HU: collection of scientific works. Kharkov, KhNAHU, 56, 26 – 29. **5.** *Shevchenko, V. A.* (2012). Validation of the simulation results of nonparametric statistics methods. Vestnik NTU «KPI». Kharkov, NTU «KPI», Issue. 34, 75 – 79.**6.** *Shevchenko, V. A.* (2012). Distribution of students into typological groups by means of cluster analysis, depending on the factors influencing the students progress, Collection of scientific works of the international scientific and methodological conference "The problems of Integration of National Higher Educational Institutions into the European Educational Medium". V. 2 "Modern Approaches Concerning Higher Education Quality Provision". Kharkov, KhNAHU, 120 – 123.**7.** *Shevchenko, V. A.* (2013). Predicting student performance using the methods of cluster analysis. Expert assessments of the elements of the educational process: Materials XI Interuniversity scientific and practical conference. Kharkov: NUA, 112 – 115.**8.** *Meteshkin, K. A., Shevchenko, V. A.* (2012). Vague understanding of students clusterization results. Public information and computer integrated technologies : collection of scientific works. Kharkov, KhAI, 56, 201 – 208.**9.** *Shevchenko, V. A.* (2015). Graphical representation of the function of student typological group membership, depending on the academic performance. Vestnik NTU «KPI». Kharkov, NTU «KhPI», 14 (1123), 15 – 20.**10.** *Shevchenko, V. A.* (2013). Checking the effectiveness of teaching students using the method of nonparametric statistics. Vestnik KhNAHU: collection of scientific works. Kharkov, KhNAHU, 60, 18 – 21.

*Надійшла (received) 29.04.2015*

**УДК 004.91**

***О. В. БІСІКАЛО,*** д-р техн. наук, проф., декан, ВНТУ, Вінниця;
***А. І. ЛІСОВЕНКО,*** аспірант, ВНТУ, Вінниця;
***О. В. ЯХИМОВИЧ,*** магістр, ВНТУ, Вінниця;
***С. С. ТРАЧЕНКО,*** студент, ВНТУ, Вінниця

# ВИЗНАЧЕННЯ ЗМІСТОВНИХ ОЗНАК ТЕКСТУ НА ОСНОВІ АНАЛІЗУ ЗВ'ЯЗКІВ МІЖ ЛЕКСИЧНИМИ ОДИНИЦЯМИ

Визначено змістовні ознаки і характеристики англомовного тексту на основі дослідження зв'язків між лемами та синсетами, що розпізнано лінгвістичними пакетами. Результати у вигляді списків ключових слів, елементів онтологій та змістовних кластерів понять отримано на прикладі «Address by President of the Russian Federation 2013/2014». Проведене дослідження було здійснено за допомогою пакетів DKPro Core та NLTK.

**Ключові слова:** лема, синсет, зв'язок, ключові слова, стоп-слово, елемент онтології, гіпероніми, кластер, DKPro, NLTK.

**Вступ.** Вилучення знань з природно-мовних текстів стає одним з найбільш актуальних напрямів досліджень в комп'ютерній лінгвістиці завдяки надшвидкому збільшенню обсягів електронної інформації та рівня її доступності через мережу Інтернет. Найбільш відомі дослідження в цьому напряму проведено на основі статистичного аналізу закономірностей розподілу слів природно-мовного тексту. Більш релевантні результати досягаються за допомогою додаткового лінгвістичного аналізу тексту – найкращу ситуацію маємо з практично завершеним морфологічним аналізом, значно слабші результати демонструє синтаксичний аналіз речень, а в семантичному аналізі і до цього часу не вирішено низку проблемних питань. Такий підхід пояснюється не тільки легкістю фіксації окремого символу/слова у текстовому файлі, але й, у більшій мірі, пануючими лінгвістичними концепціями, що надають слову основоположне значення.

Альтернативним шляхом до розв'язання проблеми вилучення знань з текстової інформації може стати підхід до формалізації методів образного аналізу та синтезу природно-мовних конструкцій [1]. Відмінністю підходу є перенесення акцентів з аналізу окремих лексичних одиниць на аналіз зв'язків між цими одиницями як наслідок результатів моделювання асоціативного образного мислення людини. Внаслідок цього з'являється можливість а) введення додаткового образного рівня аналізу та синтезу до традиційної лінгвістичної тріади морфологія–синтаксис–семантика [2] та б) технологічної підтримки цього рівня завдяки реалізованим у сучасних лінгвістичних пакетах функціям визначення синтагматичних та парадигматичних зв'язків між словами/словоформами/лемами речення. Потребують подальшого розвитку та прозорої перевірки формальні методи визначення ключових слів [3] на основі запропонованого підходу, у т.ч. з залученням синсетів з *WordNet*.

**Мета роботи.** Метою роботи є формальне визначення змістовних ознак і характеристик англомовного тексту на основі досліджень зв'язків між лемами та синсетами на прикладі «Address by President of the Russian Federation 2013/2014». Розпізнавання лексичних одиниць та зв'язків між ними необхідно здійснити за допомогою лінгвістичних пакетів *DKPro Core* та *NLTK*.

**Методика експериментів.** Для проведення експерименту було обрано тексти «Presidential Address to the Federal Assembly 2013» [4] та «Presidential Address to the Federal Assembly 2014» [5]. Додатковим аргументом на користь такого вибору є визначення формальних ознак тих актуальних суспільно-політичних змін в Україні, які відбулися за останній рік і відношення до яких потенційно має відображатися у офіційних текстах звернень президента Росії 2013 та 2014 років. Певне порівняння запропонованої виключно формальної обробки текстів можна зробити з експертною оцінкою на основы частотного словника [6], яка, власне і стала поштовхом для проведення даного дослідження.

З метою технологічної реалізації експерименту було розроблено програмне забезпечення у межах двох відомих лінгвістичних пакетів – *DKPro Core* на мові *Java* [7] та *NLTK* на мові *Python* [8]. Обидва пакети належать до *Open Source* та мають у своєму складі розвинені бібліотеки з функціями, достатніми для реалізації задач дослідження.

*Java*-програму для визначення ключових слів засобами *DKPro Core* було розроблено на основі [9]. Ключові слова визначалися 2-ма способами – традиційним, на основі визначення частотного словника без стоп-слів та альтернативним, на основі зв'язків між словами/лемами тексту. Схема алгоритму роботи програми на основі зв'язків представлена на рис. 1.

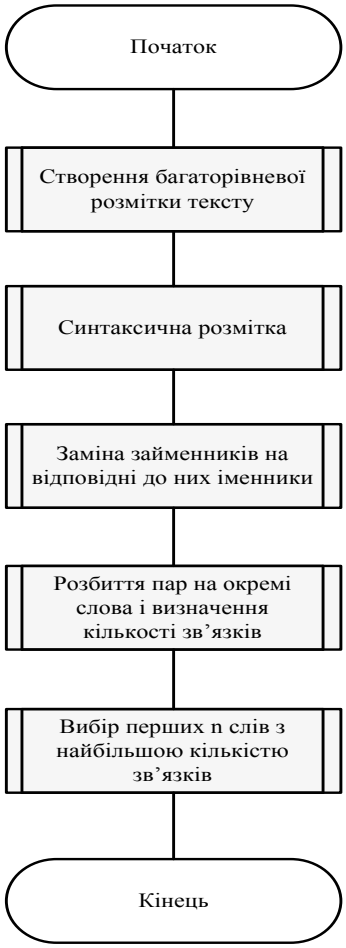


Рис. 1 – Схема алгоритму роботи програми визначення ключових слів на основі *DKPro Core*

Визначення ключових слів запропонованим алгоритмом відбувається за декількома послідовними

етапами:

а) створення багаторівневої розмітки тексту;

б) синтаксична розмітка у межах кожного речення, що враховує складні залежності між парами лем;

в) заміна займенників в отриманих парах на відповідні до них іменники;

г) розбиття пар на окремі слова (леми) і визначення загальної кількості зв'язків, що відповідають кожному такому слову;

д) вибір перших *n* слів з найбільшою кількістю зв'язків, де *n* – кількість потрібних ключових слів.

Внаслідок реалізації експерименту на основі пакету *NLTK* на мові *Python* було побудовано онтології обох обраних текстів. Побудова онтології відбувалась наступним чином: із текстів природної мови було вилучено усі *SynSet* (синсети) на основі бібліотеки [10]. Далі, із кожного синсету було вилучено усі гіпероніми, що представляються також класом *SynSet*. Після чого перевірялось, чи є в отриманому списку гіперонімів ті синсети, які було розпізнано в тексті раніше. Якщо так, то його вага (кількість входжень даного синсету як гіпероніму інших) збільшується на одиницю. На виході отримуємо відсортований список синсетів, що представляють базові поняття тексту, що було подано на вхід.

З метою порівняння для обох текстів визначалися аналогічні відсортовані списки синсетів, в яких було враховано всі зв'язки, що підтримуються класом *SynSet*, а не тільки гіперонімічні. Окрім того, шляхом фільтрування за найбільшою вагою в графічному вигляді визначалися кластери найбільш тісно пов'язаних між собою синсетів.

**Обговорення експериментальних даних.**

*а) Результати, отримані з використанням програмного забезпечення на основі пакету DKPro Core.*

Для тексту «Presidential Address to the Federal Assembly 2013» [4] отримані перші 10 ключових слів з вагами від 343 до 80 за власною розробкою – разом зі стоп-словами це: we, need, work, I, make, be, system, have, authorities, development, ask, people. З них be і have – стоп-слова. Перші 10 ключових слів разом зі стоп-словами (ваги від 143 до 32), для частотного словника: that, is, we, will, I, our, be, are, must, it, have, not, all, their, work, Russia, need, should, also, Russian, system. Стоп-слова для частотного словника: that, is, will, be, are, it, have, not, their, should, also. З результатів пошуку ключових слів видно, що при знаходження однакової кількості ключових слів власна розробка має 2 стоп-слова, а частотний словник 11. Однаковими ключовими словами є 5, проте їх ваги суттєво відрізняються we (343/111), I (138/92), work (155/44), need (157/42), system (119/32).

Окрім того, власна розробка виявила 3 найбільш вагомі словосполучення – We need (36), I ask (22) та We make (11).

Для тексту «Address by President of the Russian Federation 2014» [5] перші 10 ключових слів за власною розробкою з вагами від 287 до 54 (разом із стоп-словами): right, people, Russia, have, be, work, I, do, provide, create, support, make, this, like. Де have, be, do, this – стоп-слова. Перші 10 ключових слів зі стоп-словами (ваги від 106 до 19) для частотного словника: that, we, will, I, be, is, Russia, our, are, have, it, should, not, all, people, must, has, their, was, also, its, they, who, Russian, them, work, national, can, what, course. Стоп-слова для частотного словника: that, will, be,

is, our, are, have, it, should, not, has, their, was, also, its, they, who, them, can, what. З результатів видно, що при знаходження однаковоє кількості ключових слів власна розробка має 4 стоп-слова, а частотний словник 20. Однаковими ключовими словами, проте з різними вагами є такі 4: I (105/87), Russia (178/65), people (194/39), work (108/20).

На відміну від 2013 року власна розробка не виявила найвагоміших сталих словосполучень, проте найбільш часто в знайдених зустрічалися слова I, right та like. На думку авторів цікавим для семантичного аналізу результатом формального дослідження є власне склад перших п'ятірок ключових слів у зверненнях президента РФ до власного народу – у 2013 році йшлося про we, I, work, need, system; у 2014 році вже чуємо – I, Russia / Russian, people, work – акценти зрозумілі та доповнюють висновки з [6].

*б) Результати, отримані з використанням програмного забезпечення на основі пакету NLTK.*

Із представлених текстів «Presidential Address to the Federal Assembly 2013/2014» [4, 5] за допомогою пакету *NLTK* було відібрано ті синсети, які мають найбільшу «вагу» по входженню в список гіперонімів інших синсетів (табл. 1). Гіперонімічний зв'язок було обрано тому, що гіперонім як поняття в ставленні до інших понять виражає загальнішу сутність (родове поняття) та грає роль основи онтології тексту.

Таблиця 1 – Порівняльна таблиця гіперонімічних синсетів з «Presidential Address to the Federal Assembly 2013/2014»

| № | 2013 рік | | 2014 рік | |
|---|---|---|---|---|
| | SynSet | Вага SynSet | SynSet | Вага SynSet |
| 1 | Synset('**event**.n.01') | 98 | Synset('**event**.n.01') | 102 |
| 2 | Synset('**organization**.n.01') | 18 | Synset('person.n.01') | 71 |
| 3 | Synset('***move***.v.02') | 17 | Synset('**change**.v.01') | 26 |
| 4 | Synset('**change**.v.01') | 16 | Synset('***move***.v.02') | 19 |
| 5 | Synset('**property**.n.02') | 12 | Synset('**organization**.n.01') | 16 |
| 6 | Synset('***evaluate***.v.02') | 10 | Synset('**property**.n.02') | 13 |
| 7 | Synset('**thing**.n.12') | 8 | Synset('**thing**.n.12') | 9 |
| 8 | Synset('sum.n.01') | 6 | Synset('***decide***.v.01') | 8 |
| 9 | Synset('district.n.01') | 6 | Synset('***evaluate***.v.02') | 6 |
| 10 | Synset('adult.n.01') | 6 | Synset('mechanism.n.05') | 6 |
| 11 | Synset('method.n.01') | 5 | Synset('asset.n.01') | 6 |
| 12 | Synset('gain.n.04') | 5 | (Synset('choose.v.01') | 5 |
| 13 | Synset('***decide***.v.01') | 4 | Synset('equipment.n.01') | 5 |

На основі даних таблиці отримані синсети та ваги зв'язків між ними на рис. 2, 3 представлено графами онтологічних кластерів.

Проведений аналогічний експеримент для побудови упорядкованого списку з усіма вагами зв'язків класу *SynSet* показав, що гіперонімічний зв'язок є визначальним, оскільки 5 ключових синсетів виявилися однаковими для всіх 4-х списків –

це event, organization, change, property, thing (позначені у таблиці 1 жирним шрифтом). Тільки три дієслова move, decide та evaluate (жирний шрифт з нахилом) не увійшли до загальних списків на відміну від гіперонімічних, але в останніх також співпали.

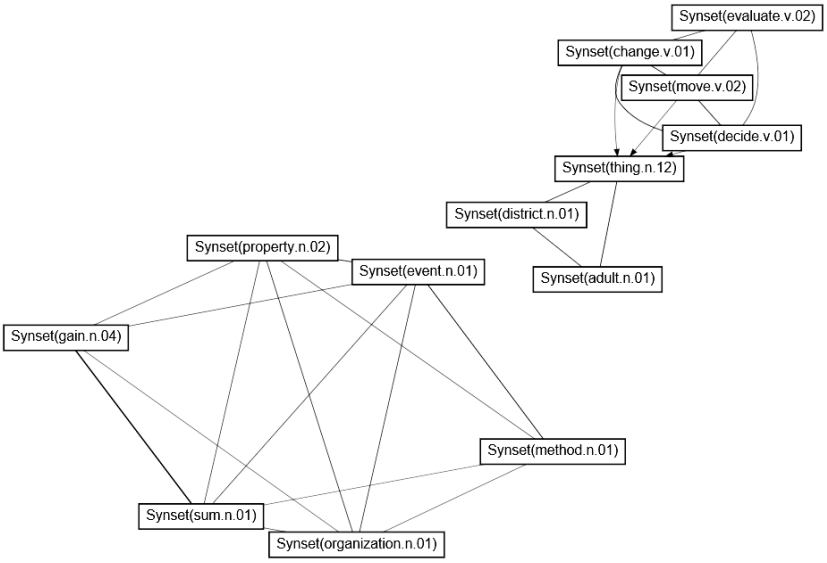


Рис. 2 – Граф зв'язків «найважчих» *SynSet* в «Presidential Address to the Federal Assembly 2013»

У більшості 13 перших елементів кожного гіперонімічного списку склали іменники, які доповнюють 4 дієслова у 2013 році та 5 дієслів у 2014 році, що також свідчить про онтологічну сутність отриманих даних таблиці 1 та кластерів з рисунків 2 та 3. Характерно, що у кожному рисунку маємо по 2 кластери з 13 синсетів, хоча для 2013 року вони розподіляються як 7 +6 синсетів, а у 2014 році – 9 + 4 синсети.

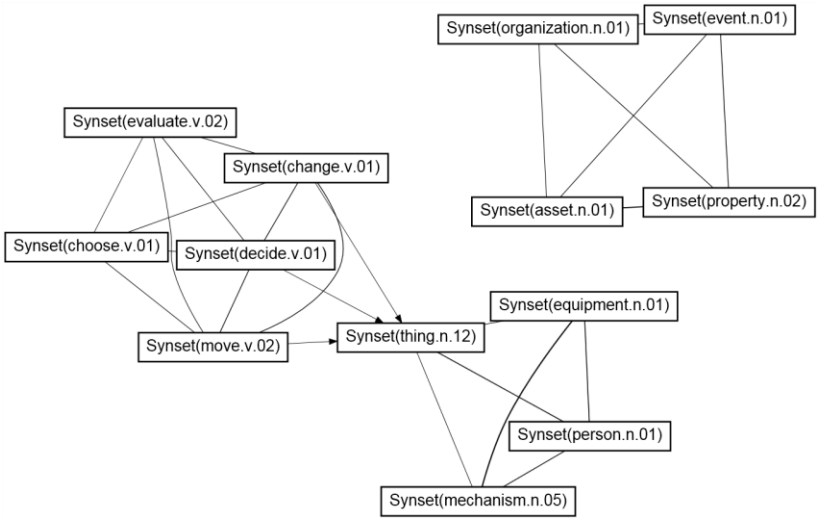


Рис. 3 – Граф зв'язків «найважчих» SynSet в «Presidential Assembly 2014»

Проте основи кожного кластеру з незмінних для двох списків 8-ми синсетів збігаються, що може свідчити про єдність авторського стилю. Тоді логічно, що переміщення акцентів викладеної в текстах інформації має знайти своє відображення у відмінностях топології кластерів.Формально з синсетів sum, district, adult, method та gain 2013 року переходимо до asset, person, mechanism, choose та equipment 2014 року – отримані дані потребують експертної інтерпретації.

**Висновки та перспективи подальших досліджень.** В даній роботі було проведено формальне визначення змістовних ознак і характеристик англомовного тексту на основі досліджень зв'язків між лемами та синсетами на прикладі «Address by President of the Russian Federation 2013/2014». З цією метою розроблено програмне забезпечення на основі лінгвістичного пакету *DKPro Core* для визначення 2-ма способами ключових слів досліджуваного тексту та проведено порівняльну побудову онтологій тексту на основі бібліотек платформи *NLTK*.

За результатами експериментів визначено, що реалізований у *DKPro Core* підхід на основі врахування зв'язків між лемами більш точно та більш повно визначає списки ключових слів у порівнянні з відомим методом на основі частотного словника тексту. Зокрема, у запропонований спосіб кількість стоп-слів англомовного тексту менша в 5 разів від відомого, а ваги однакових ключових слів суттєво більші (у 1,5–6 разів). Додатково власна розробка може виявляти найбільш вагомі словосполучення тексту, що, у сукупності, є формальною підставою для подальшої експертної інтерпретації щодо змісту та акцентів текстової інформації.

За допомогою програмного забезпечення, написаного на мові Python у додаток до бібліотек платформи *NLTK* було створено таблицю порівняння гіперонімів та побудовано зв'язані графи досліджуваних офіційних текстів. Виявлено, що гіперонімічний зв'язок є визначальним для кластерного представлення онтології загальної лексики тексту, що дозволяє у першому наближенні не враховувати усі ваги зв'язків класу *SynSet*. За складом та топологією кластерів отримано потенційні формальні ознаки єдності авторського стилю, а також змін тональності викладеної інформації, проте цей висновок потребує більш масштабних експериментів.

Подальшого дослідження також потребують порівняння отриманих списків ключових слів з авторськими на різних категоріях текстів. Корисно було б також застосувати інші типи відношень, наприклад, метонімічні для побудови онтологій тексту.

**Список літератури:** **1**. *Бісікало О. В*. Формальні методи образного аналізу та синтезу природномовних конструкцій : монографія [Текст] / *О. В. Бісікало* // – Вінниця : ВНТУ, 2013. – 316 с. – ISBN 978-966-641-528-1. **2.** *Бісікало О. В.* Формальне введення образного рівня до традиційної лінгвістичної тріади морфологія–синтаксис–семантика [Текст] / *О. В. Бісікало, І. В. Богач* // Бионика интеллекта. – 2013. – № 2 (81). – С. 27–30. **3.** *Бісікало О. В.* Метод визначення ключових слів англомовного тексту на основі DKPro Core [Текст] / *О. В. Бісікало, О. В. Яхимович* // Технологический аудит и резервы производства: Информационные технологии. – 2015. – Том 1, № 2(21). – С. 26–30. **4**. Address by President of the Russian Federation [Electronic resource]. – Available at: \www/URL: http://en.kremlin.ru/events/president/news/19825. – 12.12.2013. **5**. Address by President of the Russian Federation [Electronic resource]. – Available at: \www/URL: http://en.kremlin.ru/events/president/news/47173. – 04.12.2014. **6.** *Matlack, Carol*. To Understand Putin, Try Counting His Words [Electronic resource]. – Bloomberg Businessweek, December 11, 2014. – Available at: \www/URL: http://www.bloomberg.com/bw/articles/2014-12-11/counting-how-many-times-putin-said-russia. **7.** Natural Language Processing: Integration of Automatic and Manual Analysis [Electronic resource]. – Technischen Universität Darmstadt, 2014. – Available at: \www/URL: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf. – 21.04.2015. **8**. *Bird, Steven*. Natural Language Processing with Python Analyzing Text with the Natural Language Toolkit [Electronic resource] / *Steven Bird, Ewan Klein, Edward Loper*. – O'Reilly, – 2010. Available at: \www/URL: http://victoria.lviv.ua/html/fl5/NaturalLanguageProcessingWithPython.pdf. **9.** *Gurevych, I.* Darmstadt Knowledge Processing Repository Based on UIMA [Electronic resource] / *I. Gurevych, M. Muhlhauser, Ch. Muller, J. Steimle, M. Weimer, T. Zesch*. – February 9, 2007. – Available at: \www/URL: https://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2007/gldv-uima-ukp.pdf. – 21.04.2015. **10.** *Banerjee, Satanjeev and Pedersen, Ted*. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet, Lecture Notes In Computer Science. – Vol. 2276, Pp. 136-145, 2002. – ISBN 3-540-43219-1.

**Bibliography (transliterated):** **1.** *Bisikalo O. V.* (2013). Formal'ni metody obraznoho analizu ta syntezu pryrodno-movnykh konstruktsij : monohrafiia [Formal methods imagery analysis and synthesis of natural language constructions: monograph]. Vinnitsa, VNTU, 316. ISBN 978-966-641-528-1. **2.** *Bisikalo O. V., Bohgach I. V.* (2013).

Formal'ne vvedennia obraznoho rivnia do tradytsijnoi linhvistychnoi triady morfolohiia–syntaksys–semantyka [The formal introduction of the traditional figurative linguistic triad morphology-syntax-semantics]. Bionics intelligence, 2 (81), 27-30 . **3.** *Bisikalo O. V, Yahimovich O. V.*(2015). Metod vyznachennia kliuchovykh sliv anhlomovnoho tekstu na osnovi DKPro Core [The method of determining keywords at English text based on the DKPro Core]. Technology Audit and Reserves Production. Information Technology., Vol. 1 № 2 (21), 26-30. **4.**Address by President of the Russian Federation. Available at: http://eng.kremlin.ru/transcripts/6402. **5**. Address by President of the Russian Federation. Available at: http://eng.kremlin.ru/news/6889. **6.** *Matlack, Carol*. (2014) To Understand Putin, Try Counting His Words. Bloomberg Businessweek. Available at: http://www.bloomberg.com/bw/articles/2014-12-11/counting-how-many-times-putin-said-russia. **7.** Natural Language Processing: Integration of Automatic and Manual Analysis. Darmstadt. Technischen Universität. 2014. Available at: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf. **8**. *Steven Bird, Ewan Klein, Edward Loper*. (2010) Natural Language Processing with Python Analyzing Text with the Natural Language Toolkit. O'Reilly.Available at: http://victoria.lviv.ua/html/fl5/NaturalLanguageProcessingWithPython.pdf. **9.** *Gurevych I, Muhlhauser M., Muller Ch., Steimle J., Weimer M., Zesch T*. (2007) Darmstadt Knowledge Processing Repository Based on UIMA. Available at: https://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2007/gldv-uima-ukp.pdf. – 21.04.2015. **10.** *Banerjee, Satanjeev and Pedersen, Ted*. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. Lecture Notes In Computer Science, 2276, 136-145. ISBN 3-540-43219-1.

*Поступила (received) 26.04.2015*

**УДК 004.8: 681.51**

***О. В. ГЕРАСІНА,*** канд. техн. наук, доц., ДВНЗ «Національний гірничий університет», Дніпропетровськ

# АЛГОРИТМИ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ ДЛЯ ПРОГНОЗУВАННЯ ПРОЦЕСІВ ГІРНИЧО-МЕТАЛУРГІЙНОГО ВИРОБНИЦТВА

Запропоновано для підвищення точності прогнозування процесів гірничо-металургійного виробництва використовувати адаптивні фільтри-апроксиматори на основі нечіткої кластеризації, а також проводити налаштування їх параметрів за допомогою методів глобальної оптимізації. Визначено ефективність запропонованого підходу на прикладі прогнозування технологічних процесів крупного дроблення і доменного виробництва.

**Ключові слова:** прогнозування, фільтр-апроксиматор, нечітка логіка, кластеризація, глобальна оптимізація, крупне дроблення, доменне виробництво.

**Вступ.** З позицій керування складними об'єктами керування (ОК) є динамічні об'єкти з нестаціонарними параметрами, нелінійними залежностями і стохастичними змінними. До них відносяться технологічні процеси доменного виробництва (ДВ) і рудопідготовки (процеси дроблення, здрібнювання руд), витрати на які складають значну частину собівартості гірничо-металургійного виробництва [1, 2]. Тому актуальним є вирішення задач прогнозування цих процесів, що дозволяє підвищити якість управління за рахунок підвищення точності оцінки їх стану.

Підвищення якості керування процесом крупного дроблення (ККД) на гірничо-збагачувальних комбінатах призводить до поліпшення якості наступного за ним процесу здрібнювання, і як наслідок – продуктів збагачення, що неможливо без ефективного прогнозування.

Одним з основних показників ДВ є тепловий стан доменної печі, оцінювати