Վ. ԲՐՅՈՒՍՈՎԻ ԱՆՎԱՆ ՊԵՏԱԿԱՆ
ՀԱՄԱԼՍԱՐԱՆ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ В. БРЮСОВА
BRUSOV STATE UNIVERSITY

ԲԱՆԲԵՐ
Վ. ԲՐՅՈՒՍՈՎԻ ԱՆՎԱՆ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆԻ
ВЕСТНИК ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА ИМЕНИ В. БРЮСОВА
BULLETIN OF BRUSOV STATE UNIVERSITY

ԼԵԶՎԱԲԱՆՈՒԹՅՈՒՆ ԵՎ ԲԱՆԱՍԻՐՈՒԹՅՈՒՆ

ЛИНГВИСТИКА И ФИЛОЛОГИЯ

LINGUISTICS AND PHILOLOGY

1 (58)

Վ. ԲՐՅՈՒՍՈՎԻ ԱՆՎԱՆ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆԻ
«ԼԻՆԳՎԱ» ՀՐԱՏԱՐԱԿՉՈՒԹՅՈՒՆ

ԵՐԵՎԱՆ - 2021

ՀՏԴ 81՝374:004.656
DOI: 10.51307/18293107/laph/21.58-203

ԷԼԵԿՏՐՈՆԱՅԻՆ ՍՐԲԱԳՐԻՉԻ ՏՎՅԱԼՆԵՐԻ ՇՏԵՄԱՐԱՆԻ ԲԱՌԱՊԱՇԱՐԻ ՁԵՎԱՅԻՆ ՆԿԱՐԱԳՐՈՒԹՅԱՆ ԽՆԴԻՐՆԵՐԻ ՄԱՍԻՆ

ՄԵՐԻ ՍԱՐԳՍՅԱՆ

Հիմնաբառեր՝ *Էլեկտրոնային սրբագրիչ, սրբագրման համակարգ, համակարգչային լեզվաբանություն, լեզվի ձևային նկարագրություն, տվյալների շտեմարան, ծրագրաշար, ձևային նկարագրության սկզբունքներ, հայերենի ձևային նկարագրություն*

Սրբագրման ցանկացած համակարգի հաջող գործարկման հիմքում ոչ միայն հմտորեն մշակված ծրագրաշարն է, այլև տվյալների հարուստ շտեմարանը: Բացի այդ՝ սրբագրիչի տվյալների շտեմարանում ընդգրկվելիք բառապաշարը պետք է ձևային նկարագրության ենթարկել: Լեզվի ձևայնացման գործընթացը բարդ և պատասխանատու աշխատանք է, ի հայտ են գալիս բազմաթիվ խնդիրներ՝ կապված ոչ միայն ձևայնացման սկզբունքների մշակման ճշգրտության հետ, այլև լեզվի կառուցվածքային տիպի, ձևաբանական փոփոխությունների ենթարկվող բառերի հոլովման և խոնարհման հարացույցների առանձնահատկությունների հետ և այլն: Լեզվի ձևային նկարագրության սկզբունքները մշակելիս պետք է հաշվի առնել տվյալ լեզվի կառուցվածքային առանձնահատկությունները: Ձևայնացման լիարժեքությունն ապահովելու համար պակաս կարևոր չէ նաև այն նպատակի հստակ սահմանումը, որին ծառայելու է ձևայնացումը: Հաշվի առնելով ժամանակակից հայերենի ձևայնացման առանձնահատկությունները և դրանցով պայմանավորված դժվարությունները՝ սույն հոդվածի շրջանակներում քննության ենք ենթարկում այս բազմազանությունը ապահովող ձևային նկարագրության խնդրահարույց հարցերը, ներկայացնում դրանց լուծման ուղիները, սկզբունքները, որոնք ընկած են մեր մշակած էլեկտրոնային սրբագրիչի տվյալների շտեմարանի ստեղծման հիմքում:

Թվային տեխնոլոգիաների դարաշրջանում էլեկտրոնային տեքստը դարձել է մեր կյանքի անբաժանելի մասը: Էլեկտրոնային տեքստի առկայությունն էլ իր հերթին նպաստեց ոչ միայն էլեկտրոնային տարատեսակ ընթերցիչների ի հայտ գալուն, այլև ասպարեզ եկան էլեկտրոնային տեքստը կազմելու, խմբագրելու, սրբագրելու տարատեսակ գործիքներ: Վերջին շրջանում մեծ տարածում են գտել էլեկտրոնային սրբագրիչները: Հայերեն էլեկտրոնային առկա և գործող սրբագրիչներին,

դրանց թերացումներին, առավելություններին և հեռանկարներին անդրադարձել ենք մեր նախորդ հոդվածներում (Սարգսյան 2018: 598-604, Sargsyan 2018: 21-24): Սույն հոդվածում կներկայացնենք հայալեզու տեքստերի էլեկտրոնային նոր սրբագրիչի տվյալների շտեմարանի բառապաշարի ձևային նկարագրության խնդիրները:

«Հայերեն էլեկտրոնային սրբագրման համակարգ» գիտահետազոտական թեմայի[1] շրջանակներում ստեղծել ենք սրբագրման մի համակարգ, որի միջոցով փորձ է արվում ամբողջական և լիարժեք կերպով սրբագրել արդի արևելահայերենով գրված ցանկացած տեքստ, ինչպես նաև օգտագործելով վեբ տեխնոլոգիաների բոլոր հնարավորությունները, հայերեն էլեկտրոնային սրբագրման համակարգը հասանելի դարձնել հանրության լայն շրջանակներին՝ ապահովելով համակարգից հեշտությամբ օգտվելու հնարավորություն:

Սրբագրման ցանկացած համակարգի հաջող գործարկման հիմքում ոչ միայն հմտորեն մշակված ծրագրաշարն է, այլև տվյալների հարուստ շտեմարանը: Բացի այդ՝ սրբագրիչի տվյալների շտեմարանում ընդգրկվելիք բառապաշարը պետք է ձևային նկարագրության ենթարկել: Լեզվի ձևայնացման գործընթացը բարդ և պատասխանատու աշխատանք է, ի հայտ են գալիս բազմաթիվ խնդիրներ՝ կապված ոչ միայն ձևայնացման սկզբունքների մշակման ճշգրտության հետ, այլև լեզվի կառուցվածքային տիպի, ձևաբանական փոփոխությունների ենթարկվող բառերի հոլովման և խոնարհման հարացույցների առանձնահատկությունների հետ և այլն: Լեզվի ձևային նկարագրության սկզբունքները մշակելիս պետք է հաշվի առնել տվյալ լեզվի կառուցվածքային առանձնահատկությունները: Ձևայնացման լիարժեքությունն ապահովելու համար պակաս կարևոր չէ նաև այն նպատակի հստակ սահմանումը, որին ծառայելու է ձևայնացումը:

Ժամանակակից հայերենի ձևային նկարագրությունը հանգամանալից աշխատանք է պահանջում: Խնդրահարույց հարցեր են առաջացնում մեր լեզվի հատկապես հետևյալ առանձնահատկությունները.

1. Ժամանակակից հայերենին բնորոշ են քերականական կարգերի առատությունը, ինչպես նաև դրանց արտահայտման միջոցների բազմազանությունը: Ինչպես գիտենք, հայերենում միևնույն քերականական իմաստը կարող է արտահայտվել մի քանի տարբերակային ձևերով:

---

1 Հետազոտությունն իրականացվել է ՀՀ ԿԳՆ գիտության կոմիտեի տրամադրած ֆինանսավորմամբ՝ 19YR-6B070 *ծածկագրով գիտական թեմայի շրջանակներում:*

2. Հոլովման և խոնարհման հարացույցներում, նորմատիվ ձևերից բացի, առկա են անկանոնություններ, որոնք շեղվում են հոլովման և խոնարհման ընդհանուր օրինաչափություններից: Դրանց համար անհրաժեշտ է հստակ սահմանումներ մշակել:
3. Ժամանակակից հայերենը գերազանցապես կցական լեզու է, սակայն և՛ խոնարհման, և՛ հոլովման հարացույցներում առկա են անջատականության և թեքականության դրսևորումներ, որը ձևայնացման գործընթացը որոշակիորեն դժվարացնում է:

Դժվար չէ նկատել, որ վերոնշյալ հանգամանքները մեծացնում են ձևային նկարագրության ծավալը՝ ստեղծելով տարբերակային ձևերի առատություն:

Բառապաշարի ձևային նկարագրությունը հնարավորություն է տալիս բացահայտելու ինչպես կանոնավոր կազմությունների, այնպես էլ բոլոր տարբերակային ձևերի, շեղումների ու անկանոնությունների ամբողջական պատկերը՝ հաճախականության տվյալներով և դրանց ավտոմատ վերլուծության ու սերման հնարավորություններով: Այսինքն՝ էլեկտրոնային շտեմարանը ներկայացնում է արդի հայերենի բառապաշարի բառամիավորների ձևաբանական բնութագիրը:

Հայերենի բառապաշարը ձևայնացնելու և համապատասխան էլեկտրոնային շտեմարան ստեղծելու համար լուծվել են հետևյալ **խնդիրները.**

1. ուսումնասիրելով արդի լեզվաբանության մեջ ընդունված լեզվի ձևային նկարագրության սկզբունքներն ու եղանակները՝ մշակվել են հայերենի բառապաշարի ձևային նկարագրության ընդհանրական սկզբունքներ,
2. որոշվել է բառապաշարի մոտավոր ծավալը, որի ձևային նկարագրությունը նախատեսվում է իրականացնել (այժմ շտեմարանում մուտքագրված է շուրջ 200 000 գլխաբառ, դրանցից յուրաքանչյուր գոյականի համար 10 թեքված ձև, դերանվան համար՝ 12, բայի համար՝ 160, ըստ անհրաժեշտության նշված են նաև զուգաձևությունները),
3. բառարաններից առանձնացվել և մշակվել են համապատասխան բառամիավորները (շուրջ 200 000 գլխաբառ),
4. կատարվել է ընտրված բառամիավորների ձևային նկարագրություն ձևաբանական տեսակետից,
5. կատարվել է վերոնշյալ բառերի ձևաբանական վերլուծություն,
6. առանձնացնելով լեզվական նյութը՝ մշակվել են համապատասխան էլեկտրոնային շտեմարանի կազմության սկզբունքները,
7. ստեղծվել է էլեկտրոնային շտեմարանի կառուցվածքը/կաղապարը,

8. օգտագործելով վեբ տեխնոլոգիաների հնարավորությունները՝ ստեղծվել է համացանցային ինքնուրույն կայքում գործող էլեկտրոնային շտեմարան՝ armspell.am վեբ հասցեի ներքո, որը, սրբագրման խնդիրը լուծելուց բացի, ունի մեծածավալ տվյալների հենք (Սարգսյան 2021):

Համակարգն ունի կառավարման վահանակ (այսուհետ՝ ԿՎ), որը ծառայում է համակարգի շտեմարանում նոր բառեր ավելացնելու, ինչպես նաև վրիպակները, հնարավոր սխալները շտկելու համար: ԿՎ-ը բաղկացած է մի քանի էջերից: Առաջին էջը նախատեսված է բառերի մուտքագրման համար, որի համար օգտագործվում է OCR (Optical character recognition) տեխնոլոգիա, որը հնարավորություն է տալիս շտեմարանում բառեր մուտքագրել և՛ մեծ խմբաքանակով, և՛ առանձին-առանձին: Երկրորդ էջը նախատեսված է արդեն մուտքագրված բառերն ըստ խոսքի մասերի դասակարգելու համար: Այս էջում առկա է նաև գլխաբառերի հնարավոր թեքված տարբերակները թե՛ մուտքագրելու, թե՛ խմբագրելու, թե՛ հեռացնելու հնարավորություն:

Կայքի օգտատերերը սրբագրման համակարգը կիրառելիս որոնման դաշտում կարող են մուտքագրել ստուգվող տեքստը (մինչև 300 բառ) և ստուգել այդ տեքստի բառերի ուղղագրական և քերականական սխալները: Սխալ գտնելու դեպքում համակարգը սխալն ուղղելու 6-10 տարբերակ է առաջարկում՝ հիմնվելով բառի ձևային նմանության վրա, ընդ որում՝ առաջարկում է նաև տվյալ բառի թեքված տարբերակներից:

Համակարգը ուղղագրական և քերականական սխալների ստուգումը կատարում է շտեմարանում առկա բառերի հետ համեմատելու միջոցով: Եթե ստուգվող բառը բացակայում է շտեմարանից, ապա այն համակարգի կողմից համարվում է ուղղագրական սխալ ունեցող բառ: Սխալ ունեցող բառը գտնելուց հետո համակարգը համեմատում է այն շտեմարանում առկա բառերի հետ և գտնում ձևային առումով ամենամոտ բառերը՝ ներկայացնելով այն օգտատիրոջը՝ որպես հնարավոր ճիշտ տարբերակ: Բառերի համեմատությունը կատարվում է տեքստերի նմանության Լևենշտեյնյան հեռավորության չափման միավորի հիման վրա (Levenshtein distance):

Հայերենի էլեկտրոնային սրբագրիչի տվյալների շտեմարանը կազմված է հետևյալ աղյուսակներից՝ ***բառույթ, խոսքի մաս, հոգնակի թվի տիպ, հոգնակի թիվ, թեքման տիպ, հոգնակի թվի թեքման տիպ***:

Բառույթ -բառի ուղիղ ձևն է, մուտքագրվում է շտեմարան թե՛ առանձին, թե՛ խմբաքանակով, մուտքագրելիս կրկնությունները ավտոմատ կերպով հեռացվում են,

Խոսքի մաս- մեկից ավելի ընտրելու հնարավորությամբ ընտրովի դաշտ է՝ գոյական, ածական, թվական, դերանուն, բայ, մակբայ, կապ, շաղկապ, վերաբերականներ, ձայնարկություններ- կարող է մնալ չլրացված,

Հոգնակի թվի տիպ- մուտքագրվում է համապատասխան տիպի պիտակը, որի հիման վրա ավտոմատ կազմվում և լրացվում է **Հոգնակի թիվ** դաշտը,

Հոգնակի թիվ – կարող է չլրացվել,

Թեքման տիպ – մուտքագրվում է համապատասխան տիպի պիտակը, որի հիման վրա ավտոմատ կազմվում և լրացվում են թեքված ձևերը՝ ամեն մեկը իր տիպին բնորոշ դաշտերի քանակով,

Հոգնակի թվի թեքման տիպ – մուտքագրվում է համապատասխան տիպի պիտակը, որի հիման վրա ավտոմատ կազմվում և լրացվում են հոգնակի թվով թեքված ձևերը՝ ամեն մեկը իր տիպին բնորոշ դաշտերի քանակով:

Համակարգը ավտոմատ կազմում է մուտքագրված բառերի թեքված ձևերը՝ ըստ մուտքագրված պիտակի: Օրինակելի նմուշները պատրաստի կազմություններ են, որոնք նախապես մուտքագրվում են համակարգ: Կարող են լինել նույն բառույթի մի քանի տիպեր: Երկրորդ, երրորդ և ավել տիպերը կազմելու համար **Թեքման տիպ** դաշտում ավելացվում է նոր պիտակ մուտքագրելու դաշտ, որը մուտքագրելուց հետո կազմվում են համապատասխան ձևերը:

Բոլոր դաշտերը կարող են խմբագրվել: Ավտոմատ կազմված ձևերը կարելի է խմբագրել՝ պահպանելով նոր ձևը կամ ուղղելով համակարգի սխալ կազմած ձևը, դաշտերը կարող են մնալ նաև չլրացված կամ դատարկ:

Օգտատիրոջ մուտքագրած տեքստը սրբագրելիս համակարգը որոնում և համապատասխանեցում է անում ոչ միայն նախապես շտեմարան մուտքագրված **Բառույթ** դաշտի հետ, այլև տիպային պիտակերի հիման վրա կազմված/ գեներացված համապատասխան ձևերի հետ:

Թեքման տիպի պիտակները տեղեկություն են հաղորդում տվյալ բառույթի խոսքիմասային պատկանելության, թեքման տիպի, դրա ենթատիպի մասին:

Եզակի թվով բառույթի հոգնակի թիվը կազմելու համար **Հոգնակի թիվ** դաշտին կից **Հոգնակի թվի տիպ** դաշտում (կարող է ավելացվել նոր պիտակ մուտքագրելու դաշտ) մուտքագրվում է հոգնակի թվի ուղիղ/ելակետային ձևը կազմելու պիտակը: Այդ պիտակով կազմվում է տվյալ բառույթի հոգնակի ձևը:

Հոգնակի թվի թեքված ձևերը կազմելու համար **Հոգնակի թվի թեքման տիպ** դաշտում լրացվում է համապատասխան պիտակը, որի հիման վրա կազմվում են համապատասխան հոգնակի ձևերը: Դաշտը կարող է մնալ չլրացված:

Ըստ վերոնշյալ սկզբունքների՝ միայն գոյականը եզակի թվում ունի տասնմեկ տիպ, տիպերի հիմնական մասն ունեն ենթատիպեր, ենթատիպերի ընդհանուր թիվը հիսուն է: Նկատի են առնվել նաև անկանոն ձևերը:

Այսպիսով՝ հաշվի առնելով ժամանակակից հայերենի ձևայնացման առանձնահատկությունները և մշակելով դրանք նկարագրելու ուրույն համակարգ՝ կարող ենք փաստել, որ հայերեն էլեկտրոնային նոր սրբագրիչի շտեմարանը լիարժեք կերպով արտացոլում է հայերենի արդի վիճակը, քանի որ ներառում է ոչ միայն հայերենի հիմնական բառաֆոնդը, թեքվող ձևերի տարբերակային ձևերը, այլև՝

- հատկապես վերջին շրջանում ստեղծված մեծաթիվ նոր բառեր և նորաբանություններ,
- օտարաբանություններ,
- հատուկ անուններ (աշխարհագրական անուններ, անձնանուններ, ազգանուններ),
- հապավումներ,
- համառոտագրություններ:

Կայքից օգտվողների թիվը որևէ սահմանափակում չի ունենա: Ծառայելով հասարակության լայն շրջանակներին՝ այն **կարող է օգտակար լինել**.

ա. արևելահայերի համար ցանկացած տեքստի սրբագրման ժամանակ (գիտական, հրապարակախոսական, գեղարվեստական և այլն),

բ. արևմտահայերի համար, որոնք, ուղղագրական համակարգերի տարբերություններով պայմանավորված, ոչ միշտ են կարողանում ինչպես ճիշտ ուղղագրությամբ արևելահայերեն տեքստեր գրել, այնպես էլ առհասարակ լիարժեք կերպով կիրառել հայերենի քերականական կանոնները:

գ. օտարերկրացիների համար, որոնք բավարար լեզվական գիտելիքներ չունենալու պատճառով չեն կարող կառուցել իրենց գրավոր խոսքը պատշաճ մակարդակով:

Այսպիսով՝ կայքը և էլեկտրոնային նոր սրբագրիչը կարող է կիրառել ցանկացած մարդ՝ անկախ սեռից, տարիքից, կրթական մակարդակից, բնակության վայրից:

Ծրագրաշարը մշակված է այնպես, որ հետագայում հնարավոր է ընդլայնել համակարգի հնարավորությունը՝ ընդգրկելով շարահյուսական, կետադրական և ոճական սխալների ուղղումը: Այն կարելի է գործակցելի դարձնել նաև օտարալեզու սրբագրման համակարգերի հետ, որոնց միջոցով կարելի է կատարել հայերեն տեքստերի սրբագրում (http://www.stars21.com/spelling/armenian_spell_checker.html, https://www.spellchecker.net/eastern_and_western_spell_checker.html, https://addons.mozilla.org/en-US/firefox/addon/armenian-spell-checker-diction/,):

Էլեկտրոնային նոր սրբագրիչը կնպաստի ինչպես լեզվաբանական հետազոտությունների բնագավառում համակարգչային տեխնիկայի կիրառությանը, համակարգչային ծրագրերի միջոցով տեքստերում եղած լեզվական սխալների վերացմանը, այնպես էլ լայն առումով հայերենի՝ որպես համացանցային լեզվի հաջող գործառությանը:

ԳՐԱԿԱՆՈՒԹՅԱՆ ՑԱՆԿ
REFERENCES

1. Սարգսյան Մ., Հայերեն էլեկտրոնային սրբագրման առկա համակարգերի քննություն, Բանբեր Երևանի Վ. Բրյուսովի անվան պետական լեզվահասարակագիտական համալսարանի, 1(44), Երևան, 2018, էջ 598-604: Sargsyan M., Hayeren e'lektronayin srbagrman ar'ka hamakargeri qnnowt'yown, Banber Er&ani V. Bryowsovi anvan petakan lezvahasarakagitakan hamalsarani, 1(44), Er&an, 2018, e'j 598-604:
2. Sargsyan M., Some Observations on Armenian Electronic Proofreading Systems, "World Science", №6 (34), Vol. 8, June 2018, Warsaw, Poland, pp. 21-24.
3. Սարգսյան Մ., Հայերեն էլեկտրոնային սրբագրման համակարգ, https://armspell.am/hy, 2021: Sargsyan M., Hayeren e'lektronayin srbagrman hamakarg, https://armspell.am/hy, 2021:

МЕРИ САРГСЯН - О ПРОБЛЕМАХ ФОРМАЛЬНОГО ОПИСАНИЯ СЛОВАРЯ БАЗЫ ДАННЫХ ЭЛЕКТРОННОГО СПЕЛЛЧЕКЕРА

Ключевые слова: *электронная проверка орфографии, система корректуры, компьютерная лингвистика, формальное описание языка, база данных, программное обеспечение, принципы формального описания, формальное описание армянского языка*

Основой любой успешной системы проверки орфографии является не только умело разработанное программное обеспечение, но и богатая база данных. Кроме того, лексика, которая будет включена в базу данных

спеллчекера, должна быть описана формально. Процесс формального описания языка - сложная, ответственная работа, возникает множество проблем, связанных не только с точностью выработки принципов формального описания, но и со структурным типом языка, с особенностями парадигм склонения и спряжения слов, с грамматическими изменениями. Разрабатывая принципы формального описания языка, следует учитывать структурные особенности языка. Для обеспечения полноты формального описания не менее важно четко определить цель, для которой будет служить формальное описание. Учитывая особенности формального описания современного армянского языка и вызванные ими трудности, в рамках данной статьи мы рассматриваем проблемные вопросы формального описания, представляем пути их решения, принципы, лежащие в основе создания электронной базы данных армянского спеллчекера.

MERI SARGSYAN - ON THE ISSUES OF FORMAL DESCRIPTION OF THE DATABASE VOCABULARY OF THE ELECTRONIC SPELLCHECKER

Keywords: *Electronic spellchecker, proofreading system, computer linguistics, formal description of language, database, software, principles of formal description, formal description of the Armenian language*

The basis of any successful spellcheck system is not only skillfully developed software but also a rich database. In addition, the vocabulary to be included in the spellchecker database must be described formally. The process of formal description of a language is a difficult and challenging task, there are many problems related not only to the accuracy of the elaboration of the principles of formal description but also to the structural type of the language, to the peculiarities of the paradigms of declension and conjugation of words. Developing the principles of the formal description of a language, one should take into account the structural features of the language. To ensure the completeness of the formal description, it is equally important to clearly define the purpose for which the formal description will serve. Taking into account the peculiarities of the formal description of the Modern Armenian language and the difficulties related to them, in the framework of this article we examine the problematic issues of the formal description, present the ways of solving them, the principles underlying the creation of the electronic database for our spellchecker.

Ներկայացվել է՝ 08.09.2021
Գրախոսվել է՝ 07.09.2021