АЎТАМАТЫЧНАЕ ВЫЯЎЛЕННЕ ТЭХНІЧНАЙ ЛЕКСІКІ ПРЫ СКЛАДАННІ ІНФАРМАЦЫЙНА-ПОШУКАВАГА ТЭЗАЎРУСА

М. Д. Мурашка[1)], **М. У. Буракова**[2)]

[1)] *Гомельскі дзяржаўны тэхнічны ўніверсітэт імя П. В. Сухога, пр-т Кастрычніка, 48, 246746, г. Гомель, Беларусь, burakova_75@bk.ru*
[2)] *Гомельскі дзяржаўны тэхнічны ўніверсітэт імя П. В. Сухога, пр-т Кастрычніка, 48, 246746, г. Гомель, Беларусь*

У дакладзе асноўная ўвага скіравана на праблему аўтаматычнага выяўлення тэхнічнай лексікі пры складанні інфармацыйна-пошукавага тэзаўруса (ІПТ). Крыніцай збору для выяўлення тэхнічных тэрмінаў паслужылі навукова-тэхнічныя тэксты і спецыяльныя слоўнікі, мадэляванне якіх дае магчымасць выявіць шэраг колькасных і якасных параметраў тэхнічных тэрмінаў, што дазваляе скласці тэрміналагічную базу для адбору патэнцыяльных адзінак ІПТ. Распрацавана методыка выяўлення тэхнічнай лексікі ў ІПТ з улікам выдзеленых колькасных параметраў тэрміналогіі, якая дазваляе максімальна выкарыстоўваць гатовы прадукт – тэхнічную тэрміналогію, зафіксаваную ў навукова-тэхнічных тэкстах і слоўніках, і павысіць якасць ІПТ.

Ключавыя словы: аўтаматычнае выяўленне; інфармацыйна-пошукавы тэзаўрус; тэхнічныя тэрміны; мадэляванне.

Адметнасцю навукова-тэхнічнага прагрэсу ў наш час з’яўляецца паскораны рост аб’ёму інфармацыйнага патоку. Спыніць рост колькасці інфармацыі немагчыма. Спецыялізацыя даследаванняў пашыраецца, што прыводзіць да павелічэння аб’ёмаў друкаваных матэрыялаў. Цяжкасць арыентацыі ў вялікай колькасці недастаткова ўпарадкаваных матэрыялаў часта прыводзіць да дубліравання даследаванняў, якое суправаджаецца незваротнымі стратамі часу, матэрыяльных сродкаў і інтэлектуальных рэсурсаў. Таму вылучэнне пошуку інфармацыі ў самастойную праблему з’яўляецца выключна важнай неабходнасцю.

На сучасным этапе развіцця аўтаматызаваных сістэм навукова-тэхнічнай інфармацыі рашэнне праблемы інфармацыйнага пошуку магчыма толькі пры ўмове стварэння спецыяльных слоўнікаў – інфармацыйна-пошукавых тэзаўрусаў (ІПТ). Без тэзаўрусаў аўтаматызаваныя сістэмы апрацоўкі інфармацыі не могуць у поўным аб’ёме і якасна выканаць складаныя аперацыі па аналізе зместу навукова-тэхнічных тэкстаў. Пры гэтым сучасныя інфармацыйна-пошукавыя сістэмы (ІПС) патрабуюць удасканалення структуры тэзаўрусаў.

Мэтай даклада з’яўляецца спроба распрацоўкі і даследавання метаду аўтаматычнага выяўлення тэхнічнай лексікі пры складанні інфармацыйна-пошукавага тэзаўруса. У сувязі з гэтым была пастаўлена задача стварэння

тэрміналагічнага банка даных (ТБД) для вучэбнага руска-беларускага ІПТ тэхнічных тэрмінаў, які будзе прызначаны для студэнтаў тэхнічнага ўніверсітэта. Работа з ІПТ такога віду з'яўляецца неабходнай часткай вучэбнага працэсу на занятках па дысцыпліне «Беларуская мова (прафесійная лексіка)» і будзе спрыяць арганізацыі і прадстаўленню тэхнічнай тэрміналогіі і спецыяльнай лексікі для фарміравання лексічнай кампетэнцыі навучэнцаў.

З'яўленне ІПТ непарыўна звязана з развіццём аўтаматызаваных інфармацыйных сістэм (АІС). Першапачаткова мэтай стварэння ІПТ з'яўлялася павышэнне паказчыкаў якасці пошуку інфармацыі ў АІС. У адпаведнасці з гэтай мэтай прызначэнне ІПТ заключалася ў наступным:

1) забяспечваць індэксаванне дакументаў і запытаў сродкамі дэскрыптарнай мовы шляхам замены ключавых слоў адпаведнымі дэскрыптарамі;

2) адлюстроўваць парадыгматычныя адносіны, якія існуюць паміж лексічнымі адзінкамі пэўнай галіны навукі або тэхнікі;

3) служыць сродкам кантролю і нармалізацыі лексікі канкрэтнай галіны тэхнікі, забяспечваць адзінае і фармалізаванае прадстаўленне інфармацыі ў ІПС.

На практыцы, у інструктыўна-метадычнай літаратуры, існуе вялікая блытаніна пры вызначэнні паняцця ІПТ (тэзаўрус). Тэзаўрусам часам называюць любую класіфікацыю, любы рубрыкатар ці нават спіс. Тым не менш неабходна адрозніваць ІПТ ад камп'ютарных слоўнікавых спісаў узаемаразмяшчэння тэрмінаў у дакументах, якія часта ў літаратуры называюць аўтаматызаванымі тэзаўрусамі; ад спісаў прадметных загалоўкаў і ключавых слоў, калі ў іх не вызначаны семантычныя адносіны паміж тэрмінамі.

Праведзены намі аналіз навуковых прац па вызначэнні ІПТ дае магчымасць сцвярджаць, што інфармацыйна-пошукавы тэзаўрус – гэта структураваны слоўнік для кантролю лексікі, у якім відавочна і сістэмна вызначаюцца асноўныя семантычныя адносіны (эквівалентнасці, іерархічныя і асацыятыўныя) паміж тэрмінамі натуральнай мовы [1–4].

Тэзаўрусы прымяняюцца ў якасці інструмента тэрміналагічнага кантролю ў працэсе аналізу і індэксавання дакументаў і інфармацыйных запытаў, а таксама ў працэсе аўтаматызаванага пошуку інфармацыі. Функцыянальная роля тэзаўруса ў ІПС заключаецца ў прад'яўленні высокіх патрабаванняў да якасці падрыхтоўкі тэзаўруса, ад ступені дасканаласці якога ў асноўным залежыць эфектыўнасць пошуку.

Распрацоўка методыкі адбору лексікі ў ІПТ з улікам выдзеленых колькасных параметраў тэрміналогіі дазваляе максімальна выкарысто-

ўваць гатовы прадукт – тэхнічную тэрміналогію, зафіксаваную ў навукова-тэхнічных тэкстах і спецыяльных слоўніках. Методыка даследавання заключаецца ў сістэмным падыходзе да аналізу лексікі, выкарыстанні метадаў мадэлявання, прымяненні апарату тэорыі графаў і тэорыі мностваў для фармальнага апісання задач, змястоўнай (істотнай) інтэрпрэтацыі вынікаў.

Так, існуе некалькі метадаў вылучэння тэхнічнай лексікі ў ІПТ:

✓ *метад частотнасці, заснаваны на тым, што тэхнічныя тэрміны часта сустракаюцца ў тэкстах, звязаных з канкрэтнай тэмай або галіной ведаў;*

✓ *метад марфалагічнага аналізу, заснаваны на аналізе граматычных характарыстык слоў, такіх як склон, лік, час і інш.;*

✓ *метад аналізу кантэксту, заснаваны на аналізе кантэксту, у якім ужываецца слова;*

✓ *метад машыннага навучання, заснаваны на выкарыстанні алгарытмаў машыннага навучання для аўтаматычнага выдзялення тэхнічнай лексікі;*

✓ *метад семантычнага аналізу, заснаваны на аналізе значэння слова і яго сувязей з іншымі словамі ў сказе;*

✓ *метад камбінаванага аналізу, які выкарыстоўвае камбінацыю розных вышэйпералічаных метадаў.*

Кожны з названых метадаў мае свае перавагі і свае недахопы і можа быць выкарыстаны ў залежнасці ад канкрэтнай задачы і ўмоў. Напрыклад, калі неабходна хутка вылучыць тэхнічныя тэрміны з невялікай колькасці тэкстаў, то можна прымяняць метад частотнасці або метад марфалагічнага аналізу. Калі ж трэба вылучыць тэхнічныя тэрміны з вялікага корпуса тэкстаў з высокай дакладнасцю, то найбольш карысна выкарыстоўваць метад машыннага навучання.

Асобна кожны метад ці ў спалучэнні можа быць ужыты для вылучэння пэўных тыпаў тэхнічных тэрмінаў, што дазваляе атрымаць больш поўны і дакладны тэзаўрус.

Працэс пабудовы тэзаўруса незалежна ад метаду ўключае наступныя этапы:

1) папярэдні адбор лексічных адзінак (складанне спісаў ключавых слоў, слоўнікаў);

2) пабудова класаў умоўнай эквівалентнасці (прывядзенне лексічных адзінак да адпаведнай стандартнай формы);

3) выяўленне зададзеных семантычных адносін.

З улікам вышэйназваных агульнапрынятых лексікаграфічных прынцыпаў і палажэнняў намі быў распрацаваны алгарытм работы з вучэбным

матэрыялам (падручнікамі, вучэбнымі дапаможнікамі, вучэбна-метадычным комплексам, слоўнікам і даведачным дапаможнікам) з мэтай адбору і апісання тэхнічнай лексікі ў межах дысцыплін прафесійнага цыкла, якія вывучаюцца на пэўным этапе навучання. Трэба адзначыць, што ў адпаведнасці з вучэбным планам дысцыпліна «Беларуская мова (прафесійная лексіка)» выкладаецца на розных курсах (1–4) першай ступені вышэйшай адукацыі, а значыць, авалодванне тэрміналогіяй электратэхнікі будзе залежаць ад узроўню падрыхтоўкі навучэнца і вызначэння аб'ёму спецыяльных паняццяў, фарміравання ўяўленняў пра тэрміналагічную сістэму прадметнай галіны, якая вывучаецца на пэўным курсе.

Стварэнне ТБД тэхнічных тэрмінаў адбывалася ў адпаведнасці з наступнымі этапамі:

1) адбор крыніц на базе рускамоўных вучэбных дапаможнікаў, тэрміналагічных слоўнікаў і даведнікаў адпаведнай тэхнічнай галіны і яе раздзелаў;

2) логіка-паняційны аналіз адабранага матэрыялу з мэтай фарміравання корпуса слоўніка ў адпаведнасці з абранай структурай;

3) пошук адпаведнага эквівалента слова рускай мовы ў беларускай;

4) складанне тэрміналагічных запісаў артыкулаў тэрмінаў, дэфініцый на рускай і беларускай мовах.

Вынік такой работы – складанне лексікаграфічных артыкулаў, ілюстрацыяй з якіх можа быць наступны:

ЭЛЕКТРАТЭХНІКА
ЭЛЕКТРЫЧНЫЯ ЛАНЦУГІ

(1) Активная мощность || Актыўная магутнасць

(2) деф.: *средняя за период мощность цепи, характеризующая среднюю скорость необратимого преобразования электрической энергии в тепловую, световую, механическую, химическую и другие виды энергии* [5] || *сярэдняя за перыяд магутнасць ланцуга, якая характарызуе сярэднюю хуткасць незваротнага пераўтварэння электрычнай энергіі ў цеплавую, светлавую, механічную, хімічную і іншыя віды энергіі.*

(1) Активное сопротивление || Актыўнае супраціўленне

(2) деф.: *сопротивление r резистивного элемента переменному току, в котором электрическая энергия преобразуется в тепло* [5] || *супраціўленне r рэзістыўнага элемента пераменнаму току, у якім электрычная энергія пераўтвараецца ў цяпло.*

(1) Выпрямитель || Выпрамнік

(2) деф.: *устройство для преобразования переменного тока в постоянный* [5] || *прыстасаванне для пераўтварэння пераменнага току ў пастаянны.*

(1) Газоразрядные приборы || Газаразрадныя прыборы

(2) деф.: *ионные приборы, действие которых основано на электрическом разряде в газе или парах металла* [5] || *іонныя прыборы, дзеянне якіх заснавана на электрычным разрадзе ў газе або парах металу.*

(1) Стационарное электрическое поле || Стацыянарнае электрычнае поле

(2) деф.: *электрическое поле в проводнике, если ток в проводнике с течением времени не изменяется* [6] || *электрычнае поле ў правадніку, калі ток у правадніку з цягам часу не змяняецца.*

(1) Схема замещения цепи || Схема замяшчэння ланцуга

(2) деф.: *схема электрической цепи, которую составляют для расчета режима работы цепи* [6] || *схема электрычнага ланцуга, якую складаюць для разліку рэжыму працы ланцуга.*

Такім чынам, прадстаўленая распрацоўка аўтаматызаванай сістэмы аналізу тэхнічнай тэрміналогіі, якая ажыццяўляе пабудову лексічнай сеткі тэрміналогіі, дае аналіз лексічнай сеткі, вылічэнне шэрагу колькасных характарыстык і параметраў тэрміналагічнай лексікі, дазваляе выявіць і прадставіць аб’ектыўны адбор лексікі ў тэзаўрус, вырашае шэраг тэарэтычных і прыкладных задач адносна аналізу і карэкцыі як тэрміналагічных слоўнікаў, так і інфармацыйна-пошукавага тэзаўруса.

У сувязі з вялікай працаёмкасцю стварэння тэрміналагічных слоўнікаў усё большае прымяненне ў апошні час знаходзяць аўтаматызаваныя метады падрыхтоўкі слоўнікаў. Аўтаматызацыя лексікаграфічных работ прывяла да стварэння тэрміналагічных банкаў даных, у якіх з’яўляецца магчымасць назапашваць і хутка апрацоўваць вялікія аб’ёмы тэрміналагічнай інфармацыі. Стварэнне ТБД патрабуе вялікіх рэсурсаў, а змяненне складу даных (фармату) ТБД звязана з вялікімі аб’ёмамі работ. Таму надзвычай важным з’яўляецца папярэдняе даследаванне складу тэрміналагічнай інфармацыі.

Бібліяграфічныя спасылкі

1. *Кобрин Р. Ю.* Терминосистема как информационный язык. К проблеме построения ИПС на естественном языке // Семиотические проблемы языков науки, терминологии и информатики : сборник / отв. А. Г. Волков. М. : Изд-во Моск. ун-та, 1971. С. 640–756.

2. *Колчинский М. Л.* Автоматизированная система информационного обслуживания «Сетка» // Информ. бюл. выставки-смотра «НТИ-74». 1974. № 3. С. 25–28.

3. *Копылов В. А.* Построение автоматизированных информационно-поисковых систем. М. : Энергия, 1974.

4. *Королев Э. И.* О типологии языков автоматизированных информационных систем // Современное состояние теории и практики машинного перевода и автоматизации информационных процессов / редкол.: Ю. Н. Марчук (отв. ред.) [и др.]. М. : ВЦП, 1977. С. 73–87.

5. *Бензарь В. К.* Словарь-справочник по электротехнике, промышленной электронике и автоматике. Минск : Выш. шк., 2009.

6. *Матвиенко В. А.* Основы теории цепей : учеб. пособие для вузов. Екатеринбург : УМЦ УПИ, 2016.