ՀՏԴ 004.02

Տեքստի ինքնատիպության աստիճանի գնահատման նախապատրաստում օգտակար բառերի ուղիղ ձևի զանգվածի ձևավորման միջոցով

Պետրոսյան Գևորգ
Սահակյան Ռուստամ

Հանգուցային բառեր. *տեքստի գնահատման նախապատրաստում, տեքստի ինքնատիպության աստիճան, օգտակար բառեր, բառերի զանգված, ինքնատիպության աստիճանի գնահատման համակարգ, «stop» բառեր*

Ժամանակակից տեխնոլոգիական առաջընթացը հնարավորություն է տալիս կրկնօրինակելու հետազոտական աշխատանքները ավելի դյուրին, քան երբևէ նախկինում, քանի որ յուրաքանչյուր օր միլիոնավոր օգտատերեր ներբեռնում են ինտերնետ միջավայր իրենց հետազոտական աշխատանքները հանրային և անձնական օգտագործման նպատակներով:

Նշված համատեքստում կարևոր տեղ է զբաղեցնում հետազոտական աշխատանքների ինքնատիպության աստիճանի որոշման հարցը:

Սույն աշխատանքի շրջանակներում առաջարկվում է ինքնատիպության աստիճանի գնահատման համակարգի համար հայերեն տեքստի նախապատրաստման մոտեցում՝ օգտակար բառերի ուղիղ ձևի զանգվածի ձևավորման միջոցով: Այն կարևոր տեղ է զբաղեցնում հետազոտական աշխատանքների ինքնատիպության աստիճանի գնահատման համակարգի մշակման գործում: Նման մոտեցման ներդրումը կնպաստի հետազոտական աշխատանքների ինքնատիպության աստիճանի բարձրացմանը:

Նախաբան

Հետազոտական աշխատանքի ինքնատիպություն ասելով հասկանում ենք մտավոր գործունեության յուրահատուկ արդյունք, որն իր մեջ ներառում է նոր մտքեր, հայացքներ, տեսություններ և մտահանգումներ [5, 146]:

Տարբեր հեղինակների կողմից գրված հետազոտական աշխատանքները երբեմն ունենում են որոշակի բովանդակային ընդհանրություններ: Այդ ընդհանրությունները կարող են առաջանալ ոչ միտումնավոր, բայց

երբեմն էլ նպատակադրված են, քանի որ այլ հեղինակի հետազոտական աշխատանքի բովանդակությունը որպես սեփական ներկայացնելու համար պահանջվում են տեխնիկական պարզ գործողություններ:

Հետազոտողի կողմից իր աշխատանքում առանց հղումների փոխառության մեծ մասը հիմնականում իրականացվում է կանխամտածված՝ նպատակ ունենալով առանց սեփական ռեսուրսների և ջանքերի կիրառման որոշակի մտավոր արտադրանք ունենալը [6, 169]:

Գիտահետազոտական նկատառումներից ելնելով՝ նման երևույթի կանխարգելումը ունի էական նշանակություն, իսկ դրա լուծման ուղիներից մեկը հետազոտական աշխատանքի տեքստի ինքնատիպության աստիճանի գնահատման ավտոմատ համակարգերի ստեղծումն է [8, 1]: Եթե աշխատանքները մեծածավալ չեն, դա կարող է ստուգվել անզեն աչքով, սակայն այն դեպքերում, երբ հետազոտական աշխատանքները մեծ ծավալ են ունենում, անհրաժեշտ է կիրառել աշխատանքի ինքնատիպության գնահատման ավտոմատ համակարգեր [7, 1]:

Ինքնատիպության աստիճանի գնահատման (ԻԱԳ) համակարգերում կարևոր տեղ է զբաղեցնում տեքստերի նախապատրաստման ենթահամակարգը, քանի որ դրա միանշանակությունից է կախված վերջնական արդյունքը [4, 4]:

Հիմնախնդրի նկարագրությունը

Տեքստերի նախապատրաստման փուլում կարևոր խնդիր է տվյալ լեզվի առանձնահատկությունները հաշվառելը: Ի տարբերություն ինքնատիպության աստիճանի գնահատման համակարգի մյուս երկու ենթահամակարգերի՝ բովանդակության գնահատման և արդյունքների ամփոփման, տեքստի նախապատրաստման փուլը, կախված տվյալ լեզվի առանձնահատկություններից, կարող է ստանալ ամբողջովին այլ ընթացք [4, 4]: Հետևաբար, մի լեզվով գրված տեքստի նախապատրաստման համար մշակված մոտեցումները մեկ այլ լեզվով գրված տեքստի նկատմամբ կիրառելիս չեն հանգեցնի ցանկալի արդյունքի:

Հայերենը, լինելով քերականորեն համեմատաբար բարդ լեզու, աչքի է ընկնում ձևաբանական մի շարք տարբերություններով՝ համեմատած այլ լեզուների հետ: Այդ տարբերությունների օրինակներից են հոլովի և թվի քերականական կարգերը: Եզակի թիվը հակադրվում է հոգնակիին, ուղղական հոլովը՝ մյուս հոլովներին: Սակայն կան լեզուներ, որոնք հոլովի քերականական կարգ չունեն, կամ հոլովների քանակը ավելի քիչ է՝ ի հակադրություն հայերենի հոլովների բազմազանության (ռուսերեն, ֆրանսերեն և այլն) [2, 5]:

Խնդրի դրվածքը

Ներկայացված աշխատանքի նպատակն է մշակել մոտեցում, որի հիման վրա կկատարվի հայերենով գրված տեքստի նախապատրաստում ինքնատիպության աստիճանի գնահատմանը օգտակար բառերի ուղիղ ձևի զանգվածի ձևավորման միջոցով:

Արդյունքում ստացված բառերի ուղիղ ձևի զանգվածը հնարավոր կլինի կիրառել նախագծված ԻԱԳ համակարգում:

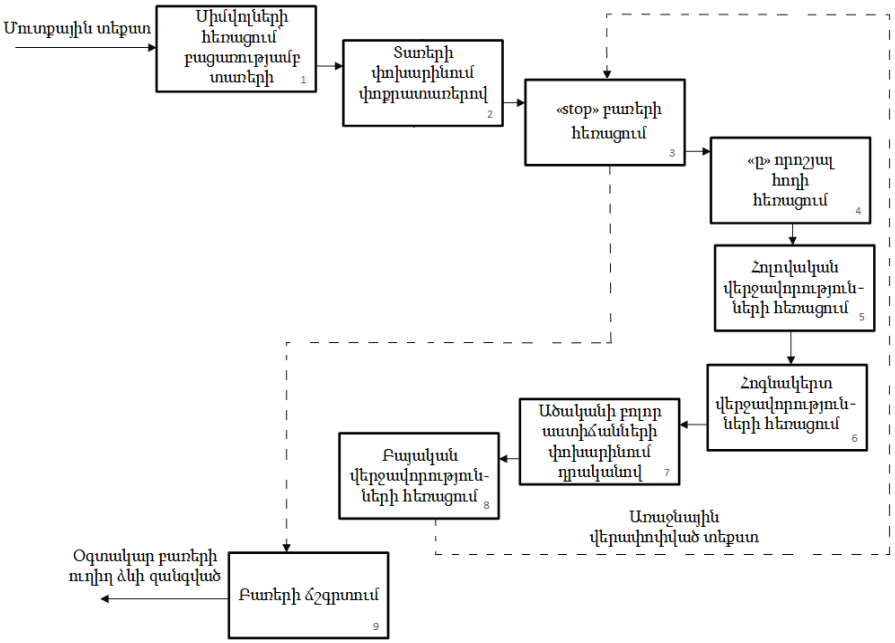
Մոտեցման նկարագրությունը

Հայերենում բառի քերականական ձևերը կազմվում են՝ բառին ավելացնելով առանձին մասնիկներ, որոնցից յուրաքանչյուրն արտահայտում է մի քերականական կարգ. օրինակ «սեղաններից» բառաձևի մեջ «սեղան» բառին ավելացնում ենք երկու մասնիկ, որոնցից առաջինը հոգնակիություն արտահայտելու համար (-ներ), իսկ երկրորդը՝ բացառական հոլովի համար: Նման դեպքերում լեզուն կոչվում է կցական: Ստացվում է, որ բառի հիմքը իր կազմով անփոփոխ է մնում մասնիկավորման ժամանակ. նա ստանում է մասնիկներ՝ առանց փոփոխության ենթարկելու իր հիմքային հնչյունները [1, 589]:

Մի շարք դեպքերում քերականական մեկ մասնիկը կարող է արտահայտել քերականական մեկից ավելի իմաստներ: Օրինակ, «գրի՛ր» բառի մեջ -իր մասնիկն արտահայտում է բայական խոսքիմասային իմաստ, հրամայական եղանակ, եզակի թիվ, երկրորդ դեմք [3, 223]:

Սա նույնպես բարդացնում է լեզվի կառուցվածքը, ի տարբերություն այն դեպքերի, երբ յուրաքանչյուր մասնիկ միայն մեկ քերականական իմաստ է արտահայտում:

Հաշվի առնելով հայերենի այս և նմանատիպ այլ առանձնահատկությունները՝ առաջարկվում է մոտեցման քայլերի հետևյալ հաջորդականությունը (նկար 1):



Նկար 1. Մոտեցման քայլերի նկարագրություն

Մոտեցման համապատասխան քայլերի նկարագրությունը

Մինչև մոտեցման բուն քայլերին անցնելը անհրաժեշտ է ստեղծել «stop» բառերի զանգված և նրա մեջ ավելացնել հետևյալ խոսքի մասերին պատկանող բոլոր բառերը՝ դերանուն (ինչպես նաև դերանվան բոլոր տարաձևիմք հոլովաձևերը), կապ, շաղկապ, վերաբերական, ձայնարկություն և օժանդակ բայեր՝ իրենց բոլոր ժամանակաձևերով և դեմքերով և «պիտի», «պետք է» եղանակիչ մասնիկները: Սրանք այն բառերն են, որոնք նախադասության անդամ չեն դառնում (բացառությամբ դերանվան) և տեքստի՝ ինքնատիպության աստիճանի գնահատման նախապատրաստման սույն մոտեցման մեջ տեքստի բովանդակության վրա ազդեցություն չունեն:

Ստորև ներկայացված են տեքստի նախապատրաստման բուն քայլերը.

1) Սիմվոլների հեռացում՝ բացառությամբ տառերի:
2) Տառերի փոխարինում փոքրատառերով:
3) «stop» բառերի հեռացում:

4) «Ը» որոշյալ հոդի հեռացում:

5) Հոլովական վերջավորությունների հեռացում.

Ա) բացառական հոլով

- հեռացնել բառավերջի «ից» տառերի զույգը (նախքան «ից»-ը հեռացնելը հեռացնել «յից» տառերի եռյակը, քանի որ «ա»-ով և «ո»-ով վերջացող բառերի բացառական հոլովը կազմելիս ավելանում է «յ»),
- բառավերջի «ուց» տառերի զույգը դարձնել «ի»,

Բ) գործիական հոլով

- հեռացնել բառավերջի «ով» տառերի զույգը (նախքան «ով»-ը հեռացնելը հեռացնել «յով» տառերի եռյակը, քանի որ «ա»-ով և «ո»-ով վերջացող բառերի գործիական հոլովը կազմելիս ավելանում է «յ»),
- բառավերջի «ությամբ» վերջավորությունը դարձնել «ություն»,
- բառավերջի «մամբ» վերջավորությունը դարձնել «ում»,

Գ) տրական հոլով

- հեռացնել «ին», «ուն», «վան» վերջավորությունները,

Դ) սեռական հոլով

- հեռացնել բառավերջի «ի» տառը,
- բառավերջի «ու» տառը դարձնել «ի» ,
- հեռացնել բառավերջի «վա» տառերի զույգը,
- հեռացնել բառավերջի «ոջ» տառերի զույգը,
- բառավերջի «ության» վերջավորությունը դարձնել «ություն»:

6) Հոգնակերտ վերջավորությունների հեռացում՝ եր, ներ (հեռացնել նաև «երն» «ներն» վերջավորությունները, քանի 5-րդ քայլում տեքստի բառերից հեռացվել էր միայն «ը» որոշյալ հոդը):

7) Ածականի բաղդատական և գերադրական աստիճանների փոխարինում դրական աստիճանով.

Ա) բաղդատական աստիճան

- հեռացնել «ավելի», «պակաս», «նվազ», «նույնքան» բառերը,

Բ) գերադրական աստիճան

- հեռացնել «ամենա» նախածանցը,
- հեռացնել «ագույն» վերջածանցը,
- հեռացնել «ամենից», «բոլորից» բառերը:

8) Հեռացնել բայական վերջավորությունները.

Ա) անդեմ բայ

- հեռացնել «ել», «ալ» վերջավորությունները (անորոշ դերբայ),
- հեռացնել «ելիս», «ալիս» վերջավորությունները (համակատար

դերբայ),
- հեռացնել «ած», «ացած», «եցած» վերջավորությունները (հարակատար դերբայ),
- հեռացնել «ող», «ացող», «եցող» վերջավորությունները (ենթակայական դերբայ),

Բ) դիմավոր բայ

ա) սահմանական եղանակ
- հեռացնել «ում» վերջավորությունը (անկատար ձևաբայ),
- հեռացնել «ելու», «ալու» վերջավորությունները (ապակատար ձևաբայ),
- հեռացնել «ացել» վերջավորությունը (վաղակատար ձևաբայ) (վաղակատար ձևաբայի «ել» վերջավորությունը համընկնում է անորոշ դերբայի «ել» վերջավորության հետ, որը արդեն հեռացրել ենք),
- հեռացնել «ա» վերջավորությունը (ժխտական ձևաբայ) (ժխտական ձևաբայի «ի» վերջավորությունը համընկնում է սեռական հոլովի «ի» հոլոման հետ, որը արդեն հեռացրել ենք),

բ) ըղձական, ենթադրական, հարկադրական եղանակ
- հեռացնել «եմ», «ես», «ենք», «եք», «են», «ամ» ,«աս», «ա» «անք», «աք», «ան» վերջավորությունները,

գ) հրամայական եղանակ
- հեռացնել «իր», «եք» վերջավորությունները:

9) Բառերի ճշգրտում: Եթե տեքստում առկա են 5 և ավելի տառերից կազմված բառեր, որոնց տառերի քանակների տարբերությունը 1 է, և բառերից մեկը պարունակում է մյուսի բոլոր տառերը, ապա այդ բոլոր բառերը փոխարինել ավելի շատ տառեր ունեցող բառով:

Մոտեցման մեջ որպես մուտքային տվյալ հանդես է գալիս «Մուտքային տեքստը»: 1-ին և 2-րդ քայլերի կատարումից հետո ստացված տեքստը բաղկացած կլինի միայն բառերից, որոնց բոլոր տառերը փոքրատառ են: 3-րդ քայլում տեքստից հեռացվում են այն բառերը, որոնք առկա են նախապես ստեղծված «stop» բառերի զանգվածում:

4-8-րդ քայլերում տեքստի բառերից հեռացվում են վերջավորությունները, «ը» որոշյալ հոդը, ինչպես նաև ածականները բերվում են դրական աստիճանի: Այնուհետև կրկին իրականացվում է 3-րդ քայլը (նշված է կետագծերով), որից հետո տեղի է ունենում տեքստի բառերի ճշգրտում: Մշակված մոտեցման արդյունքում որպես ելքային տվյալ ստացվում է մուտքային տեքստի «Օգտակար բառերի ուղիղ ձևի զանգված»:

Եզրակացություն

Աշխատանքի շրջանակներում ներկայացված է ինքնատիպության աստիճանի գնահատման համակարգի համար տեքստի նախապատրաստման մոտեցում՝ օգտակար բառերի ուղիղ ձևի զանգվածի ձևավորման միջոցով:

Ներկայացված մոտեցումը նախագծված համակարգում փորձարկվելուց հետո կարող է ենթարկվել բարելավման՝ հաշվի առնելով իրական տեքստերի հետ աշխատանքի առանձնահատկությունները և արդյունքների վերլուծությունը:

Արդյունքում ստացված բառերի ուղիղ ձևի զանգվածը հնարավոր կլինի կիրառել նախագծված ինքնատիպության աստիճանի գնահատման համակարգում:

Գրականություն

1. Աղայան Է. Բ., Լեզվաբանության ներածություն, երրորդ հրատարակություն, Երևան, 1967, 639 էջ:
2. Բաղիկյան Խ. Գ., Ժամանակակից հայոց լեզվի ձևաբանության տեսության և գործնական աշխատանքների ուսումնական ձեռնարկ, Երևան, 2010, 264 էջ:
3. Եզեկյան Լ., Հայոց լեզու, Երևան, Երևանի պետական համալսարանի հրատարակչություն, 2007, 402 էջ:
4. Սահակյան Ռ. Ռ., Պետրոսյան Գ. Ա., Հետազոտական աշխատանքների ինքնատիպության աստիճանի գնահատման համակարգի նախագծում, Հայաստանի ճարտարագիտական ակադեմիայի լրաբեր, 2022, h. 19:
5. Арутюнов Э. К., Улитин И. Н., К законодательному вопросу проверки уникальности (оригинальности) текста гуманитарных научных работ, Научная периодика проблемы и решения, Том 7, № 3, 2017, с. 144-150.
6. Хачецуков З. М., Проверка на оригинальность научных текстов: Вопросы теории и практики, Гуманитарий Юга России, № 1, 2014, с. 166-179.
7. Ranti E. P., Andysah P. U. S., Examination of Document Similarity Using Rabin-Karp Algorithm, International Journal of Recent Trends in Engineering and Research, 2017.
8. Shukla P., PlagCaps: Prediction of Plagiarised Text on a Corpus Dataset using Deep Learning Algorithms, 2021.

Подготовка к определению степени уникальности текста на основе формирования массива прямой формы полезных слов

Петросян Геворг
Саакян Рустам

Резюме

Ключевые слова: *подготовка текста к оценке, степень уникальности текста, полезные слова, массив слов*

Современные технологические достижения делают копирование научно-исследовательских работ проще, чем когда-либо, поскольку каждый день миллионы пользователей загружают в интернет-среду свои исследовательские работы для публичного и личного использования.

Большая часть безссылочных заимствований в работе, как правило, исследователем делается преднамеренно, с целью получения определенного интеллектуального продукта без использования собственных ресурсов и усилий.

Исходя из научно-исследовательских позиций, профилактика такого явления имеет существенное значение, а одним из способов его решения является создание автоматических систем оценки степени уникальности текста исследовательской работы. Если работы не масштабные, это можно проверить невооруженным глазом, но в случаях, когда исследовательские работы имеют большой объем, необходимо применять автоматические системы оценки уникальности работы.

В системах оценки степени уникальности важное место занимает подсистема подготовки текста, ведь от ее однозначности зависит конечный результат.

В рамках данной работы предлагается подход к подготовке текста для системы оценки уникальности путем формирования массива полезных слов прямой формы. Реализация такого подхода будет способствовать повышению степени уникальности исследовательских работ.

На этапе подготовки текстов важной задачей является учет особенностей языка, на котором написан текст. В отличие от других подсистем системы оценки уникальности, этап подготовки текста может протекать совершенно по-разному в зависимости от специфики конкретного языка. Поэтому подходы, разработанные для подготовки текста, написанного на одном языке, не приведут к желаемому результату при применении к тексту, написанному на другом языке.

Армянский, будучи грамматически относительно сложным языком, отличается рядом морфологических отличий по сравнению с другими языками.

В разработанном подходе «Исходный текст» является входной информацией, а «Массив прямых форм полезных слов» — выходной информацией.

После апробации представленного подхода в проектируемой системе возможна его доработка с учетом особенностей работы с реальными текстами и анализа полученных результатов. Полученный массив слов прямой формы можно будет использовать в разработанной системе оценки уникальности текста.

Preparation for the Determination of Text's Uniqueness Degree based on the Formation of Direct Form Massive of Useful Words

Petrosyan Gevorg
Saakyan Rustam

Summary

Key words: *preparation for the text determination, degree of text uniqueness, useful words, massive of words*

Modern technological advancements make it easier than ever to duplicate scientific research papers, as millions of users daily upload their research papers for public and personal uses.

Most of the unreferenced borrowings, used by the researcher in his work, are generally done deliberately, with the aim of having a certain intellectual product without using his own resources and efforts.

Based on scientific research considerations, the prevention of such a phenomenon has an essential importance and one of the ways to prevent it is the creation of automatic systems for evaluating the text's degree of uniqueness.

If the work is small-scale, it can be checked with the naked eye, but if the research work has a large volume, it is necessary to apply automatic systems for evaluating the uniqueness of the work.

In the systems for evaluating the degree of uniqueness, the text preparation subsystem occupies an important place, because the final result

depends on its ambiguity.

Within the framework of this work, an approach to text preparation for the system of evaluating the uniqueness is proposed by forming a direct form massive of useful words. Implementation of such an approach will contribute to increasing the degree of uniqueness of research works.

At the stage of text preparation, an important task is to take into account the specifications of the given language. Unlike other sub-systems of the uniqueness evaluation system, the stage of text preparation can take a completely different course depending on the specifics of the given language. Therefore, the approaches developed for the preparation of a text written in one language will not lead to the desired result when applied to a text written in another language.

Armenian, being a grammatically complex language, is distinguished by a number of morphological differences compared to other languages (of which are the grammatical categories of gender and number).

In the developed approach, the “Source Text” is the input data, and the "Useful Words Direct Form Massive” is the output data.

After testing the presented approach, it can be improved in the designed system by taking into account the peculiarities of working with real texts and analyzing the outcomes. The resulting direct form massive of words will be possible to use in the system.

Ներկայացվել է 14.10.2022թ.
Գրախոսվել է 15.10.2022 թ.
Ընդունվել է տպագրության 25.11.2022 թ.