԱԳՈՐԾՎԱԾ ԳՐԱԿԱՆՈՒԹՅԱՆ ՑԱՆԿ

Бородкин Я. Н., Милов Л. В., О некоторых аспектах автоматизации текстологического исследования, в кн.: «Математические методы в историко-экономических и историко-культурных исследованиях». Наука, М., 1977, стр. 235—280.

Виноградов В. В., Проблемы авторства и теория стилей, М., 1961.

Мелихов А. Н. Ориентированные графы и конечные автоматы, М., Наука, 1971.

Шмидт З. М. «Текст» и «История» как базовые категории в сб.: «Новое в зарубежной лингвистике, вып. VIII», М., 1978, стр. 89—110.

Ոտոտյան Ռ. Լ., ժամանակակից հայերենի անվանական բառափոխման ձևային նըկարագրություն, «Լեզվի և ոճի հարցեր, 6-րդ պրակ» Երևան, 1982։

Ոտոտյան Ռ. Լ., ժամանակակից հայերենի բայական ձևակազմություն, «Բառ, նախադասություն, տեքստ» Երևան, 1987։

Maschinell Auswertung sprach-historischer Quellen. Ein Ber Zur computerunterstützten Analys der Flextonsmorphologie des Frühneuhochat (Lenciers W.,Wegera K —P., Berg E., et al—Tübingen: Nimeger, 1982—XI, 236 s.