**DOI:** https://doi.org/10.52027/18294685-hga2023.sp

ՀՈՎՀԱՆՆԻՍՅԱՆ ԳԱՅԱՆԵ, ԱՅԴԻՆՅԱՆ ՍՐԲՈՒՀԻ
*Երևան, Հայաստան*

# ԹՎԱՅՆԱՑՎԱԾ ԳՐԱԿԱՆՈՒԹՅՈՒՆԻՑ ՄԻՆՉԵՎ ԼԵԶՎԱԿԱՆ ՇՏԵՄԱՐԱՆՆԵՐ. ԴՐԱՆՑ ՆՇԱՆԱԿՈՒԹՅՈՒՆԸ, ՀԵՌԱՆԿԱՐՆԵՐՆ ՈՒ ՄԱՐՏԱՀՐԱՎԵՐՆԵՐԸ

**Բանալի բառեր՝** լեզվական կորպուս, լեզվական շտեմարան, կորպուսային լեզվաբանություն, հաշվողական լեզվաբանություն, բնական լեզվի մշակում, տեքստերի հավասարեցում, ծանոթագրում, զուգահեռ կորպուսներ, թվային ժառանգություն:

**Keywords:** language corpus, linguistic database, corpus linguistics, computational linguistics, natural language processing, text alignment, annotation, parallel corpora, digital heritage.

Բնական լեզվի թվային մշակումն ունի մի քանի տասնամյակի պատմություն, և համաշխարհային մակարդակով արդեն կուտակվել է զգալի փորձ ու գիտելիք. կան հսկայական քանակությամբ գրավոր և բանավոր խոսքի թվայնացված նյութեր: Այդ նյութերը, որոնք հիմնականում հավաքագրված են որոշակի լեզվական և արտալեզվական նկարագրիչների համապատասխան, ընդունված է կոչել **լեզվական կորպուս կամ շտեմարան:**

21-րդ դարասկզբին կորպուսային լեզվաբանության հիմնադիր համարվող Ջոն Սինքլերը կորպուսը բնորոշել է որպես էլեկտրոնային տեքստերի հավաքածու, որը ընտրված է արտաքին որոշ չափանիշների համաձայն՝ հնարավորինս ներկայացնելու լեզուն կամ լեզվի տվյալ տարբերակը, և որը կծառայի որպես տվյալների աղբյուր լեզվաբանական հետազոտությունների համար [Sinclair, 2005, 1-16]: Այս իմաստով, լեզվական շտոմարանները մասնակիորեն իրականացնում են նույն գրադարանների գրավոր տեքստեր հավաքագրելու և դասակարգելու գործառույթը, գումարած այդ տեքստերի մեջ պարունակված գիտելիքի և հենց լեզվի տարբեր հատկանիշների հետ աշխատելու հնարավորությունը, որը տալիս է դրանց թվայնացված ներկայացումը: Ի դեպ, բնական լեզվի մշակումը հենց իրականացվում է այդ լեզվի տեքստերում պարունակված գիտելիքի տարամակարդակ շերտերը վեր հանելու և տարբեր նպատակներով գործածելու համար:

Այսպիսով, կորպուսը լեզվական տարամակարդակ միավորների՝ տվյալների որոնողական բազա է, որը օգտագործվում է լեզվաբանական հետազոտությունների, լեզվի ուսուցման, մեքենայական թարգմանության բարելավման, և ընդհանրապես՝ մեքենային տարբեր նպատակներով բնական լեզուն մշակել ուսուցանելու համար: Կորպուսները կարող են լինել ինչպես ծանոթագրված (ինչն իրենից ենթադրում է հատույթավորում, խոսքիմասային պիտակավորում, շարահյուսական վերլուծություն և այլն), այնպես էլ չծանոթագրված՝ «հում» տեքստերի հավաքածու: Դրանք լինում են մեկ լեզվով, կամ երկլեզու ու բազմալեզու:

«Կորպուսը ուշագրավ բան է ո՛չ այնքան այն պատճառով, որ բնական տեքստերի հավաքածու է, այլ այն հատկությունների շնորհիվ, որոնք նա ձեռք է բերում լավ նախագծված և խնամքով կառուցված լինելու դեպքում:»,- հետաքրքիր բնորոշում է տալիս Սինքլերը՝ մատնանշելով կորպուսների գիտելիքի և տեղեկատվության իմաստով կիրառական բարձր արժեքը գրագետ հավաքագրված լինելու պարագայում [Նույն տեղում, 5]:

Շատ նշանավոր գիտնականներ, ինչպիսիք են Մ. Հալիդեյը, Ջ. Լիչը, Դ. Բայբերը, Ս. Յոհանսոնը, Գ. Ֆրենսիսը, Ս. Հանսթոնը, Ս. Կոնրադը, Մ. Մաքքարթին և Ա. Ստեֆանովիչը նպաստել են ժամանակակից կորպուսային լեզվաբանության զարգացմանը: Գրվել են բազմաթիվ աշխատություններ, ու ոլորտը գնալով ավելի հավակնոտ է դարձել ու հիմա քայլում է ժամանակակից տեխնոլոգիաներին համընթաց՝ օգտվելով դրանց ընձեռած աճող հնարավորություններից:

Դեռ ժամանակին Սինքլերը մատնանշում էր, որ անհրաժեշտ է լեզվի ուսումնասիրման նոր դիտանկյուն՝ կապված մասնավորապես համակարգչային տեխնոլոգիաների առաջընթացով ու հաշվողական լեզվաբանության զարգացումներով [Sinclair, Carter, 2004, 10-23]:

Լեզվական շտեմարանների ստեղծման պատմության համառոտ ակնարկը կօգնի ընթերցողին ինքնուրույն զուգահեռներ անցկացնել թվային տեխնոլոգիաների և բնական լեզվի ամբարած գիտելիքի փոխակերպման միտումների միջև: Այդ զուգահեռները նաև կօգնեն հասկանալ լեզվական շտեմարանների դերը արհեստական բանականության, կամ, ինչպես մենք ենք գերադասում կոչել՝ արհեստական մտքի ստեղծման գործում: Չէ՞ որ բնական լեզվի տեքստերում ամբարված է լեզվակիր հանրության հավաքական գիտակցությունը՝ իր ժամանակային և տարածական բազմազանության մեջ: Ավելին, մեր վերջին հետազոտությունները նվիրված են լեզվակիր անձի, որպես գիտելիքի և գիտակցության, կոնկրետ ունակություններ կրողի: Անձն իբրև լեզվաստեղծ միավոր դիտարկելը մեծ հեռանկարներ է բացում լեզվի մեքենայական մշակման բնագավառում [Hovhannisyan, 2022, 79-95]:

Առաջին լեզվական կորպուսը ԱՄՆ Բրաունի համալսարանում ստեղծված BROWN կորպուսն է՝ ստեղծված 1961 թվականին հրատարակված գրական արձակ տեքստերից: [Stefanowitsch, 2020, 30-31]

Ավելի ուշ ի հայտ եկան դրանց ձևավորման սկզբունքները, ապա սկսեցին ներառել գրավորից բացի նաև բանավոր տեքստային նյութեր: Այսօր ամերիկյան և բրիտանական անգլերենի ամենահայտնի կորպուսներից են COCA-ն, COHA-ն, BNC-ն և MICASE-ը: Ժամանակակից ամերիկյան անգլերենի կորպուսը (COCA) ամերիկյան անգլերենի ամենամեծ ազատ հասանելի կորպուսն է՝ բաղկացած ավելի քան 1 միլիարդ բառից: COCA-ն թողարկվել է 2008 թվականին, տեքստերը թվագրվում են 1990-2019 թվականներին [COCA կորպուսի կայքէջ]: Պատմական ամերիկյան անգլերենի կորպուսը (COHA) տարժամանակյա անգլերենի ամենամեծ հավաքածուն է, որի տեքստերը թվագրվում են 1810-ից 2009 թվականը, ունի ավելի քան 400 միլիոն բառ: BYU/BNC-ն ժամանակակից բրիտանական անգլերենի 100 միլիոնանոց կորպուս է, որը նույնպես պարունակում է գրավոր և բանավոր տեքստեր: Այն ստեղծվել է Oxford University Press-ի կողմից 1980-1990-ականներին [BNC կորպուսի կայքէջ]: Միչիգանի ակադեմիական խոսակցական անգլերենի կորպուսը (MICASE)՝ ավելի քան 1,8 միլիոն տառադարձված տեքստերի հավաքածու է, որ ստեղծել են Միչիգանի համալսարանի հետազոտողներն ու ուսանողները [MICASE կորպուսի կայքէջ]:

Որպես հանրային գիտակից գործունեության փաստագրված արդյունքներ, և որպեսզի կորպուսը համարվի հավաստի գործիք, կիրառելի լինի, պետք է այն հավաքագրել հետևյալ մի քանի սկզբունքով. **ներկայացուցչականություն**, **հավասարակշռություն, ծանոթագրում, որակ** [Atkins, Clear & Ostler, 1992, 1-16]: Բացի այդ, լեզվական մտածողության նման, կորպուսը պետք է շարունակ թարմացվի, հակառակ դեպքում որոշ ժամանակ անց չի արտացոլի լեզվի արդի վիճակը, կհնանա: Նման փորձի օրինակը կա, և կարելի է դրանից սովորել, հետագայում փակուղային վիճակներից խուսափելու համար:

Լեզվական շտեմարանների կամ կորպուսների ստեղծման, պահպանման և զարգացման պայմանների ապահովումը մեծ ռեսուրսներ և կազմակերպչական աշխատանք է պահանջում: Արդեն առաջին իսկ հետազոտություններից պարզ է, որ հատկապես մեծ է պահանջարկը հայերենի բազմալեզու կամ առնվազն երկլեզու զուգահեռ ռեսուրսներ ունենալու, որը հայերենի նման փոքրաթիվ խոսողներ ունեցող լեզվի համար գլոբալ թվային քաղաքակրթությանը ինտեգրվելու ամենաարագ ուղիներից է: Այս պահանջները բավարարելու համար բազմալեզու զուգահեռ կորպուսների կառուցումը դառնում է բնական լեզվի մշակմամբ զբաղվող (NLP) մեր առաջին գիտական խմբերի ուշադրության առարկան: Ի տարբերություն միալեզու կորպուսների, աշխարհում հասանելի երկլեզու կամ բազմալեզու զուգահեռ կորպուսների թիվը սահմանափակ է, և, քանի որ այդ կարգի կորպուսի հավաքագրումը շատ ռեսուրսներ է պահանջում, ոչ բոլորն են ազատ հասանելի: Բացի այդ, այսպիսի ռեսուրսներ հավաքագրվում են հիմնականում այն լեզուների համար, որոնք հայտնի են խոսողների մեծ քանակով ու հատկացված ռեսուրսների մասշտաբով [Steingrímsson, Loftsson, & Way, 2020, 182-190]:

ԲԼՄ ռեսուրսների առկայության առումով, ինչպես նշվեց, հայերենը ետ է մնում: Իհարկե, կան մի քանի նախաձեռնություններ, որոնք աշխատել են հայերենի թվայնացման տարբեր խնդիրների վրա: Հայերենի առաջին թվայնացված կորպուսի՝ ԱՐԵՎԱԿ-ի (Արևելահայերենի ազգային կորպուս) նախագիծը գործարկվել է 2006 թվականին, սակայն այն 2009 թվականից հետո չի

թարմացվել: 2009 թ. մարտի դրությամբ ԱՐԵՎԱԿ-ն ընդգրկում է մոտ 110 մլն. բառամթերք [ԱՐԵՎԱԿ կորպուսի կայքէջ]: Մյուս նախաձեռնությունը «Հայերենի ծառադարան»-ն է: «… նախագիծը մաս է կազմում Universal Dependencies (UD) – «Համընդհանուր կախվածություններ» հարթակի, որը, տիպաբանական և կիրառական հետազոտական նպատակներով, քերականությունների ձևայնացման և տեքստերի լեզվաբանական ծանոթագրման միասնական մեթոդ է առաջարկում բնական լեզուների մեքենական մշակման համար» [Յավրումյան Մ., Դանիելյան Ա., 2020]: Ծառադարանը պարունակում է շուրջ 100 հազար բառամթերք և 5 հազար շարահյուսական ծառ [Հայերենի ծառադարանի կայքէջ]:

Մյուս մեզ հայտնի հետազոտական խումբը՝ «Ազգային արագացման հանրային նախաձեռնությունը», «իրականացնում է հայերեն լեզվով մեքենայական սարքերի հետ հաղորդակցության գործիքների՝ արհեստական ձայնի սինթեզի TTS (Text To Speech), խոսքի ճանաչման ASR (Automatic Speech Recognition), մեքենայական թարգմանության NMT (Neural Machine Translation) մշակումը» [Ազգային արագացման հանրային նախաձեռնությյան կայքէջ]: Այս խումբը ևս առաջիններից է, որ համագործակցության հուշագիր է ստորագրել ԲՊՀ հետ՝ նախադեպ ստեղծելով ՀՀ համալսարանական միջավայրում բնական լեզվի կայուն-շարունակական մշակման աշխատանքներ ապահովելու տեսանկյունից: Հարկ է, որ այս նախաձեռնությունը պետական հովանավորության տակ առնվի, որպեսզի գիտական, կրթական և արտադրական ջանքերի արդեն կայացած միավորումը երկրի կենսական խնդրի լուծմանն անհրաժեշտ մասշտաբներն ու արդյունքներն ապահովի:

Այսօր դժվար է գտնել մի հեղինակավոր համալսարան, որն իր լեզվական թեկուզ փոքր կորպուսները, ասել է, թե սեփական գիտելիքի և հետազոտական ռեսուրսների զարգացող բազան չունենա: Վերևում թվարկված բոլոր խոշոր կորպուսները ևս ստեղծվել են ու զարգացվում են համալսարաններում: Հայերենում, հատկապես, նմանատիպ հետազոտությունների պահանջարկը շատ մեծ է, իսկ մասնավոր ձեռնարկությունների ուժերով արվող ուսումնասիրությունները միայն փոքր առևտրային նպատակներ են հետապնդում: Մշակութային ժառանգության, հումանիտար և սոցիալական գիտությունների բնագավառի հետազոտողի համար ունենալ լեզվի տարբեր կորպուսներ ու ենթակորպուսներ նշանակում է հնարավորություն ստեղծել մեքենայական միջլեզվական թարգմանությունների, տեքստերից տարաբնույթ ռազմավարական, արժեքաբանական, տեղեկատվական և այլ գիտելիքի զտման, խոսքային ներգործման աշխատանքների բարելավման, լեզուների ուսուցման (կորպուսների տարաբնույթ կիրառություններ կարելի է թվել թե սովորողի, թե սովորեցնողի համար [Cobb, Boulton 2015, 478-497]), լեզվախոսքային գիտակցության տեքստերի բազմաշերտ իմաստագործառական քննություն կատարելու և ընդհանրապես, լեզվաբանական հետազոտությունների արդիականացման համար, լեզվամշակութային ժառանգության և դրա ընթացիկ թվային նկարագրերի ստեղծման, պահպանման, և գլոբալ ինտեգրման համար [Biber, Douglas 2006, 243-257]:

Բնական լեզվի մշակումը անգլերենի նման հարուստ լեզվում արդեն լուծել է որոշակի համակարգային խնդիրներ, որոնց հարմարեցումը հայերենին զուգահեռման միջոցով կարող է արագացնել սկզբնական աշխատանքները: Սակայն երկլեզու տեքստերի հավասարեցումը բառերի, նախադասությունների մակարդակով, ծանոթագրումը և այլ անհրաժեշտ գործընթացներ այնուամենայնիվ, աշխատատար են և էական նշանակություն ունեն զուգահեռ կորպուսների ձևավորման գործում [Devlin et al, 2019]: Զուգահեռ կորպուսների պարագայում, տեքստերի ծանոթագրում կարող է և չարվել, սակայն առնվազն մեկ հարթության վրա հավասարեցում պետք է արվի: Թարգմանաբանական հետազոտությունների, լեզվի ուսուցման, ինչպես նաև ավելի մեծ ծավալների մեքենայական թարգմանության բարելավման նպատակով, հավասարեցումը կարող է լինել միակողմանի, երկկողմանի կամ երկուսի համադրությունը:

Վիճակագրական մեքենայական թարգմանությունը մեկն է այն կիրառություններից, որտեղ կորպուսները զգալի արդյունք են տվել [Eisele, 2006]: Դրա հաջողության հիմնական գործոնը բարձրորակ տվյալներն են: Իսկ բարձրորակ թարգմանական նյութի մշակումն ու կայուն ապահովումը հնարավոր է աշխատանքի վրա հենված կրթական ծրագրեր իրականացնող համալսարաններում: Նշված փորձառական աշխատանքներ մենք արդեն իրականացնում ենք Բրյուսովի

համալսարանում մեր արտաքին գործընկերների աջակցությամբ: Իհարկե, կան տեքստերի հատույթավորման և հավասարեցման ծրագրեր, որոնք կարագացնեին բնօրինակի և թարգմանության հավասարեցումը, բայց դրանք հայերենի համար չեն: Մեր գործիքները ինքներս պիտի ստեղծենք: Մեկտեղելու համար առկա ռեսուրսները, մենք օգտագործում ենք արդեն առկա երկլեզու տեքստեր՝ գիտակրթական, հետազոտական կենտրոններում, գրադարաններում արդեն թվայնացված ռեսուրսներից:

Այս առումով, Հայաստանի Ազգային գրադարանի թվայնացված ռեսուրսների ներդրումը հայերենի երկլեզու շտեմարանների ստեղծման գործին անգնահատելի կլինի, միաժամանակ կբարձրանա գրադարանային ռեսուրսների կիրառական արժեքը:


## SUMMARY

The Republic of Armenia has traditionally had a solid culture of preserving written heritage. However, in the last few decades, due to various geopolitical circumstances, we have lagged behind the global advancement of digitization, information, and communication technologies in terms of natural language processing for the Armenian language, access to the expanding opportunities of systematising and utilising the collated enormous amount of literature. This paper presents the current global state of natural language processing and the related methodological problems of the field in our country. The process of dataset development and consumption, from the primary stage of text digitising collected in physical databases, libraries, and archives, to the very use of the knowledge and information stored in these texts for various scientific, educational and practical purposes, requires a clear scientific-methodological program introducing not only the importance and perspectives of the field but will also ensure the conditions for its sustainable development, including the processing of written (and spoken) digitized language data and the retrieval and “industrialization” of the information and knowledge contained in the language corpora.


## ՕԳՏԱԳՈՐԾՎԱԾ ԳՐԱԿԱՆՈՒԹՅԱՆ ՑԱՆԿ


1. Atkins, S., Clear, J. and Ostler, N., Corpus design criteria // Literary and Linguistic Computing 7/1, Oxford University Press, Oxford, 1992, pp. 1-16.
2. Biber, D., Representativeness in corpus design // Corpus linguistics: Readings in a widening perspective, London, 2004, pp. 174-97.
3. Biber, D., University Language: A corpus-based study of spoken and written registers, John Benjamins Publishing, Amsterdam, 2006, 261 p.
4. Cobb T., Boulton A., Classroom applications of corpus analysis // Cambridge Handbook of Corpus Linguistics, Cambridge University Press, Cambridge, 2015, pp. 478-497.
5. Devlin, J., Chang, M., Lee, K., & Toutanova, K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Association for Computational Linguistics, Minneapolis, 2019, pp. 4171-4186. [Electronic publication] URL: https://aclanthology.org/N19-1423, 29.11.2022.
6. Eisele A., Parallel Corpora and Phrase-Based Statistical Machine Translation for New Language Pairs via Multiple Intermediaries // In Proceedings of the Fifth International Conference on Language Resources and Evaluation, European ELRA, Genoa, 2006, pp. 845-848 [Electronic publication] URL : http://www.lrec-conf.org/proceedings/lrec2006/pdf/643_pdf.pdf, 29.11.2022.
7. Frankenberg-Garcia, A., Compiling and using a parallel corpus for research in translation // Babel: international journal of translation, Vol. XXI (1), John Benjamin Publishing, Amsterdam, 2009, pp. 57-71.
8. Hovhannisyan, G., The Architecture of Language Personality // Individual and Contextual Factors in the English Language Classroom, Springer, Cham, 2022, pp. 79-95.
9. Sinclair, J. & Carter R., Trust the Text // Language, Corpus and Discourse, Routledge, London, 2004, 224 pages.
10. Sinclair, J., Corpus and Text: Basic Principles // Developing Linguistic Corpora: A Guide to Good Practice, Oxbow Books, Oxford, 2005, pp. 1-16.
11. Sinclair, J., Corpus, Concordance, Collocation, Oxford University Press, Oxford, 1991, 179 pages.

**12.** Stefanowitsch, A., Corpus linguistics: A guide to the methodology // Textbooks in Language Sciences 7, Language Science Press, Berlin, 2020, 481 pages.

**13.** Steingrímsson, S., Loftsson, H., & Way, A., Effectively Aligning and Filtering Parallel Corpora under Sparse Data Conditions // ACL, 2020, pp. 182-190. [Electronic publication] URL: https://aclanthology.org/2020.acl-srw.25, 29.11.2022.

**14.** Յավրումյան Մ., Դանիելյան Ա., Համընդհանուր կախվածություններ և հայերենի ծառադարա-նը // Լրաբեր հասարակական գիտությունների, ՀՀ ԳԱԱ «Գիտություն» հրատարակչություն, Երևան, 2020, էջ 231-244:

**15.** Հայերեն ծառադարանի կայքէջ [էլեկտրոնային ռեսուրս] URL : http://armtreebank.yerevann.com/, [29.11.2022].

**16.** Արևակ կորպուսի կայքէջ [էլեկտրոնային ռեսուրս] URL: http://www.eanc.net/am/composition/, [29.11.2022].

**17.** Ազգային արագազման նախաձեռնության կայքէջ [էլեկտրոնային ռեսուրս] URL: https://hayq.ican24.net/hyenmt/index.php, [29.11.2022].

**18.** Միչիգանի ակադեմիական խոսակցական անգլերենի կորպուսի կայքէջ [էլեկտրոնային ռե-սուրս] URL : https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase, [29.11.2022].

**19.** BNC Կորպուսի կայքէջ [էլեկտրոնային ռեսուրս] URL: https://www.english-corpora.org/bnc/, [29.11.2022].

**20.** COCA կորպուսի կայքէջ [էլեկտրոնային ռեսուրս] URL: https://www.english-corpora.org/coca/, [29.11.2022].

**21.** COHA կորպուսի կայքէջ [էլեկտրոնային ռեսուրս] URL: https://www.english-corpora.org/coha/, [29.11.2022].