# Безпека життєдіяльності

УДК 681.3:355

О.С. Андрощук

*Національна академія Державної прикордонної служби України ім. Б. Хмельницького, Хмельницький*

## МОДЕЛЬ ПОБУДОВИ ТЕЗАУРУСА ПРЕДМЕТНОЇ СФЕРИ УПРАВЛІННЯ ПІДРОЗДІЛАМИ В ОСОБЛИВИХ СИТУАЦІЯХ

*У статті подано комплексний підхід до побудови предметно-орієнтованого тезауруса предметної сфери управління підрозділами, органами, управліннями Державної прикордонної служби України на основі інтеграції лінгвістичного аналізу текстів, що регламентують процес управління, і результатів об'єктно-орієнтованого моделювання процесів управління в особливих ситуаціях. Використання розробленого предметно-орієнтованого тезауруса в процесах формування рекомендацій щодо прийняття рішень надає можливість забезпечити більшу повноту і точність пошуку рішень.*

***Ключові слова:*** *моделювання, аналіз, тезаурус, база знань, особлива ситуація.*

### Вступ

**Постановка проблеми.** Сучасний етап розвитку інтегрованої інформаційно-телекомунікаційної системи "Гарт" Державної прикордонної служби України (ДПСУ) характеризується принципово новими вимогами до процедур збору, збереження, обробки та передачі інформації, що випливають з обмеження часу на прийняття обґрунтованого рішення, і слабоформалізованістю та неформалізованістю завдань, які вирішуються. Поява засобів автоматизації управління на базі нових інформаційних технологій надає можливість покласти на ЕОМ такі функції: розпізнавання тих чи інших особливих ситуацій, класифікація цих ситуацій, знаходження варіантів рішень відповідно до мети, видача рекомендацій та їх обґрунтування. Реалізація на ЕОМ відповідних функцій здійснюється в межах концепції побудови інтелектуальних систем підтримки прийняття рішень (СППР), а теоретичним фундаментом є інженерія знань [1].

Одним з *актуальних* питань у цьому напрямку є розробка тезауруса предметної сфери (ПС) процесу управління підрозділами, органами, управліннями (далі – підрозділами), що подаються як організаційно-технічні системи (ОТС) під час подій, надзвичайних, кризових ситуацій (далі – особливих ситуацій).

**Аналіз останніх досліджень і публікацій.** У роботах, присвячених дослідженню семантики інформаційних технологій [2 – 5], зазначено, що універсальної автоматизованої технології розробки семантичного тезауруса не існує. Це пов'язано з труднощами і неоднозначностями подання природно-мовних описів ПС. Особливістю тезаурусного способу подання знань є інтуїтивний характер виявлених відношень, нечіткість у визначенні сили зв'язку між термінами.

**Мета статті** – визначення підходів до вирішення даного завдання з урахуванням особливостей оперативно-службової діяльності ДПСУ.

### Виклад основного матеріалу

Терміном "тезаурус" зазвичай позначається список лексичних одиниць (словоформ, канонічних форм слова, основ, словосполучень), між якими задано смислові зв'язки як ієрархічного (родовидового), так і неієрархічного (синонімія, антонімія, асоціація) типів [2]. Тезаурус, таким чином, містить основні семантичні категорії ПС та встановлені між ними парадигматичні (обумовлені предметно-логічними, а не мовними чинниками) відношення.

У запропонованій згідно з концепцією [1] інтелектуальній СППР розробка тезауруса має важливе значення, оскільки тезаурус є основою баз знань (БЗ), які включають прецеденти. Тезаурус потрібен для побудови індексної структури БЗ прецедентів і розширеного пошуку в БЗ, що надає можливість знаходити необхідні описи прецедентів. Унаслідок специфіки процесів ідентифікації і прийняття рішень у ДПСУ в особливих ситуаціях (ОС) процес розробки тезауруса залишає сфери невизначеності для деяких зв'язків між термінами в списку зв'язків тезауруса, тому в результаті ми маємо нечіткий тезаурус.

Тезаурус ПС процесу управління підрозділами пропонується подати у вигляді ієрархічного семантичного графа $\mathrm{Th}(\mathrm{T}, \mathrm{R})$, у вершинах якого знаходяться семантичні категорії (терміни) – T, а ребра визначають відношення $\mathrm{R}^{z}(\mathrm{t_i}, \mathrm{t_j})$ між семантичними категоріями $\mathrm{t_i}, \mathrm{t_j} \in \mathrm{T}$, що формалізуються у вигляді матриці суміжності $Z \in Z\{\mathrm{F}, \mathrm{A}, \mathrm{H}, \mathrm{S}\}$, де $\{\mathrm{F}, \mathrm{A}, \mathrm{H}, \mathrm{S}\}$ – типи семантичних відношень: F – функціональної зв'язаності; H – ієрархії (спадкоємності); A – асоціативної



© О.С. Андрощук

зв'язаності; S – семантичної еквівалентності.

Термін – це одиниця якої-небудь конкретної природної або штучної мови (слово, словосполучення, абревіатура, символ, поєднання слова і буквсимволів, поєднання слова та цифр-символів), що в результаті стихійно складеної або особливо свідомої колективної домовленості має спеціальне термінологічне значення, яке може бути виражене або у словесній формі, або в тому чи іншому формалізованому вигляді і достатньо точно та повно відображає основні суттєві на даному рівні розвитку науки і техніки ознаки відповідного поняття [4].

Тезауруси зазвичай використовують фахівці, які працюють із документальними інформаційно-пошуковими системами [5], завдяки широкому використанню Інтернет технологій.

Розглянемо детальніше етапи розробки тезауруса ПС з управління підрозділами в ДПСУ на прикладі ОС.

Спочатку розглянемо формування словника термінів ПС на основі лінгвістичного аналізу текстів. У більшості робіт, які присвячені цьому питанню [2, 3], пропонується для розробки словника та тезауруса ПС використовувати лінгвістичний аналіз текстів і побудову матриць схожості – відмінності понять за участю експертів. Кількість рядків та стовпців матриці дорівнює кількості слів-понять у словнику ПС. Очевидно, що для реальних обсягів словників ПС (від декількох сотень до тисяч понять) такий метод є дуже трудомістким і слід використовувати спеціальні математичні та програмні методи і засоби аналізу текстів. При розробці словника ПС використовуються основні поняття лінгвістичної статистики, які містяться в роботах з природно-мовного опису ПС [2, 4, 6].

Першим теоретичним результатом у галузі статистичного аналізу тексту вважається емпіричний закон, встановлений Ципфом, що називається законом частот слів. Закон пов'язує гіперболічною залежністю частоту слова, що зустрічається в тексті, з рангом цього слова в списку, впорядкованому щодо зменшення частот:

$$f(k, r) = pkr^{-\gamma}, \qquad (1)$$

де f – частота слова в тексті; k – загальна кількість слів у тексті; r – ранг слова (кількість словоформ або слів, які зустрілися f та більше разів, що визначає порядковий номер слова у впорядкованому щодо зменшення частотної функції словнику); p і $\gamma$ – параметри, визначені Ципфом. Він припустив, що для природних мов у першому наближенні коефіцієнт p = 0,1, $\gamma$ = 1, у зв'язку з чим формула (1) приймає вигляд

$$f(k, r) = 0,1/r. \qquad (2)$$

Проте, подальші дослідження текстів у різних природно-мовних системах (різних мовах) не підтвердили точного співвідношення (2) для знайдених Ципфом коефіцієнтів. Мандельброт [6] запропонував іншу формулу для опису "закону частот слів", в якій було враховано названі невідповідності шляхом введення нового параметра v:

$$f(k, r) = pk(r + v)^{-\gamma}. \qquad (3)$$

Графіки залежності f та r на рис. 1 демонструють закони Ципфа і Мандельброта, які зазначають, що добуток частоти використання слів та рангового порядку є приблизно константою.

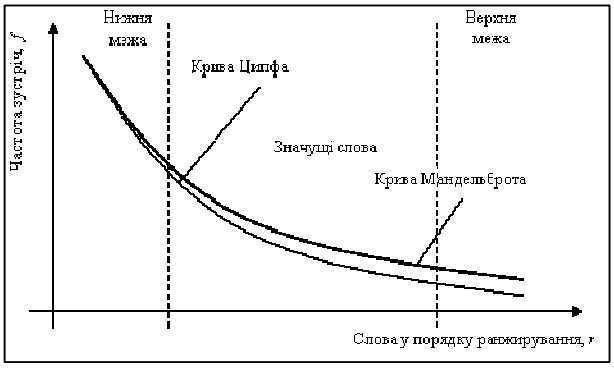


Рис. 1. Графіки залежності частоти слів у тексті від рангу слів

Упровадження закону Ципфа стосовно автоматизованого текстового аналізу здійснив Лун [6]. Його припущення полягає в тому, що ці частоти можуть використовуватися, щоб витягнути слова і пропозиції подання документа. Лун використовував це як нульову гіпотезу, щоб визначати дві межі значущості слів – верхню та нижню (рис. 1).

Найбільш простим прикладом статистичного аналізу текстів є обчислення частоти використання (зустрічі) окремих слів із подальшим поданням цього тексту відповідним частотним списком слів (словником).

У більшості систем лінгвістичного аналізу [2, 4, 5] було виявлено, що розрахунок частот зустрічі слів без урахування контексту і різноманітних семантичних зв'язків між окремими словами і словосполученнями насправді є малоефективним інструментом. Зважаючи на це, більшість робіт останніх років з автоматичної обробки текстів статистичними методами базувалася не на розрахунках частот окремих слів, а на обчисленнях коефіцієнтів статистичної подібності (асоціації) між словами і документом та класифікації термінів і документів. Ці підходи реалізовано у низці програмних комплексів автоматизованого аналізу текстів.

У процесі досліджень у результаті аналізу текстів із використанням Text Analyst 2.0 було складено словники термінів у таких ПС: здійснення прикордонного контролю у пунктах пропуску; здійснення прикордонної служби на ділянці відділу прикордонної служби; управління прикордонними підрозділами у ОС, що виникають. На рис. 2 наведено скрин-шот з результатами аналізу програми Text Analyst. Для ана-

лізу було використано тексти описів ОС і документів, що регламентують процес управління під час надзвичайної ситуації. Одержаний словник $Dic(T_{TX})$ є однією зі складових при формуванні тезауруса ПС.

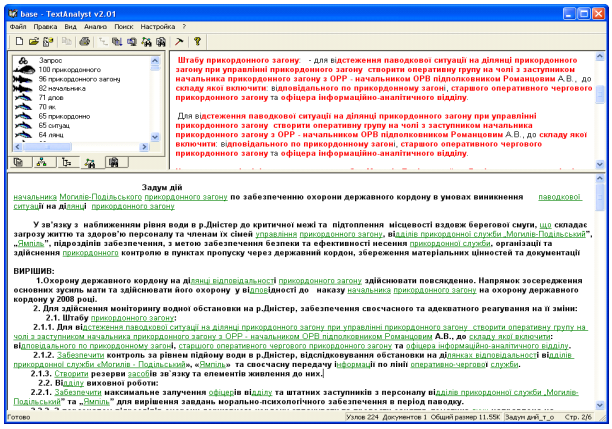

Рис. 2. Результат аналізу тексту з використанням Text Analyst

У табл. 1 наведено фрагмент словника ПС управління Могилів-Подільським прикордонним загоном під час повені 26 липня 2008 року, одержаний з використанням Text Analyst.

Таблиця 1

Фрагмент словника предметної сфери з управління прикордонними підрозділами

| № | Батьківський термін | Частота | Вага | Термін підлеглий |
|---|---|---|---|---|
| 1 | прикордонний | 88 | 99 | відділ |
| 2 | паводок | 30 | 82 | підтоплення |
| 3 | кордон | 90 | 87 | рубіж |

Отже, формується деяка матриця типу "термін – документ – вага", що відображає внесок кожного з термінів словника ПС у розкриття змісту кожного з включених у розгляд документів. Фрагмент такої матриці наведено на рис. 3. Тут $w_{ij}$ – вага і-го терміну в описі j-го документа.

Множина записів матриці "термін – документ" формує індексну базу даних, призначену для пошуку регламентуючих документів процесу управління (рішення) в ОС.

| | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ | $T_7$ |
|---|---|---|---|---|---|---|---|
| $D_1$ | 0,27 | 0,05 | 0,15 | 0,46 | 0,00 | 0,12 | 0,00 |
| $D_2$ | 0,57 | 0,22 | 0,00 | 0,13 | 0,00 | 0,26 | 0,45 |
| $D_3$ | 0,24 | 0,45 | 0,00 | 0,49 | 0,00 | 0,10 | 0,00 |
| $D_4$ | 0,34 | 0,28 | 0,88 | 0,00 | 0,00 | 0,00 | 0,00 |
| $D_5$ | 0,00 | 0,00 | 0,00 | 0,00 | 0,11 | 0,00 | 0,87 |
| $D_6$ | 0,00 | 0,00 | 0,00 | 0,00 | 0,96 | 0,15 | 0,00 |
| $D_7$ | 0,26 | 0,00 | 0,00 | 0,21 | 0,00 | 0,54 | 0,00 |

Рис. 3. Приклад матриці "термін – документ"

Для більш глибокого семантичного аналізу ПС запропоновано ввести додаткові способи виокремлення процедурних знань. Отже, виявляється можливим урахувати функціональні відношення між поняттями, прив'язавши їх до контексту конкретного процесу управління. Такі відношення пропонується встановлювати в процесі моделювання процесу управління.

Розглянемо формування словника термінів ПС процесу управління підрозділами в ОС на основі результатів об'єктно-орієнтованого моделювання.

Мова об'єктної моделі є поєднанням метамови UML [7], яка функціонує у сфері термінології та позначення вузьких понять моделювання, і природної мови, що є основою словника ПС. У словнику моделі визначаються та містяться всі терміни, які підвищують ступінь розуміння ПС і виключають ризик виникнення розбіжностей при її обговоренні. Таким чином формується множина термінів $T_M$ – повна множина інформаційних елементів, що складається з множини сутностей $T_{Mc}=\{t_{Mc}\}$ та множини атрибутів $T_{Ma}=\{t_{Ma}\}$; $T_M = T_{Mc} \cup T_{Ma}$.

Об'єктна модель містить інформацію щодо синонімічних ієрархічних відношень і тематичних асоціацій між поняттями, які пропонується використовувати разом із частотними характеристиками понять при створенні тезаурусу ПС.

Для формування словника ПС слід здійснити виокремлення сутності з використанням стандартних засобів генерації звітів про результати об'єктного моделювання. Це надасть можливість, крім статистичної, морфологічної та синтаксичної інформації, урахувати функціональні зв'язки між наявними індексними термінами, іменами класів та об'єктів, а також їх атрибутами на діаграмах моделі. Робота над словником підтримується впродовж усього процесу моделювання. На рис. 4 показано приклад формування відношень між сутностями в моделі організаційного управління і взаємодією керівників підрозділу (органу, управління) при виникненні ОС (діаграма класів UML).

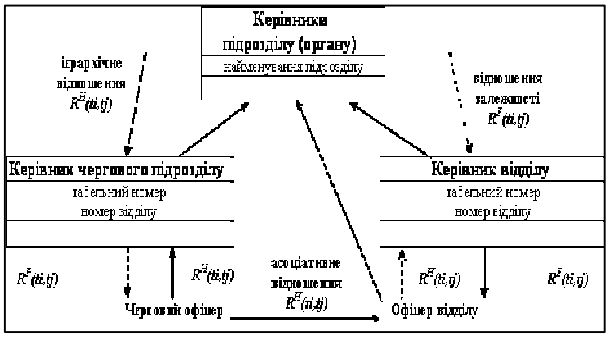


Рис. 4. Приклад формування відношень між сутностями моделі організаційного управління

Отже, для побудови тезауруса ПС управління підрозділами в ОС пропонується модель семантичного аналізу, особливістю якої є інтеграція в предметно-орієнтованому тезаурусі понять, що використовуються в текстах опису ПС, текстах описів ОС і

моделях процесів управління. Модель семантичного аналізу є відображенням вигляду E: M→Th, що перетворює множину інформаційних елементів моделі тексту $M = \left\{T, R^Z\right\}$ у множину термінів і семантичних відношень тезауруса $Th = \left(T, R^Z, W\right)$, де T – терміни; $R^Z$ – характер відношень між термінами; W – вага відношень термінів у тезаурусі моделі.

На основі моделі розроблено алгоритм семантичного аналізу, основними етапами якого є: складання словника термінів ПС; формування матриці схожості термінів на основі лінгвістичного аналізу текстової маси; формування матриці схожості термінів за наслідками моделювання процесу, що досліджується; об'єднання цих матриць і встановлення вагомостей відношень між термінами.

За частотою зустрічі в описі моделі процесу двох будь-яких термінів, пов'язаних одним із певних типів відношень, формується ваговий коефіцієнт, що характеризує силу їх зв'язаності. Для визначення "сили зв'язку" використовується теорія нечіткої логіки, диференціюючи коефіцієнт зв'язаності понять залежно від частоти зустрічі цих понять на основі нечітких правил [8].

За наслідками об'єктно-орієнтованого моделювання будується множина триплетів виду $t_i\ F\ t_j$, де F – предикат, що визначає відношення між $t_i$ і $t_j$ (із перерахованих вище типів відношень), за якими будується матриця відношень для двовимірного матричного аналізу.

Використання об'єктно-орієнтованих моделей (логічна модель, діаграма класів UML) надає можливість аналізувати документи в контексті функціонування системи управління. У табл. 2 наведено фрагмент словника $Dic(T_M)$, складеного на основі цієї діаграми.

Результат об'єднання словника лінгвістичного аналізу $V_{TX}\left(T_{TX}\right)$ і словника об'єктно-орієнтованого аналізу $V_M\left(T_M\right)$ будемо називати об'єднаним словником $V(T)$. Множина T є множиною термінів ПС та утворюється шляхом об'єднання термінів множини $T_{TX}$ і $T_M$ за формулою:

$$T = T_{TX} + T_M - T_{TXM}. \qquad (4)$$

Таблиця 2

Фрагмент словника моделі предметної сфери

| Термін основний | Категорія | Термін підлеглий |
|---|---|---|
| керівник | клас | черговий |
| начальник відділу | клас | офіцер відділу |
| найменування відділу | атрибут | відділ |

Однакові терміни, які належать множині термінів різних словників, замінюються одним. В об'єднаний словник терміни словників $V\left(T_{TX}\right)$ і $V\left(T_M\right)$ входять зі своїми вагами, а для однакових термінів $T_{TX.M}$ будується сумарна вагова функція.

Проведення класифікації термінів породжує групи асоційованих термінів, придатних до включення в тезаурус. Групи при цьому можуть бути непов'язаними одна з одною або, навпаки, між ними може бути визначено відношення. Якщо за своєю природою відношення між групами термінів належать до родового типу, то одержують ієрархії термінів; в інших випадках групи можуть бути впорядковані у вигляді двовимірної семантичної мережі. Кластер семантичних відношень фактично формує структуру спрямованого ациклічного графа, що є фрагментом тезауруса. Відношення між термінами понять переносяться з об'єктно-орієнтованої моделі [9]. На рис. 5 наведено приклад семантичної мережі термінів тезауруса, одержаний за наслідками кластерного аналізу і моделювання [9].

Одержана семантична мережа є наочним засобом обговорення при автоматизованій побудові тезауруса. Експерти переглядають та обговорюють граф тезауруса, ідентифікуючи відношення між індексними термінами і, можливо, вводячи нові терміни, які не були виокремлені при первинному застосуванні частотного аналізу. Експертам може бути запропоновано виразити ступінь семантичної зв'язності між термінами тезауруса за допомогою деякої лінгвістичної змінної

$$L_{ij} = \left\{l_{ij}^1, ..., l_{ij}^s\right\}.$$

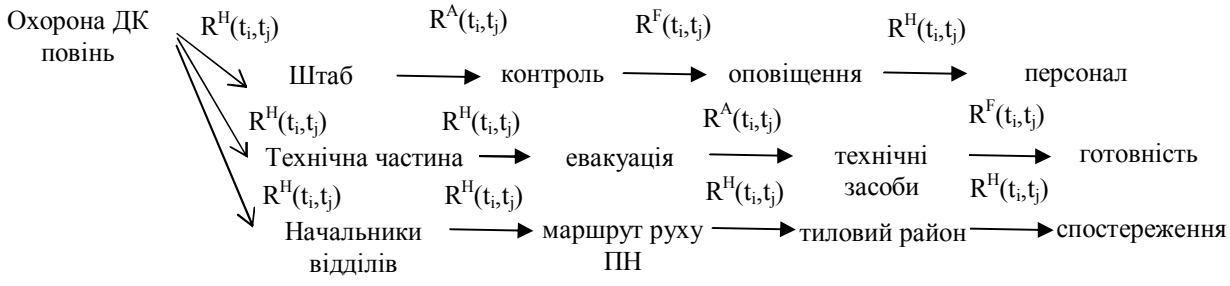


Рис. 5. Приклад семантичної мережі термінів тезауруса

Графічне подання тезауруса є ефективним засобом отримання знань від експертів. Проте, семантична мережа для опису ОС, що виникають у реальних складних динамічних об'єктах, містить десятки термінів і є дуже складною для сприйняття експертами. Тут необхідно використовувати формальну модель тезауруса у вигляді матриці семантичної суміжності термінів ПС $W = \left\| w_{ij} \right\|$, яка є квадратною матрицею, що проіндексована по обох осях множиною індексних термінів Т:

$$w_{ij} = \begin{cases} w_{ij}(R^{\rho}(t_i, t_j)), & \exists t_i R^{\rho} t_j, \rho = \{H, A, S, F\} \\ 0, & \text{в іншому випадку} \end{cases}. \quad (5)$$

Цій матриці ставиться у відповідність зважений орграф тезауруса $\mathrm{Th}\left(T, R^{z}\right)$, множиною вершин якого є терміни $t \in T$, а ребра орграфу відображають наявність або відсутність семантичної зв'язності між термінами тезауруса, з вагою ребра $\left(t_i, t_j\right)$, яке дорівнює $w_{ij}\left(R^{\rho}\left(t_i, t_j\right)\right)$.

## Висновки

У роботі подано модель побудови тезауруса предметної сфери управління підрозділами в ОС у ДПСУ на основі інтеграції лінгвістичного аналізу текстів, які регламентують процес управління в ОС, та результатів об'єктного-орієнтованого моделювання процесів управління в ОС. Класифікація сутностей, виявлених за допомогою об'єктного-орієнтованого моделювання процесів управління з формуванням словника надасть можливість, крім статистичної, морфологічної і синтаксичної інформації, урахувати функціональні зв'язки між окремими поняттями, що входять в опис ОС.

Використання відповідного предметно-орієнтованого тезауруса в процесах формування рекомендацій щодо прийняття рішень надає можливість забезпечити більшу повноту і точність пошуку рішень.

**Напрямком подальших досліджень** слід вважати семантичне ототожнення термінів опису ОС, яке можна використати при розробці алгоритму пошуку прецедентів у БЗ інтелектуальних СППР.

## Список літератури

*1. Вертузаєв М.С. Розробка концепції підтримки прийняття рішень при управлінні складними державними системами в особливих умовах на основі інженерії знань / М.С. Вертузаєв, О.С. Андрощук // Збірник наукових праць. – Хмельницький: Видавництво Національної академії Державної прикордонної служби України ім. Б. Хмельницького, 2008. – № 41, Ч. ІІ. – С. 45-47.*

*2. Филиппович Ю.Н. Семантика информационных технологий: Опыты словарно-тезаурусного описания / Ю.Н. Филиппович, А.В. Прохоров. – М.: МГУП, 2002. – 368 с.*

*3. Quillian M.R. Semantic memory. In Semantic Information Processing (Minsky M., eds) / M.R. Quillian. – Cambridge, MA: MIT Press. – P. 227-270.*

*4. Пиотровский Р.Г. Текст, машина, человек / Р.Г. Пиотровский. – Л.: Наука, 1975. – 327 с.*

*5. Организация работы с документами / В.А. Кудряев и др. – М. : ИНФРА-М, 1999. – 575 с.*

*6. Мандельброт Б. Теория информации и психолингвистика: теория частот слов / Б. Мандельброт // Математические методы в социальных науках: сб. статей под ред. П. Лазарсфельда и Н. Генри. – М.: Прогресс, 1973. – С. 42-47.*

*7. Буч Г. Язык UML : руководство пользователя / Г. Буч, Д. Рамбо, А. Джекобсон; [пер. с англ.]. – М.: ДМК, 2000. – 432 с.*

*8. Нечеткие множества в моделях управления и искусственного интеллекта / [А.Н. Аверкин, И.З. Батыршин, А.Ф. Блишун та ін.]; под ред. Д.А. Поспелова. – Наука, 1986. – 312 с.*

*9. Андрощук О.С. Структурні моделі складових інтелектуальної системи підтримки прийняття рішень / О.С. Андрощук // Збірник наукових праць Національна академії Державної прикордонної служби України ім. Б. Хмельницького. – Хмельницький, 2008. – № 45. Ч. ІІ. – С. 25-29.*

*Надійшла до редколегії 28.10.2010*

**Рецензент:** д-р техн. наук, проф. М.С. Вертузаєв, Інститут Служби зовнішньої розвідки, Київ.

### МОДЕЛЬ ПОСТРОЕНИЯ ТЕЗАУРУСА ПРЕДМЕТНОЙ ОБЛАСТИ УПРАВЛЕНИЯ ПОДРАЗДЕЛЕНИЯМИ В ОСОБЕННЫХ СИТУАЦИЯХ

О.С. Андрощук

*В статье подан комплексный подход к построению тезауруса предметной сферы управления подразделениями, органами, управлениями Государственной пограничной службы Украины на основе интеграции лингвистического анализа текстов, которые регламентируют процесс управления, и результатов объектно-ориентированного моделирования процессов управления в особенных ситуациях. Использование разработанного тезауруса в процессах формирования рекомендаций относительно принятия решений предоставляет возможность обеспечить большую полноту и точность поиска решений.*

***Ключевые слова:*** *моделирование, анализ, тезаурус, база знаний, особенная ситуация.*

### MODEL OF CONSTRUCTION THESAURUS OF SUBJECT SPHEREOF MANAGEMENT BY SUBSECTIONS IN THE SPECIAL SITUATIONS

O.S. Androshchyk

*In the article complex approach is given to construction of thesaurus of subject sphere of management by subsections, organs, managements of Government boundary service of Ukraine on the basis of integration of linguistic analysis of texts, which regulate the process of management, and results of the object-oriented design of management processes in the special situations. The use of developed thesaurus in the processes of forming of recommendations in relation to the decision-making gives possibility to provide greater plenitude and exactness of search of decisions.*

***Keywords:*** *design, analysis, thesaurus, knowledges base, special situation.*