УДК 811.161.3’42

А. М. ПАЛУЯН, А. С. КОРЖ
(г. Гомель, УА “Гомельскі дзяржаўны ўніверсітэт імя Францыска Скарыны”)

ПАРАЎНАЛЬНА-СУПАСТАЎЛЯЛЬНЫ АНАЛІЗ ЭНТРАПІІ МОВЫ МАСТАЦКІХ ТЭКСТАЎ

У артыкуле выкладзены некаторыя аспекты эксперыментальнага падыходу вылічэння энтрапіі мовы мастацкіх тэкстаў, прыводзяцца эксперыментальныя даныя, якія дэманструюць вынікі. Прапануецца лінгва-матэматычная мадэль для аналіза структуры тэксту, пабудаваная на аснове фундаментальнага закона захавання сумы інфармацыі і энтрапіі з выкарыстаннем формулы Шэнана, а таксама даецца параўнальна-супастаўляльны аналіз энтрапійных характарыстык мовы твораў на гістарычную тэму двух беларускіх аўтараў – Уладзіміра Караткевіча і Леаніда Дайнекі. Прапануемая метадалогія заснавана на сістэмным, шматузроўневым падыходзе да пабудовы складанай іерархічнай сістэмы мовы.

Тэрмін энтрапія (стар.-гр. – вяртанне, пераўтварэнне) сёння шырока выкарыстоўваецца ў шэрагу навук. Упершыню гэта паняцце як характарыстыку меры выпадковасці ўвёў ва ўжытак у XIX ст. нямецкі навуковец Р. Клаўзіус [1]. Энтрапію як ступень хаосу, бязладдзя апісаў у 1948 г. К. Шэнан, які, характарызуючы сістэму перадачы даных, выявіў наяўнасць шуму, што непазбежна вядзе да памылкі (скажэння) паведамлення [2, с. 125]. Р. Арнхейм даследаваў энтрапію як меру няпэўнасці [3], якая вылічваецца па формуле верагоднасці з’яўлення ў тэксце элементаў пэўнага кода (сімвалаў, літар). Інфармацыйная каштоўнасць перадаваемага паведамлення залежыць ад ступені нечаканасці яго зместу, энтрапія – ускосны паказчык, які дазваляе ацаніць сэнсавую інфарматыўнасць [4].

Энтрапія цесна звязана з аб’ёмам інфармацыі, заключанай у даных: чым больш яны няпэўныя, тым больш інфармацыі патрабуецца для іх апісання. У дакладных навуках пры ацэнцы эфектыўнасці кадзіравання даных яна вылічваецца па пэўнай формуле [5]. У лінгвістыцы энтрапія разглядаецца як ступень дыскурсіўнай інфармацыі, што ўстанаўліваецца па колькасці лексем. Энтрапія мовы – гэта статыстычная функцыя тэксту на канкрэтнай мове або самой мовы, якая вызначае колькасць інфармацыі на адзінку тэксту [1].

Энтрапія мовы вылічваецца як:

$$S = -\sum_i p_i \ln p_i,$$

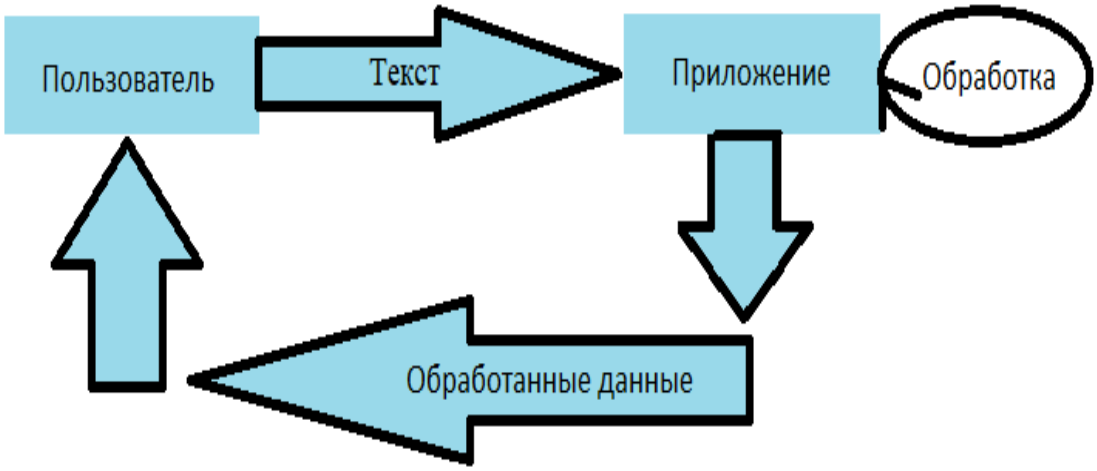
дзе p_i – верагоднасць з’яўлення канкрэтнага сімвала ў тэксце, вызначаемая адносінамі колькасці канкрэтных сімвалаў у тэксце да колькасці ўсіх сімвалаў у тэксце.

Інфармацыйная энтрапія можа прымаць значэнне да ln n, дзе n – колькасць унікальных сімвалаў тэксту. У выпадку вылічэння энтрапіі тэкстаў n будзе роўна колькасці літар у алфавіце канкрэтнай мовы. Для беларускай мовы максімальнае значэнне S будзе роўна ln 32.

Велічыню S можна вызначыць для кожнага канкрэтнага выпадку перадачы інфармацыі, напрыклад, для літаратурнага твора ці запісанай вуснай гаворкі. Тэкст з меншай энтрапіяй трактуецца як больш інфарматыўны ў тым сэнсе, што ён перадае даверануюяму інфармацыю з меншымі скажэннямі. Напрыклад, самым інфарматыўным з’яўляецца тэкст, які складаецца толькі з адной літары, паколькі яго энтрапія роўная нулю. У звычайным разуменні такі тэкст будзе несці, наадварот, мінімальную інфармацыю.

Калі б выкарыстанне асобных сімвалаў адбывалася незалежна, то дастаткова было б абысціся энтрапіяй першага парадку $S^{(1)}$. Насамрэч, у тэксце літары размешчаны не проста так, а з сэнсам, што выяўляецца ва ўпарадкаванні пар літар па частаце іх выкарыстання. Таму даводзіцца разглядаць энтрапіі вышэйшых парадкаў. З іншага боку, і ў мяжы $S^{(2)}$ няма асаблівай неабходнасці, паколькі "памяць" аб мінулых суседзях літары губляюць ужо пры пераходзе да іншага слова. Такім чынам, дастаткова вызначыць $S^{(k)}$ для велічыні k, роўнай сярэдняй даўжыні слова ў лексіконе дадзенай мовы. На жаль, нават для даўжыні $k = 5$ пэўна ацаніць верагоднасці спалучэнняў з 5 літар немагчыма, паколькі мінімальны аб'ём тэксту для такой ацэнкі павінен пераўзыходзіць 10^{10}, а такая даўжыня супастаўляльная з даўжынёй увогуле ўсіх твораў мастацкай літаратуры на адной мове. Таму ў даследаваннях разлічваюць $S^{(1)}$ і $S^{(2)}$.

Для аналізу інфармацыйнай энтрапіі мовы мастацкіх тэкстаў была распрацавана праграма на мове праграмавання Python, якая аналізуе тэкст на беларускай мове, зададзены карыстальнікам. Дадатак выводзіць інфармацыю, што датычыцца інфармацыйнай збыткоўнасці мовы, такую як частата літар, пар літар, камутатыўнасць, абсалютная камутатыўнасць (малюнак 1).



Малюнак 1 – Схема работы праграмы

Для дэманстрацыі работы праграмы было праведзена параўнанне мовы мастацкіх тэкстаў на гістарычную тэму: аповесці Уладзіміра Караткевіча "Сівая легенда", "Цыганскі кароль", "Ладдзя роспачы" і раманы Леаніда Дайнекі "Меч князя Вячкі" і "Жалезныя жалуды".

На першым этапе адбываецца перадача ў праграму тэкставага файла ад карыстальніка, перадача ажыццяўляецца шляхам уводу карыстальнікам назвы файла. Увод ажыццяўляецца з пашырэннем txt. Шлях да візуалізацыі тэксту пачынаецца з папкі, у якой знаходзіцца праграма.

Другі этап уяўляе сабой апрацоўку зададзенага файла: 1) тэкст пераўтвараецца ў радок і пераходзіць у ніжні рэгістр; 2) з радка выдаляюцца ўсе спецыяльныя сімвалы і пераносы; 3) апрацоўваецца тэкст, даныя перадаецца карыстальніку ў тэкставай і візуалізаванай форме.

На трэцім этапе карыстальнік атрымлівае апрацаваную інфармацыю: графік частаты сімвалаў беларускай мовы, верагоднасць іх сустрэчы, колькасць галосных, зычных, усяго сімвалаў (галосныя + зычныя + апостраф), велічыня інфармацыйнай энтрапіі, велічыня інфармацыйнай збыткоўнасці, пары літар, якія сустракаюцца найбольш часта, часцей сустракаемыя пары літар, дзе першая літара "т", часцей сустракаемыя пары літар, дзе другая літара "т", камутатыўнасць "т", абсалютная камутатыўнасць "т".

Пры напісанні праграмы былі выкарыстаныя бібліятэкі: nltk для спрашчэння падліку значэнняў і іх пар шляхам іх аб'яднання па значэнні; string, math, якія з'яўляюцца ўбудаванымі модулямі Python: string выкарыстоўваўся для перадачы тэксту зададзенага карыстальнікам, math для падлікаў розных функцый; re для выкарыстання рэгулярных функцый; matplotlib для выведу атрыманых даных у выглядзе графіка.

На аснове атрыманых даных па частаце літар можна зрабіць выснову, што ў дадзенай выбарцы частата літар практычна аднолькавая, самай частай літарай з'яўляецца літара "а", яе частата ў сярэднім 15,7 %. Энтрапія практычна аднолькавая, аднак яна залежаць ад абранай мовы і для ўсіх тэкстаў на дадзенай мове будзе прыблізна аднолькавая. Выключэннем могуць стаць тэксты, у якіх не выкарыстоўваюцца некаторыя сімвалы з мовы. Камутатыўнасць у мове тэкстаў Уладзіміра Караткевіча мае большае значэнне, чым у мове тэкстаў Леаніда Дайнекі. Абсалютная камутатыўнасць таксама вышэйшая ў творах Уладзіміра Караткевіча.

Частата пар літар таксама практычна аднолькавая, аднак у творах Леаніда Дайнекі сустракаецца больш пар літар, чым у творах Уладзіміра Караткевіча. Частата пар літар з першай "т" і з другой "т" у сярэднім аднолькавая. З першай літарай "т" самай частай парай з'яўляецца пара "та", якая мае сярэднюю частату 36,69 % у творах У. Караткевіча і 33,6 % у творах Л. Дайнекі.

Далей было праведзена параўнанне частот літарных пар з такімі адметнымі сімваламі беларускай мовы, як апостраф і ў (у нескладовае).

На аснове частотнасці літарных пар з апострафам можна зрабіць выснову, што ў творах Леаніда Дайнека павелічэнне верагоднасцяў асобных пар ідзе раўнамерна. Разам з гэтым, ва Уладзіміра Караткевіча пара апостраф+я ва ўсіх прааналізаваных творах сустракаецца ў больш чым палове выпадкаў.

На аснове частоты літарных пар з другім апострафам можна зрабіць выснову, што хоць частата ва Уладзіміра Караткевіча і стала больш раўнамернай, але ў творы "Сівая легенда" ўсё яшчэ бачная велізарная розніца ў суседніх частотах. Варта таксама адзначыць, што разнастайнасць пар у мове твораў Леаніда Дайнекі значна вышэйшая, чым у творах Уладзіміра Караткевіча.

На аснове частоты літарных пар з першым ў (у нескладовым) можна зрабіць выснову, што ва ўсіх выпадках частоты пар раўнамерна павялічваюцца, за выключэннем пары "ўс", якая мае відавочную перавага над астатнімі і займае, як мінімум, трэць усіх пар з першым ў (у нескладовым).

Варта адзначыць, што ў творах Леаніда Дайнекі разнастайнасць пар з другім апострафам і з першым ў (у нескладовым) вышэйшая, чым у мове твораў Уладзіміра Караткевіча.

На аснове частаты літарных пар з другім ў (у нескладовым) можна зрабіць выснову, што іх частоты і рангі практычна ідэнтычныя, за выключэннем мовы твора Уладзіміра Караткевіча "Ладдзя Роспачы".

На аснове прыведзенага вышэй частотнага аналізу літарных пар можна зрабіць агульную выснову па мове твораў Уладзіміра Караткевіча і Леаніда Дайнекі: у сярэднім частоты літар і іх рангі ў творах абодвух аўтараў падобныя, аднак ва Уладзіміра Караткевіча часам прыкметныя ўсплескі частот канкрэтных пар. Разам з гэтым, разнастайнасць літарных пар у мове твораў Леаніда Дайнекі прыкметна вышэйшая, чым ва Уладзіміра Караткевіча.

Такім чынам, намі было даследавана паняцце энтрапіі мовы праведзены аналіз мовы мастацкіх тэкстаў з пункту гледжання частаты літар і пар літар, якія ў іх сустракаюцца. На падставе атрыманых даных можна зрабіць вывад, што энтрапія аднолькавых па спецыфіцы тэкстаў (у нашым выпадку мастацкіх тэкстаў на гістарычную тэму) мала адрозніваецца.

Спіс выкарыстанай літаратуры

1 Orphal J. Rudolph Clausius (1822–1888) and His Concept of Mathematical Physics // Ann. Phys. 2022. Vol. 535, № 1. Art. № 2200065. DOI: 10.1002/andp.202200065

2 Shannon C.E. A Mathematical Theory of Communication // Bell Syst. Tech. J. 1948. Vol. 27, № 3. P. 379–423.

3 Arnheim R. Order and Complexity in Landscape Design // Arnheim R. Towards the Psychology of Art. Berkeley: University of California Press, 1966. P. 123–135.

4 Коротаев С.М. Энтропия и информация – универсальные естественно-научные понятия. URL: https://textarchive.ru/c-2201579 (дата звароту: 16.01.2025).

5 Шелестюк Е.В., Щетинкина Е.А. Стохастичность и энтропия в лингвистике // Вестн. Челяб. гос. ун-та. 2023. № 2(472). С. 150–165.

The article describes some aspects of the experimental approach to calculating the entropy of literary texts, and provides experimental data demonstrating the results. A linguistic and mathematical model for analyzing the structure of a text is proposed, based on the fundamental law of conservation of the sum of information and entropy using the Shannon formula, and a comparative analysis of the entropy characteristics of the language of works on a historical topic by two Belarusian authors– Vladimir Korotkevich and Leonid Daineko, is given. The proposed methodology is based on a systematic, multi-level approach to building a complex hierarchical language system.