УДК 519.7:007.52; 519.711.3

І.Д. Вечірська, Г.Г.Четвериков
ХНУРЕ, м.Харків, Україна, ira_se@list.ru

МАТЕМАТИЧНІ АСПЕКТИ ПОБУДОВИ ЛАНЦЮГІВ ЛЕКСИЧНИХ ОДИНИЦЬ

В статті досліджено математичні аспекти опису побудови ланцюгів лексичних одиниць на основі алгебри скінченних предикатів. Проведено аналіз особливостей побудови, наведено застосування методу знаходження степеня лінійного логічного перетворення для вилучення характерних слів словарної статті. Наведено аналіз та перспективи результатів дослідження.

ЛЕКСИЧНА ОДИНИЦЯ, ЛАНЦЮГ ЛЕКСИЧНИХ ОДИНИЦЬ, СЕМАНТИЧНА КЛАСИФІКАЦІЯ, ЛІНІЙНЕ ЛОГІЧНЕ ПЕРЕТВОРЕННЯ, ЯДРО ЛІНІЙНОГО ЛОГІЧНОГО ПЕРЕТВОРЕННЯ

Вступ

Серед розмаїття задач автоматичної обробки природномовних об'єктів і на сьогоднішній день залишаються невирішеними ще багато актуальних питань. Здебільшого це пов'язано з нездатністю об'єктивно описати суб'єктивні стани людського мозку [1]. Для того щоб побудувати адекватні природномовні системи, необхідно автоматизувати не тільки синтаксис тієї чи іншої природної мови, але й її семантику. А ця проблема, не зважаючи на різні підходи до формалізації семантичних задач, залишається до кінця невирішеною [2-4]. Обумовлено це з одного боку недостатньо глибокими дослідженнями з опису процесів обробки природномовної інформації людським мозком, а з іншого – недостатніми формальними засобами для опису семантичних відношень.

Академік НАНУ Широков В.А. відзначає: "... інтенсифікація інформаційних процесів спонукає до прискорення розробки різноманітних методик формалізації мови, а також створення все потужніших методів лексикографування явищ предметного світу". Виділено чотири проблеми сучасної лексикографії. Перша з них пов'язана з необхідністю постійного оновлення словникових структур, оскільки природна мова залишається своєрідним "живим організмом", який постійно розвивається, змінюється та оновлюється. Друга проблема стосується безпосереднього переведення словникової інформації з "твердого" книжного вигляду до комп'ютерного. Третя проблема пов'язана з онтологіями як засобом для опису декларативних знань за допомогою формальних логік для знаходження прихованих семантичних структур на великих словникових масивах. І, нарешті, четверта проблема полягає у створенні гнучких і потужних лексикографічних засобів для підтримки саме багатомовного лексикографування [5].

Важливою для вирішення зазначених вище другої і третьої проблем лексикографування є задача розмітки словників. На шляху її розв'язання одним з основних завдань є побудова класифікаторів семантичних структур. Наприклад, у [5] визначено чотири типи відношень, кожен з яких розбивається на підкласи: відношення семонімії (синоніми, антоніми, пароніми, омоніми); відношення словотвору (однокореневі слова); тезурусні відношення (рід-вид, частина-ціле, комплекс-елемент, причина-наслідок); відношення асоціацій і аналогій (асоціатори і аналогеми).

Досить вдалі спроби семантичної класифікації було проведено за допомогою програм ПроСеКа [6]: було розроблено для надання користувачеві можливості задавати, редагувати та аналізувати семантичні відношення між лексичними одиницями у вигляді ланцюгів, елементи яких пов'язані відношенням «тлумачиться через», зберігати ці дані у формі, орієнтованій на комп'ютер. На відміну від неї, програма «Побудова гіперланцюгів» є автоматизованою, тобто робота з побудови може виконуватися користувачем, безпосередньо самою програмою або ж пропонується комбінований пошук. Крім цього, програма дозволяє також будувати ланцюги за відношенням синонімії [7]. При побудові програмних засобів неодноразово виникали запитання, відповідь на які можна було б чекати лише внаслідок формальної постановки задачі, знаходження методу пошуку, який скоригував би критерій закінчення побудови ланцюгів. Тому важливо розробити математичний інструментарій для побудови ланцюгів лексичних одиниць, який в свою чергу надалі використовувався б для задач семантичної класифікації.

Як правило, задачі, пов'язані з семантикою, носять здебільшого дослідницький характер. Тому для побудови відповідних методів і моделей доцільно використовувати поняття та принципи досить високого рівня абстракції.

У своїх дослідженнях автори відштовхувались від понять множини та відношення, тобто використовували апарат математичної логіки, а також алгебру скінченних предикатів.

Таким чином, метою статті є розвиток засобів моделювання природномовних об'єктів для класифікації семантичних структур шляхом формального опису етапів побудови ланцюгів лексичних одиниць.

1. Математичний опис об'єкту

Побудова ланцюгів лексичних одиниць (ЛЛО) здійснюється наступним чином: шляхом аналізу словникових дефініцій виділяються типи диференціальних сем, характерних для формул тлумачення лексичних одиниць, і встановлюються типові структури усіх компонентів. Побудову такого ланцюга вважатимемо закінченою, якщо серед знайдених лексичних одиниць знову з'явиться вихідне слово.

На рис. 1 графічно зображена схема побудови ЛЛО будь-якої природної мови, де x — вихідне слово, x_{11} — перше характерне слово з тлумачення (синонімів) слова x, x_{1m} — останнє характерне слово з тлумачення (синонімів) слова x, x_{21}^{11} - перше характерне слово з тлумачення (синонімів) слова x_{11}, x_{2n}^{11} — останнє характерне слово з тлумачення (синонімів) слова x_{11}, x_{21}^{1m} — перше характерне слово з тлумачення (синонімів) слова x_{1m}, x_{2q}^{1m} — останнє характерне слово з тлумачення (синонімів) слова x_{1m} і так далі. Під "характерним" розуміємо слово з правої частини словарної статті тлумачного словника, яке безпосередньо характеризує вихідну лексичну одиницю, має те ж концептуальне значення (включає такі ж семантичні компоненти).

Таким чином, можна записати, що x_{ij} — це j-те характерне слово на i-ому рівні ЛЛО (на лексичну одиницю, до якої відноситься тлумачення, вказує верхній індекс слова), де j — порядковий номер в тлумаченні (знаходженні синонімів) лексичної одиниці.

Далі спробуємо записати побудову ЛЛО за допомогою теорії лінійних логічних перетворень.

Апарат алгебри скінченних предикатів (АСП) ґрунтується на понятті відношення. А оскільки за допомогою відношень можна описувати об'єкти будь-якої природи, то говорять про універсальність алгебри. Від понять образу і прообразу множини переходимо до поняття лінійного логічного перетворення.

Образом множини $P \subseteq M$ відносно відображення K називається множина $Q \subseteq N$, що складається з усіх образів предметів, які належать до множини P:

$$Q = \{y \mid \exists x \in M(K(x,y) \wedge P(x)) = 1\}.$$

Лінійне логічне перетворення запишемо наступним виразом: $Q(y) = \exists x \in M(K(x,y) \wedge P(x))$, де предикат $K(x,y)$ визначений на декартовому добутку $M \times N$, тобто описує зв'язок між елементами множини M та відповідними елементами множини N. Іншими словами, лінійним логічним перетворенням називається відображення, якщо для нього виконуються властивості адитивності та однорідності.

Далі введемо поняття прообразу множини та, за аналогією, виведемо з нього поняття дуального лінійного логічного перетворення.

Прообразом множини $Q \subseteq N$ відносно відображення K називається множина $P \subseteq M$, що складається з усіх прообразів предметів, які належать до множини Q:

$$P = \{x \mid \exists y \in N(K(y,x) \wedge Q(y)) = 1\}.$$

Дуальне лінійне логічне перетворення запишемо наступним виразом:

$$P(x) = \exists y \in N(K(y,x) \wedge Q(y)),$$

де предикат $K(y,x)$ визначений на декартовому добутку $N \times M$, тобто описує зв'язок між елементами множини N та відповідними елементами множини M.

Отже, процес побудови ЛЛО за допомогою теорії лінійних логічних перетворень, (схематично зображений на рис. 2), можна записати таким чином: беремо слово; всі слова тлумачення (множина M_2). Виділяємо характерні слова, отримуємо

$$P(x_2) = \exists x_1 \in M_1(K_1(x_1,x_2) \wedge P(x_1)).$$

Далі встановлюємо зв'язки між елементами множини $P(x_1)$ та їх словами-тлумаченнями (множина M_2), формуємо ядро лінійного логічного перетворення $K_2(x_2,x_3)$, отримуємо $P(x_3) = \exists x_2 \in M_2(K_2(x_2,x_3) \wedge P(x_2))$. Встановивши зв'язки, перевіряємо, чи зв'язані між собою елементи y_j з x_i. Побудова закінчується за виконання умови: $x_n = x_i$, якщо $\exists x_i, x_n\, K_i(x_i,x_n) = 1$, $i = \overline{1,n-1}$.

Отже, отримуємо загальний вигляд лінійного логічного перетворення для пошуку n-ого елементу ланцюга лексичних одиниць:

$$P(x_n) = \exists x_{n-1} \in M_{n-1}(K_{n-1}(x_{n-1},x_n) \wedge P(x_{n-1})).$$

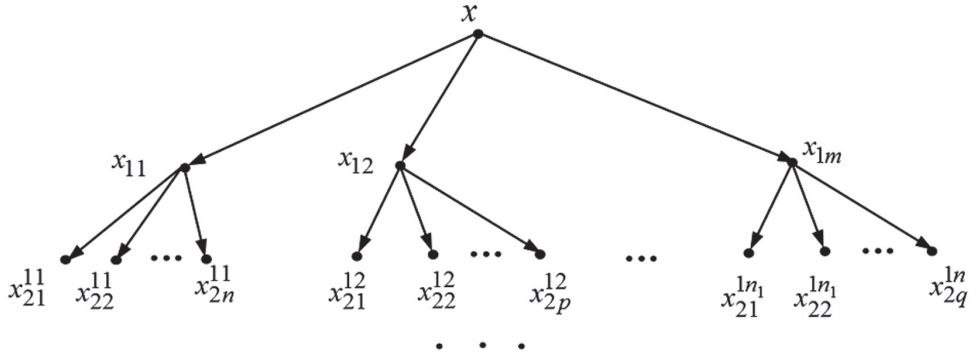

Рис. 1. Схема побудови ланцюга лексичних одиниць

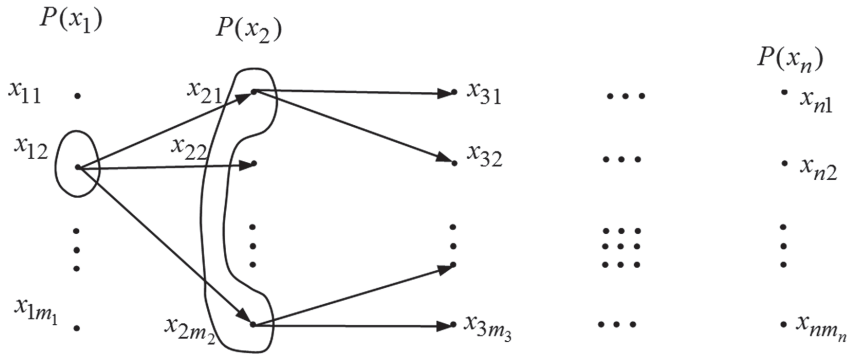

Рис. 2. Схема побудови ЛЛО за допомогою лінійних логічних перетворень

2. Виділення характерних слів в тлумаченнях

Цікавою є задача виділення характерних слів в тлумаченнях. В розглянутих вище системах побудови ЛЛО цю задачу розв'язує дослідник. Це і є тим етапом побудови ланцюга, через який отримуємо системи автоматизовані, а не автоматичні. Далі за допомогою засобів АСП опишемо підхід до вирішення цієї проблеми, що ґрунтується на схемах синтаксичного підпорядкування та методі знаходження n-ого лінійного логічного перетворення.

Відомо, що в граматиці для наглядного подання структури речень використовуються схеми синтаксичного підпорядкування. В них слова речення поєднуються в синтаксичні пари стрілками, що називаються дугами синтаксичного підпорядкування. Слово, з якого виходить дуга, називається головним словом пари, а слово, в яке вона входить, — залежним словом. Коренем речення називається слово, в яке не входить жодна з дуг. Схема синтаксичного підпорядкування, у якій дуги не перетинаються, а корінь не покривається ні однією з дуг, називається проективною.

Далі коротко наведемо метод знаходження n-ого лінійного логічного перетворення.

Лінійне логічне перетворення задає перетворення однієї підмножини значень змінної x з областю визначення M, яку задано предикатом $P(x)$, у відповідну підмножину значень змінної y з областю визначення N, що задано предикатом $Q(y)$.

В роботах [8], [9] було проведено дослідження дій над лінійними логічними перетвореннями, а саме знаходження степеня лінійного логічного перетворення (рис.1). Виведено формулу для знаходження n-ого степеня лінійного логічного перетворення $Q^{(n)}(y)$ та дуального йому $P^{(n)}(x)$:

$$Q^{(n)}(y) = \bigwedge_{i=1}^{n} K_i Q(y), \text{ де } K_i = K = K(x,y)K(y,x),$$

$$P^{(n)}(x) = \bigwedge_{i=1}^{n} K_i' P(x), \text{ де } K_i' = K' = K(y,x)K(x,y).$$

Метод знаходження степеня лінійного логічного перетворення $Q^{(n)}(y)$ можна розбити на наступні етапи. Спочатку необхідно знайти матрицю K, яка є суперпозицією ядер лінійних логічних перетворень з $P(x)$ в $Q(y)$ і, відповідно, з $Q(y)$ в $P'(x)$: $K = K(x,y)K(y,x)$.

Нехай ядро лінійного перетворення можна представити виразом $K(x,y) = \left|a_{ij}\right|_{\substack{i=\overline{1,m} \\ j=\overline{1,n}}}$. Тоді матриця ядра дуального йому лінійного логічного перетворення має вигляд $K(y,x) = \left|a_{ji}\right|_{\substack{i=\overline{1,m} \\ j=\overline{1,n}}}$.

Таким чином, можна зробити висновок, что n-а степінь лінійного логічного перетворення ($n \geq 1$) залежить від виду матриці K. А матриця K, в свою чергу, залежить тільки від області визначення змінної x і не залежить від області визначення змінної y. Звідки випливає, що крок, на якому степінь лінійного логічного перетворення в подальших діях не змінюється, безпосередньо залежить від розмірності області визначення змінної x.

В роботах [8], [9] було доведено твердження, що якщо для знаходження степеня лінійного логічного перетворення на двох послідовних кроках значення перетворення повторюється, то це значення буде повторюватись також і на наступних кроках. Тобто якщо при знаходженні n-ого степеня лінійного логічного перетворення було отримано однакові результати на n-ому та $n-1$-ому кроках, то цей результат отримаємо також і на наступних $n+1$-ому, $n+2$-ому і т.д. кроках. Тоді таке лінійне перетворення і є шуканим.

Розглянемо тепер застосування методу знаходження n-ого лінійного логічного перетворення для виділення характерних слів в тлумаченнях при побудові ЛЛО.

Наприклад, розглянемо речення: "*Словник — це систематизований перелік соціалізованих мовних форм.*"

На рис. 3 схематично зображено пошук n-ого лінійного логічного перетворення, яке відображає всі характерні слова тлумачення.

Позначимо слова речення наступними змінними: x_1 — "це"; x_2 — "систематизований"; x_3 — "перелік"; x_4 — "соціалізованих"; x_5 — "мовних"; x_6 — "форм"; x_7 — "словник".

Нехай $x_i \in M$, $i = \overline{1,7}$. Тоді

$M = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\} = \{$*це, систематизований, перелік, соціалізованих, мовних, форм, словник*$\}$.

Аналогічно формуємо множину $N = \{y_i\}$, ($i = \overline{1,6}$) з усіх слів правої частини словникової статті. Нехай

$$P(x) = x_3^{перелік} \vee x_6^{форм}.$$

Ядро лінійного логічного перетворення формують правила побудови схем синтаксичного підпорядкування:

$$K(x,y) = x_3^{перелік} y_2^{систематизований} \vee x_3^{перелік} y_6^{форм} \vee$$
$$\vee x_6^{форм} y_4^{соціалізованих} \vee x_6^{форм} y_5^{мовних} \vee x_7^{словник} y_3^{перелік}.$$

Знаходимо

$$Q(y) = K(x,y)P(x) = y_2^{систематизований} \vee y_4^{соціалізованих} \vee$$
$$\vee y_5^{мовних} \vee y_6^{форм}.$$

Далі знаходимо

$$P^{(1)}(x) = K(y,x)Q(y) = K(y,x)K(x,y)P(x) = K'P(x),$$
$$P^{(1)}(x) = x_3^{перелік} \vee x_6^{форм} \vee x_7^{словник};$$
$$Q^{(1)}(y) = K(x,y)P^{(1)}(x) = K(x,y)K(y,x)Q(y) = KQ(y),$$
$$Q^{(1)}(y) = y_2^{систематизований} \vee y_3^{перелік} \vee y_4^{соціалізованих} \vee$$
$$\vee y_5^{мовних} \vee y_6^{форм};$$
$$P^{(2)}(x) = K(y,x)Q'(y) = K(y,x)K(x,y)P'(x) = K'K'P(x),$$
$$P^{(2)}(x) = x_3^{перелік} \vee x_6^{форм} \vee x_7^{словник};$$
$$Q^{(2)}(y) = K(x,y)P^{(2)}(x) = K(x,y)K(y,x)Q^{(1)}(y) = KKQ(y)$$
$$Q^{(2)}(y) = y_2^{систематизований} \vee y_3^{перелік} \vee y_4^{соціалізованих} \vee$$
$$\vee y_5^{мовних} \vee y_6^{форм}.$$

Таким чином, оскільки значення лінійного логічного перетворення на двох кроках поспіль співпадають , то за критерієм закінчення роботи методу знаходження n-ого лінійного логічного перетворення ми отримали шукану множину всіх характерних слів правої частини словникової статті.

Висновки

Отже, в статті проведено аналіз процесу побудови ланцюгів лексичних одиниць. Для його формалізації обрано апарат алгебри скінченних предикатів, а саме лінійні логічні перетворення. За допомогою лінійних логічних перетворень було здійснено математичний опис самого процесу побудови. Визначено критерій закінчення побудови ЛЛО.

Крім того, застосування методу знаходження n-ого лінійного логічного перетворення та правил побудови схем синтаксичного підпорядкування дозволило формалізувати процес виділення характерних слів в словниковій статті для подальшої побудови ланцюга лексичних одиниць. Це дозволяє провести автоматизацію процесу побудови ЛЛО без втручання користувача. Звичайно, варто дослідити, чи вилучення характерних слів з правої частини словарної статті не забере у програмної системи більше часу, ніж у людини. Проте тут очевидна корисність для великих словарних статей. Крім того, чіткий критерій закінчення пошуку дозволяє уникнути

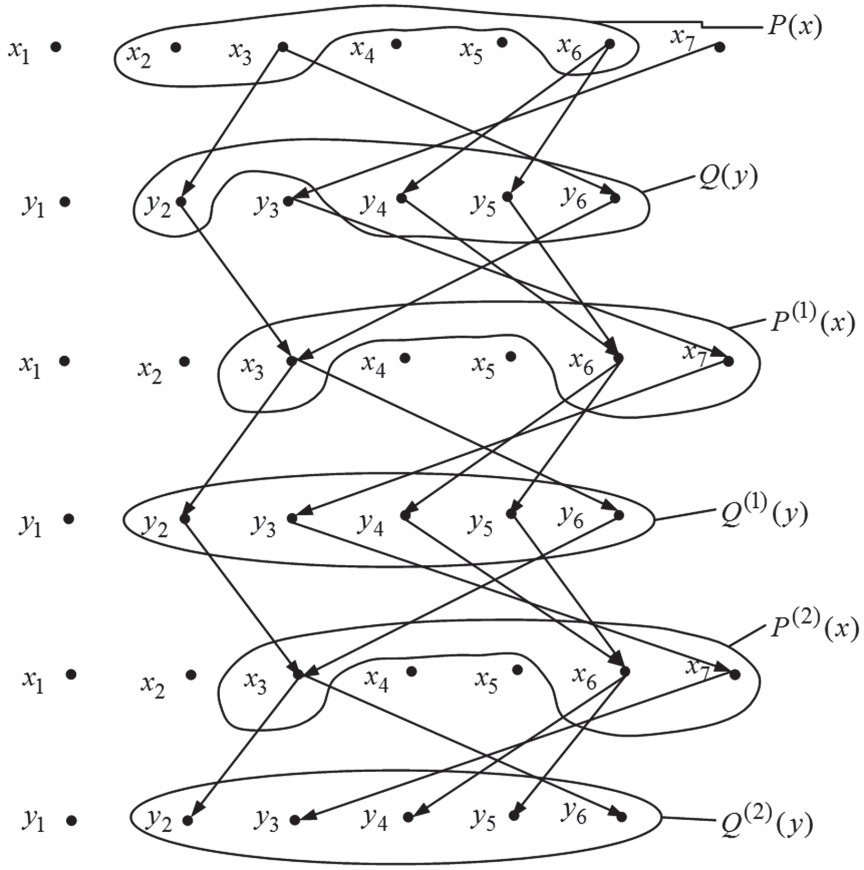

Рис. 3. Знаходження n-ого лінійного логічного перетворення для пошуку характерних слів тлумачень

неточності пошуку (людині легко пропустити якесь з значущих слів) та позбавляє зайвих кроків.

Застосування математичного апарату лінійних логічних перетворень дозволило подати побудову ЛЛО не лише у формульному, а й в більш наглядному, схематичному вигляді. Це дає змогу дослідникам глибше аналізувати сам процес побудови, узагальнити клас задач, які можна ров'язувати розглянутими методами, в перспективі застосовувати ланцюги лексичних одиниць для семантичної класифікації.

Так, було виявлено, що завжди елементів множини M на один більше, ніж елементів множини N. Необхідно дослідити, як це може вплинути на кількість кроків для знаходження всіх характерних слів, тобто на швидкість знаходження. Для цього слід розглянути, як за таких умов зміниться матриця ядра вихідного та результуючого лінійного логічного перетворення та знайти закономірність в її обчисленнях.

Дослідження показали, що відсутні відображення типу $x_i \to y_i$, тобто в матриці лінійного логічного перетворення по діагоналі будуть стояти нулі. Слід дослідити обчислення саме для таких матриць.

Як розвиток досліджень, множиною N можна задавати слова не лише одного речення, а й всієї правої частини словникової статті. Тоді ядро лінійного логічного перетворення відображатиме всі зв'язки правої і лівої частин словарної статті.

Можна також задавати різні правила для запису ядра лінійного логічного перетворення не лише семантичного підпорядкування. Звичайно, чим більше обмежень у правила, більше значень змінних у ядра лінійного логічного перетворення, тим швидше знаходяться характерні слова. Цікаво також дослідити, чи відобразиться на швидкості розв'язку попереднє перетворення непроективної схеми синтаксичного підпорядкування в проективну.

Для задач опрацювання природної мови дуже важко знайти якусь кількісну характеристику чи закономірність. Хоча у кожній природній мові є свої властивості. Важливо їх виявити і кількісно охарактеризувати. Так, В.А. Широковим було пораховано структури видових комплексів дієслова української мови [10].

Деяку кількісну оцінку побудови різних ланцюгів лексичних одиниць дає відповідне оцінювання методу знаходження n-ого лінійного логічного перетворення, оскільки кількість кроків, за які знаходиться кінцевий результат, прямо залежить від розмірності матриці K. Дослідження предметного простору ядра лінійного логічного перетворення наведено в [11].

Список літератури: **1**. *Бондаренко, М.Ф.* Мозгоподобные структуры: справочное пособие. Том первый [Текст] / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко – К.: Наукова думка, 2011. – 460 с. **2**. *Широков, В.А.* Очерк основных принципов квантовой лингвистики [Текст] /В.А.Широков // Бионика интеллекта: науч.-техн. журнал. – 2007. – № 1(66). – С. 25-32. **3**. *Морковкин, В. В.* О единицах лексической системы [Текст] / В.В. Морковкин // Лексика и лексикография: Сб. науч. трудов / Отв. ред. Ю. Г. Коротких, А. М. Шахнарович. М., 1992. С. 127–134. **4**. *Бондаренко, М.Ф.* Концепції уніфікації інформаційно-інтелектуальних технологій в системах мовлення [Текст] / М.Ф. Бондаренко, З.Д. Коноплянко, Г.Г. Четвериков // Біоніка інтелекту: наук.-техн. журнал. – 2011. – № 3 (77). – С.150 – 156. **5**. *Широков В.А.* Комп'ютерна лексикографія [Текст] / В.А.Широков – Київ: Науково-виробниче підприємство «Видавництво "Наукова думка" НАН України», 2011. – 351с. **6**. *Рафаєва, А.В.* Програму семантичної класифікації лексики ПроСеКа [Текст] / А.В. Рафаєва // Прикладна лінгвістика та лінгвістичні технології: MegaLing-2009: Зб. наук. праць / НАН України. Укр.мовн.-інформ.фонд, Таврійськ. нац. Ун-т ім. В.І.Вернадського/ За ред. В.А. Широкова. – К.: Довіра, 2009. – 527 с. **7**. *Федорова, Т.Н.* Автоматизация построения цепочек лексических единиц на примере украинских народных сказок [Текст] / Т.Н. Федорова // Материалы международной научной конференции "Горизонты прикладной лингвистики и лингвистических технологий" (MegaLing'2011). – 20-26 сентября 2011, Украина, Киев, http://megaling.ulif.com.ua/ **8.** *Вечирская, И.Д.* О методе вычисления линейных логических преобразований [Текст] / И.Д. Вечирская // Бионика интеллекта: науч.-техн. журнал. – 2007. – № 2 (67). – С.65 – 68. **9.** *Вечирская И.Д.* О методе нахождения n-ого линейного логического преобразования [Текст] / И.Д. Вечирская, Ю.П. Шабанов–Кушнаренко // Искусственный интеллект. – Донецк: Институт проблем искусственного интеллекта. – 2007. – № 3. – С. 382-389. **10**. *Широков, В.А.* Елементи лексикографії [Текст] / В.А. Широков – Київ: Видавництво "Довіра", 2005. – 303с. **11**. *Вечірська, І.Д.* Дослідження розмірності предметного простору в задачах моделювання об'єктів у вигляді реляційних мереж [Текст] /І.Д. Вечірська // Біоніка інтелекту: наук.-техн. журнал. – 2009. – № 2 (71). – С.31 – 35.

Надійшла до редколегії 24.05.2012

УДК 519.7:007.52; 519.711.3

Математические аспекты построения цепочек лексических единиц / И.Д. Вечирская, Г.Г. Четвериков// Бионика интеллекта: науч.-техн. журнал. – 2012. – № 2 (79). – С. 84–88.

Исследован процесс построения цепей лексических единиц. Приведено его формальное описание с помощью теории линейных логических преобразований. Применение метода нахождения *n*-ого линейного логического преобразования и правил построения схем синтаксического подчинения позволило формализовать процесс выделения характерных слов толкования для автоматического построения цепочек лексических единиц.

Ил.: 3. Библиогр.: 11 назв.

УДК 519.7:007.52; 519.711.3

The mathematical aspects of building of chain of lexical units / I.D. Vechirska, G.G. Chetverikov// Bionics of Intelligense: Sci. Mag. – 2012. – № 2 (79). – P. 84–88.

The process of building chains of lexical units has been investigated. Its formal description have been presented using the theory of linear logical transformation. The application of the finding method of the n-th power of linear transformations and the construction rules of the schemes for the syntactic subordination allowed to formalize The process of selection of characteristic words of interpretation for automatical building chains of lexical units.

Fig.: 3. Ref.: 11 items.