ISSN 1816-0301 (print)
УДК 004.912:025.045

Паступіла ў рэдакцыю 19.01.2018
Received 19.01.2018

С. І. Лысы[1], Г. Р. Станіславенка[2], Ю. С. Гецэвіч[1]

[1]Аб'яднаны інстытут праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі, Мінск, Беларусь
[2]Цэнтр даследаванняў беларускай культуры, мовы і літаратуры Нацыянальнай акадэміі навук Беларусі, Мінск, Беларусь

АЎТАМАТЫЗАВАНАЯ ГЕНЕРАЦЫЯ АЛФАВІТНА-ПРАДМЕТНАГА ПАКАЗАЛЬНІКА УНІВЕРСАЛЬНАЙ ДЗЕСЯТКОВАЙ КЛАСІФІКАЦЫІ НА БЕЛАРУСКАЙ МОВЕ

Анатацыя. Універсальная дзесятковая класіфікацыя (УДК) з'яўляецца міжнароднай класіфікацыйнай сістэмай, якая адпавядае найбольш істотным патрабаванням да класіфікацый: міжнароднасць, універсальнасць, пашыральнасць. Табліцы УДК былі перакладзены і апублікаваны цалкам ці часткова на больш чым 40 мовах, а выкарыстоўваецца УДК прыкладна ў 130 краінах свету. На тэрыторыі Беларусі УДК дзейнічае на працягу апошніх 50 гадоў. Аднак толькі ў 2016 г. з'явілася афіцыйнае выданне УДК на беларускай мове. Алфавітна-прадметны паказальнік (АПП), які складае больш за чвэрць выдання, быў падрыхтаваны пры дапамозе алгарытму, які аўтаматызаваў працэс яго стварэння.

У артыкуле падрабязна апісваецца падыход да аўтаматызаванага стварэння АПП беларускамоўнага выдання УДК. Разглядаюцца патрабаванні да АПП, параўноўваюцца фарматы АПП, выкарыстаныя ў выданнях УДК розных краін. Таксама апісваюцца электронныя лінгвістычныя рэсурсы, неабходныя для аўтаматызаванай генерацыі АПП, прыводзіцца падрабязны пакрокавы алгарытм. У якасці падцвярджання працаздольнасці і карэктнасці працы дадзенага алгарытму прыводзіцца распрацаваны праграмны прататып сістэмы генерацыі АПП УДК. Дадзены прататып быў непасрэдным чынам выкарыстаны пры стварэнні АПП першага выдання Універсальнай дзесятковай класіфікацыі на беларускай мове.

Ключавыя словы: апрацоўка электронных тэкстаў, сістэмы класіфікацыі, беларуская мова, Універсальная дзесятковая класіфікацыя, алфавітна-прадметны паказальнік

Для цытавання. Лысы, С. І. Аўтаматызаваная генерацыя алфавітна-прадметнага паказальніка Універсальнай дзесятковай класіфікацыі на беларускай мове / С. І. Лысы, Г. Р. Станіславенка, Ю. С. Гецэвіч // Информатика. – 2018. − Т. 15, № 2. – С. 45–54.

S. I. Lysy[1], H. R. Stanislavenka[2], Yu. S. Hetsevich[1]

[1]The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus
[2]Center for the Belarusian Culture, Language and Literature Researches of the National Academy of Sciences of Belarus, Minsk, Belarus

AUTOMATED ALPHABETIC SUBJECT INDEX GENERATION FOR UNIVERSAL DECIMAL CLASSIFICATION IN BELARUSIAN

Abstract. Universal Decimal Classification (UDC) is an international classification system that corresponds to the essential classification requirements: internationalism, universality, expansiveness. UDC Tables were translated and published fully or partially in more than 40 languages. UDC is used in nearly 130 countries. In Belarus UDC has been used for the last 50 years. But the official edition of UDC in the Belarusian language was released only in 2016. An alphabetic subject index (ASI) that makes more than a quarter of the edition was prepared with the help of an algorithm, which automated its creation.

In the main part of the article an approach to automated ASI generation for Belarusian UDC edition is described in detail. Authors review ASI requirements, compare ASI formats used in different countries. Electronic linguistic resources needed for the automated generation of ASI are described in the article. The detailed step-by-step algorithm is provided. As confirmation of efficiency and correctness of the algorithm, a program prototype of ASI generation system was developed and described in the article. This prototype was directly used in ASI creation for the first edition of UDC in the Belarusian language.

Keywords: electronic texts processing, classification systems, the Belarusian language, Universal Decimal Classification, alphabetic-subject index

For citation. Lysy S. I., Stanislavenka H. R., Hetsevich Yu. S. Automated alphabetic subject index generation for Universal Decimal Classification in Belarusian. *Informatics*, 2018, vol. 15, no. 2, pp. 45–54 (in Belarusian).

Уводзіны. У XIX ст. бельгійскія бібліёграфы Поль Атлé (фр. *Paul Otlet*) і Анры Лафантэ́н (фр. *Henri La Fontaine*) дапрацавалі «Дзесятковую класіфікацыю» Мэлвіла Дзьюі (англ. *Melvil Dewey*), якую апошні стварaў для Бібліятэкі Кангрэса ў ЗША, і заклалі пачатак развіцця Універсальнай дзесятковай класіфікацыі [1].

На сённяшні дзень распрацаваная бельгійцамі класіфікацыя адпавядае найбольш істотным патрабаванням да класіфікацый – міжнароднасці і ўніверсальнасці. Міжнароднасць падцвярджаецца тым фактам, што табліцы УДК былі перакладзены і апублікаваны цалкам ці часткова на больш чым 40 мовах, а выкарыстоўваецца УДК прыкладна ў 130 краінах свету. Універсальнасць класіфікацыі заключаецца ў дзесятковым прынцыпе яе пабудовы: дзяленні аднаго класа на дзесяць (ці менш) падкласаў. Дзесятковы прынцып дазваляе практычна неабмежавана пашыраць яе шляхам дадавання новых лічбаў да ўжо існуючых, не ламаючы ўсёй сістэмы ў цэлым, і з дапамогай шматлікіх сродкаў і прыёмаў індэксіравання выкарыстоўваць УДК для сістэматызацыі і наступнага пошуку разнастайных крыніц інфармацыі – ад вузкатэматычных выданняў спецыяльнай дакументацыі да вялікіх галіновых і шматгаліновых даведачна-інфармацыйных фондаў. Гэта дазваляе наблізіцца да рашэння праблемы пошуку дадзеных у вялікім інфармацыйным патоку і іх сістэматызацыі.

На тэрыторыі Беларусі УДК была ўведзена ў якасці адзінай сістэмы класіфікацыі для тэхнічных бібліятэк і органаў навукова-тэхнічнай інфармацыі ў 1963 г. [2]. На працягу наступных 30 гадоў канкурэнцыю УДК стваралі такія сістэмы класіфікацыі, як ББК (Бібліятэчна-бібліяграфічная класіфікацыя) і АКЛ (Адзіная класіфікацыя літаратуры). Аднак пасля распаду СССР і атрымання Беларуссю незалежнасці УДК замацавалася ў якасці асноўнай інфармацыйна-пошукавай сістэмы ў нашай краіне. У 1993 г. Нацыянальная кніжная палата Беларусі для ўнутранага карыстання пераклала на беларускую мову і выдала скарочаныя табліцы УДК [2]. Нягледзячы на тое, што больш за дзве сотні ўстаноў у Беларусі выкарыстоўвалі і выкарыстоўваюць у сваёй працы УДК, да 2016 г. афіцыйнага выдання на беларускай мове не існавала. Аўтары дадзенага артыкула ў супрацоўніцтве з Нацыянальнай бібліятэкай Беларусі ўдзельнічалі ў падрыхтоўцы выдання перакладу табліц УДК на беларускую мову [3] і, у прыватнасці, займаліся стварэннем алфавітна-прадметнага паказальніка да УДК, які з’яўляецца неад’емнай яго часткай і важным элементам у працэсе пошуку інфармацыі па табліцах УДК.

З улікам вялікага аб’ёму табліц УДК і іх перыядычнага дапаўнення і ўдасканалення хуткае і якаснае стварэнне АПП, а таксама яго дынамічная актуалізацыя могуць быць магчымымі толькі пры дапамозе сучасных інфармацыйных тэхналогій. Па гэтых прычынах было вырашана распрацаваць алгарытм, які дазваляе з масіву запісаў УДК аўтаматызавана згенераваць масіў адпаведнасцяў «тэрміналагічная адзінка – каардынацыйны код/індэкс». Такім чынам, перад аўтарамі паўстала задача распрацоўкі алгарытму пераўтварэння тэкстаў табліц УДК у АПП. Ён традыцыйна ўяўляе сабой спіс слоў, адсартаваных у алфавітным парадку разам з індэксамі УДК, паводле якіх можна адшукаць той ці іншы тэрмін у табліцах УДК.

Дадзены алгарытм і распрацаваны на яго аснове прататып сістэмы генерацыі АПП пакліканы паспрыяць у дасягненні наступных мэт:

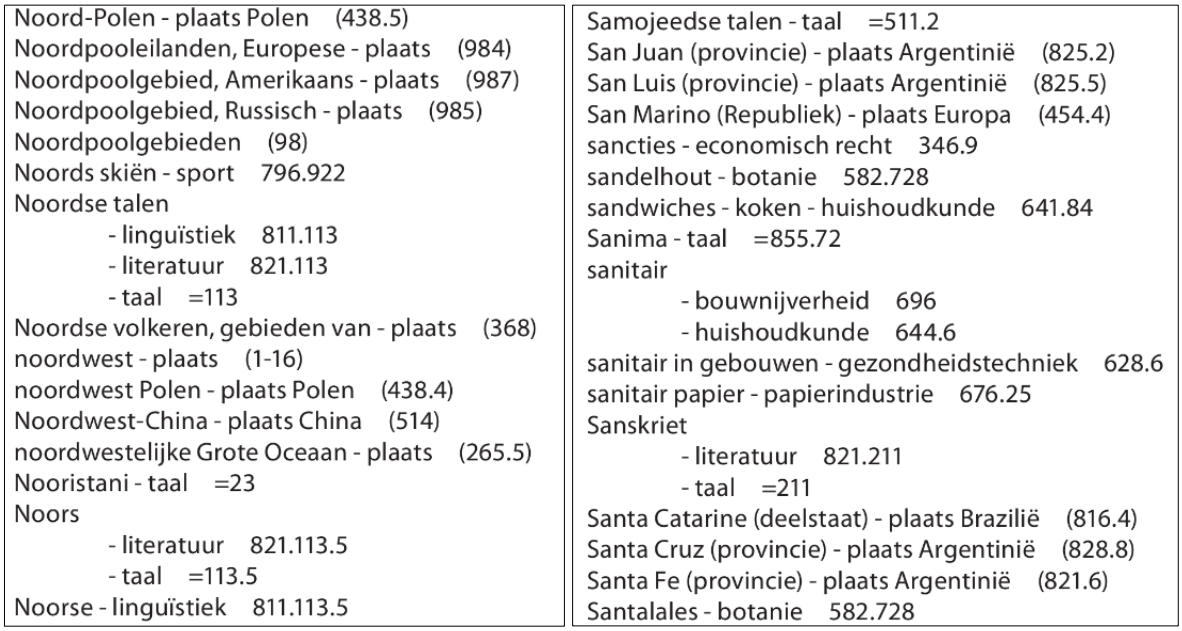
– стварэння першага выдання табліц УДК на беларускай мове;

– папаўнення электронных лінгвістычных рэсурсаў спецыялізаванай і вузкатэрміналагічнай лексікай, якая выкарыстоўваецца ва УДК;

– удасканалення алгарытмаў апрацоўкі электронных тэкстаў вялікіх аб’ёмаў на беларускай мове.

Паняцце алфавітна-прадметнага паказальніка і патрабаванні да яго. Алфавітна-прадметны паказальнік з’яўляецца важным складнікам сістэмы УДК і арыентаваны на выкананне каардынацыйнай і пошукавай функцыі. АПП выкарыстоўваецца для спрашчэння і паскарэння тэрміналагічнага і прадметнага пошуку ў межах асноўных табліц УДК у працэсе індэксавання літаратуры для забеспячэння максімальнай уніфікацыі і адзінства індэксавання.

АПП уяўляе сабой масіў адпаведнасцяў «тэрміналагічная адзінка – каардынацыйны код/індэкс» (мал. 1). Кожная адпаведнасць уяўляе сабой асобную прадметную рубрыку.



Мал. 1. Фрагменты алфавітна-прадметнага паказальніка галандскага выдання УДК [4]

Варта адзначыць, што фармат галандскага АПП не з’яўляецца адзіным магчымым. Напрыклад, А. А. Сербін у артыкуле [5] прыводзіць наступны спіс спосабаў падачы прадметных рубрык:

1. Каардынацыйная рубрыка тлумачальнага характару, дзе тлумачэнне тэрміну асвятляецца з пункту гледжання вобласці прымянення.

Прыклад: *Ацэталі (хімія) 547–316*

2. Каардынацыйная рубрыка простай формы семантычнага прадстаўлення.

Прыклад: *Электроніка 621.3*

3. Каардынацыйная рубрыка аспектнага характару, у якой указваецца аспект прымянення дадзенага тэрміну.

Прыклад:

Кракадзілы:

– (Заалогія) 598.14

– (Палеазаалогія) 568.14

– (Паляванне) 639.14

4. Каардынацыйна-арыентаваная рубрыка з аспектна-прадметнай дэталізацыяй, дзе асвятляецца аспектнасць і прадметнасць аспекта дадзенага тэрміну.

Прыклад:

Тэрмадынаміка

– Атмасферы (метэаралогія) 551.511.33

– Біялагічных працэсаў (біяфізіка) 577.31

– Зямлі (геафізіка) 550.36

– Хімічная (фіз. хімія) 544.3

Па разглядзе фарматаў АПП, выкарыстаных у галандскім [4] і рускім выданнях УДК, для беларускамоўнага выдання было вырашана выкарыстоўваць каардынацыйныя рубрыкі аспектнага характару для слоў, якія адносяцца адначасова да некалькіх абласцей прымянення, ва ўсіх жа астатніх выпадках – падаваць каардынацыйныя рубрыкі тлумачальнага характару (мал. 2).

Жалезабетон *(будаўніцтва)* 693.5
Жалоба *(антрапалогія)* 393.7
Жаніцьба *(антрапалогія)* 392.5
Жанр
(мастацтва) 7.041
(кіно) 791.2, 791.22, 791.223
(літаратура) 82-1/-9, 82-94, 82-95
Жанчына
(антрапалогія) 392.6
(медыцына) 618.16, 618.17, 618.18
Жарганізм *(мовазнаўства)* 81`373.48
Жарт *(антрапалогія)* 398.25, 398.6
Жаўтушка *(заалогія)* 598.296
Жах *(кіно)* 791.221.9
Жвачныя
(заалогія) 599.735
(сельская гаспадарка) 636.2
Жвір *(будаўніцтва)* 691.22

Снарад
(фізіка) 531.55
(тэхналогіі) 623.4, 623.46, 629.76, 629.762
Снег
(геалагічныя навукі) 551.322, 551.578.4, 556.12
(тэхналогіі) 624.14
(прамысловасць) 685.36
Снегіровыя сойкі *(заалогія)* 598.293
Снэпшот *(фатаграфія)* 77.055
Совападобныя *(заалогія)* 598.27, 598.279, 598.279.4
Сода *(хімія)* 661.3
Сойка *(заалогія)* 598.293
Сок *(хімія)* 547.914, 663.8, 663.81
Сокалападобныя *(заалогія)* 598.27, 598.279, 598.279.3
Солад *(хімія)* 663.032
Соль
(хімія) 54-38, 549.451.1, 549.89, 661.74, 661.8, 664.41
(геалагічныя навукі) 552.53, 553.63, 553.78
(прыкладное мастацтва) 745.56

Мал. 2. Фрагменты алфавітна-прадметнага паказальніка беларускага выдання УДК [3]

Для карэктнай працы алгарытму аўтаматызаванай генерацыі АПП і паслядоўнага выканання патрабаванняў да АПП было неабходна распрацаваць адмысловыя лінгвістычныя рэсурсы.

Электронныя лінгвістычныя рэсурсы для аўтаматызаванай генерацыі АПП. АПП з'яўляецца алфавітна ўпарадкаваным спісам паняццяў, якія сустракаюцца ў класах асноўных табліц УДК. Гэтыя паняцці прыводзяцца ў пачатковай форме разам з кодам адпаведных класаў УДК. Для карэктнага функцыянавання алгарытму аўтаматызаванай генерацыі АПП было неабходна атрымаць або распрацаваць адмысловыя лінгвістычныя і тэматычныя рэсурсы, сярод якіх:

– спіс стоп-класаў (класаў УДК, якія не выкарыстоўваюцца пры фарміраванні АПП);

– спіс стоп-слоў (слоў, прысутных у табліцах УДК, якія не павінны выкарыстоўвацца пры фарміраванні АПП);

– спіс адпаведнасцяў «код класа – дамен»;

– слоўнік лексікі, які змяшчае ўсе словы, якія сустракаюцца ва УДК, з іх пачатковымі формамі, пазнакамі націску і граматычнымі характарыстыкамі.

У той час, як спіс стоп-класаў быў атрыманы ад Кансорцыума УДК (*UDC Consortium*) (мал. 3, *а*), астатнія тры рэсурсы былі адмыслова распрацаваны. У спіс стоп-слоў увайшлі словы, якія маюць агульны характар і не валодаюць асаблівай тэрміналагічнай спецыфікай. Гэта такія словы, як «асноўны», «агульны», «другі», «кожны» і інш. Спіс адпаведнасцяў «код класа – дамен» прызначаны для фарміравання рубрык аспектнага і тлумачальнага характару, якія прадугледжваюць указанне вобласці прымянення таго ці іншага тэрміну. Дадзены спіс таксама фарміраваўся ўручную і складаўся са скарочанага кода класа і адпаведнай вобласці прымянення, або дамена (мал. 3, *б*).

7.049
7.06
7.08
7.094
711.1
72.01/.05
725
725.22
725.89
726.9
728.8
736

а)

72 (архітэктура)
73 (пластычнае мастацтва)
74 (прыкладное мастацтва)
75 (жывапіс)
76 (графіка)
77 (фатаграфія)
78 (музыка)
79 (забавы)
80 (мовазнаўства)
81 (мовазнаўства)
82 (літаратура)

б)

Мал. 3. Фрагменты спісаў: *а)* стоп-класаў; *б)* адпаведнасцяў «код класа – дамен»

Для атрымання электроннага слоўніка лексікі УДК былі здзейснены наступныя крокі:

1. Складанне спісу ўнікальных слоў асноўных табліц УДК з іх кантэкстамі пры дапамозе вэб-сэрвіса «Падлік частотнасці слоў» (http://corpus.by/WordFrequencyCounter, http://ssrlab.by/1457).

2. Вылучэнне слоў, якія адсутнічаюць у даступных электронных граматычных слоўніках беларускай мовы. Для гэтай мэты быў выкарыстаны інтэрнэт-сэрвіс «Праверка правапісу» (http://corpus.by/SpellChecker, http://ssrlab.by/3334).

3. Ручное складанне спісу адпаведнасцяў «слова – пачатковая форма» для слоў, адсутнічаючых у наяўных граматычных слоўніках (мал. 4, *а*).

4. Ручная пастаноўка націскаў у складзеным спісе (мал. 4, *б*).

5. Удасканаленне слоўніка: генерацыя парадыгмаў слоў па пачатковай і некалькіх ускосных формах з атрыманнем граматычных характарыстык (http://corpus.by/ WordParadigm Generator, http://ssrlab.by/5047) (мал. 4, *в*).

2248	Жанчына-вайсковец
2249	Жанчыны-вайскоўцы
2250	+
2251	Жанчына-маці
2252	Жанчыны-маці
2253	+
2254	Жанчына-юрыст
2255	Жанчыны-юрысты
2256	+
2257	жаўтушка
2258	жаўтушкі
2259	+
2260	жужаль
2261	жужалі
2262	+
2263	Жужуй
2264	+
2265	жук-алень
2266	жукі-алені
2267	+
2268	жук-насарог
2269	жукі-насарогі

а)

3357	кулі=к-саро+ка
3358	кулікі=-саро+кі
3359	+
3360	куліна+рна
3361	+
3362	культу=рна-выхава+ўчы
3363	культу=рна-выхава+ўчае
3364	+
3365	культу=рна-мо+ўны
3366	культу=рна-мо+ўныя
3367	культу=рна-мо+ўных
3368	+
3369	кумі+н
3370	+
3371	кумква+т
3372	+
3373	куні=цападо+бны
3374	куні=цападо+бныя
3375	+
3376	ку=рападо+бны
3377	ку=рападо+бныя
3378	+

б)

5044	+
5045	Во=дападрыхто+ўка_NNIFO
5046	Во=дападрыхто+ўкі_NNIFG
5047	Во=дападрыхто+ўцы_NNIFD
5048	Во=дападрыхто+ўку_NNIFA
5049	Во=дападрыхто+ўкай_NNIFI
5050	Во=дападрыхто+ўкаю_NNIFS
5051	Во=дападрыхто+ўцы_NNIFR
5052	Во=дападрыхто+ўкі_NNIFPO
5053	Во=дападрыхто+вак_NNIFPG
5054	Во=дападрыхто+ўкам_NNIFPD
5055	Во=дападрыхто+ўкі_NNIFPA
5056	Во=дападрыхто+ўкамі_NNIFPI
5057	Во=дападрыхто+ўках_NNIFPR
5058	+
5059	Во=жыкападо+бны_JJMO
5060	Во=жыкападо+бнага_JJMG
5061	Во=жыкападо+бнаму_JJMD
5062	Во=жыкападо+бны_JJMA
5063	Во=жыкападо+бнага_JJMU
5064	Во=жыкападо+бным_JJMI
5065	Во=жыкападо+бным_JJMR
5066	Во=жыкападо+бная_JJFO

в)

Мал. 4. Фрагменты распрацаваных слоўнікаў: *а)* адпаведнасці «слова – пачатковая форма»; *б)* адпаведнасці «слова – пачатковая форма» з указаннем націскаў; *в)* поўныя парадыгмы слоў з указаннем націскаў і граматычных характарыстык

Атрыманыя электронныя лінгвістычныя рэсурсы зрабілі магчымай не толькі тэарэтычную распрацоўку алгарытму аўтаматызаванай генерацыі АПП УДК на беларускай мове, але і яго практычную рэалізацыю ў адмысловым праграмным прататыпе.

Алгарытм аўтаматызаванага стварэння алфавітна-прадметнага паказальніка УДК на беларускай мове. Алгарытм дае магчымасць сфарміраваць АПП УДК па тэкстам яе табліц і арыентаваны на працу з беларускім выданнем УДК, але аддзеленасць лінгвістычных і тэматычных рэсурсаў, якія выкарыстоўваюцца ў алгарытме, ад самога алгарытму дазваляе адаптаваць яго для працы з УДК на іншых мовах. Дадзены алгарытм заснаваны на падыходах, апісаных аўтарамі ў артыкулах [6, 7].

Уваходныя дадзеныя алгарытму: поўны тэкст асноўных табліц УДК T_{UDC}.

Рэсурсы алгарытму:

– тэкставы файл F_{sc}, які змяшчае спіс стоп-класаў;

– тэкставы файл F_{sw}, які змяшчае спіс стоп-слоў;

– тэкставы файл F_{dic}, які змяшчае характэрную для УДК лексіку з указаннем пэўных граматычных характарыстык;

– тэкставы файл F_{dom}, які змяшчае спіс адпаведнасцяў «код класа – дамен».

Уваход:

Крок 1. Загрузка рэсурсаў. Адбываецца загрузка файлаў F_{sc}, F_{sw}, F_{dic}, F_{dom} з адмысловымі рэсурсамі ў памяць камп'ютара, фарміруюцца адпаведныя спісы.

Крок 1.1. Фарміраванне спісу стоп-класаў. Адбываецца загрузка файла са спісам стоп-класаў F_{sc}. Фарміруецца спіс $L_{sc} = <sc_1, \ldots, sc_A>$, дзе sc_a – a-ты стоп-клас, $a = 1, ..., A$.

Крок 1.2. Фарміраванне спісу стоп-слоў. Адбываецца загрузка файла са спісам стоп-слоў F_{sw}. Фарміруецца спіс $L_{sw} = <sw_1, \ldots, sw_B>$, дзе sw_b – b-е стоп-слова, $b = 1, ..., B$.

Крок 1.3. Фарміраванне спецыялізаванага слоўніка. Адбываецца загрузка файла-слоўніка F_{dic}. Фарміруецца спіс $L_{dic} = <<w_1, wa_1, wc_1, wi_1>, \ldots, <w_C, wa_C, wc_C, wi_C>>$, дзе w_c – c-е слова слоўніка, wa_c – націск c-га слова слоўніка, wc_c – катэгорыя c-га слова слоўніка, wi_c – пачатковая форма c-га слова слоўніка, $c = 1, ..., C$.

Крок 1.4. Фарміраванне спісу прыналежнасці класаў даменам. Адбываецца загрузка файла F_{dom}. Фарміруецца спіс $L_{dom} = <<cl_1, dom_1>, \ldots, <cl_D, dom_D>>$, дзе cl_d – d-ты клас спісу, dom_d – дамен, які адпавядае d-му класу спісу, $d = 1, ..., D$.

Крок 2. Фарміраванне спісу класаў УДК. Уваходны тэкст T_{UDC} разбіваецца на асобныя запісы – класы УДК; у кожным запісе вылучаецца код класа і апісанне класа. Такім чынам на аснове ўваходнага тэксту T_{UDC} фарміруецца спіс $L_{UDC} = <Cl_1, \ldots, Cl_N>$, дзе $Cl_n = <Not_n, Cap_n>$, Not_n – n-ты код класа, Cap_n – n-е апісанне класа, $n = 1, ..., N$.

Крок 3. Апрацоўка спісу класаў УДК. Ствараецца спіс адпаведнасцяў «слова – мноства кодаў класаў» L_{res}, у які будуць заносіцца вынікі апрацоўкі. Кожны элемент Cl_n спісу L_{UDC} праходзіць крокі 3.1–3.6.

Крок 3.1. Фільтрацыя паводле спісу стоп-класаў. Адбываецца вызначэнне, ці прыналежыць код класа Not_n спісу стоп-класаў L_{sc}. Калі прыналежнасць выяўлена, то адбываецца пераход да наступнага элементу Cl_{n+1} і кроку 3.1 (пры $n = N$ – да кроку 4). Іначай – да кроку 3.2.

Крок 3.2. Вылучэнне слоў. У апісанні класа Cap_n вылучаюцца ўсе сімвальныя паслядоўнасці, якія адпавядаюць шаблону будовы слова Pt_w. Выкарыстоўваючы сінтаксіс рэгулярных выразаў PCRE (http://www.pcre.org/original/doc/html/pcrepattern.html), дадзены шаблон можна прадставіць наступным чынам:

$$Pt_w = [set1][set1set2]^*,$$

дзе $set1$ – мноства сімвалаў, з якіх можа пачынацца слова, $set2$ – мноства сімвалаў, з якіх можа складацца, але не можа пачынацца слова, $set1set2 = set1 \cup set2$. У склад мноства $set1$ уваходзяць літары беларускага алфавіту, у склад мноства $set2$ – злучок, апостраф, сімвалы націскаў і інш.

Вылучаныя ў апісанні класа Cap_n словы заносяцца ў спіс $W_{cap} = <wrd_1, ..., wrd_M>$, дзе wrd_m – m-е вылучанае слова, $m = 1, ..., M$.

Крок 3.3. Нармалізацыя слоў. Кожнае слова wrd_m са спісу W_{cap} прыводзіцца да нармалізаванай электроннай формы (словы прыводзяцца да ніжняга рэгістра; адбываецца ўніфікацыя апострафаў, замена літары «ў» на «у» ў пачатку слова і інш.). Вынікам дадзенай апрацоўкі з'яўляецца спіс нармалізаваных слоў $W'_{cap} = <wrd'_1, ..., wrd'_M>$, дзе wrd'_m – m-е нармалізаванае слова, $m = 1, ..., M$.

Крок 3.4. Фільтрацыя паводле спісу стоп-слоў. Для кожнага слова са спісу W'_{cap} адбываецца вызначэнне, ці прыналежыць яно спісу стоп-слоў L_{sw}. Калі прыналежнасць не выяўлена, то слова заносіцца ў спіс $W''_{cap} = <wrd''_1, ..., wrd''_K>$, дзе wrd'_k – k-е дапушчальнае слова, $k = 1, ..., K$.

Крок 3.5. Фільтрацыя паводле часціны мовы і прывядзенне да пачатковай формы. Для кожнага слова са спісу W''_{cap} адбываецца вызначэнне часціны мовы wc і пачатковай формы wi паводле спісу характэрнай для УДК лексікі L_{dic}. Калі бягучае слова з'яўляецца назоўнікам, прыметнікам або дзеепрыметнікам, то яго пачатковая форма wi заносіцца ў спіс $W'''_{cap} = <wrd'''_1, \ldots, wrd'''_J>$, дзе wrd'_j – j-е дапушчальнае слова ў пачатковай форме, $j = 1, ..., J$.

Крок 3.6. Занясенне ў выніковы спіс. Кожнае слова са спісу W'''_{cap} заносіцца ў выніковы спіс L_{res}. Калі ў выніковым спісе L_{res} бягучае слова яшчэ не сустракаецца, то слова дадаецца ў выглядзе новага элемента, у адпаведнасць якому ставіцца масіў з адным элементам – кодам

апрацоўваемага класа Not_n. Калі ж слова ўжо занесена ў выніковы спіс L_{res}, то код апрацоўваемага класа Not_n заносіцца ў масіў кодаў класаў, адпаведных апрацоўваемаму слову.

Крок 4. Прысвойванне даменаў. Адбываецца перабор элементаў спісу L_{res}. У кожным элеменце L_{res} да кожнага кода класа вызначаецца тэматычны дамен *dom* паводле спісу L_{dom}. Коды класаў, дамены якіх супадаюць, групуюцца ў адзін масіў.

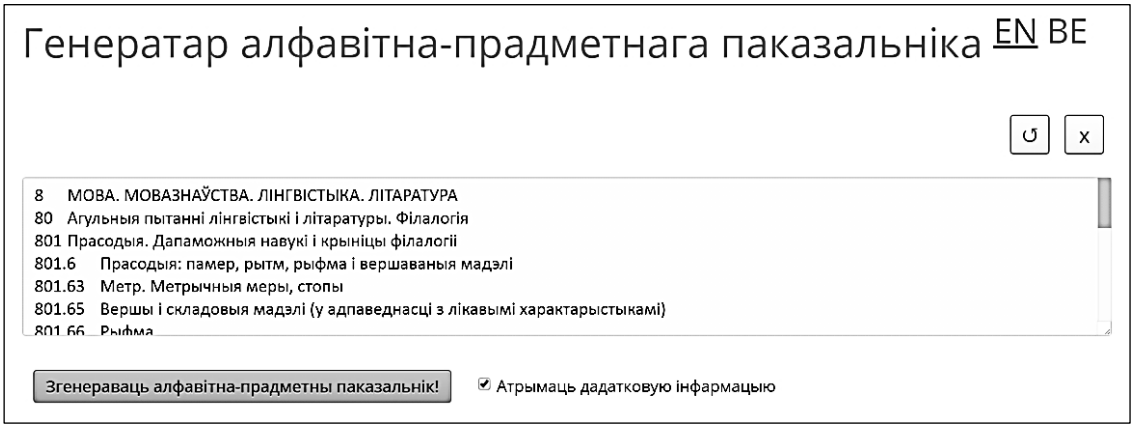
Такім чынам, вынікам з'яўляецца спіс $L_{res} = <<wrd_1, ent_1>, ..., <wrd_R, ent_R>>$, дзе wrd_r – *r*-е слова выніковага спісу, $r = 1, ..., R$; у сваю чаргу, $ent_r = <<dom_1, Lnot_1>, ..., <dom_S, Lnot_S>>$, дзе dom_s – *s*-ты дамен, да якога адносіцца слова wrd_r, $s = 1, ..., S$. І ўрэшце $Lnot_s = <not_1, ..., not_T>$, дзе not_t – *t*-ты код класа, адпаведнага дамену dom_s, да якога адносіцца слова wrd_r, $t = 1, ..., T$.

Крок 5. Фарміраванне выніку. Адбываецца сартыроўка выніковага спісу L_{res} паводле беларускага алфавіту і прывядзенне да тэкставага фармату з неабходным фарматаваннем. Вынік T_{ASI} выводзіцца на экран і захоўваецца ў адпаведным файле на серверы.

Канец алгарытму.

У выніку выканання апісанага алгарытму на аснове табліц УДК на беларускай мове фарміруецца АПП. Варта адзначыць, што прапанаваны алгарытм з'яўляецца прыдатным для пашырэння на іншыя мовы, бо лінгвістычныя і тэматычныя рэсурсы, выкарыстаныя ў ім, аддзелены ад алгарытму і могуць быць распрацаваны для іншай мовы.

Прататып сістэмы генерацыі алфавітна-прадметнага паказальніка УДК. Для апрабацыі, тэставання і нагляднасці працаздольнасці апісанага вышэй алгарытму быў распрацаваны прататып сістэмы генерацыі АПП УДК у форме вэб-сэрвіса «Генератар алфавітна-прадметнага паказальніка», які даступны для вольнага выкарыстання ў Інтэрнэце па адрасе http://corpus.by/AlphabeticalSubjectIndexGenerator. Дадзены сэрвіс дае магчымасць канвертаваць тэкст табліц УДК у АПП (мал. 5).



Мал. 5. Інтэрнэт-сэрвіс «Генератар алфавітна-прадметнага паказальніка»

Для карэктнай працы вэб-сэрвіса табліцы УДК павінны быць пададзены ў наступным фармаце: класы УДК аддзелены адзін ад аднаго пераводам радка, у межах аднаго запісу код класа аддзелены ад апісання класа табуляцыяй (мал. 5). Сэрвіс дае магчымасць карыстальніку не толькі генераваць АПП на падставе ўведзенага тэксту, але і адсочваць пэўную дадатковую інфармацыю датычна працэсу апрацоўкі. Для гэтага патрэбна адзначыць пункт «Атрымаць дадатковую інфармацыю», сэрвіс прадставіць інфармацыю пра словы, якія не атрымалася апрацаваць з-за адсутнасці той ці іншай неабходнай інфармацыі ў базе, а таксама падасць спіс знойдзеных ва ўваходным тэксце амографаў. У табліцы прадстаўлены прыклад выніковага АПП, згенераванага на падставе фрагменту табліц УДК пры дапамозе вэб-сэрвіса «Генератар алфавітна-прадметнага паказальніка».

Прыклад працы вэб-сэрвіса «Генератар алфавітна-прадметнага паказальніка»

Фрагмент асноўных табліц УДК	АПП, згенераваны на аснове фрагмента УДК
8 МОВА. МОВАЗНАЎСТВА. ЛІНГВІСТЫКА. ЛІТАРАТУРА 80 Агульныя пытанні лінгвістыкі і літаратуры. Філалогія 801 Прасодыя. Дапаможныя навукі і крыніцы філалогіі 801.6 Прасодыя: памер, рытм, рыфма і вершаваныя мадэлі 801.63 Метр. Метрычныя меры, стопы 801.65 Вершы і складовыя мадэлі (у адпаведнасці з лікавымі характарыстыкамі) 801.66 Рыфма 801.67 Стансы, строфы, куплеты, вершы (у паэме) 801.7 Дапаможныя філалагічныя дысцыпліны 801.8 Філалагічныя і лінгвістычныя крыніцы. Зборнікі тэкстаў	**А** Адпаведнасць *(мовазнаўства)* 801.65 **В** Верш *(мовазнаўства)* 801.65, 801.67 Вершаваны *(мовазнаўства)* 801.6 **З** Зборнік *(мовазнаўства)* 801.8 **К** Крыніца *(мовазнаўства)* 801.8 Куплет *(мовазнаўства)* 801.67 **Л** Лікавы *(мовазнаўства)* 801.65 Лінгвістыка 8, 80 Лінгвістычны *(мовазнаўства)* 801.8 Літаратура 8, 80 **М** Мадэль *(мовазнаўства)* 801.6, 801.65 Мера *(мовазнаўства)* 801.63 Метр *(мовазнаўства)* 801.63 Метрычны *(мовазнаўства)* 801.63 Мова 8 Мовазнаўства 8 **П** Памер *(мовазнаўства)* 801.6 Паэма *(мовазнаўства)* 801.67 Прасодыя *(мовазнаўства)* 801.6 Пытанне 80 **Р** Рытм *(мовазнаўства)* 801.6 Рыфма *(мовазнаўства)* 801.6, 801.66 **С** Складовы *(мовазнаўства)* 801.65 Станс або Стансы *(мовазнаўства)* 801.67 Стапа *(мовазнаўства)* 801.63 Страфа *(мовазнаўства)* 801.67 **Т** Тэкст *(мовазнаўства)* 801.8 **Ф** Філалагічны *(мовазнаўства)* 801.8 Філалогія 80 **Х** Характарыстыка *(мовазнаўства)* 801.65

Прыклад працы прататыпа сістэмы генерацыі АПП на падставе табліц УДК на беларускай мове дэманструе працаздольнасць распрацаваных алгарытмаў. Карэктнасць працы алгарытму была праверана рэдакцыйнай калегіяй выдання УДК на беларускай мове. Праграмная рэалізацыя дапамагла значна скараціць час на выданне, звёўшы працу рэдакцыйнай калегіі па распрацоўцы АПП да працэсу вычыткі і карэктыроўкі.

Заключэнне. У артыкуле прапанавана ідэя магчымасці фарміравання алфавітна-прадметных паказальнікаў для розных структураваных сістэм праз выкарыстанне алгарытмаў апрацоўкі электронных тэкстаў машынным чынам. Гэтая ідэя праілюстравана распрацаваным і апісаным у артыкуле алгарытмам аўтаматызаванай генерацыі АПП да беларускага выдання УДК. У аснову алгарытму былі пакладзены метады і падыходы да аўтаматызаванай апрацоўкі

вялікіх аб’ёмаў электронных тэкстаў і распрацоўкі лінгвістычных і тэматычных рэсурсаў, шырока прадстаўленыя на інтэрнэт-платформе апрацоўкі тэкстаў і маўлення www.Corpus.by.

На аснове апісанага ў артыкуле алгарытму быў распрацаваны прататып сістэмы генерацыі АПП, які знайшоў непасрэднае прымяненне ў падрыхтоўцы макета АПП першага беларускамоўнага выдання УДК [3].

Спіс выкарыстаных крыніц

1. McIlwaine, I. C. The Universal Decimal Classification: a guide to its use / I. C. McIlwaine. – The Hague : UDC Consortium, 2007. – 278 p.

2. Інструментарый індэксатара і яго прымяненне ў бібліятэках Беларусі / Нацыянальная бібліятэка Беларусі ; склад. С. А. Пугачова. – Мінск : Нацыянальная бібліятэка Беларусі, 2016. – 191 с.

3. Універсальная дзесятковая класіфікацыя: звыш 10 000 асноўных і дапаможных класаў / Аб’яднаны інстытут праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі, Нацыянальная бібліятэка Беларусі ; рэдкал.: Ю. С. Гецэвіч [і інш.]. – Мінск : Нацыянальная бібліятэка Беларусі, 2016. – 370 с.

4. Universele Decimale Classificatie: beknopte nederlandse uitgave / ed. G. J. A. Riesthuis. – The Hague : UDC Consortium, 2013. – 412 p.

5. Сербин, О. О. Подготовка УДК на украинском языке / О. О. Сербин // Библиосфера. – 2013. – № 2. – С. 69–73.

6. Станіславенка, Г. Р. Выкарыстанне камп’ютарна-лінгвістычных сродкаў для перакладу ўніверсальнай дзесятковай класіфікацыі дамена «тэатр» з англійскай на беларускую мову і генерацыя алфавітна-прадметнага паказальніка / Г. Р. Станіславенка, Ю. С. Гецэвіч, С. І. Лысы // Актуальные вопросы германской филологии и лингводидактики : материалы XX Междунар. науч.-практ. конф. / Брест. гос. ун-т имени А. С. Пушкина ; редкол.: Е. Г. Сальникова [и др.]. – Брест, 2016. – С. 264–266.

7. Станиславенко, А. Г. Этапы подготовки первого издания УДК на белорусском языке / А. Г. Станиславенко, С. И. Лысы, Ю. С. Гецевич // Информация в современном мире : докл. Междунар. конф., Москва, 25–26 октября 2017 г. / ВИНИТИ РАН. – М., 2017. – С. 297–303.

References

1. McIlwaine I. C. *The Universal Decimal Classification: a guide to its use*. The Hague, UDC Consortium, 2007, 278 p.

2. Puhachova S. A. (ed.) Instrumientaryj indeksatara i jaho prymianiennie ŭ biblijatekach Bielarusi. *Toolkit of indexers and its use in the libraries of Belarus.* Minsk, National Library of Belarus, 2016, 191 p. (in Belarusian).

3. Hetsevich Yu. S., Puhachova S. A., Stanislavenka H. R., Kuzminich T. V., Narejka A., Hetsevich S. A. (eds.) Univiersaĺnaja dziesiatkovaja klasifikacyja: zvyš 10 000 asnoŭnych i dapamožnych klasaŭ. *Universal Decimal Classification: more than 10 000 main and auxiliary classes.* Minsk, National Library of Belarus, 2016, 370 p. (in Belarusian).

4. Riesthuis G. J. A. (ed.) *Universele Decimale Classificatie: beknopte nederlandse uitgave*. The Hague, UDC Consortium, 2013, 412 p. (in Dutch).

5. Sierbin O. O. Podhotovka UDK na ukrainskom jazykie. *Preparation of UDC in Ukrainian*. Bibliosfiera, 2013, no. 2, pp. 69–73 (in Russian).

6. Stanislavenka H. R., Hetsevich Yu. S., Lysy S. I. Vykarystannie kamp’iutarna-linhvistyčnych srodkaŭ dlia pierakladu ŭniviersaĺnaj dziesiatkovaj klasifikacyi damiena «teatr» z anhlijskaj na bielaruskuju movu i hienieracyja alfavitna-pradmietnaha pakazaĺnika [Using computational linguistic tools for translation of Universal Decimal Classification («theater» domain) from English into Belarusian and generation of Alphabetical Subject Index]. Aktuaĺnyje voprosy hiermanskoj filolohii i linhvodidaktiki [*Relevant questions of German philology and linguistics : XX International Scientific and Practical Conference Proceedings*]. Brest, 2016, pp. 264–266 (in Belarusian).

7. Stanislavenka H. R., Lysy S. I., Hetsevich Yu. S. Etapy podhotovki piervoho izdaniia UDK na bielarusskom jazykie [Stages of preparation of the first edition of UDC in Belarusian]. Informaciia v sovriemiennom mirie : doklady Miežđunarodnoj konfieriencii [*Information in the Modern World: Proceedings of the International Conference*]. Moscow, 2017, pp. 297–303 (in Russian).

Інфармацыя пра аўтараў

Лысы Станіслаў Іосіфавіч – аспірант, Аб'яднаны інстытут праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі (вул. Сурганава, 6, 220012, Мінск, Рэспубліка Беларусь). E-mail: stanislau.lysy@gmail.com

Станіславенка Ганна Рыгораўна – аспірант, Цэнтр даследаванняў беларускай культуры, мовы і літаратуры Нацыянальнай акадэміі навук Беларусі (вул. Сурганава, 1, 220072, Мінск, Рэспубліка Беларусь). E-mail: hanna.stanislavenka@gmail.com

Гецэвіч Юрый Станіслававіч – кандыдат тэхнічных навук, Аб'яднаны інстытут праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі (вул. Сурганава, 6, 220012, Мінск, Рэспубліка Беларусь). E-mail: yuras.hetsevich@newman.bas-net.by

Information about the authors

Stanislau I. Lysy – PhD student, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus (6, Surhanava Str., 220012, Minsk, Republic of Belarus). E-mail: stanislau.lysy@gmail.com

Hanna R. Stanislavenka – PhD student, Center for the Belarusian Culture, Language and Literature Researches of the National Academy of Sciences of Belarus (1, Surhanava Str., 220072, Minsk, Republic of Belarus). E-mail: hanna.stanislavenka@gmail.com

Yuryj S. Hetsevich – PhD (Engineering), The United Institute of Informatics Problems of the National Academy of Sciences of Belarus (6, Surhanava Str., 220012, Minsk, Republic of Belarus). E-mail: yuras.hetsevich@ newman.bas-net.by