



**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**

ԼԵՃՎԱԲԱՆՈՒԹՅՈՒՆ ԵՎ ԲԱՆԱՍԻՐՈՒԹՅՈՒՆ

ЛИНГВИСТИКА И ФИЛОЛОГИЯ

LINGUISTICS AND PHILOLOGY

1 (58)

**Վ. ԲՐՅՈՒՍՈՎԻ ԱՆՎԱՆ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆԻ
«ԼԻՆԳՎԱ» ՀՐԱՏԱՐԱԿԶՈՒԹՅՈՒՆ**

ԵՐԵՎԱՆ - 2021

**ԲԱՌԱՊԱՇԱՐԻ ՁԵՎԱՎՈՐՄԱՆ ՍԿԶԲՈՒՆՔՆԵՐԸ
ՀՈՒՋԱՐՏԱՀԱՅՏՉԱԿԱՆՈՒԹՅԱՆ ՎԵՐՀԱՆՄԱՆ ԾՐԱԳՐԱԿԱԶՄՈՒՄ
ԱՆՆԱ ԱՎԵՏԻՍՅԱՆ**

Հիմնաբառեր՝ հաշվողական լեզվաբանություն, բառարանահեն մոդելում, կորպուսահեն մոդելում, հուզական բառապաշար, բառաֆոնդ, հուզական լիցք, հուզարտահայտչականության վերհանման ծրագրակազմ

Վերջին տարիներին հաշվողական լեզվաբանություն (ՀԼ) եզրույթը առավել հաճախ օգտագործվում է կիրառական ծրագրային ապահովման համատեքստում, ինչը պայմանավորված է ոչ միայն համացանցում տեքստային տեղեկատվության արագ աճով, այլև բնական լեզվի ավտոմատ մշակման անհրաժեշտությամբ (Большакова 2011: 90):

Լեզվաբանության այս ուղղությունն աղերսվում է տարբեր գիտակարգերի՝ մաթեմատիկայի, համակարգչային գիտության և արհեստական բանականության հետ: Այդ իսկ պատճառով այն հայտնի է նաև *մեքենական լեզվաբանություն* կամ *բնական լեզվի ավտոմատ մշակում* անվամբ: Այս միջգիտակարգի հիմքում ընկած է Նոամ Չոմսկու բնական լեզվի կառուցվածքի ձևայնացման շուրջ կատարած հետազոտությունները (Большакова 2011: 90):

Ներկայումս հաշվողական լեզվաբանության շրջանակում առաջարկվել են բազմաթիվ նորարարական մեթոդներ և գաղափարներ, սակայն ոչ բոլորն են գործնական կիրառություն ստացել: Օրինակ, դեռևս հնարավոր չէ իրականացնել հայերեն առցանց հաղորդակցության հուզարտահայտչականության վերհանում, ինչի առաջին քայլը համապատասխան բառապաշարի հավաքագրումն է:

Սույն հետազոտության նպատակն է ներկայացնել այն սկզբունքները, որոնք կիրառվում են հուզարտահայտչականության վերհանման նպատակով անգլերենի բառապաշարի հավաքագրման համար, ինչը կարող է հետագայում հիմք դառնալ հայերեն համապատասխան բառապաշարի ընտրության համար:

Հետազոտության արդիականությունը պայմանավորված է տեղեկատվական տեխնոլոգիաների զարգացմանը զուգընթաց առցանց հաղորդակցության շրջանակների ընդլանմամբ, որը խթանում է

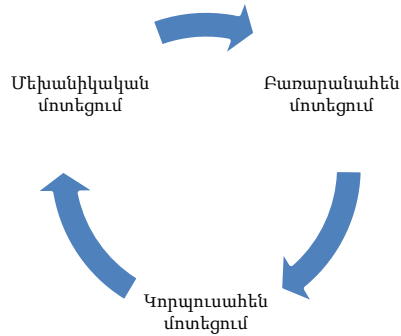
հուզականության տարբեր դրսևորումներ պարունակող հրապարակումների տարածումը: Նման հրապարակումների ավտոմատ վերհանումն այսօր կենսական անհրաժեշտություն է դարձել, քանի որ դրանք բացասական ազդեցություն են ունենում հասարակության վրա:

Սույն ուսումնասիրության առարկան բնական լեզվով հրապարակումների բառաֆոնդի ուսումնասիրությունն է: Նման հրապարակումների բառային վերլուծությունը կարևորվում է, քանի որ դրական կամ բացասական հույզեր արտահայտող լեզվական միավորները հուզարտահայտչականության վերհանման գործընթացում հույզերի առաջնային ցուցիչներն են:

Մինչ այժմ կատարված հետազոտություններում հույզերն արտահայտող բառերը դիտարկվել են որպես կարծիք պարունակող միավորներ: Դրական բառերը, ինչպիսիք են *գեղեցիկ*, *հիասքանչ* կամ *անկրկնելի* արտահայտում են ցանկալի իրադրություն կամ դրական որակական հատկանիշներ, մինչդեռ բացասական բառերը, ինչպիսիք են *վատ*, *սարսափելի* կամ *անտանելի*՝ անցանկալի իրադրություն կամ բացասական որակական հատկանիշներ: Առանձնակի ուշադրություն են գրավում հուզականություն պարունակող բառակապակցությունները և դարձվածքները, որոնք շատ հաճախ են հանդիպում սոցիալական ցանցերում կատարված հրապարակումներում (Liu 2015: 189): Նման միավորների օրինակ է *յոթերորդ երկնքում լինել* արտահայտությունը, որը նշանակում է *կատարյալ երջանկության մեջ լինել* (<https://bit.ly/3cKdBHE>):

Ըստ Բ. Լիուի հուզարտահայտչական բառերը բաժանվում են 2 հիմնական խմբի. հիմնական և համեմատական: Բոլոր վերոբերյալ օրինակները հիմնական տեսակի խմբում են: Համեմատական տեսակը, որը կիրառվում է համեմատական և գերադրական կարծիքներ արտահայտելիս, ընդգրկում է այնպիսի խոսքի մասեր, որոնք հնարավոր է ներկայացնել նման քերականական կարգով: Դրանք են ածականների (օրինակ՝ սարսափելի, ավելի սարսափելի, ամենասարսափելի) և մակբայների (դաժանաբար, ավելի դաժանաբար, ամենադաժանաբար) համեմատական և գերադրական աստիճանները (Liu 2015: 189):

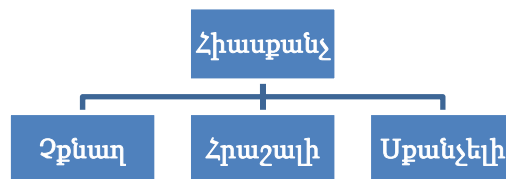
Բինգ Լիուն առանձնացնում է հուզական բառապաշարի հավաքագրման 3 մոտեցում (Liu 2015: 189), որոնք ներկայացված են գծապատկեր 1-ում: Հարկ է նշել, որ հուզական բառապաշարի հավաքագրման գործընթացում յուրաքանչյուր մոտեցում ուրույն տեղ է զբաղեցնում, քանի որ դրանք փոխկապակցված են և լրացնում են միմյանց աշխատանքը:



Բառապաշարի հավաքագրման *մեխանիկական* մոտեցումը ոչ միայն պահանջում է ֆիզիկական աշխատանք, այլև ժամանակատար է: Այդ իսկ պատճառով նախապատվությունը տրվում է այս մոտեցմանը միայն այն պարագայում, երբ անհրաժեշտություն է ստեղծվում ստուգել մեխանիկական տարբերակով կազմված բառապաշարի ճշգրտությունը:

Հուլական նման բառաֆոնդը կարող է առաջացնել մի շարք խնդիրներ, մասնավորապես այն կարող է ընդգրկել այնպիսի միավորներ, որոնց հուլականությունը կախված է ոլորտից կամ համատեքստից: Բացի այդ, կարող են լինել այնպիսի չեզոք իմաստով բառեր կամ բառակապակցություններ, որոնք որոշակի համատեքստում կարող են հուլականություն ստանալ (Liu 2015: 190):

Բառապաշարի հավաքագրման բառարանահեն մոտեցումը ենթադրում է բառարանների կիրառություն, որոնք յուրաքանչյուր բառի համար ցուցակագրում են հոմանիշներ և հականիշներ: Այս մոտեցման հիմնական մեթոդն այն է, որ նախ հավաքագրվում են մի քանի հիմնական հուլարտահայտչական բառեր, այնուհետև բառապաշարն ընդլայնվում է՝ համապատասխան հոմանիշների և հականիշների հավելմամբ: Օրինակը ներկայացված է գծապատկեր 2-ում:



Նման բառարանների օրինակներ են General Inquirer (GI), SentiWordNet-ը և այլն;

General Inquirer (GI) առանձնացրել է արժեքականության երկու կարգեր՝ դրական և բացասական: Դրական արժեք վերագրվել է 1,915 բառի, որոնց թվում ընդգրկված չեն այդ կամ համաձայնություն արտահայտող բառային միավորները: (<http://www.wjh.harvard.edu/~inquirer/homecat.htm>)

Օրինակը ներկայացված է գծապատկեր 3-ում: (<http://www.wjh.harvard.edu/~inquirer/Yes.html>)

Գծապատկեր 3

Yes

List shows first 100 entries. Total number of entries in this category:

20

Entries for this category are shown with all tags assigned and sense definitions:

ABSOLUTE#2

H4Lvd Strong Ovrst Yes SureLw LY 73% adv: "Absolutely"--without qualification, certainly, totally

AGREE#1

H4Lvd Pstv Affil Yes IAV PowGain PowTot SUPV 96% verb: To be of one mind, harmonize in opinion or feeling, give consent, assent

ALRIGHT

H4 Yes Intrj

Բացասական արժեք վերագրվել է 2291 բառի, որոնց թվում նույնպես ներառված չեն ոչ կամ մերժում արտահայտող բառային միավորները: (<http://www.wjh.harvard.edu/~inquirer/homecat.htm>)

Օրինակը ներկայացված է գծապատկեր 4-ում: (<http://www.wjh.harvard.edu/~inquirer/No.html>)

No

List shows first 100 entries. Total number of entries in this category:

7

Entries for this category are shown with all tags assigned and sense definitions:

DISAGREE

H4Lvd Ngtnv Hostile ComForm No LAV PowCon PowTot SUPV verb: To differ

MEAN#5

H4Lvd Ovrst No Negate NotLw LY 1% idiom-adv: "By no means"--not at all

NAY

H4 No Intrj

SentiWordNet-ը, որն առաջինը նկարագրվել է Էսուլիի և Սեբաստիանիի կողմից, ընդգրկում է հոյգեր արտահայտող բառապաշար, որը ընդլայնում է WordNet-ը¹ համալրելով այն հուզական տեղեկատվությամբ: (Esuli, Sebastiani 2006: 417):

SentiWordNet-ում հոյգերի վերագրման գործընթացը հստակ տարրորոշված չէ և իրականացվում է WordNet-ի յուրաքանչյուր սինսեթին² ավելացնելով երեք հուզական միավոր: Յուրաքանչյուր սինսեթ ունի երեք միավոր: Դրանք են

1. Pos(s): սինսեթների դրական միավորը,
2. Neg(s): սինսեթների բացասական միավորը,
3. Obj(s): սինսեթների չեզոք միավորը: (Joshi, Bhattacharyya, Ahire 2017: 89)

Բառարանահեն մոտեցման մեկ այլ փորձված մեթոդ են առաջարկում Զ. Կեմպը և մի շարք այլ հեղինակներ (Kamps, Marx, Mokken, De Rijke 2004: 1115-1116): Նրանք կիրառում են WordNet -ի

¹ WordNet-ը անգլերեն բառապաշարի հսկայական շտեմարան է, որտեղ ընդգրկված են 166,000-ից ավելի բառաձևեր և իմաստային զույգեր: Այն ընդգրկում է մի շարք իմաստային հարաբերություններ, մասնավորապես հոմանիշներ, հակահիշներ, տեսականիշներ և այլն:

² WordNet-ը կիրառում է հոմանիշների շարքերը՝ սինսեթերը, բառային իմաստները նկարագրելու համար (Miller 1995: 40):

տարրերի միջև հեռավորության վրա հիմնված մեթոդը: Այն ենթադրում է տրված ածականների հուզականության դիրքորոշման որոշումը: Համաձայն նրանց « t_1 և t_2 բառերի միջև հեռավորությունը $d(t_1, t_2)$ այն ուղին է, որը կապում է t_1 -ը՝ t_2 -ին WordNet-ում»: (Kamps, Marx, Mokken, De Rijke 2004: 1115-1116): Դա ածականի հուզական դիրքորոշումը որոշվում է երկու հիմնական հուզականություն պարունակող բառերի, օրինակ՝ good և bad ածականների, հարաբերական հեռավորությամբ (Kamps, Marx, Mokken, De Rijke 2004: 1115-1116): Ածականի հուզական լիզքը հաշվարկելու համար Ջ. Կեմպը կիրառել է ստորև ներկայացված բանաձև 1-ում (1) (Kamps, Marx, Mokken, De Rijke 2004: 1115-1116):

(1)

$$SO(t) = \frac{(d(t, bad) - d(t, good))}{d(good, bad)}$$

Հարկ է նշել որ t դրական է, եթե միայն $SO(t) > 0$, այլապես այն բացասական է: $SO(t)$ բացարձակ արժեքը ներկայացնում է հուզականության ուժգնությունը:

Համաձայն Ն. Կաշի և Մ. Կիթսուրեգավայի՝ վեր էջի աղյուսակը կարող է պարունակել սյունակներ, որոնք հստակ տարանջատում են համապատասխան տեքստի դրական և բացասական դիրքորոշումները: Նման տեղեկատվությունը կարելի է կիրառել ընդարձակ վեր էջերի հուզական միավորներ պարունակող նախադասություններն առանձնացնելու համար: Այս գործընթացը հիմնված է համապատասխանաբար դրական և բացասական նախադասությունների շարքում ածականական արտահայտությունների վիճակագրության վրա (Kaji, Kitsuregawa 2007: 1075-1076):

Չնայած գոյություն ունեն բառարանահեն շատ մոտեցումներ, սակայն դրանց անկախ ուսումնասիրությունը կարծես թե անհնարին է: Այդ իսկ պատճառով երբեմն դժվար է նշել, թե այս մեթոդներից որոնք են արդյունավետ և նպատակահարմար: Ընդհանուր առմամբ, կարող ենք փաստել, որ բառարանահեն մոտեցման առավելությունն այն է, որ *մեծածավալ տվյալներից* (big data) հնարավոր է հեշտ և արագ ընդգծել հուզականություն պարունակող բառերը և դրանց հուզական լիզքը:

Հարկ է նշել, որ հնարավոր են որոշ բացթողումներ, որոնց հայտնաբերման և շտկման համար անհրաժեշտ կլինի մեխանիկական մոտեցման կիրառությունը: Բառարանահեն մոտեցման թերությունն այն է, որ բառերի հուզական լիզքը որոշելիս դուրս է մղվում համատեքստի դերը:

Կորպուսահեն մոտեցումը իրականացվում է 2 եղանակով՝ (Liu 2015: 193-194):

- հիմնական հոյգեր արտահայտող բառերի միջոցով բացահայտվում են այլ հոյգեր արտահայտող բառերը և դրանց հուզական դիրքորոշումները ոլորտային կորպուսից (domain specific corpus): (Liu 2015: 193):
- ոլորտային կորպուսի կիրառման միջոցով համընդհանուր կիրառություն ունեցող հուզական բառապաշարը համապատասխանեցվում է նոր բառապաշարի հետ: Գործնականում այս խնդիրը կարող է լինել ավելի բարդ, քան պարզապես ոլորտին բնորոշ հուզականություն պարունակող բառապաշարի ձևավորումը, քանի որ մի համատեքստում բառը կարող է արտահայտել դրական հոյգ, մեկ այլ համատեքստում՝ բացասական (Liu 2015: 194): Հարկ է նշել, որ կորպուսահեն մոտեցումը կարող է կիրառվել համընդհանուր կիրառություն ունեցող հոյգերի բառապաշար ձևավորելու համար, սակայն լայնածավալ կորպուսի բացակայության դեպքում ավելի նպատակահարմար է կիրառել բառարանահեն մոտեցումը, քանի որ բառարանները ընդգրկում են բոլոր բառերը (Liu 2015: 194):

Հուզական բառերի վերհանումը կորպուսում կարող է իրականացվել երկու եղանակով (Liu 2015: 194)՝

1. շաղկապների միջոցով՝ հիմնվելով դրանց լեզվաբանական օրենքների և կանոնների վրա: (Liu 2015: 194):
2. կարծիք արտահայտող ասույթներում շարահյուսական հարաբերությունների քննությամբ: (Liu 2015: 194):

Առաջին եղանակը առաջարկել են Վ. Հաթզիվասիլոգլուն և Կ. Ռ. Մկկոունը: (Hatzivassiloglou, McKeown 1997: 174-176) Կանոններից մեկը, որը վերաբերում է **և** շաղկապին, ենթադրում է, որ դրանով կապակցված ածականները պետք է ունենան նույն հուզական դիրքորոշումը: Օրինակ՝ «Այս մեքենան գեղեցիկ և ընդարձակ է:» նախադասության մեջ, եթե *գեղեցիկ* ածականը պարունակում է դրական հոյգ, ապա կարող ենք ենթադրել, որ *ընդարձակ* ածականը նույնպես դրական պետք է լինի: Նման կանոնները կիրառվում են նաև այլ շաղկապների հետ, մասնավորապես՝ **կամ, բայց, կամ կամ և ո՛չ ո՛չ**: Այս եղանակը կոչվում է հուզականության կայունություն: Հեղինակները նաև կիրառել են մեքենական ուսուցման քայլերը, որպեսզի պարզեն երկու ածականների հուզական դիրքորոշումները (Liu 2015: 194):

Երկրորդ եղանակը, որի հիմքում ընկած են շարահյուսական հարաբերությունները, ի սկզբանե նախատեսված է եղել կարծիքների

թիրախների վերհանման համար (Liu 2015: 195): Բ. Լիուն բերում է հետևյալ օրինակը (Liu 2015: 141)՝

The software is amazing.

Վերորբերյալ օրինակում *amazing* ածականը, բնորոշելով գոյականը կամ գոյականական բառակապակցությունները (տվյալ դեպքում software գոյականը), այն դարձնում է հուզականության ասպեկտ:³

Ամփոփելով կատարված հետազոտությունը՝ կարող ենք եզրակացնել, որ գոյություն ունի հուզական բառարանի ձևավորման երեք սկզբունք: Այդ սկզբունքներն անմիջականորեն փոխկապակցված են և դրանց կիրառության միջոցով հնարավոր է ստեղծել հայերեն առցանց հաղորդակցության համար նախատեսված ուրույն հուզական բառապաշար: Նման բառապաշարի առկայությունը թույլ կտա ճշգրիտ եղանակով վերհանել առցանց հրապարակումների հուզարտահայտչականությունը:

ԳՐԱԿԱՆՈՒԹՅԱՆ ՑԱՆԿ REFERENCES

1. Большакова Е.И., Компьютерная лингвистика: методы, ресурсы, приложения, Москва, 2011. Bol'shakova E.I., Komp'yuternaja lingvistika: metody, resursy, prilozhenija, Moskva, 2011.
2. Esuli A., Sebastiani F., SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining, Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Italy, 2006
3. Hatzivassiloglou V., McKeown K. R., Predicting the Semantic Orientation of Adjectives. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-1997). 1997.
4. Joshi A., Bhattacharyya P., Ahire S., Sentiment Resources: Lexicons and Datasets, A Practical Guide to Sentiment Analysis, Springer International Publishing AG, 2017

³ Ասպեկտը կատեգորիա, հատկանիշ կամ թեմա է, որի մասին տեքստում խոսվում է: (<https://monkeylearn.com/blog/aspect-based-sentiment-analysis/>) Ասպեկտների վրա հիմնված հուզարտահայտչականության վերլուծությունը տեքստի վերլուծության մի տեսակ է, որը դասակարգում է տեքստերը ըստ ասպեկտների և վերհանում յուրաքանչյուր ասպեկտի առնչվող հոյզերը: (<https://www.wonderflow.ai/blog/challenges-in-aspect-based-sentiment-analysis-absa>)

5. Kaji N., Kitsuregawa M.. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-2007), Prague 2007.
6. Kamps J., Marx M., Mokken R. J., De Rijke M., Using Wordnet to Measure Semantic Orientation of Adjectives. In Proceedings of LREC-2004.
7. Liu B., Sentiment Analysis, Mining Opinions, Sentiments, and Emotions, Cambridge University Press, 2015
8. Miller, G. A., WordNet: A lexical database for English. Communications of the ACM, 1995

INTERNET SOURCES

9. <https://bit.ly/3cKdBHE> (01.09.2021)
10. <http://www.wjh.harvard.edu/~inquirer/homecat.htm> (01.09.2021)
11. <http://www.wjh.harvard.edu/~inquirer/Yes.html> (01.09.2021)
12. <http://www.wjh.harvard.edu/~inquirer/No.html> (01.12.2021)
13. <https://monkeylearn.com/blog/aspect-based-sentiment-analysis/> (01.12.2021)
14. <https://www.wonderflow.ai/blog/challenges-in-aspect-based-sentiment-analysis-absa> (01.12.2021)

АННА АВETИСЯН - ПРИНЦИПЫ ФОРМИРОВАНИЯ СЛОВАРНОГО ЗАПАСА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ АНАЛИЗА ТОНАЛЬНОСТИ

Ключевые слова: *Вычислительная лингвистика, словарный подход, корпусный подход, лексика настроений, лексика, программное обеспечение для определения настроений, ориентация настроений*

Анализ тональности - неотъемлемая часть компьютерной лингвистики, целью которой является выявление полярности текста. Этот процесс основан на выявлении тональной лексики, поскольку именно она является основным проводником эмоциональной конъюнктуры.

Таким образом, основная цель данной статьи - представить принципы генерации лексики тональности, которая в основном осуществляется с помощью трех основных принципов: ручного, словарного и корпусного подходов. Каждый подход имеет свою собственную функцию и может быть использован с целью системной компиляции тональной лексики. Эти подходы взаимосвязаны и наилучшим образом дополняют друг друга.

Развитие информационных технологий в наивысшей степени определяет актуальность данного исследования, так как тональная лексика способствует распространению постов в социальных сетях.

ANNA AVETISYAN - THE PRINCIPLES OF SENTIMENT LEXICON GENERATION IN SENTIMENT ANALYSIS SOFTWARE

Keywords: *Computational Linguistics, Dictionary-Based Approach, Corpus-Based Approach, Sentiment Lexicon, Lexicon, Sentiment Detection Software, Sentiment Orientation*

Sentiment analysis (SA) is an essential part of computational linguistics that aims at disclosing the polarity of the text. This process is mainly based on singling out sentiment lexicon since words and phrases are the main means of conveying sentiments.

The main objective of this article is to present the principles of sentiment lexicon generation. The latter is mainly carried out through three approaches, namely manual approach, dictionary-based approach, and corpus-based approach. Each approach has its own peculiarities and can be used to make sentiment lexicon compilation accurate and relevant. Those approaches are intertwined and they complement each other perfectly.

The topicality of the research is mainly determined by the development of information technology that promotes the dissemination of online content pertaining to several sentiments.

Ներկայացվել է՝ 10.09.2021

Գրախոսվել է՝ 09.09.2021

Երաշխավորվել է ԲՊՀ Անգլերենի ամբիոնի կողմից՝ 27.10.2021