УДК 81’322.3

М.В. Давидов
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ГРАМАТИЧНА КОРЕКЦІЯ РЕЧЕНЬ ІЗ ВИКОРИСТАННЯМ ГРАФУ ВЗАЄМОВИКЛЮЧНИХ ГІПОТЕЗ

Описано метод автоматичної корекції речень української мови, оснований на граматиці залежностей. Метод використовує граф взаємовиключних гіпотез для усунення неоднозначності. У проведеному експерименті з неоднозначними реченнями правильно виправлено 37 % речень, що більше за 14 % правильно виправлених із використанням засобів перевіряння орфографічних помилок.

Ключові слова: опрацювання мови, граматика залежностей, автоматична корекція помилок.

A method for automatic correction of Ukrainian sentences is introduced. The method is based on dependency grammar and utilizes mutually exclusive hypothesis graph for word sense disambiguation. 37 % of ambiguous sentences which were correctly corrected as opposed to 14 % corrected by spell checker.

Key words: natural language processing, dependency grammars, automatic error correction.

Вступ

Задача автоматизації корекції речень виникає під час розроблення засобів розпізнавання тексту із зображень [1], у засобах голосового набору тексту [2], у засобах комп’ютерного перекладу та допоміжних засобах, під час набору речень у текстових процесорах, засобах роботи з електронними каталогами [3] тощо.

Задача автоматичної корекції речень полягає у виборі множини виправлень речення, які в сукупності забезпечують правильне стилістично, граматично та орфографічно речення, яке підходить за змістом до тексту. У загальному формулюванні ця задача належить до повних задач штучного інтелекту, оскільки вимагає розуміння суті тексту для вибору правильної множини виправлень. У простих випадках задачу корекції можна розв’язати з використанням наближених методів, основаних на статистичному аналізі, застосуванні правил, граматичному аналізі речень.

Сьогодні найрозвиненішою відкритою технологією граматичної корекції україномовного тексту є програма перевіряння граматики LanguageTool, яка може працювати надбудовою для OpenOffice. Програма LanguageTool виконує частковий розбір речення та містить множину правил для усунення поширених граматичних помилок.

У статті поставлена задача автоматичної корекції помилки у реченні, тобто задача вибору виправленого речення без участі людини, що є логічним розвитком допоміжних інформаційних технологій коригування тексту.

Аналіз відомих досліджень

Проблема виправлення слів у тексті належить до проблем опрацювання природної мови і вже довгий час досліджується науковцями. Серед помилок, які можуть трапитись у тексті, виділяють помилки, які утворюють відоме слово, і помилки, що утворюють слово, якого немає у словнику [4].

Серед вже відомих підходів до виправлення речень найпоширеніші підходи з використання словника словоформ, із застосуванням словника найпоширеніших помилок [5], і гібридні методи.

Метод, оснований на використанні словника словоформ, полягає у пошуку найбільш подібних слів у словнику і заміні слова на найподібніше зі словника. При цьому найчастіше використовуються модифікації відстані Левенштейна [6] зі статистично розрахованими параметрами [7] для

встановлення міри відмінності слів. Цей підхід дає хороші результати у разі формування підказки для користувачів текстового процесора, але виявився малоефективним для автоматичного виправлення речень у випадку, коли є декілька однаково добрих варіантів виправлення слова.

Подальше вдосконалення інформаційної технології граматичної корекції тексту вимагає застосування методів граматичного та семантичного аналізу текстів [8].

Класичним підходом до синтаксичного аналізу вважається підхід із застосуванням формальних граматик. Формальна граматика – це спосіб опису формальної мови, тобто виділення деякої підмножини з множини всіх скінченних рядків, складених із символів деякого скінченного алфавіту. Для синтаксичного аналізу речень усіх мов, крім полісинтетичних, алфавітом можуть бути всі допустимі слова мови, а скінченними рядками, які розглядаються, є речення мови. Полісинтетичні мови відрізняються складними правилами морфології, що унеможливлює побудову універсального словника.

Розрізняють граматики структури речення (Phrase structure grammars), які розробив американський вчений Ноам Хомський [9, 10], та граматики залежностей (Dependency grammars), що ввів Люсьє Теньєр [11]. Крім того, для моделювання речень у чітко визначеній предметній області добре себе зарекомендували орієнтовані семантичні мережі, використані, зокрема, у роботах Т.К. Вінцюка [12].

Сучасні системи синтаксичного аналізу складаються зі словників, модуля морфологічного аналізу, бази граматичних правил та модуля синтаксичного та семантичного аналізу [13]. При цьому якісний синтаксичний аналіз неможливий без семантичного аналізу, оскільки для побудови правильного дерева синтаксичного розбору необхідно не лише зняти омографію слів, але і правильно трактувати слово у зв'язку з іншими словами речення [14].

Проте не досліджено, якою мірою засоби граматичного аналізу можуть знизити відсоток помилок автоматичного виправлення слів тексту.

Постановка задачі

Серед можливих варіантів постановки задачі автоматичного виправлення помилок у тексті виділяють два основних: виправлення штучних помилок, згенерованих випадково, і виправлення помилок, зроблених у певних обставинах, зокрема під час введення тексту з клавіатури, розпізнавання мовлення, оптичного розпізнавання тексту. Для визначеності введемо модель введення слів користувачем *M*, яка враховуватиме можливість зробити ту чи іншу помилку. Процес виправлення окремого слова w полягає у виборі одного слова з множини слів, запропонованих системою виправлення за словником $D(w) = \{w_1, w_2, ..., w_n\}$.

Під задачею автоматичного виправлення слів будемо розуміти задачу побудови функції виправлення $C : (w, S, C, M) \to W$, яка для заданого помилково написаного слова w, речення S, до якого входить помилково написане слово, та контексту C повертає слово з множини допустимих слів мови *W*, так, щоб максимізувати імовірність отримати правильний варіант виправлення.

Досліджується ефективність автоматичного виправлення помилок у тексті із застосуванням таких методів:

1) вибір першого з варіантів, запропонованих системою корекції орфографії;

2) вибір варіанта, запропонованого системою корекції орфографії, з наданням переваги словам, які вже траплялися у тексті у такій самій або іншій формі;

3) вибір варіанта корекції, який дає максимальну вагу в разі застосування зваженої граматики залежностей;

4) вибір варіанта корекції, який дає максимальну вагу у разі застосування зваженої граматики залежностей у поєднанні з наданням переваги словам, які вже траплялися у тексті у такій самій або іншій формі.

Для корекції орфографії використано систему корекції spell-uk, яка має відкриті вихідні коди, що дає змогу використати її для досліджень. Для граматичного аналізу речень розроблено метод на основі зваженого графу взаємовиключних гіпотез, описаний нижче. Задача визначення того, яке слово є помилковим, у статті не розглядається.

Основна частина

Побудова сучасної технології виправлення слів тексту вимагає врахувати модель введення слів користувачем *M*, модель речення, тобто граматичну модель *G*, та модель контексту. У роботі використано граматичну модель на основі залежностей та статистичну модель контексту, яка враховує лише те, чи слово повторюється у контексті у тій чи іншій граматичній формі, чи ні.

Розроблення граматичної моделі речень є важкою задачею, яка ускладнюється омографією та неоднозначністю конструкцій мови [15]. Для ефективного розбору речень використано зважений граф взаємовиключних гіпотез, що дало змогу об'єднати варіанти семантики слів та варіанти їх виправлення в одну структуру.

Зважений граф взаємовиключних гіпотез (ЗГВГ) – це орієнтований граф $G = (V,E)$, доповнений розбиттям множини вершин на підмножини $H = \{H_1, H_2, \ldots, H_N\}$. Кожна вершина графу є гіпотезою, а гіпотези, які належать до однієї групи, є взаємовиключними. Зв'язок між гіпотезами з різних груп задається дугами, а між гіпотезами, які належать до однієї групи, дуги відсутні. Ваги вершин та дуг задаються функціями $\alpha : V \to \Re$ та $\beta : E \to \Re$ відповідно. Вибір однієї гіпотези з кожної групи та побудову породженого підграфу називають конкретизацією ЗГВГ. Приклад ЗГВГ та однієї з його конкретизацій наведено на рис. 1.

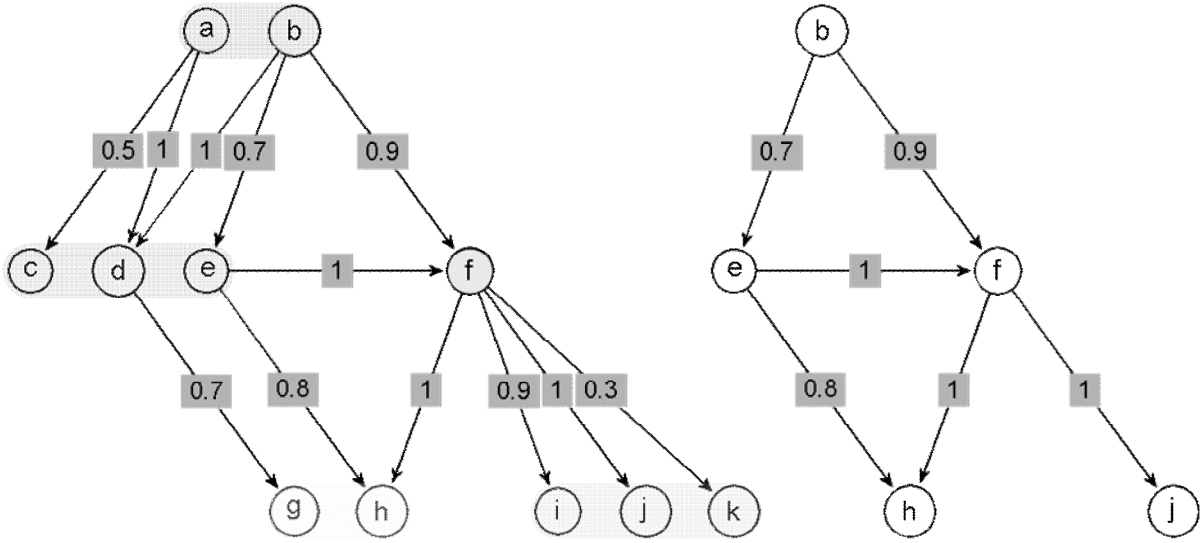

Рис. 1. Зважений граф взаємовиключних гіпотез (зліва) та одна з його конкретизацій (справа). У наведеному прикладі групами гіпотез є $H_1 = \{a,b\}$, $H_2 = \{c,d,e\}$, $H_3 = \{f\}$, $H_4 = \{g,h\}$, *та* $H_5 = \{i,j,k\}$

У разі моделювання речення у вигляді ЗГВГ галуження максимальної ваги його конкретизацій утворюють варіанти ймовірного розбору речення. Використана модель речення основана на граматиці залежностей та враховує можливе трактування слів та їх роль у реченні.

Розбір речення складається з двох етапів. На першому відбувається морфологічний аналіз слів, створення гіпотез та визначення можливих граматичних категорій слова. На другому етапі формується модель речення на основі ЗГВГ.

Використання словника spell-uk дає змогу отримати для слова української мови базову форму та набір тегів, які застосовуються для словозміни в словнику. Словник spell-uk може використовуватися з різними системами перевіряння орфографії, зокрема з бібліотекою OpenOfficeSpellDictionary [16]. Ця бібліотека надає засоби швидкого пошуку основної форми слова та тегів, які вказують на можливі варіанти словозміни. Також доступні функції пошуку варіантів можливої помилки під час написання слова.

Для кожного слова речення виконується пошук можливих варіантів у словнику spell-uk, а отримані основна форма слова та теги використовуються для визначення граматичних категорій роду (M, F, N), числа (SG, PL), особи (p1, p2, p3, p-), відмінку (c1-c7), частини мови (NOUN,

PRONOUN, VERB, ADJ, ADV, NEGATION, PUNCT, CONJ, NUMERAL, PARTICLE, ADVPART, ADJPART, PREPOS, HELPWORD), часу (PAST, PRESENT, FUTURE), та виду (PERFECT, SIMPLE). Для кожного слова речення формується група гіпотез, яка складається з можливих значень слова. До таких можливих значень належать як варіанти слова з урахуванням омографії, так і варіанти змісту і ролі в реченні, яке слово може відігравати. Наприклад, слово “жаль” може бути як частиною вставного словосполучення “на жаль”, так і окремим іменником “жаль”. З гіпотез усіх слів речення формується зважений граф взаємовиключних гіпотез.

У ході синтаксичного аналізу може виявитися, що вибрати правильний варіант розбору речення можна лише з врахуванням теми і контексту мовлення. При цьому вага окремих гіпотез для слова може змінюватися залежно від того, чи відповідає слово тематиці мовлення і словам у сусідніх реченнях. На рис. 2 наведено граф зв’язків речення "Робота зроблени вчасно". Речення містить помилку, і визначити, якому виправленню надати перевагу – “Робота зроблена вчасно”, чи “Робота зроблено вчасно”, можна лише з врахуванням контексту та тематики мовлення. У наведеному прикладі слову “робот” надано вагову перевагу, і синтаксичний розбір речення “Робота зроблено вчасно” має більшу вагу (рис. 3).

Для обгрунтування надання переваги словам, які повторюються у контексті, проведено низку експериментів із текстовими корпусами за напрямами “законодавство”, “технічна література”, “художня література”. Експерименти проведено з метою встановлення імовірності того, чи слово буде використано у тій самій, чи в іншій формі в уривку заданої довжини. Довжина уривку вимірюється у словах. Під контекстом C розумітимемо множину слів тексту навколо заданого слова w.

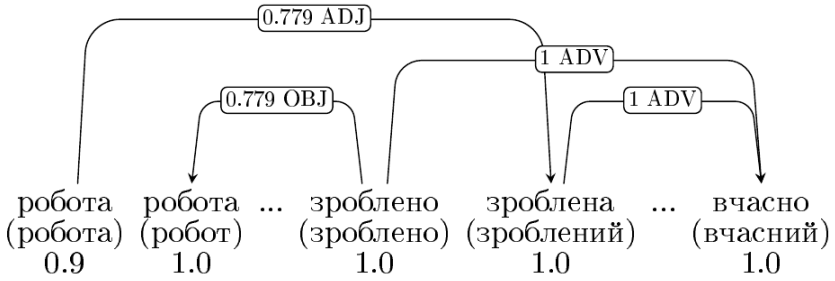

Рис. 2. Граф зв’язків речення "Робота зроблени вчасно" з деякими варіантами виправлення

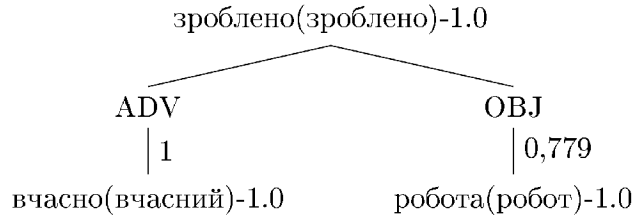

Рис. 3. Синтаксичний розбір речення з рис. 2 із наданням переваги слову "робот"

Розраховують два показники повторення слів $p_r = P(w \in C)$ та $p_b = P(Base(w) \in Base(C))$, де $Base(w)$ позначає базову форму слова w зі словника, а $Base(C)$ – множину всіх базових форм слів із контексту C.

Результати експериментів, проведених із текстами трьох типів, а саме законодавства, наукових книжок, художніх книжок, наведено на рис. 4. Отримані для наукової літератури та законодавства показники повторення слів вищі, ніж для художньої літератури, вказують на те, що статистичний аналіз сусідніх речень для цих видів творів може виявитися продуктивнішим, ніж для творів художньої літератури.

Знаючи імовірність повторення слова p_r та імовірність повторення базової форми слова p_b, можна розрахувати апостеріорну суб'єктивну імовірність виправлення w_i за теоремою Байєса:

$$P(w_i \mid w,S,C,M) \sim P(w \mid w_i,S,C,M) \cdot P(w_i \mid S,C,M). \quad (1)$$

Приймаючи припущення, що процес створення помилки не залежить від навколишніх слів, а залежить лише від процесу введення тексту користувачем, отримаємо

$$P(w \mid w_i,S,C,M) = P(w \mid w_i,M). \quad (2)$$

Крім того, можливість виявити слово w_i у реченні S у контексті C не залежить від моделі створення помилки M : $P(w_i \mid S,C,M) = P(w_i \mid S,C)$.

Приймаючи припущення про незалежність контексту та поточного речення, отримаємо

$$P(w_i \mid S,C) \sim P(S,C \mid w_i) \cdot P(S,C) = P(S \mid w_i) \cdot P(C \mid w_i) \cdot P(S,C). \quad (3)$$

З (1–3) випливає остаточна формула розрахунку імовірності виправлення

$$P(w_i \mid w,S,C,M) \sim P(w \mid w_i,M) \cdot P(S \mid w_i) \cdot P(C \mid w_i). \quad (4)$$

Беручи за відомий показник контексту лише показник того, чи слово в контексті повторювалось у такій самій, чи в іншій формі, розраховуємо значення множника, яке відповідає за використання слова у контексті:

$$P(C \mid w_i) = \begin{cases} p_b, якщо Base(w_i) \in Base(C) \\ 1 - p_b, якщо Base(w_i) \notin Base(C). \end{cases}$$

Імовірність граматично правильного речення $P(S \mid w_i)$ розраховується за показником ваги максимального галуження $W(w_i,S)$ у графі взаємовиключних гіпотез, побудованому за граматикою залежностей за формулою $P(S \mid w_i) \sim e^{W(w_i,S)}$.

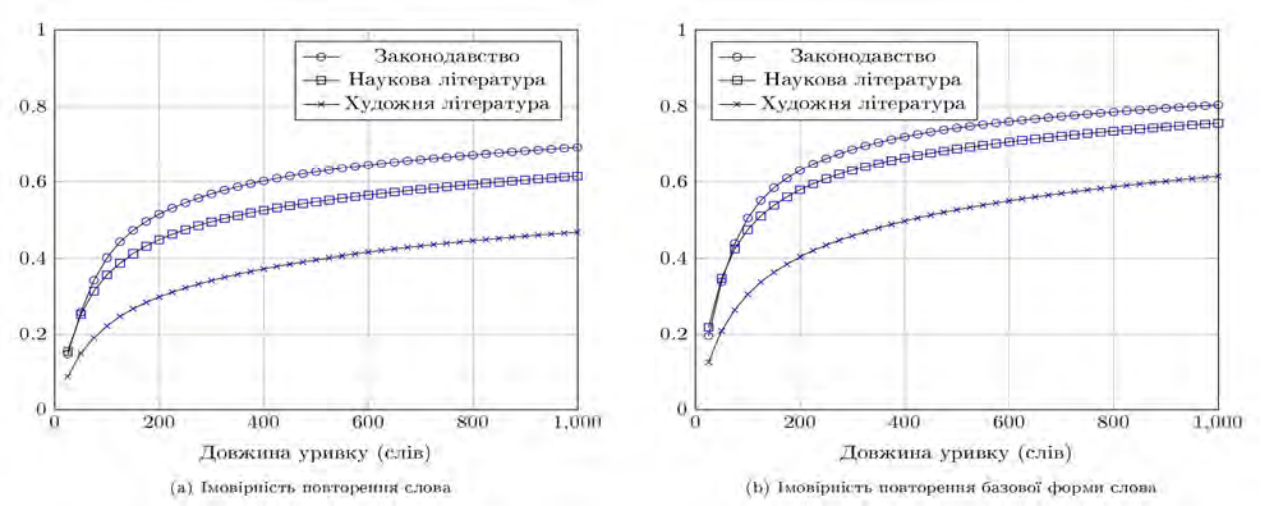

Рис. 4. Імовірність повторення випадково вибраного слова у тій самій (a) або іншій граматичній формі (b) залежно від довжини уривку для різних текстових баз

Результати експериментів

Для експериментального порівняння ефективності методів корекції слів поставлено експеримент на корпусі художніх та корпусі наукових текстів. З текстових баз випадковим вибором одержано по 2000 речень, а у кожному реченні довільно змінено одну літеру в не відомому наперед

слові. Після цього речення опрацьовано засобами корекції орфографії. Позитивним результатом корекції вважався збіг виправленого речення й оригіналу. Під час тестування не враховувалися речення, які мали лише один варіант виправлення, тобто для яких не потрібно було застосовувати додаткових засобів аналізу. Таких речень було 36 % у художніх текстах та 51 % у наукових.

Застосування штучного методу створення помилки дає змогу уникнути необхідності врахування моделі помилки, тобто $P(w \mid w_i, M) \sim 1$.

У табл. 1 наведено досліджені методи і відповідні їм методи розрахунку суб'єктивної апостеріорної імовірності застосування виправлення w_i з множини виправлень, запропонованих системою перевіряння орфографії. У табл. 2 наведено відсоток правильно виправлених речень і часові показники методів.

Менший відсоток правильно виправлених речень наукових текстів пов'язаний із використанням скорочень та абревіатур, які не перелічені у словнику spell-uk.

Таблиця 1

Досліджувані методи та формули розрахунку імовірності виправлення

№	Метод	Формула підрахунку суб'єктивної апостеріорної імовірності виправлення
1	Випадковий вибір	$P(w_i \mid w,S,C,M) \sim 1$
2	З урахуванням граматичної моделі	$P(w_i \mid w,S,C,M) \sim P(S \mid w_i)$
3	З урахуванням моделі контексту	$P(w_i \mid w,S,C,M) \sim P(C \mid w_i)$
4	З урахуванням моделі контексту і граматичної оделі	$P(w_i \mid w,S,C,M) \sim P(S \mid w_i) \cdot P(C \mid w_i)$

Таблиця 2

Відсоток правильно виправлених речень і часові показники методів виправлення

№	Метод	Відсоток правильно виправлених речень			Середній час опрацювання одного речення, мс
		Художні тексти	Наукові тексти	Середній показник	
1	Випадковий вибір	16 %	11 %	14 %	26,0
2	З урахуванням граматичної моделі	21 %	22 %	22 %	30,0
3	З урахуванням моделі контексту	37 %	27 %	32 %	26,5
4	З урахуванням моделі контексту і граматичної моделі	41 %	33 %	37 %	30,5

Висновки

У ході проведених робіт розроблено метод автоматичної орфографічної та граматичної корекції речень із врахуванням контексту та граматичної моделі речення. Розроблені засоби корекції та тестові дані розміщені в системі контролю версії github за адресою https://github.com/mdavydov/UkrParser та доступні для завантаження за ліцензією GPLv3.

Проведені експерименти показали доцільність застосування статистичної моделі контексту та граматичної моделі речення на основі залежностей для автоматичного виправлення помилок.

Подальші дослідження будуть спрямовані на вдосконалення засобів семантичного аналізу та засобів видобування знань із текстів. Такими знаннями, корисними для граматичного аналізу та перекладу, можуть бути спеціальні терміни, скорочення, власні назви.

1. Bassil Y. Ocr post-processing error correction algorithm using google's online spelling suggestion / Youssef Bassil, Mohammad Alwani // LACSC – Lebanese Association for Computational Sciences: Journal of Emerging Trends in Computing and Information Sciences. – ISSN 2079-8407. – Vol. 3. – No. 1. – January 2012. – 9 p. 2. Робейко В.В. Розпізнавання спонтанного мовлення на основі акустичних композитних моделей слів у реальному часі / В.В. Робейко, М.М. Сажок // Штучний інтелект. – 2012. – № 4. – С. 253–267. 3. Ярмолюк Р.С. Основні типи та джерела помилок у записах електронного каталогу / Р.С. Ярмолюк // Вісник Нац. у-ту "Львівська політехніка". Інформаційні системи та мережі. – № 689 (2010). – С. 348–357. 4. How difficult is it to develop a perfect spell-checker? A cross-linguistic analysis through complex network approach / M. Choudhury, M. Thomas, A. Mukherjee, A. Basu, and N. Ganguly // Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing, 2007. – P. 81–88. 5. Lee J.S.Y. Automatic Correction of Grammatical Errors in Non-native English Text / John Sie Yuen Lee // Massachusetts Institute of Technology. – 2009. – PhD Thesis. 6. Владимир И. Левенштейн. Двоичные коды с исправлением выпадений, вставок и замещений символов / // Доклады Академии Наук СССР. – 163 (4): С. 845–848. 7. Whitelaw C. Using the web for language independent spellchecking and autocorrection / Casey Whitelaw, Ben Hutchinson, Grace Y. Chung, and Gerard Ellis // In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 – Volume 2 (EMNLP '09). – Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, p. 890 899. 8. Большаков И. А. Проблемы автоматической коррекции текстов на флективных языках. Итоги науки и техн. Сер. Теор. вероятн. Мат. стат. Теор. кибернет., 28, ВИНИТИ, М., 1988, 111–139. 9. Chomsky N. Syntactic Structures / Mouton & Co. – Feb. 1957. – 117 p. 10. Chomsky N. Aspects of the Theory of Syntax / Cambridge, Massachusetts: MIT Press – 1965. – 251 p. 11. Tesniere L. Elements de syntaxe structurale / L. Tesniere // Paris: Klincksieck. – 1966. – 670 p. 12. Винцюк Т. К. Анализ, распознавание и интерпретация речевых сигналов. – К.: Наук. думка, 1987. – 262 с. 13. Кагиров И.А. Автоматический синтаксический анализ русских текстов на основе грамматики составляющих / И. А. Кагиров, А. Б. Леонтьева // Известия ВУЗов. Сер, Приборостроение. – СПб.: Издание Санкт-Петербургского государственного института точной механики, 2008. – Том 51, № 11. – С.47–56. 14. Лингвистический энциклопедический словарь / гл. ред. В. Н. Ярцева. – М.: Советская энциклопедия, 1989. – 688 с. 15. Гаршина В.В. Разработка лингвистического парсера русского языка / В. В. Гаршина, Ю. А. Богоявленская // Вестник Вгу, Серия: Системный анализ и информационные технологии. – 2012, № 2. – С. 174–182. 16. Open Office Spell Dictionary [Эл. ресурс].– Режим доступа: http://www.jamochamud.org/docs/org/dts/ spell/dictionary/OpenOfficeSpellDictionary.html. – перевірено 17.03.2014.