УДК 004.519.7 (045)

А.І. ВАВІЛЕНКОВА[*]

МЕТОДОЛОГІЧНІ ОСНОВИ АВТОМАТИЧНОГО АНАЛІЗУ ЛОГІКО-ЛІНГВІСТИЧНИХ МОДЕЛЕЙ ТЕКСТОВИХ ДОКУМЕНТІВ

[*]Національний авіаційний університет, Київ, Україна

Анотація. *Виявлено основні проблеми на шляху побудови формальних моделей текстових документів. Сформовано логіко-лінгвістичну модель тексту, що складається з лінгвістичної та семантико-синтаксичної компонент. Запропоновано алгоритм аналізу логіко-лінгвістичних моделей текстових документів, в результаті роботи якого відновлюється текст на природній мові.*
Ключові слова: *логіко-лінгвістична модель, обробка текстових документів, інформаційний пошук, формальний опис, інформаційні технології.*

Аннотация. *Выявлены основные проблемы на пути построения формальных моделей текстовых документов. Сформирована логико-лингвистическая модель текста, которая состоит из лингвистической и семантико-синтаксической компонент. Предложен алгоритм анализа логико-лингвистических моделей текстовых документов, в результате работы которого восстанавливается текст на естественном языке.*
Ключевые слова: *логико-лингвистическая модель, обработка текстовых документов, информационный поиск, формальное описание, информационные технологии.*

Abstract. *The article reveals the main problems on the way of formal models constructing of text documents. It is formed a logical-linguistic model of the text, which consists of linguistic and semantic-syntactic components. It is proposed an algorithm of analysis of logical-linguistic models of text documents due to which restores text on natural language.*
Keywords: *logical-linguistic model, processing of text documents, information search, formal description, information technology.*

1. Вступ

Розвиток та удосконалення інформаційних технологій, зростання об’єму інформації, перехід до суспільства знань – все це зробило інформаційні комп’ютерні технології потужним інструментом підвищення продуктивності виробництва, економічного зростання, створення нових засобів комунікації. Абсолютно всі інформаційні технології у тій чи іншій мірі використовують методи обробки текстової інформації. Саме тому існує необхідність розробити якісний формальний апарат, який дозволив би уникнути неоднозначності при пошуку текстових документів, а також аналізувати текстову інформацію за єдиним принципом.

На сьогодні найбільш результативною технологією роботи зі знаннями вважається Data Mining, що об’єднує у собі широкий математичний інструментарій та останні досягнення у сфері інформаційних технологій [1]. В основу Data Mining покладена концепція шаблонів, що відображають фрагменти багатоаспектних відношень у даних. Такі шаблони представляють собою закономірності, пошук яких обмежений певними наборами розподілу значень показників, що аналізуються [2]. Незважаючи на величезну кількість програмних продуктів, які здійснюють сьогодні аналітичну обробку електронних текстів, глибинний рівень знань досі залишається прихованим. Це пов’язане з відсутністю формальних засобів здійснення семантичного та лінгвістичного аналізу текстів.

2. Постановка задачі

Усі методи здійснення інформаційного пошуку поділяються на статистичні, методи пошуку за семантичними мережами та комбіновані методи.

© Вавіленкова А.І., 2015

Основною ідеєю статистичних методів є визначення ваги кожного слова у документі. Для них притаманна якісна математична модель, що дозволяє отримати правильні оцінки релевантності для документів. Недоліком статистичних методів є те, що вони не враховують змістовне навантаження текстів та тексту запиту. Статистичні методи лежать в основі роботи пошукових машин Google, Yandex, Yahoo та ін.

Методи пошуку за семантичними мережами використовують дані, представлені у вигляді онтологій, а пошук відбувається шляхом задання властивостей шуканого об'єкта. Такі методи враховують змістовне навантаження, проте застосовувати їх можна тільки для таких електронних документів, які містять семантичний опис контенту.

Комбіновані методи, окрім статистичних, використовують методи семантичного аналізу текстів. Саме до цієї групи методів пошуку інформації відносяться подальші дослідження.

Формальною моделлю представлення знань, що враховує зміст речень природної мови, є логіко-лінгвістична модель [3]. Тому, якщо побудувати формальну змістовну модель тексту будь-якої тематики та структури, то можна буде аналізувати електронні текстові документи за змістом, вилучати з них знання, порівнювати їх.

Основною проблемою на шляху побудови логіко-лінгвістичної моделі тексту є виявлення лінгвістичних правил написання документів та опис їх на формальній мові. У цій сфері проведено багато досліджень як лінгвістами (це роботи таких вчених, як Гальперіна І.Р., Лайонза Дж., Кобозева І.М.), так і технічними спеціалістами в галузі комп'ютерної лінгвістики (Широков В.А., Ланде Д.В., Леонтьєва Н.Н., Шемакін Ю.І.). Проте питання про створення єдиної методики аналізу текстових документів досі залишається відкритим.

3. Аналіз логіко-лінгвістичних моделей текстових документів

Аналіз логіко-лінгвістичних моделей текстових документів представляє собою складний процес отримання інформації про структуру та зміст тексту, що розглядається, на основі виявлення закономірностей і тенденцій синтаксичної, семантичної та лексичної побудови тексту.

Логіко-лінгвістична модель текстового документа – це абстрактна модель, яка об'єднує в собі основні властивості тексту та його складових частин, відображає основні взаємозв'язки між структурними компонентами, представляє собою впорядковану четвірку та масив логіко-лінгвістичних моделей речень природної мови, що входять до тексту.

Лінгвістична складова формального опису тексту:

$$t =< CQ, F, B, A >, \tag{1}$$

де T – множина текстів;

$t \in T$ – конкретний електронний текст із всієї множини текстів;

$CQ = \{cq_1, ..., cq_i, ..., cq_n\}$ – множина існуючих типів текстів, $i = \overline{1, n}$, n – кількість типів;

$F = \{f_1, ..., f_j, ..., f_m\}$ – множина складних синтаксичних частин тексту, $j = \overline{1, m}$, m – кількість складних синтаксичних частин;

B – текстова база, що складається з набору ключових слів тексту та взаємопов'язаних пропозицій і яку можна представити у вигляді трійки: $B =< K, SJ, D >$, K – множина ключових слів тексту, SJ – множина ключових словосполучень тексту S_j, $j = \overline{1, m}$, D – множина пропозицій;

$A = \{a_1, ..., a_k, ..., a_q\}$ – множина абзаців тексту, $k = \overline{1, q}$, q – кількість абзаців.

Кожен абзац у свою чергу описується четвіркою: $a_k =< H, Y, R, KG >$, $H = \{1, 2\}$ – множина типів зв'язків між реченнями (ланцюговий чи паралельний); $Y = \{1, 2, 3, 4, 5\}$ –

множина типів тематичних прогресій, що вжиті у абзаці $a_k \in A$; $R = \{1,2,3,4,5,6,7\}$ – множина рематичних домінант в абзаці $a_k \in A$; KG – одновимірний масив засобів когезії, що використовуються у даному абзаці [4].

Семантико-синтаксична складова формального опису тексту:

$$t' = \bigwedge_{g=1}^{N(t)} L_g(S_g), \tag{2}$$

де $L_g(S_g)$ – логіко-лінгвістична модель речення S_g, $g = \overline{1, N(t)}$;

$N(t)$ – кількість речень у тексті t.

Логіко-лінгвістична модель речення має вигляд [5]:

$$L(S) = \bigwedge_{\mu=1}^{\nu(S)} L_\mu(S), \tag{3}$$

де $L_\mu(S)$ – простий предикат, що описує частину речення S, яка відображає закінчений зміст;

$\mu = \overline{1, \nu(S)}$, $\nu(S)$ – кількість частин речення S, які відображають закінчений зміст.

Простий предикат записується у вигляді формули

$$L_\mu(S) = P(x_1, c(x_1), x_3, x_2, c(x_2), z, v(p), w(z)), \tag{4}$$

де $X(S)$ – множина сутностей, що входять до речення S;

X_1 – множина суб'єктів, що входять до речення S, $X_1 \subseteq X(S)$;

$c(x_1)$ – кортеж характеристик суб'єкта x_1:

$$c(x_1) = [\, c_k(x_1) \,\big|\, k = \overline{1, m_1(x_1)}\,];$$

$m_1(x_1)$ – кількість характеристик суб'єкта x_1;

$X_2(x_1)$ – множина об'єктів, над якими виконує дію суб'єкт x_1, $x_1 \in X_1$, $X_2(x_1) \subseteq X(S)$;

$c(x_2)$ – кортеж характеристик об'єкта x_2:

$$c(x_2) = [\, c_l(x_2) \,\big|\, l = \overline{1, m_2(x_2)}\,];$$

$m_2(x_2)$ – кількість характеристик об'єкта x_2;

$P(x_1, x_2)$ – множина відношень між суб'єктом x_1 та об'єктом x_2, $x_1 \in X_1$, $x_2 \in X(x_1)$, $x_1 \neq x_2$;

$X_3(x_1)$ – множина об'єктів, пов'язаних з суб'єктом x_1, $x_1 \in X_1$, $X_3(x_1) \subseteq X(S)$, $x_3 \in X(x_1)$, $x_1 \neq x_3$;

$Z(x_1, x_2, p)$ – множина об'єктів p-го відношення між суб'єктом x_1 та об'єктом x_2, $p \in P(x_1, x_2)$, $x_1 \in X_1$, $x_2 \in X(x_1)$;

$c(z)$ – кортеж характеристик об'єктів p-го відношення між суб'єктом x_1 та об'єктом x_2:

$$c(z) = [\, c_q(z) \,\big|\, q = \overline{1, m_3(z)}\,];$$

$m_3(z)$ – кількість характеристик об'єктів p-го відношення між суб'єктом x_1 та об'єктом x_2;

$v(p)$ – кортеж параметрів p-го відношення між суб'єктом x_1 та об'єктом x_2, $p \in P(x_1, x_2)$, $x_1 \in X_1$, $x_2 \in X(x_1)$:

$$v(p) = [\, v_i(p) \,\big|\, i = \overline{1, m(p)} \,];$$

$m(p)$ – кількість параметрів цього відношення;

$w(z)$ – кортеж параметрів z-го об'єкта p-го відношення між суб'єктом x_1 та об'єктом x_2, $p \in P(x_1, x_2)$, $x_1 \in X_1$, $x_2 \in X(x_1)$, $z \in Z(x_1, x_2, p)$:

$$w(z) = [\, w_j(z) \,\big|\, j = \overline{1, n(z)} \,];$$

$n(z)$ – кількість параметрів z-го об'єкта.

Таким чином, модель (1) – (2) містить вичерпну інформацію про текст та зв'язки у ньому. Побудова такої логіко-лінгвістичної моделі для довільного типу тексту дає змогу перейти до аналізу текстової інформації, порівняння текстів за змістом, пошуку протиріч та збігів.

Результатом проведення аналізу логіко-лінгвістичної моделі текстового документа є відновлений текст. Нехай у відповідність деякому тексту поставлена його логіко-лінгвістична модель. Лінгвістична та семантико-синтаксична складова логіко-лінгвістичної моделі заданого тексту:

$$t_1 = < cq_i, F_1, B_1, A_1 >,$$

$$t_1' = \begin{Bmatrix} P_{[1]} & x_{1[1]} & c(x_{1[1]}) & x_{3[1]} & x_{2[1]} & c(x_{2[1]}) & z_{[1]} & v(p_{[1]}) & w(z_{[1]}) \\ P_{[2]} & x_{1[2]} & c(x_{1[2]}) & x_{3[2]} & x_{2[2]} & c(x_{2[2]}) & z_{[2]} & v(p_{[2]}) & w(z_{[2]}) \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ P_{[g1]} & x_{1[g1]} & c(x_{1[g1]}) & x_{3[g1]} & x_{2[g1]} & c(x_{2[g1]}) & z_{[g1]} & v(p_{[g1]}) & w(z_{[g1]}) \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ P_{[N(t_1)]} & x_{1[N(t_1)]} & c(x_{1[N(t_1)]}) & x_{3[N(t_1)]} & x_{2[N(t_1)]} & c(x_{2[N(t_1)]}) & z_{[N(t_1)]} & v(p_{[N(t_1)]}) & w(z_{[N(t_1)]}) \end{Bmatrix}.$$

Тут $g_1 = \overline{1, N(t_1)}$ – номер речення у тексті, $N(t_1)$ – загальна кількість речень у тексті t_1.

У процесі аналізу логіко-лінгвістичних моделей використовується база правил формування зв'язків між складними частинами тексту, а також між моделями $L_g(S_g)$, $g = \overline{1, N(t)}$ безпосередньо. Аналіз логіко-лінгвістичних моделей текстових документів потрібно здійснювати за чітко визначеним алгоритмом (рис. 1).

1. Аналізується перший параметр лінгвістичної складової моделі cq_i. Так як тип тексту визначає його структуру, а також стилістичні, семантичні та синтаксичні особливості, то в залежності від значення cq_i для тексту будуть характерні певні граматичні особливості, на які буде звертатися увага при подальшому аналізі. Тоді можна сказати, що існує такий оператор $Q_i(r_i)$, який ставить у відповідність конкретному значенню змінної cq_i із множини можливих значень CQ вектор граматичних параметрів r_i:

$$Q_i(r_i) : CQ \to cq_i,$$

де $CQ = \{cq_1, ..., cq_i, ..., cq_n\}$ – множина існуючих типів текстів, $i = \overline{1, n}$, n – кількість типів.

2. Фіксується кількість складних частин електронного документа f_j, $j=\overline{1,m}$, m – кількість складних синтаксичних частин.

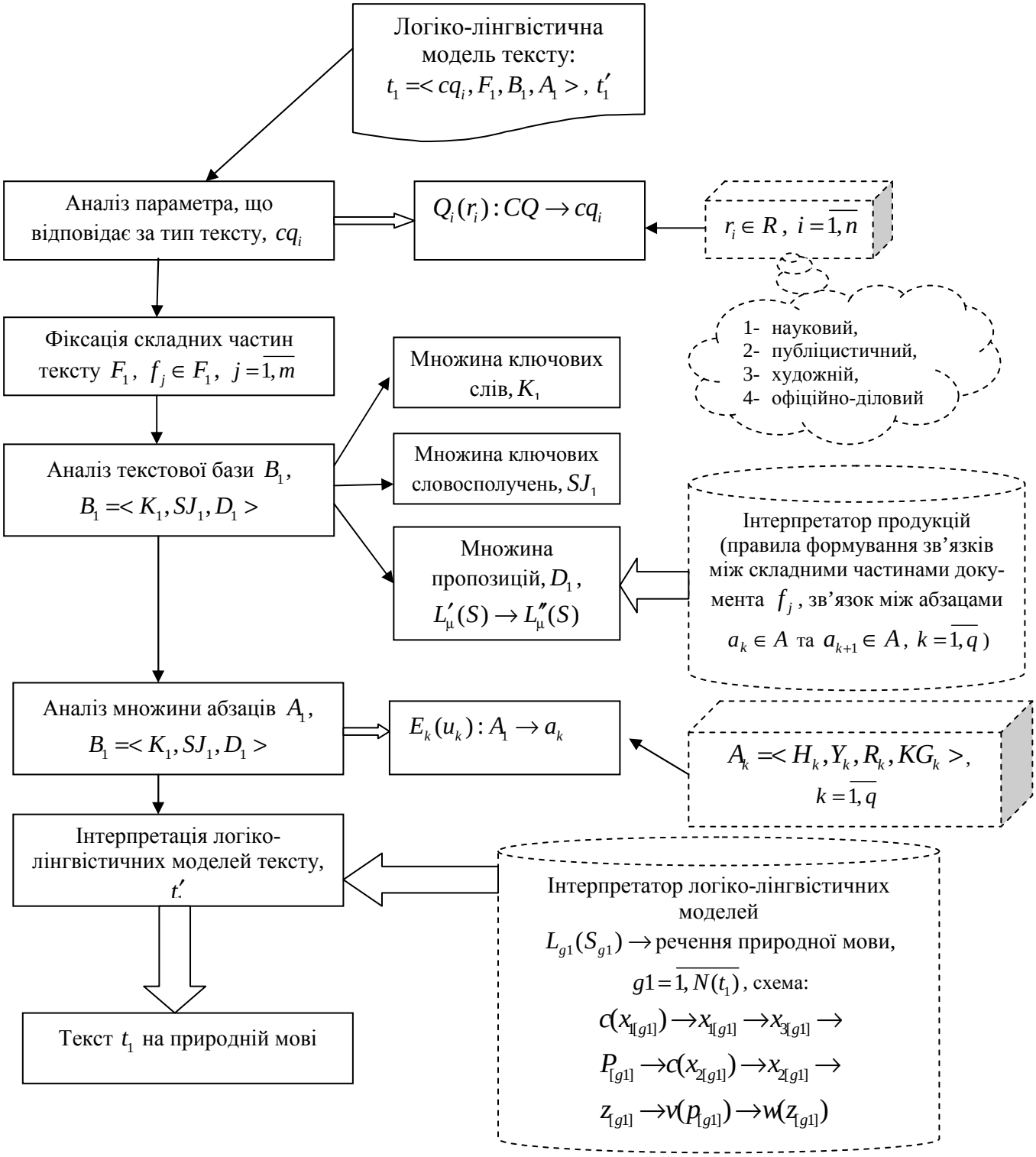

Рис. 1. Алгоритм аналізу логіко-лінгвістичних моделей текстових документів

3. На відміну від двох попередніх параметрів, текстова база B_1 є однією із значущих змінних для формування змістовного портрета документа. Множина ключових слів, мно-

жина ключових словосполучень, а також множина пропозицій формують основу для вилучення знань з електронного документа.

Для аналізу множини пропозицій застосовується інтерпретатор продукцій, який працює циклічно. У кожному циклі він переглядає правила формування зв'язків між складними частинами документа f_j із множини заданих в моделі F_1, щоб з'ясувати, які посилання із заданої множини D_1 збігаються з відомими на даний момент фактами з робочої пам'яті.

Після вибору правило спрацьовує, його висновок заноситься в робочу пам'ять, а цикл повторюється спочатку.

Тобто в результаті роботи інтерпретатора продукцій шукаються такі моделі $L'_{\mu}(S)$, що належать до абзацу $a_k \in A$, які за змістом передують або з яких випливають моделі $L''_{\mu}(S)$ з абзацу $a_{k+1} \in A$, $k = \overline{1,q}$: $L'_{\mu}(S) \to L''_{\mu}(S)$.

1. Аналізується множина абзаців A_1, у кожному з них ($a_k \in A_1$) відмічається тип зв'язку між реченнями H_k, тип тематичної прогресії Y_k, тип рематичної домінанти R_k та засоби когезії KG_k. Таким чином, буде існувати такий оператор $E_k(u_k)$, який ставить у відповідність кожному абзацу a_k із множини A_1 вектор параметрів u_k, що формують зміст відповідного абзацу:

$$E_k(u_k): A_1 \to a_k \text{ або } E_k(u_k): A_1 \to < H_k, Y_k, R_k, KG_k >.$$

2. Застосовується інтерпретатор логіко-лінгвістичних моделей речень природної мови. Він перетворює модель $L_{g1}(S_{g1})$, $g1 = \overline{1, N(t_1)}$ у речення природної мови шляхом синтезу простих предикатів $L_{\mu}(S)$ (з урахуванням вектора граматичних параметрів r_i, набору пропозицій $L'_{\mu}(S) \to L''_{\mu}(S)$ та вектора параметрів u_k) за такою схемою:

$$c(x_{1[g1]}) \to x_{1[g1]} \to x_{3[g1]} \to P_{[g1]} \to c(x_{2[g1]}) \to x_{2[g1]} \to z_{[g1]} \to v(p_{[g1]}) \to w(z_{[g1]}).$$

Запропонований алгоритм дозволяє автоматично аналізувати логіко-лінгвістичні моделі текстів різних типів та довільної складності. Створена база правил формування зв'язків між складними частинами тексту, а також в середині абзаців є неодмінною складовою функціонування алгоритму, через те що саме вона дає можливість вилучати зміст з текстової інформації.

4. Висновки

Логіко-лінгвістична модель (1) – (2) є засобом, що дозволяє формалізувати тексти природної мови за єдиним принципом. У свою чергу, аналіз таких моделей дає змогу зробити зворотну операцію – відновити текст. Формальний апарат трансформації тексту у логіко-лінгвістичну модель і навпаки виступає єдиним засобом автоматизації процесу обробки текстової інформації. Створення автоматичної системи лінгвістичного аналізу електронних документів спростовує такі гіпотези щодо складності текстів, як:

– чим більша кількість термінів у тексті, тим складніший він для перекладу;

– чим складніше дерево предикатної структури, синтаксис тексту, тим складніше парсинг тексту;

– складність тексту прямо пропорційна середній довжині слова та середній довжині речення.

Усі вищенаведені гіпотези сформульовані, виходячи з статистичного аналізу природно-мовних текстів. Змістовна ж компонента аналізу електронних документів містить у

собі поєднання бази правил, що створена на основі досліджень лінгвістів, з методами обробки масивів текстової інформації.

СПИСОК ЛІТЕРАТУРИ

1. Методы и модели анализа данных: OLAP и Data Mining / [Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.]. – СПб.: БХВ-Петербург, 2007. – 384 с.
2. Кобозева И.М. Лингвистическая семантика / Кобозева И.М. – М.: Эдиториал УРСС, 2000. – 352 с.
3. Вавіленкова А.І. Логіко-лінгвістичні моделі речень як засіб порівняння текстових документів за змістом / А.І. Вавіленкова // Математичні машини і системи. – 2012. – № 1. – С. 166 – 173.
4. Вавіленкова А.І. Проект комп'ютерної технології лінгвістичного аналізу електронних документів / А.І. Вавіленкова // International Scientific Journal Acta Universitatis Pontica Euxinus. Special number. – Варна, 2014. – С. 388 – 394.
5. Вавіленкова А.І. Теоретичні основи аналізу електронних текстів / Вавіленкова А.І., Ланде Д.В., Литвиненко О.Є. – К.: НАУ, 2014. – 250 с.

Стаття надійшла до редакції 22.07.2014