

ОБЪЕДИНЕННЫЙ ИНСТИТУТ ПРОБЛЕМ ИНФОРМАТИКИ
НАЦИОНАЛЬНОЙ АКАДЕМИИ НАУК БЕЛАРУСИ

ИНФОРМАТИКА

Том 15, № 2, апрель-июнь 2018

Ежеквартальный научный журнал

Издается с января 2004 г.

Учредитель – Объединенный институт проблем информатики
Национальной академии наук Беларуси

Г л а в н ы й р е д а к т о р

Александр Васильевич Тузиков – генеральный директор Объединенного института проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь

Р е д а к ц и о н н а я к о л л е г и я

- | | |
|---------------------------|---|
| М. Я. Ковалев | – Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь (<i>заместитель главного редактора</i>) |
| С. В. Абламейко | – Белорусский государственный университет, Минск, Республика Беларусь |
| В. В. Анищенко | – Компания «СофтКлуб», Минск, Республика Беларусь |
| П. Н. Бибило | – Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь |
| М. Н. Бобов | – ОАО «АГАТ – системы управления» – управляющая компания холдинга «Геоинформационные системы управления», Минск, Республика Беларусь |
| А. Б. Долгий | – Высшая инженерная школа Бретани, Нант, Франция |
| А. Н. Дудин | – Белорусский государственный университет, Минск, Республика Беларусь |
| А. А. Карпов | – Институт информатики и автоматизации Российской академии наук (СПИИРАН), Санкт-Петербург, Российская Федерация |
| С. Я. Килин | – Президиум Национальной академии наук Беларуси, Минск, Республика Беларусь |
| В. В. Краснопрошин | – Белорусский государственный университет, Минск, Республика Беларусь |
| А. М. Крот | – Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь |
| С. В. Кругликов | – Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь |
| С. П. Кундас | – Белорусский национальный технический университет, Минск, Республика Беларусь |

- Н. А. Лиходед** – Белорусский государственный университет, Минск, Республика Беларусь
- П. П. Матус** – Институт математики Национальной академии наук Беларуси, Минск, Республика Беларусь
- А. А. Петровский** – Белорусский государственный университет информатики и радиоэлектроники, Минск, Республика Беларусь
- В. А. Складар** – Университет Авейру, Португалия
- Ю. С. Сотсков** – Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь
- А. Л. Стемпковский** – Институт проблем проектирования в микроэлектронике Российской академии наук, Москва, Российская Федерация
- Ю. С. Харин** – Научно-исследовательский институт прикладных проблем математики и информатики Белорусского государственного университета, Минск, Республика Беларусь
- А. Ф. Чернявский** – Институт прикладных физических проблем им. А. Н. Севченко Белорусского государственного университета, Минск, Республика Беларусь
- В. Н. Ярмолик** – Белорусский государственный университет информатики и радиоэлектроники, Минск, Республика Беларусь

ИНФОРМАТИКА

Том 15, № 2, апрель-июнь 2018

Ответственный за выпуск *С. С. Мойсейчик*
Редактор *Г. Б. Гончаренко*
Корректор *А. А. Михайлова*
Компьютерная верстка *О. Б. Бутевич*

Сдано в набор 04.05.2018. Подписано в печать 11.06.2018.
Формат 60×84 1/8. Бумага офсетная. Гарнитура Таймс. Ризография.
Усл. печ. л. 12,8. Уч.-изд. л. 12,5. Тираж 60 экз. Заказ 6.

Государственное научное учреждение «Объединенный институт проблем информатики Национальной академии наук Беларуси».
Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий № 1/274 от 04.04.2014.
ЛП № 02330/444 от 18.12.13.
Ул. Сурганова, 6, 220012, Минск.

ISSN 1816-0301 (print)

СОДЕРЖАНИЕ

ОБРАБОТКА СИГНАЛОВ, ИЗОБРАЖЕНИЙ И РЕЧИ

Ковалев В. А., Козловский С. А., Калиновский А. А. Генерация искусственных рентгеновских изображений грудной клетки с использованием генеративно-состязательных нейронных сетей.....	7
Tuyet D. V., Ablameyko S. V. A model for medical diagnostic laboratory: multimedia database approach.....	17
Рыбенков Е. В., Петровский Н. А. Синтез FPGA-архитектур банков фильтров на основе блочной лестничной факторизации в алгебре кватернионов (часть 1).....	29
Лысы С. І., Станіславенка Г. Р., Гецэвіч Ю. С. Аўтаматызаваная генерацыя алфавітна-прадметнага паказальніка універсальнай дзесяткавай класіфікацыі на беларускай мове.....	45
Залесский Б. А., Троцкий Ф. С. Параллельная версия детектора экстремальных особых точек изображений.....	55

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Ерофеенко В. Т. Модели базисных электромагнитных волн в биизотропной среде с пространственной дисперсией	64
Романчак В. М. Субъективное оценивание вероятности	74

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

Кругликов С. В., Бекиш А. Р. Обобщенная структура имитационной модели групповой деятельности операторов АСУ.....	83
Червяков Н. И., Коляда А. А., Коляда Н. А., Кучуков В. А., Протасеня С. Ю. Нейронные сети конечного кольца на основе редукционной схемы позиционно-модулярного кодового преобразования	98

ISSN 1816-0301 (print)
УДК 004.912:025.045

Паступіла ў рэдакцыю 19.01.2018
Received 19.01.2018

С. І. Лысы¹, Г. Р. Станіслаўска², Ю. С. Гецэвіч¹

¹*Аб'яднаны інстытут праблем інфарматыкі Нацыянальнай акадэміі
наук Беларусі, Мінск, Беларусь*

²*Цэнтр даследаванняў беларускай культуры, мовы і літаратуры Нацыянальнай акадэміі
наук Беларусі, Мінск, Беларусь*

Аўтаматызаваная генерацыя алфавітна-прадметнага паказальніка універсальнай дзесятковай класіфікацыі на беларускай мове

Анотацыя. 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 падцверджання працаздольнасці і карэктнасці працы дадзенага алгарытму прыводзіцца распаўсюджаны 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. Станіслаўска, Ю. С. Гецэвіч // Информатика. – 2018. – Т. 15, № 2. – С. 45–54.

S. I. Lysy¹, H. R. Stanislavenska², Yu. S. Hetsevich¹

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., Stanislavenska 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 Атлэ (фр. *Paul Otlet*) і Анры Лафонтэн (фр. *Henri La Fontaine*) дапрацавалі «Дзесятковую класіфікацыю» Мэлвіла Дзьюі (англ. *Melvil Dewey*), якую апошні ствараў 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 дадання 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 даведчна-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 скарачаныя 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.

Noord-Polen - plaats Polen (438.5)	Samojeedse talen - taal =511.2
Noordpooleilanden, Europese - plaats (984)	San Juan (provincie) - plaats Argentinië (825.2)
Noordpoolgebied, Amerikaans - plaats (987)	San Luis (provincie) - plaats Argentinië (825.5)
Noordpoolgebied, Russisch - plaats (985)	San Marino (Republiek) - plaats Europa (454.4)
Noordpoolgebieden (98)	sancties - economisch recht 346.9
Noords skiën - sport 796.922	sandelhout - botanie 582.728
Noordse talen	sandwiches - koken - huishoudkunde 641.84
- linguïstiek 811.113	Sanima - taal =855.72
- literatuur 821.113	sanitair
- taal =113	- bouwnijverheid 696
Noordse volkeren, gebieden van - plaats (368)	- huishoudkunde 644.6
noordwest - plaats (1-16)	sanitair in gebouwen - gezondheidstechniek 628.6
noordwest Polen - plaats Polen (438.4)	sanitair papier - papierindustrie 676.25
Noordwest-China - plaats China (514)	Sanskriet
noordwestelijke Grote Oceaan - plaats (265.5)	- literatuur 821.211
Nooristani - taal =23	- taal =211
Noors	Santa Catarine (deelstaat) - plaats Brazilië (816.4)
- literatuur 821.113.5	Santa Cruz (provincie) - plaats Argentinië (828.8)
- taal =113.5	Santa Fe (provincie) - plaats Argentinië (821.6)
Noorse - linguïstiek 811.113.5	Santalales - botanie 582.728

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 простаі 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 (<i>будаўніцтва</i>) 693.5	Снарад
Жалоба (<i>антрапалогія</i>) 393.7	(<i>фізіка</i>) 531.55
Жаніцьба (<i>антрапалогія</i>) 392.5	(<i>тэхналогіі</i>) 623.4, 623.46, 629.76, 629.762
Жанр	Снег
(<i>мастацтва</i>) 7.041	(<i>геалагічныя навукі</i>) 551.322, 551.578.4, 556.12
(<i>кіно</i>) 791.2, 791.22, 791.223	(<i>тэхналогіі</i>) 624.14
(<i>літаратура</i>) 82-1/-9, 82-94, 82-95	(<i>прамысловасць</i>) 685.36
Жанчына	Снегіровыя сойкі (<i>заалогія</i>) 598.293
(<i>антрапалогія</i>) 392.6	Снэпшот (<i>фатаграфія</i>) 77.055
(<i>медыцына</i>) 618.16, 618.17, 618.18	Совападобныя (<i>заалогія</i>) 598.27, 598.279, 598.279.4
Жарганізм (<i>мовазнаўства</i>) 81`373.48	Сода (<i>хімія</i>) 661.3
Жарт (<i>антрапалогія</i>) 398.25, 398.6	Сойка (<i>заалогія</i>) 598.293
Жаўтушка (<i>заалогія</i>) 598.296	Сок (<i>хімія</i>) 547.914, 663.8, 663.81
Жах (<i>кіно</i>) 791.221.9	Сокападобныя (<i>заалогія</i>) 598.27, 598.279, 598.279.3
Жвачныя	Солад (<i>хімія</i>) 663.032
(<i>заалогія</i>) 599.735	Соль
(<i>сельская гаспадарка</i>) 636.2	(<i>хімія</i>) 54-38, 549.451.1, 549.89, 661.74, 661.8, 664.41
Жвір (<i>будаўніцтва</i>) 691.22	(<i>геалагічныя навукі</i>) 552.53, 553.63, 553.78
	(<i>прыкладное мастацтва</i>) 745.56

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

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

Электронныя лінгвістычныя рэсурсы для аўтаматызаваанай 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 аўтаматызаваанай 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 скарачанага кода класа і адпаведнай вобласці прымянення, або дамена (мал. 3, б).

7.049	72 (архітэктура)
7.06	73 (пластычнае мастацтва)
7.08	74 (прыкладное мастацтва)
7.094	75 (жывапіс)
711.1	76 (графіка)
72.01/.05	77 (фатаграфія)
725	78 (музыка)
725.22	79 (забавы)
725.89	80 (мовазнаўства)
726.9	81 (мовазнаўства)
728.8	82 (літаратура)
736	

а)

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 атрыманням граматычных характарыстык (<http://corpus.by/WordParadigmGenerator>, <http://ssrlab.by/5047>) (мал. 4, в).

2248 Жанчына-вайсковец	3357 кулі=к-саро+ка	5044 +
2249 Жанчыны-вайскоўцы	3358 кулікі=-саро+кі	5045 Во=дападыхто+ўка_NNIFO
2250 +	3359 +	5046 Во=дападыхто+ўкі_NNIFG
2251 Жанчына-маці	3360 куліна+рна	5047 Во=дападыхто+ўцы_NNIFD
2252 Жанчыны-маці	3361 +	5048 Во=дападыхто+ўку_NNIFA
2253 +	3362 культу=рна-выхава+ўчы	5049 Во=дападыхто+ўкай_NNIFI
2254 Жанчына-юрыст	3363 культу=рна-выхава+ўчае	5050 Во=дападыхто+ўкаю_NNIFS
2255 Жанчыны-юрысты	3364 +	5051 Во=дападыхто+ўцы_NNIFR
2256 +	3365 культу=рна-мо+ўны	5052 Во=дападыхто+ўкі_NNIFPO
2257 жаўтушка	3366 культу=рна-мо+ўныя	5053 Во=дападыхто+вак_NNIFPG
2258 жаўтушкі	3367 культу=рна-мо+ўных	5054 Во=дападыхто+ўкам_NNIFPD
2259 +	3368 +	5055 Во=дападыхто+ўкі_NNIFPA
2260 жужаль	3369 кумі+н	5056 Во=дападыхто+ўкамі_NNIFPI
2261 жужалі	3370 +	5057 Во=дападыхто+ўках_NNIFPR
2262 +	3371 кумква+т	5058 +
2263 Жужуй	3372 +	5059 Во=жыкападо+бны_JJMO
2264 +	3373 куні=цападо+бны	5060 Во=жыкападо+бнага_JJMG
2265 жук-алень	3374 куні=цападо+бныя	5061 Во=жыкападо+бнаму_JJMD
2266 жукі-алені	3375 +	5062 Во=жыкападо+бны_JJMA
2267 +	3376 ку=рападо+бны	5063 Во=жыкападо+бнага_JJMU
2268 жук-насарог	3377 ку=рападо+бныя	5064 Во=жыкападо+бным_JJMI
2269 жукі-насарогі	3378 +	5065 Во=жыкападо+бным_JJMR
		5066 Во=жыкападо+бная_JJFO

а)

б)

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 $L_{sc} = \langle sc_1, \dots, sc_A \rangle$, дзе sc_a – a -ты стоп-клас, $a = 1, \dots, A$.

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

Крок 1.3. Фарміраванне спецыялізаванага слоўніка. Адбываецца загрузка файла-слоўніка F_{dic} . Фарміруецца спіс $L_{dic} = \langle \langle w_1, wa_1, wc_1, wi_1 \rangle, \dots, \langle w_C, wa_C, wc_C, wi_C \rangle \rangle$, 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, $c = 1, \dots, C$.

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

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 $L_{UDC} = \langle Cl_1, \dots, Cl_N \rangle$, дзе $Cl_n = \langle Not_n, Cap_n \rangle$, Not_n – n -ты код класа, Cap_n – n -е апісанне класа, $n = 1, \dots, N$.

Крок 3. Апрацоўка спісу класаў УДК. Ствараецца спіс адпаведнасцяў «слова – мноства кодаў класаў» L_{res} , у які будучы 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/pcpattern.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 заносзяцца ў спіс $W_{cap} = \langle wrd_1, \dots, wrd_M \rangle$, дзе wrd_m – m -е вылучанае слова, $m = 1, \dots, M$.

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 $W'_{cap} = \langle wrd'_1, \dots, wrd'_M \rangle$, дзе wrd'_m – m -е нармалізаванае слова, $m = 1, \dots, M$.

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 $W''_{cap} = \langle wrd''_1, \dots, wrd''_K \rangle$, дзе wrd''_k – k -е дапушчальнае слова, $k = 1, \dots, K$.

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 бягуцае слова з'яўляецца назойнікам, прыметнікам або дзеепрыметнікам, то яго пачатковая форма wi заносіцца ў спіс $W'''_{cap} = \langle wrd'''_1, \dots, wrd'''_J \rangle$, дзе wrd'''_j – j -е дапушчальнае слова ў пачатковай форме, $j = 1, \dots, J$.

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} бягуцае 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 $L_{res} = \langle \langle wrd_1, ent_1 \rangle, \dots, \langle wrd_R, ent_R \rangle \rangle$, дзе wrd_r – r -е слова выніковага спісу, $r = 1, \dots, R$; у сваю чаргу, $ent_r = \langle \langle dom_1, Lnot_1 \rangle, \dots, \langle dom_s, Lnot_s \rangle \rangle$, дзе dom_s – s -ты дамен, да якога адносіцца слова wrd_r , $s = 1, \dots, S$. І ўрэшце $Lnot_s = \langle not_1, \dots, not_T \rangle$, дзе not_t – t -ты код класа, адпаведнага дамену dom_s , да якога адносіцца слова wrd_r , $t = 1, \dots, T$.

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, адзелены ад алгарытму і могуць быць распрацаваны для іншай мовы.

Прататып сістэмы генерацыі алфавітна-прадметнага паказальніка УДК. Для апрабавання, 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	А Адпаведнасць (<i>мовазнаўства</i>) 801.65
80 Агульныя пытанні лінгвістыкі і літаратуры. Філалогія	В Верш (<i>мовазнаўства</i>) 801.65, 801.67 Вершаваны (<i>мовазнаўства</i>) 801.6
801 Прасодыя. Дапаможныя навукі і крыніцы філалогіі	З Зборнік (<i>мовазнаўства</i>) 801.8
801.6 Прасодыя: памер, рытм, рыфма і вершава-ныя мадэлі	К Крыніца (<i>мовазнаўства</i>) 801.8 Куплет (<i>мовазнаўства</i>) 801.67
801.63 Метр. Метрычныя меры, стопы	Л Лікавы (<i>мовазнаўства</i>) 801.65
801.65 Вершы і складовыя мадэлі (у адпаведнасці з лікавымі характарыстыкамі)	Лінгвістыка 8, 80 Лінгвістычны (<i>мовазнаўства</i>) 801.8
801.66 Рыфма	Літаратура 8, 80
801.67 Стансы, строфы, куплеты, вершы (у паэме)	М Мадэль (<i>мовазнаўства</i>) 801.6, 801.65 Мера (<i>мовазнаўства</i>) 801.63 Метр (<i>мовазнаўства</i>) 801.63 Метрычны (<i>мовазнаўства</i>) 801.63 Мова 8 Мовазнаўства 8
801.7 Дапаможныя філалагічныя дысцыпліны	П Памер (<i>мовазнаўства</i>) 801.6 Паэма (<i>мовазнаўства</i>) 801.67 Прасодыя (<i>мовазнаўства</i>) 801.6 Пытанне 80
801.8 Філалагічныя і лінгвістычныя крыніцы. Зборнікі тэкстаў	Р Рытм (<i>мовазнаўства</i>) 801.6 Рыфма (<i>мовазнаўства</i>) 801.6, 801.66
	С Складовы (<i>мовазнаўства</i>) 801.65 Станс або Стансы (<i>мовазнаўства</i>) 801.67 Стапа (<i>мовазнаўства</i>) 801.63 Страфа (<i>мовазнаўства</i>) 801.67
	Т Тэкст (<i>мовазнаўства</i>) 801.8
	Ф Філалагічны (<i>мовазнаўства</i>) 801.8 Філалогія 80
	Х Характарыстыка (<i>мовазнаўства</i>) 801.65

Прыклад працы прататыпа сістэмы генерацыі АПП на падставе табліц УДК на беларускай мове дэманструе працаздольнасць распрацаваных алгарытмаў. Карэктнасць працы алгарытму была правярана 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.

Заклучэнне. 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 прадстаўлены 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. Гецевіч [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. Станіслаўенка, 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. Станіслаўенка, Ю. С. Гецевіч, С. І. Лысы // Актуальныя пытанні германскай філологіі і лінгвастадыі : матэрыялы XX Міжнароднага навукова-практ. канф. / Брэст. гос. ун-т імя А. С. Пушкіна ; рэдкал.: Е. Г. Сальнікова [і др.]. – Брэст, 2016. – С. 264–266.
7. Станіслаўенка, А. Г. Этапы падготовки першага выдання УДК на беларускім мове / А. Г. Станіслаўенка, С. І. Лысы, Ю. С. Гецевіч // Інфармацыя ў сучасным свеце : докл. Міжнароднага канф., Москва, 25–26 кастрычніка 2017 г. / ВІНІТІ 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.) *Instrumentaryj indeksatara i jaho prymanienne 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. Hetsevič Yu. S., Puhachova S. A., Stanislavenska H. R., Kuzminich T. V., Narejka A., Hetsevič S. A. (eds.) *Univiersalnaja 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*. *Bibliosfera*, 2013, no. 2, pp. 69–73 (in Russian).
6. Stanislavenska H. R., Hetsevič Yu. S., Lysy S. I. *Vykarystannie kamp'jutarna-linhvistyčnych srodkaŭ dla pierakladu ŭniviersalnaj dziesiatkovaj klasifikacyi damiena «teatr» z anhlijskaj na bielaruskuju movu i hienieracyja alfavitna-pradmetnaha pakazalnika* [Using computational linguistic tools for translation of Universal Decimal Classification («theater» domain) from English into Belarusian and generation of Alphabetical Subject Index]. *Aktualnyje 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. Stanislavenska H. R., Lysy S. I., Hetsevič Yu. S. *Etapy podhotovki piervocho izdaniia UDK na bielarusskom jazykie* [Stages of preparation of the first edition of UDC in Belarusian]. *Informaciia v sovremennom mirie : doklady Mieždunarodnoj 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. Сурганова, 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. Сурганова, 1, 220072, Мінск, Рэспубліка Беларусь). E-mail: hanna.stanislaivenka@gmail.com

Гецэвіч Юрый Станіслававіч – кандыдат тэхнічных навук, Аб’яднаны інстытут праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі (вул. Сурганова, 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.stanislaivenka@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