ՀՏԴ 811.19'32

«ԱՐԵՎԵԼԱՀԱՅԵՐԵՆԻ ԱԶԳԱՅԻՆ ԿՈՐՊՈՒՍԻ» ԿԱՌՈՒՑՎԱԾՔԸ ԵՎ ԸՆԴԳՐԿՈՒՄԸ

ՖՐԻԴԱ ՀԱԿՈԲՅԱՆ

Հիմնաբառեր՝ կորպուսային լեզվաբանություն, լեզվական կորպուս, տեքստային հենք, ԱՐԵՎԱԿ, Արևելահայերենի ազգային կորպուս, ընդհանրական կորպուս, մոնիփորային կորպուս, պատմական կորպուս

Օրինաչափ է, որ լեզուն իր գոյության ընթացքում շարունակ փոփոխությունների է ենթարկվում, և ժամանակի ընթացքում դրա ուսումնասիրության նոր եղանակներ են ի հայտ գալիս, ավելին՝ երբեմն հենց լեզվի իրացումը խոսքում է միջոց դառնում լեզվական իրողությունների ուսումնասիրության, նկարագրության ու ներկայացման համար: Խոսքային օրինակների միջոցով լեզվական երևույթների քննության եղանակներից մեկը *կորպուսային լեզվաբանությունն* է՝ ժամանակակից լեզվաբանության ամենաարագ զարգացող երիտասարդ ուղղություններից մեկը: Կորպուսային լեզվաբանության բնագավառում ձեռք բերված մեծ հաջողություններն առավելապես պայմանավորված են լեզվաբանության մեջ համակարգչային տեխնիկայի լայն կիրառությամբ և այնպիսի ծրագրերի ստեղծմաբ, որոնց միջոցով համակողմանիորեն ուսումնասիրվում և ներկայացվում է լեզուն:

Հայ լեզվաբանության մեջ լեզվական կորպուս ստեղծելու ուղղությամբ տարվող աշխատանքները սահմանափակվել են միայն *Արևելահայերենի ազգային կորպուսի՝ ԱՐԵՎԱԿ-ի*[58] ստեղծմամբ: Այն ստեղծվել է տարբեր հաստատություններ ներկայացնող մի խումբ լեզվաբանների և ծրագրավորողների մասնակցությամբ՝ Կորպուս Թեքնոլոջիս (Corpus Technologies), Ռուսաստանի գիտությունների ակադեմիայի լեզվաբանության ինստիտուտ, Մոսկվայի պետական համալսարան, Ռուսաստանի պետական հումանիտար համալսարան, Երևանի պետական լեզվաբանական համալսարան, Հայաստանի գիտությունների ակադեմիայի լեզվի ինստիտուտ: ԱՐԵՎԱԿ-ի ստեղծումը, ներառյալ գրավոր և բանավոր խոսքի հավաքումը,

[58] Տե՛ս www.eanc.net/EANC/search/?interface_language=am

ֆինանսավորվել է Կորպուս Թեքնոլոջիս (ԿԹ) ընկերության կողմից, որի նպատակն է լեզվաբանական ծրագրային ապահովումը[59]:

Անշուշտ, սույն հոդվածում մեր նպատակը ոչ թե ԱՐԵՎԱԿ-ը նկարագրել-ներկայացնելն է, այլ կորպուսային լեզվաբանության զարգացման արդի փուլում այն մյուս կորպուսների համեմատությամբ դիտելը՝ առանձնացնելով կիրառության այն ոլորտները, որոնցում, այս կորպուսն օգտագործելով, կարելի է հասնել զգալի հաջողությունների: Կորպուսը որակական զարգացման նոր մակարդակի հասցնելու նպատակով արվում են նաև առաջարկություններ ու դիտողություններ: Նախ տեսնենք, թե լեզվական կորպուսները բնութագրական ինչ հատկանիշներ ունեն, և ԱՐԵՎԱԿ-ը ինչ չափով է համապատասխանում դրանց:

Եթե լեզվական կորպուս ասելով հասկանում ենք տարբեր բնույթի տեքստերի ամբողջական հավաքածու, ապա պայմանականորեն մեկից ավելի ցանկացած տեքստ կարող է կորպուս համարվել: Սակայն կան ընդհանրական մի շարք հատկանիշներ, որոնցով բնութագրվում են կորպուսներն ընդհանրապես: Այդ հատկանիշներից ամենահիմնականն ու թերևս գլխավորն այն է, որ տեքստային հենքում եղած տեքստերի միջոցով պետք է ներկայացվի լեզուն: Հետևաբար, տեքստային հենքը պետք է կազմել այնպես, որ, դրանում եղած տեքստերն ուսումնասիրելով, կարողանանանք քննել լեզվական բոլոր իրողությունները: Հենքում եղած տեքստերը սկզբունքորեն կարելի է երկու եղանակով ուսումնասիրել: Առաջին դեպքում կարող ենք քննել բոլոր տեքստերն առանձին-առանձին: Միանգամից նկատենք, որ այս սկզբունքը գործնականականում կիրառելի չէ, որովհետև մեծ ցանկության դեպքում անգամ կարելի է միայն մեռած լեզուների տեքստային բոլոր օրինակները քննել, իսկ կենդանի լեզվում տեքստերն անթիվ են: Նաև որևէ կերպ հնարավոր չէ կազմել տեքստային ամբողջական հենք. մի կողմից՝ կարող են դուրս մնալ լեզվում հազվադեպ գործածվող օրինակները, մյուս կողմից՝ ուշադրությունը բևեռելով միայն սակավ գործածական օրինակների վրա՝ կարող են հենքից դուրս մնալ նույնիսկ հաճախադեպ գործածական լեզվական օրինակները: Ստացվում է այնպես, որ ամեն դեպքում մենք կարող ենք քննել միայն սահմանափակ թվով տեքստային օրինակներ, իսկ թե ինչ չավով հնարավոր կլինի վեր հանել լեզվական բոլոր իրողությունները դրանց քննությունից, դժվար է ասել:

[59]Տե՛ս http://www.eanc.net/am/grant/

Այնուամենայնիվ, այսօրվա համակարգչային տեխնիկան հնարավորություն է տալիս ստեղծել չափազանց հարուստ տեքստային այնպիսի հենքեր, որոնք ներառում են հազարավոր տեքստեր՝ միլիոնավոր բառերով: Տեքստերի այդ վիթխարի քանակը հնարավորությունն է տալիս ավելի մեծ թվով լեզվական երևույթների հետ գործ ունենալ. դրանք համապարփակ և ամբողջական կերպով ներկայացնում են լեզվական իրողությունների հիմնական մասը: Ուստի ավելի կարևոր է ոչ թե տեքստերի քանակական շատությունը, այլ տարաբնույթ լինելն ու լեզվական հնարավորինս մեծ թվով տարբեր օրինակներ ցույց տալը: Հետևաբար, լեզվական կորպուսի առաջին և բնութագրական ամենահիմնական հատկանիշը *ներկայացուցչականությունն* է՝ մեծաթիվ օրինակների միջոցով լեզվական երևույթներ ու իրողությունները ցույց տալը[60]:

Այս առումով ԱՐԵՎԱԿ-ի տեքստային հենքը բավական հարուստ է. 2009 թ. մարտի դրությամբ ընդգրկում է մոտ 110 մլն. բառամթերք: Դրանից բացի՝ ի տարբերություն առավել լայն տարածում գտած այնպիսի կորպուսների, ինչպիսիք են *Ռուսերենի ազգային կորպուսը* կամ *Բրիտանական ազգային կորպուսը,* որոնցում տեքստերը ծավալի հետ կապված որոշակի սահմանափակումներից ելնելով են ընտրված, ԱՐԵՎԱԿ-ը պարունակում է հնարավորինս շատ մատչելի արևելահայերեն գեղարվեստական, գիտական և բանավոր տեքստեր: Այնուամենայնիվ, ժանրային հավասարակշռությունը պահպանելու նպատակով որոշ ժանրերի տեքստեր, ինչպիսիք են, օրինակ, մամուլի և օրենսդրական տեքստերը, որոնք հասանելի են համացանցում, ընդգրկված են սահմանափակ քանակով: ԱՐԵՎԱԿ-ի գրավոր խոսքի ենթակորպուսը պարունակում է 836 արձակ և չափածո գեղարվեստական տեքստեր (ներառյալ 206 թարգմանված տեքստեր), 7858 մամուլի համարներ, ինչպես նաև գիտական և պաշտոնական տեքստերի խոշոր հավաքածու[61]:

Ժամանակագրական առումով ԱՐԵՎԱԿ-ում տարբեր ժանրերի տեքստերը հավասարապես բաշխված չեն: 19-րդ դարի տեքստերը հիմնականում ընդգրկում են արձակ և չափածո գեղարվեստական

[60] Տե՛ս Հակոբյան Ֆ., Հայերենի ընդհանրական կորպուսի կազմության սկզբունքները, «Ջահուկյանական ընթերցումներ» միջազգային գիտաժողով, զեկուցումների ժողովածու, ՀՀ ԳԱԱ «Գիտություն» հրատարակչություն, Երևան, 2017, էջ 171:

[61] Տե՛ս http://www.eanc.net/am/composition/:

գրականության տեքստեր: Հայաստանի ազգային գրադարանի հետ համատեղ նախագծի շրջանակներում ավելացվել են մեծ թվով հին պարբերականների տեքստեր: Քանակական առումով մամուլի ենթակորպուսի հիմնական մասը բեռնվել է համացանցից, այդ իսկ պատճառով մեծ մասամբ ներկայացված է ժամանակակից մամուլի լեզուն: Մամուլի նման բաշխման արդյունքում գեղարվեստական գրականության և մամուլի ժանրերի հարաբերությունը վերջին տասնամյակների համար զգալի տարբերվում է կորպուսի մյուս մասերից: Ոչ գեղարվեստական տեքստերը ևս բաշխված են անհամաչափ: Գիտական տեքստերի հիմնական մասը պատկանում է խորհրդային ժամանակաշրջանին (հիմնականում 1960-ականներին և 70-ականներին), մինչդեռ որոշ օրենսդրական տեքստեր բեռնվել են համացանցից և ներկայացնում են վերջին տասնամյակների լեզուն: ԱՐԵՎԱԿ-ի կարևորագույն բաժիններից է արևելահայերեն բանավոր խոսքի ենթակորպուսը (3 մլն. բառանիշ), որն ամբողջովին մշակվել է ԱՐԵՎԱԿ-ի ջանքերով և ներկայացված է սպոնտան երկխոսություններով, պոլիլոգներով, նպատակաուղղված (task-oriented) հարցազրույցներով, հեռուստատեսային թոք-շոուներով, կինոնկարներով և այլ ձայնագրություններով: Կորպուսի վերջին տարբերակում ավելացված էլեկտրոնային հաղորդակցության տեքստերը, որոնք ընդգրկվել են բանավոր խոսքի ենթակորպուսում, իրականում միջանկյալ դիրք են գրավում բանավոր և գրավոր խոսքի միջև[62]: Այսպիսով՝ կարելի է ասել, որ ԱՐԵՎԱԿ-ը հիմնականում համապատասխանում է կորպուսների ամենաբնութագրական հատկանիշին՝ *ներկայացուցչականությանը*:

Չնայած հսկայական թվով տեքստային օրինակներ պարունակելուն՝ տեքստային հենքի բնութագրական հատկանիշներից է նաև դրա՝ *սահմանափակ ծավալ* ունենալը, սահմանափակ թվով լեզվական օրինակներ պարունակելն ու ներկայացնելը: ԱՐԵՎԱԿ-ը նույնպես սահմանափակ ծավալ ունի. ինչպես նշեցինք, 2009թ. մարտի դրությամբ ընդգրկում է մոտ 110 մլն. բառամթերք, իսկ ժամանակագրական առումով ներկայացնում է19-րդ դարի երկրորդ կեսից մինչև 2009թ. ընկած տեքստերը:

Ժամանակակից կորպուսների բնութագրական կարևոր հատկանիշներից մեկը *էլեկտրոնային եղանակով* հասանելի լինելն է:

[62] Տե՛ս http://www.eanc.net/am/composition/

Իհարկե, նախկինում՝ մինչև լեզվաբանության մեջ համակարգչային տեխնիկայի կիրառությունը և համապատասխան ծրագրերի հասանելի լինելը, տեքստային հենքերը տպագիր եղանակով էին օգտագործվում, և դրանցից որոշները մինչև հիմա էլ հենց այդպես են օգտագործվում: Ոչ մեծ թվով տեքստային հենքեր հասանելի են թվային տարբեր սարքավորումների միջոցով (խտասկավառակների և ձայնասկավառակների վրա), որոնք հիմնականում պարունակում են խոսակցական տեքստեր (Lancaster/IBM Spoken English Corpus): Նկատենք սակայն, որ օգտագործման դյուրության և հետագա հարստացման տեսանկյունից տպագիր և ձայնային հենքերի համեմատությամբ ավելի նախընտրելի են էլեկտրոնային եղանակով հասանելի տեքստային հենքերը: ԱՐԵՎԱԿ-ը ես հասանելի է էլեկտրոնային եղանակով. դրա ծրագրային ապահովումը մշակվել է *Կորպուս Թեքնոլոջիսի* կողմից, որի առաջնային նպատակներից է կորպուսի որոնման հնարավորությունները գործածողի համար բաց և մատչելի դարձնելը: ԱՐԵՎԱԿ-ի տվյալների բազայի ծրագրային ապահովումը կազմված է հետևյալ չորս հիմնական մասերից՝ քերականական վերլուծիչ, ինդեքսատոր, սպասարկիչ, գործածողի ինտերֆեյս և սպասարկու[63]: ԱՐԵՎԱԿ-ից կարելի է օգտվել՝ չգրանցվելով կամ համակարգչային որևէ ծրագիր ներբեռնելով:

Այսպիսով՝ ընդհանրացնելով կարող ենք ասել, որ կորպուսները պետք է ներկայացնեն լեզվական կառուցվածքի նկարագիրը, պարունակեն սահմանափակ թվով տեքստային օրինակներ, հասանելի լինեն էլեկտրոնային եղանակով: Արևելահայերենի ազգային կորպուսը, ինչպես տեսանք, համապատասխանում է լեզվական կորպուսների վերոնշյալ բնութագրումներին:

Կորպուսների ստեղծումը բավական բարդ և աշխատատար գործընթաց է, որը պահանջում է լեզվաբանների և ծրագրավորողների սերտ համագործակցություն: Սակայն դա ինքնանպատակ չէ, քանի որ կորպուսներն ունեն ինչպես տեսական-լեզվաբանական, այնպես էլ գործնական-կիրառական մեծ նշանակություն: ԱՐԵՎԱԿ-ը ևս, ինչպես մյուս կորպուսները, կարող է փաստական հարուստ նյութ ապահովել հետազոտողի համար: Այնուամենայնիվ, դրա կիրառական հնարավորություններն առավել ընդլանելու համար անհրաժեշտ է կորպուսում ձևային և բովանդակային զգալի փոփոխություններ

[63] Տե՛ս http://www.eanc.net/am/software/

կատարել՝ ներկայացվող տեքստերի ժամանակային ընդգրկումից սկսած՝ մինչև ծրագրային ապահովումը:

Նախ՝ ԱՐԵՎԱԿ-ը աստիճանաբար պատմական կորպուս է դառնում, քանզի դրանում ներկայացված են 19–րդ դարի երկրորդ կեսից մինչև 2009թ. եղած տեքստերը: Այսինքն՝ անցել է ավելի քան տասը տարի և կորպուսում որևէ նոր տեքստ չի ավելացվել, իսկ այդ տարիների ընթացքում թեպետ ոչ զգալի, այնուամենայնիվ լեզվում նկատելի փոփոխություններ կատարվել են: Ուստի անհրաժեշտ է շարունակել կորպուսը նորանոր տեքստերով հարստացնել՝ հասնելով նրան, որ սա վերածվի այսպես կոչված մոնիտորային կորպուսի. մեքենական եղանակով հարստացվի համացանցային զանազան աղբյուրներից:

Նպատակահարմար է առավել ընդլայնել ներկայացման ժամանակաշրջանը՝ այն դարձնելով *ընդհանրական* կորպուս, որում ներառված կլինեն հայերենի պատմական զարգացման տարբեր շրջաններում ստեղծված տեքտերը՝ ընդգրկելով գրաբարով, միջին հայերենով, հետո նոր միայն արևելահայերենով ու արևմտահայերենով ստեղծված տեքստեր: Հարկ է ուշադրություն դարձնել ոչ թե տեքտերի քանակական շատության, այլ բազմակողմանիության, լեզվական հնարավորինս բազմաբնույթ օրինակներ ներկայացնելու վրա: Ավելացնելով վերոնշյալ բաժինները՝ կարող ենք ամբողջ կորպուսը բաժանել ենթակորպուսների՝ ըստ հայերենի պատմական զարգացման տարբեր շրջաններն ընդգրկող տեքստերի: Ըստ այդմ՝ կունենանք գրաբարի, միջին հայերենի, արևելահայերենի ու արևմտահայերենի, ինչպես նաև խոսակցական լեզվի ենթակորպուսներ[64]:

Պետք է ոչ միայն նոր բաժիններ ավելացնել ԱՐԵՎԱԿ-ում, այլև եղածները կատարելագործնել: Օրինակ՝ անհրաժեշտ է զարգացնել կորպուսի պիտավորման համակարգը, որովհետև ներկայումս ԱՐԵՎԱԿ-ում իրականացվում է երեք տեսակի պիտակավորում. ա) մետատեքստային (մատենագիտական), որը կցվում է յուրաքանչյուր տեքստային միավորի, բ) բառային և ձևաբանական, որը կցվում է բառանիշերի ավելի քան 90%-ին, ինչպես նաև անգլերեն թարգմանություններ բառանիշների մոտ 85 %-ի համար, գ) կետադրության, նախադասության սահմանների, ինչպես նաև այլ օժանդակ հատկանիշների պիտակավորում[65]:

---

[64] Տե՛ս նշվ. աշխ. էջ 173

[65] Տե՛ս http://www.eanc.net/am/annotation/

Ընդհանրապես, լեզվի կառուցվածքային տարբեր միավորների համար, պայմանավորված լեզվական ուսումնասիրությունների նպատակով, կան պիտակավորման տարբեր տեսակներ, ինչպես՝ հնչունական, բառագիտական, քերականական (ձևաբանական և շարահյուսական): ԱՐԵՎԱԿ-ում, ինչպես տեսանք, կիրառվում են միայն ձևաբանական պիտակներ, որոնք, ըստ հեղինակների, արտահայտում են հայերենի խոսքի մասերի քերականական բոլոր կարգերը, սակայն երբեմն չեն համապատասխանում հայերենի նորմատիվ քերականությանը: Օրինակ՝ առանձին խոսքի մաս է համարվում նախդիրը, առանձնացվում է վեց հոլով. սեռականն ու տրականը համարվում են առանձին հոլովներ, իսկ հայցականը ուղղականից չի առանձնացվում: Ածականի համեմատության աստիճանները համարվում են քերականական կարգ, ընդ որում՝ առանձնացվում է միայն գերադրական աստիճանը, բաղդատականը՝ ոչ: Բայի եղանակներից ներկայացվում են միայն երեքը՝ ըղձական, ենթակայական, հրամայական: Ենթակայական եղանակն առանձնացնելիս հեղինակները հավանաբար նկատի են ունեցել ենթադրականը: Իբրև քերականական կարգ է առանձնացվել գոյականացումը՝ իր գոյականացված, անորոշ դերբայ, հարաբերական գոյական տեսակներով: Այստեղ էլ հավանաբար նկատի են ունեցել այլ խոսքի մասերի գոյականաբար գործածությունը[66]:

Անհրաժեշտ է նաև զարգացնել ԱՐԵՎԱԿ-ի ծրագրային ապահովման համակարգը: Օրինակ՝ ի տարբերություն մյուս կորպուսների, որոնցում եղած տեքստերի զգալի մասը կարելի է ներբեռնել տարբեր ձևաչափերով, ԱՐԵՎԱԿ-ն այպիսի հնարավորություն չի տալիս. սա զգալիորեն սահմանափակում է դրա օգտագործման հնարավորությունները: Պետք է ստեղծել նաև համակարգչային նոր ծրագրեր՝ այս կորպուսի տեքստերն ուսումնասիրելու համար, քանի որ գործնականորեն հնարավոր չէ պատկերացնել տեքստային հենքերի հետ կապված որևէ բնույթի լեզվական աշխատանք՝ առանց համակարգչային ծրագրերի կիրառության: Շնորհիվ դրանց՝ լեզվական այն աշխատանքները, որոնց կատարման համար նախկինում երկար ժամանակ էր պահանջվում, հիմա կարելի է իրականացնել հաշված վայրկյանների ընթացքում: Թեպետ ընդհանուր առմամբ՝ տարբեր գործառույթներ են կատարում, բայց և ունեն ընդհանուր հատկանիշներ

[66]http://www.eanc.net/EANC/search/frame_parts/gramsel.php?interface_language=am&search_language=armenian1&contexts_output_language=armenian1

(լեզվական տվյալների որոնման արագություն, ճշտությունը): Օրինակ՝ բոլորի միջոցով էլ կարելի է տեքստային հենքերից բառեր և արտահայտություններ որոնել, որոշել բառերի գործածության հաճախականությունը, որևէ համատեքստում գործածվելը:

ԱՐԵՎԱԿ-ի համար այսպիսի ծրագրեր դեռևս չեն ստեղծվել: Ավելին՝ այս կորպուսի մասին եղած ոչ ամբողջական պատկերացումները թույլ չեն տալիս օգտագործել այն հետազոտական աշխատանք կատարելիս: Հարկ է ընդգծել սակայն, որ, ինչպես տեսանք, ԱՐԵՎԱԿ-ն առավել զարգացնելու, ձևային և բովանդակային լուրջ փոփոխություններ կատարելու, այն հայերենի առավել ամբողջական ու համակողմանի ուսումնասիրությանը լիարժեք ծառայեցնելու խնդիր կա:

ԳՐԱԿԱՆՈՒԹՅԱՆ ՑԱՆԿ

1.Հակոբյան Ֆ., Հայերենի ընդհանրական կորպուսի կազմության սկզբունքները, «Ջահուկյանական ընթերցումներ» միջազգային գիտաժողով, զեկուցումների ժողովածու, ՀՀ ԳԱԱ «Գիտություն» հրատարակչություն, Երևան, 2017, էջ 170-178:
2.www.eanc.net/EANC/search/?interface_language=am
3.http://www.eanc.net/am/grant/
4.http://www.eanc.net/am/composition/
5.http://www.eanc.net/am/software/
6.http://www.eanc.net/am/annotation/
7.http://www.eanc.net/EANC/search/frame_parts/gramsel.php?interface_language=am&search_language=armenian1&contexts_output_language=armenian1

ФРИДА АКОПЯН- СТРУКТУРА И ВКЛЮЧЕНИЕ ВОСТОЧНОАРМЯНСКОГО НАЦИОНАЛЬНОГО КОРПУСА

Цель статьи представить Восточноармянский национальный корпус (ВАНК) с его наиболее отличительными особенностями и в текущей фазе развития корпусной лингвистики сравнить ее с другими корпусами разделяя прикладные сфере, где использование корпуса может быть весьма успешным. Чтобы сделать EANC более полным, мы делаем предложения и замечания который приведет корпус на качественно новый уровень.

FRIDA HAKOBYAN- THE STRUCTURE AND INCLUSION OF THE EASTERN ARMENIAN NATIONAL CORPUS

The purpose of the article is to present the Eastern Armenian National Corpus (EANC) with its most distinctive features and in the current phase of the development of corpus linguistics to compare it with other corpuses separating the applied areas, where the usage of the corpus can be highly successful. To make the EANC more complete we make suggestions and remarks which will lead the corpus to a qualitatively new level.