УДК 81'322.3

М.В. Давидов
Національний університет „Львівська політехніка”,
кафедра „Інформаційні системи та мережі”

МОДЕЛЮВАННЯ ЗВ'ЯЗКІВ МІЖ СЛОВАМИ РЕЧЕННЯ З ВИКОРИСТАННЯМ ВІДКРИТОГО СЛОВНИКА SPELL-UK

Описано метод моделювання зв'язків між словами речення заснований на граматиці залежностей. Для морфологічного аналізу використано відкритий словник spell-uk, що дає змогу використовувати розроблені бібліотеки у програмах із відкритими вихідними кодами. У проведеному експерименті правильно визначено зв'язки між словами у 48% речень казки І. Франка “Фарбований лис”.

Ключові слова – граматика залежностей, spell-uk, граматичний розбір.

A method for modeling dependencies between words in Ukrainian sentences is introduced. The method is based on dependency grammar. Open source dictionary ispell-uk is used for morphology analysis that makes it possible to use the developed library in open projects. The developed algorithm exhibits parsing rate about 48% on sentences of Ivan Franco's fairy tale “Painted Fox”.

Key words – dependency grammar, spell-uk, syntax parsing.

1. ВСТУП

Проблема автоматичного синтаксичного аналізу природних мов є однією з найскладніших та, одночасно, однією з найдовше досліджуваних проблем штучного інтелекту та комп'ютерної лінгвістики. Автоматичний синтаксичний аналіз застосовується в засобах комп'ютерного перекладу, автоматичного реферування, видобування даних та знань, перевіряння граматики, допомоги при написанні текстів, розпізнавання мовлення [1] тощо.

Особливістю задач, які вимагають застосування синтаксичного аналізу є те, що необхідно не лише побудувати граматично правильний розбір речення, але і зняти омографію слів та визначити, чи речення має зміст у поточному контексті чи ні.

Розроблення програм для синтаксичного аналізу української мови ускладнюється відсутністю відкритих словників, які надають інформацію про керування дієслів та можливий контекст використання слова. Більшість словників української є комерційними проектами, що перешкоджає розповсюдженню технологій синтаксичного аналізу та граматичної корекції у відкритих програмних продуктах. Одним із небагатьох словників із відкритими для вільного використання словниковими базами є словник spell-uk. Spell-uk – це проект зі створення відкритих словників для перевіряння орфографії української мови. Словники, розроблені в рамках проекту spell-uk можуть використовуватися в декількох відкритих системах перевіряння орфографії, зокрема hunspell, myspell, aspell та ispell [2].

На сьогодні найбільш розвиненою відкритою технологією граматичної корекції україномовного тексту є програма перевіряння граматики LanguageTool, яка може працювати надбудовою для OpenOffice. Програма LanguageTool на сьогодні не містить модулів синтаксичного аналізу української мови.

Актуальною залишається задача синтаксичного розбору речень української мови з використанням відкритих словників.

2. АНАЛІЗ ВІДОМИХ ДОСЛІДЖЕНЬ

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 стрічок складених із символів деякого скінченного алфавіту. Для синтаксичного аналізу речень усіх мов, крім полісинтетичних, алфавітом можуть бути всі допустимі слова мови, а скінченними стрічками, які розглядаються, є речення мови. Полісинтетичні мови відрізняються складними правилами морфології, що унеможливлює побудову універсального словника.

Розрізняють граматики структури речення (Phrase structure grammars), розроблені американським вченим Ноамом Хомським [3, 4], та граматики залежностей (Dependency grammars), введені Люсьє Теньєром [5]. Крім того, для моделювання речень у чітко визначеній предметній області добре себе зарекомендували орієнтовані семантичні мережі, використані, зокрема, у роботах Т.К. Вінцюка [6].

Граматики структури речення, такі як контекстно-вільна граматика (Context-free grammar, CFG [7]), доповнена граматика (Affix grammar [8]), стохастична контекстно-вільна граматика (Stochastic context-free grammar [9]), комбінаційна категоріальна граматика (Combinatory categorial grammar, CCG [10]), граматика головного елементу (Head Grammar, HG [11]), граматика приєднаних дерев (Tree-adjoining grammars, TAGs [12]), відрізняються чіткими правилами визначення структури речення та найліпше підходять для аналізу мов із фіксованим порядком слів у реченні.

Граматики залежностей спрощують опис мов зі змінним порядком слів у реченні. До цих граматик відносять граматику зв'язків (Link grammar [13]), розширювану граматику залежностей [14], граматику слів [15], функціонально-генеративний опис [16] та ін.

Класична контекстно-вільна граматика побудована на основі породжувальних конструкцій. При застосуванні контекстно-вільної граматики для синтаксичного розбору речень виникають такі проблеми:

1) узгодження особи, числа, відмінку та інших граматичних категорій вимагає введення великої кількості правил;
2) якщо синтаксичне дерево може бути побудовано декількома способами, немає можливості визначити, який із них імовірніший за інший;
3) більшість дієслів мови вимагають деталізації правил керування доповненнями, що так само призводить до збільшення кількості правил.

Доповнена граматика та стохастична контекстно-вільна граматика описують той самий клас мов, що і контекстно-вільні граматики, але при цьому принцип задання правил граматики відрізняється. У доповненій граматиці введені спеціальні афікси, які дають змогу описати правила узгодження слів без збільшення кількості основних правил.

Стохастична контекстно-вільна граматика вирішує питання про вибір варіанту синтаксичного розбору у випадку, коли засоби розбору з використанням контекстно-вільної граматики дають декілька альтернативних варіантів розбору речення. Для цього у стохастичній контекстно-вільній граматиці введено імовірність застосування кожного з правил.

Комбінаційна граматика категорій, граматика головного елементу та граматика приєднаних дерев розроблені незалежно, як більш виразні, але, разом із тим, обчислювально ефективні альтернативи до контекстно-вільних граматик. Відомо, що ці граматики описують один і той самий клас мов, і їх відносять до м'яко контекстно-залежних граматик [17].

У граматиці зв'язків правила задаються не у вигляді породжувальних правил “зверху до низу”, а у вигляді опису оточення кожного слова. При цьому опис оточення задається у вигляді конекторів, які йдуть у заданому порядку та вимагають зчеплення з іншими словами речення. Класична граматика зв'язків вимагає, щоб речення підлягало правилу проективності. Тобто,

щоб усі конектори можна було намалювати над словами таким чином, щоб вони не перетиналися. У такому формулюванні граматика зв'язків еквівалентна певній контекстно-вільній граматиці. Якщо відмовитися від правила проективності в граматиці зв'язків, то значно зростає обчислювальна складність алгоритму розбору та зростає кількість речень, хибно прийнятих за правильні. Відомо, що в російській та українській мові речення ділового стилю зазвичай підлягають закону проективності [18]. З іншого боку, речення розмовної мови та художньої літератури часто порушують це правило.

Сучасні системи синтаксичного аналізу складаються зі словників, модуля морфологічного аналізу, бази граматичних правил та модуля синтаксичного та семантичного аналізу [19]. 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 [20].

Використання орієнтованих семантичних мереж, які би покривали значний набір слів та речень, лишається нерозв'язаною задачею через свою громіздкість та відсутність великих відкритих розмічених мовних баз українською мовою. Відсутність таких баз перешкоджає розвитку автоматизованих методів побудови засобів синтаксичного та семантичного розбору.

3. ПОСТАНОВКА ПРОБЛЕМИ

Актуальною на сьогодні є задача розроблення інформаційної моделі речення, яка би об'єднувала можливі варіанти синтаксичного розбору та семантичної інтерпретації змісту речення, давала би змогу інтегрувати відомості про тематику тексту, суміжні речення, дискурс для побудови найбільш імовірного синтаксичного розбору та найбільш імовірної семантичної інтерпретації речення.

У статті такою моделлю речення обрано зважений граф взаємовиключних гіпотез (ЗГВГ). ЗГВГ – це орієнтований граф $G=(V,E)$, доповнений розбиттям множини вершин на групи гіпотез $H=\{H_1,H_2,\dots,H_N\}$, які не перетинаються. Кожна вершина графа є гіпотезою, а гіпотези, які належать до однієї групи, є взаємовиключними. Зв'язок між гіпотезами з різних груп задається дугами. Між гіпотезами, які належать до однієї групи, дуги відсутні через те, що такі гіпотези несумісні. Ваги вершин та дуг задаються функціями $w_V: V \to \mathbb{R}$ та $w_E: E \to \mathbb{R}$ відповідно. Вибір однієї гіпотези з кожної групи та побудова породженого підграфа називається конкретизацією ЗГВГ. Приклад ЗГВГ та одної з його конкретизацій наведено на рис. 1.

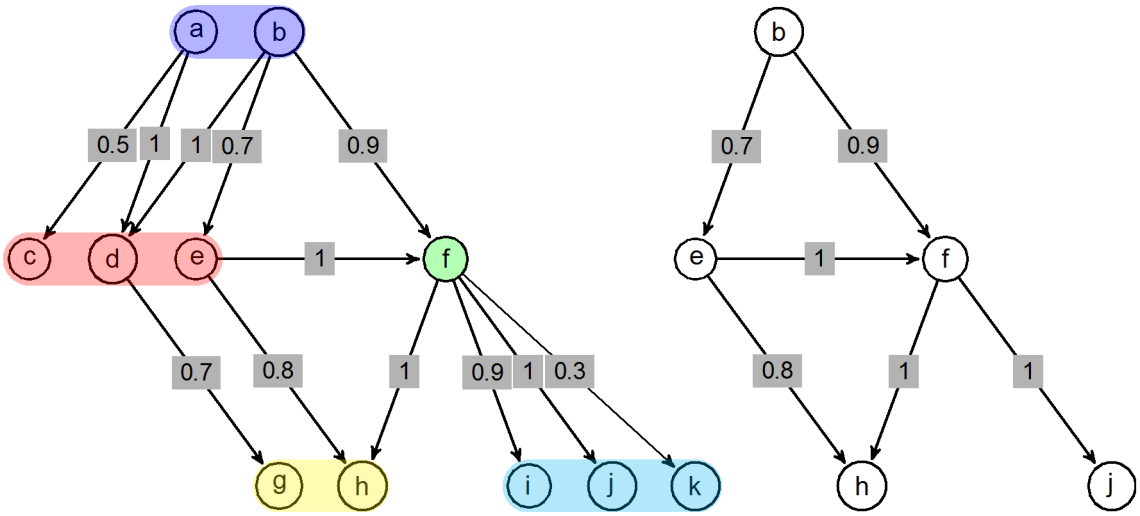

Рис. 1 Зважений граф взаємовиключних гіпотез (зліва) та одна з його конкретизацій (справа). У наведеному прикладі групами гіпотез є $H_1=\{a,b\}$, $H_2=\{c,d,e\}$, $H_3=\{f\}$, $H_4=\{g,h\}$ *і* $H_5=\{i,j,k\}$.

При моделюванні речення у вигляді ЗГВГ можливі значення омографів задаються вершинами, які належать до однієї групи, а зв'язки між членами речення задаються дугами. Задача пошуку найімовірнішої множини зв'язків між членами речення зводиться до задачі пошуку галуження максимальної ваги серед усіх конкретизацій графа.

Використана модель речення заснована на граматиці залежностей та враховує можливе трактування слів та їх роль у реченні. Розглянуто різні граматичні конструкції та наведено рекомендації, яким чином граматичні конструкції речення можуть бути інтегровані в модель на основі ЗГВГ, і яким чином може визначатися вага зв'язку між членами речення.

4. ОСНОВНА ЧАСТИНА

Розбір речення складається з двох етапів. На першому відбувається морфологічний аналіз слів, створення гіпотез та визначення можливих граматичних категорій слова. На другому етапі формується граф речення із встановленням можливих зв'язків між словами речення. Для пошуку найімовірнішої множини зв'язків між словами речення відбувається пошук галуження максимальної ваги у просторі створених гіпотез.

4.1 Морфологічний аналіз

Для морфологічного аналізу слів пошук у словнику spell-uk здійснюється з використанням бібліотеки OpenOfficeSpellDictionary [21]. Ця бібліотека надає засоби швидкого пошуку основної форми слова та тегів, які вказують на можливі варіанти словозміни. Також доступні функції пошуку варіантів можливої помилки при написанні слова.

Для кожного слова 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). У таблиці 1 наведено приклади визначення основної форми для правильно написаного слова-омографа “робота” та слова з орфографічною помилкою “зроблени”.

Таблиця 1: Приклад варіантів основної форми слова та тегів, отриманих зі словника spell-uk

Слово в реченні	Виправлене слово	Основна форма	Теги spell-uk	Визначені теги граматичних категорій
робота	робота	робота	a	SG F c1 NOUN
робота	робота	робот	efg	SG M c2 NOUN
зроблени	зробленим	зроблений	VZ	SG M N c5 ADJ
зроблени	зробленим	зроблений	VZ	PL M F N c3 ADJ
зроблени	зроблених	зроблений	VZ	PL M F N c2 c4 c6 ADJ
зроблени	зроблений	зроблений	VZ	SG M F c4 ADJ
зроблени	зроблена	зроблений	VZ	SG F c1 ADJ
зроблени	зроблені	зроблений	VZ	PL M F N c1 c4 ADJ
зроблени	зроблене	зроблений	VZ	SG N c1 c4 ADJ
зроблени	зроблену	зроблений	VZ	SG F c4 ADJ
зроблени	зроблено	зроблено	-	VERB p- PERFECT

4.2 Для кожного слова речення формується група гіпотез, яка складається з можливих значень слова. До таких можливих значень належать як варіанти слова з врахуванням омографії, так і варіанти змісту і ролі в реченні, яке слово може відігравати. Наприклад, слово “жаль” може виступати як частиною вставного словосполучення “на жаль”, так і окремим іменником “жаль”.

4.3 Встановлення зв'язків між словами речення

Встановлення зваженого зв'язку між словами речення використовується в різних моделях речень, зокрема в моделях на основі умовних випадкових полів [22]. При наявності маркованих корпусів даних вага зв'язку може вираховуватися з використанням методів машинного навчання. Через відсутність необхідних відкритих маркованих корпусів речень українською мовою вага зв'язків була визначена евристично з використанням таких правил:

1) вага зв'язку зменшується зі збільшенням віддалі між словами в реченні;
2) вага зв'язку при незвичному порядку слів у реченні менша за вагу зв'язку звичного порядку;
3) вага зв'язку зменшується, якщо слова не узгоджуються в числі, роді, відмінку, особі тощо;
4) вага зв'язку зменшується, якщо між словами стоїть одна кома, а зв'язок не є зв'язком між однорідними членами речення;
5) зв'язок між однорідними членами речення визначається прямо, сполучник впливає лише на вид зв'язку (єднальний, протиставний, розділовий);
6) для того, щоб врахувати, що деякі частини мови можуть виконувати роль інших, створюються окремі гіпотези, щоб не ускладнювати формування правил; наприклад, в уривку “Лежало два зошити. Перший був червоний, другий – зелений”, числівники перший та другий виступають у ролі іменників, і ця їхня роль позначається окремою гіпотезою;
7) зв'язок прийменника з іменником має бути сильніший за зв'язок дієслова й іменника ($\alpha > \beta$, рис. 2).

Можливі варіанти зв'язків між словами речення наведені у табл. 2.

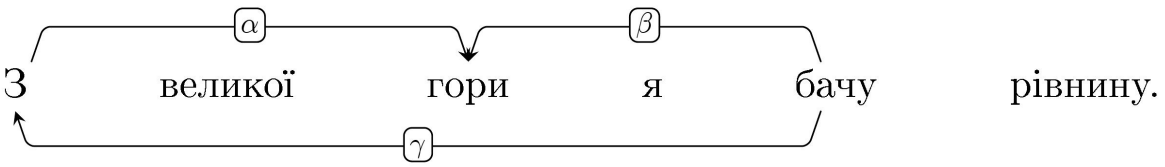

Рис. 2 Зв'язок між прийменником, іменником і дієсловом

Таблиця 2 Основні зв'язки між словами речення

До / Від	Прийменник	Іменник	Дієслово	Прикметник	Прислівник	Дієприкметник	Числівник
Прийменник	-	Група прийменник-іменник	-	-	-	-	Відношення до числа
Іменник	Відношення до матеріалу	Однорідні, прикладки	-	Ознака	-	Ознака	Номер, кількість
Дієслово	Додаток, присудок	Об'єкт, суб'єкт, додаток	Модальне дієслово-інфінітив, однорідні	Обставина	Місце, спосіб дії	-	Кількість
Прикметник	Порівняння	Ознака	-	Однорідні	-	Однорідні	-
Прислівник	Порівняння	-	-	-	Однорідні	-	-
Дієприкметник	Додаток, присудок	Об'єкт, суб'єкт, додаток	-	-	Місце, спосіб дії	Однорідні	Кількість
Числівник	Складений числівник	Кількість	-	-	-	-	Однорідні
Заперечення	Заперечення	Заперечення	Заперечення	Заперечення	Заперечення	Заперечення	Заперечення

Крім зв'язків між членами речення встановлюється також зв'язок до самого об'єкту-речення, який задається крапкою, знаком питання або окличним знаком (рис. 3). Такий зв'язок може позначати граматичні основи речення через зв'язок до присудка (ROOT) або позначати слова, або гіпотези, які можуть бути вилучені або додані до речення в результаті граматичного аналізу (OPTIONAL). У наведеному на рис. 3 прикладі розбору речення “Він намагався натиснути на жаль”, для слова “жаль” розглядається дві гіпотези: “жаль(1)” позначає іменник, “жаль(2)” позначає частину вставного словосполучення. Кома додана як один із варіантів виправлення речення з урахуванням гіпотези про те, що “на жаль” може бути вставним словосполученням. Для того, щоб була можливість виконати граматичний розбір речення без коми, додано зв'язок OPTIONAL від об'єкта-речення до коми.

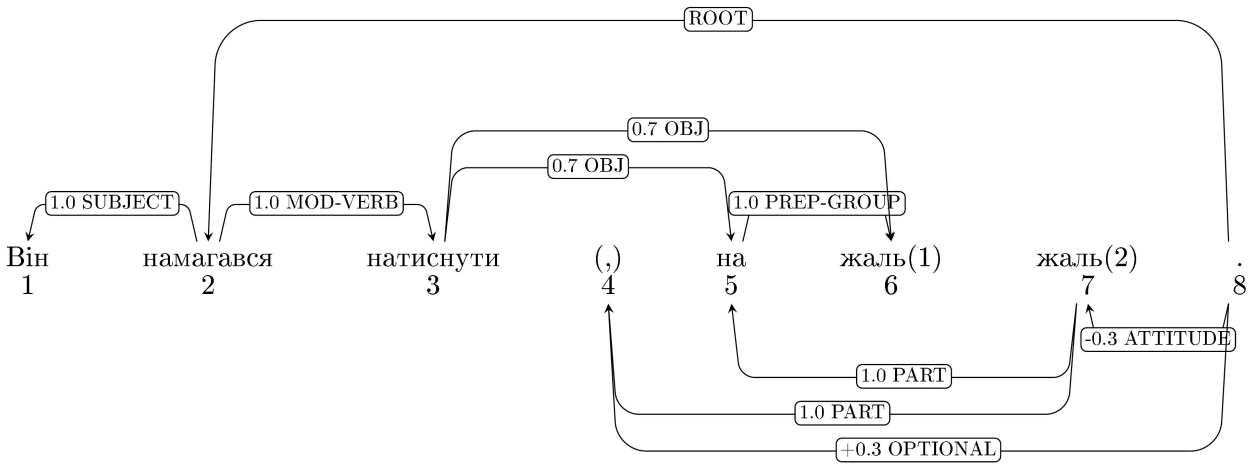

Рис. 3 Зв'язок ROOT та OPTIONAL до об'єкта-речення

У ході синтаксичного аналізу може виявитися, що обрати правильний варіант розбору речення можна лише з врахуванням теми і контексту мовлення. Для усунення неоднозначності окремих слів у реченні може бути використана ймовірнісна модель теми тексту і зв'язків між словами речення [23, 24]. При цьому вага окремих гіпотез для слова може змінюватися в залежності від того, чи відповідає слово тематиці мовлення і словам у сусідніх реченнях. На рис. 4 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. 5).

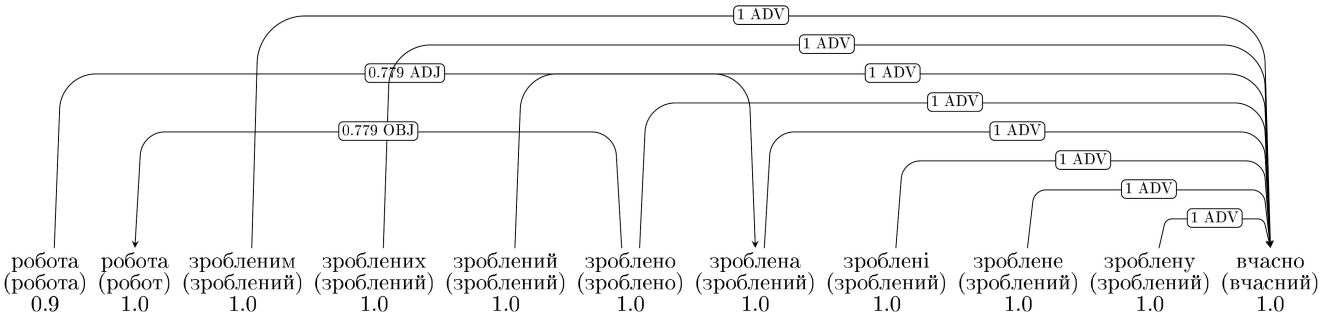

Рис. 4 Граф зв'язків речення "Робота зроблени вчасно" з варіантами виправлення

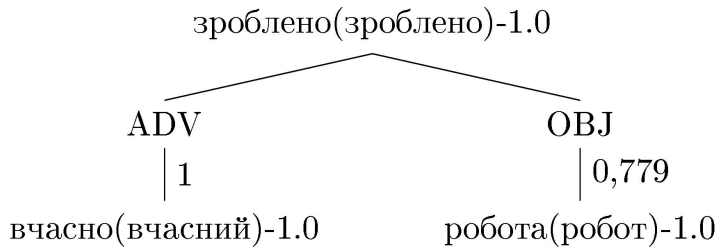

Рис. 5 Синтаксичний розбір речення з рис. 4 при наданні переваги слову "робот"

4.3 Структура програмної системи розбору речень

Схематично структуру програмної системи розбору речень наведено на рис. 6 у вигляді діаграми потоків даних. На вході система отримує текст або дискурс, у залежності від того, чи розглядається лише письмове джерело інформації, чи доступні інші джерела інформації, наприклад, результат розпізнавання голосового повідомлення. Окремі речення тесту (дискурсу) спершу опрацьовуються модулем морфологічного аналізу, який визначає можливі варіанти інтерпретації слова у випадку омографії (або омонімії при розпізнаванні речення з голосу) та визначає граматичні категорії слова в кожному з можливих варіантів. Якщо допускається наявність помилок у вхідних реченнях, морфологічний аналіз може повертати і можливі варіанти зміни слова при підозрі на орфографічну помилку.

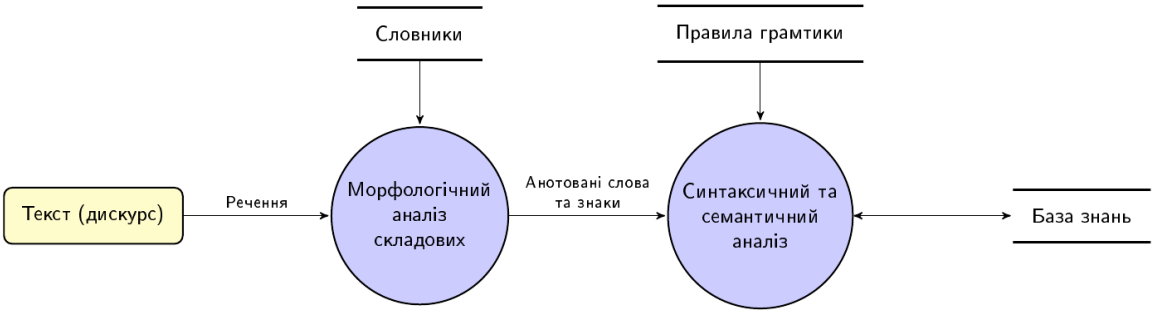

Рис. 6: Діаграма потоків даних системи синтаксичного аналізу

Модуль синтаксичного та семантичного аналізу визначає найбільш імовірні варіанти синтаксичного розбору речення і ймовірне значення висловлювання. Результат аналізу висловлювання може в подальшому використовуватися для розбору наступних речень тексту або дискурсу. Для цього оновляється база знань, яка містить як загальні знання про предметну область, так і знання, отримані при розборі попередніх частин тексту або дискурсу.

У наведеній на рис. 6 діаграмі потоків даних синтаксичний аналіз об'єднано із семантичним через те, що кількість варіантів синтаксичного розбору може експоненційно залежати від кількості слів речення, які є омографами, мають неоднозначно визначені граматичні категорії або можуть містити орфографічні помилки. Об'єднання синтаксичного та семантичного аналізу в один усуває необхідність робити граматичний розбір речень для всіх варіантів значення слів та дає можливість використати евристичні методи пошуку максимального галуження.

5. РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТІВ

Для тестування розробленого методу синтаксичного розбору використано твір І. Франка “Фарбований лис”, який містить 196 речень. Правильним вважався розбір, при якому правильно встановлено зв'язки між основними членами речення. Частки, вигуки, сполучники не містилися в синтезованому дереві розбору. Дерева синтаксичного розбору речення виводилися у файл формату TeX, з якого створювався документ для перегляду. Один із прикладів правильно розібраних речень наведено на рис. 7. За проведеними тестами правильний розбір речення зроблено для 48% речень.

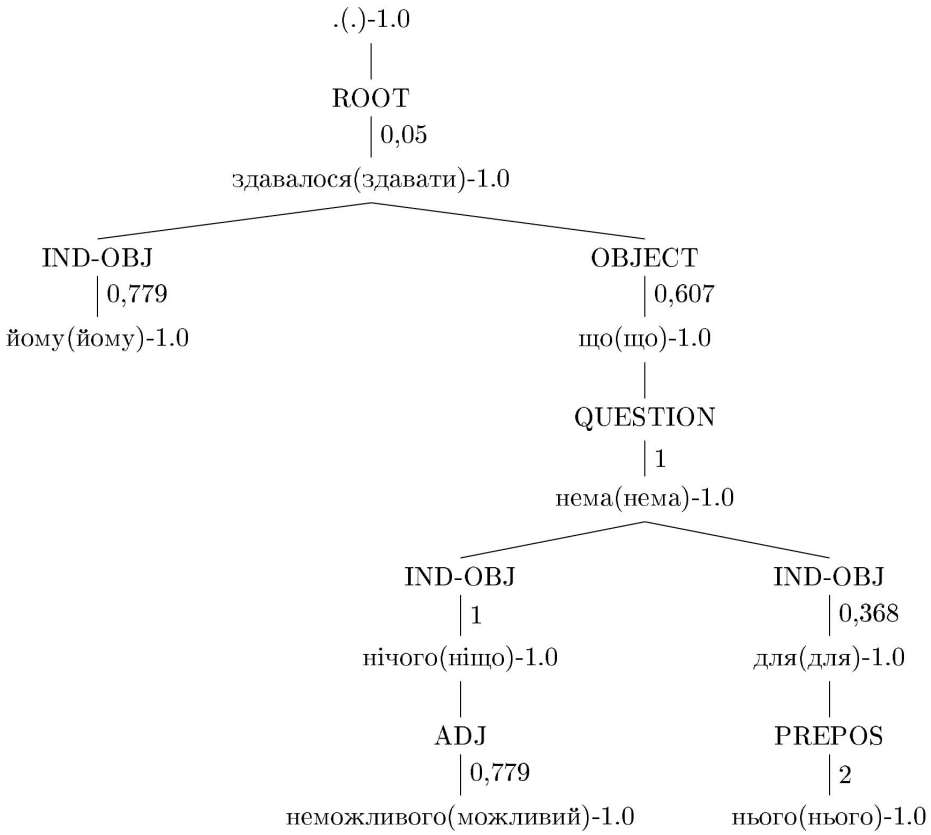

Рис. 7 Приклад синтаксичного розбору розбору речення "Йому здавалося, що нема нічого неможливого для нього"

Основними, причинами, які завадили правильному розбору інших речень були:

1) багатозначність слова “його”, яке часто використовувалося у творі;
2) складність речень твору, використання часток, вигуків;
3) відсутність тегу дієприкметника в словнику spell-uk;
4) відсутність правил керування для дієслів у словнику;
5) недосконалість системи встановлення ваг;
6) мала вага зв'язку між словами, які розділені великим вставним зворотом.

Час синтаксичного розбору тексту об'ємом 196 речень складає близько двох секунд на комп'ютері з процесором Intel i5 частотою 3,1 ГГц без розпаралелення операцій.

6. ВИСНОВКИ

У ході проведених робіт було розроблено метод синтаксичного розбору речень, який можна використовувати для визначення зв'язку між словами тексту, побудови моделей тексту, виправлення помилок. Показано можливість застосування методу разом із ваговою моделлю тематики тексту. Метод базується на використанні відкритого орфографічного словника spell-uk, що дає змогу використовувати розроблені засоби в проектах із відкритими кодами. Розроблені засоби синтаксичного розбору та тестові дані розміщені в системі контролю версії github за адресою https://github.com/mdavydov/UkrParser та доступні для завантаження за ліцензією GPLv3.

Подальші дослідження будуть спрямовані на вдосконалення засобів морфологічного аналізу, розширення словника spell-uk та виправлення недоліків, виявлених у результаті тестування методу.

1. 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.

2. А. Рисін. Сторінка проекту *spell-uk створення словників для перевірки орфографії для української мови / ел. ресурс. – реж. доступу http://code.google.com/p/spell-uk/wiki/aspell_uk. – 2010. – перевірено 24.12.2013.

3. Chomsky N. Syntactic Structures / Mouton & Co. – Feb. 1957. – 117 p.

4. Chomsky N. Aspects of the Theory of Syntax / Cambridge, Massachusetts: MIT Press - 1965. - 251 p.

5. Tesnière L. Éléments de syntaxe structurale / L. Tesnière // Paris: Klincksieck. – 1966. – 670 p.

6. Винцюк Т. К. Анализ, распознавание и интерпретация речевых сигналов. – Киев : Наук. думка,. 1987. – 262 с.

7. Berstel J. Context-Free Languages / Jean Berstel, Luc Boasson // Handbook of Theoretical Computer Science, J. van Leewen (ed.), – Vol. B. – Elsevier. – 1990. – pp. 59–102.

8. Cornelis H.A. Affix Grammars for Natural Languages / H. A. Cornelis, Koster // Attribute Grammars, applications and systems / H. Alblas and B. Melichar (Eds.). – Heidelberg, Springer, 1991 – pp. 469-484.

9. Clarence A. Probabilistic Tree Automata / A. Clarence // Proceedings of the second annual ACM symposium on Theory of computing (STOC '70). – ACM, New York, NY, USA, 1970. – pp. 198-205.

10. Steedman M. Combinatory grammars and parasitic gaps / M. Steedman // Natural Language and Linguistic Theory 5, pp. 403–439. – 1987.

11. C. Pollard. Generalized Phrase Structure Grammars, Head Grammars and Natural Language. PhD thesis, Stanford University, 1984.

12. Abeillé A. Tree Adjoining Grammar: An overview / A. Abeillé , O. Rambow // Eds., Tree Adjoining Grammars: Formalisms, Linguistic Analyses and Processing. – 2000. – p. 1–68.

13. Sleator D. Parsing English with a link grammar / D. Sleator, D. Temperley // Carnegie Mellon University: Technical Report, no CMU-CS-91-196. – 1991. – 14 p.

14. Gasser M. Towards synchronous Extensible Dependency Grammar / Michael Gasser // Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation. – Barcelona, 2011. – 8 p.

15. Word Grammar: New Perspectives on a Theory of Language Structure / Kensei Sugayama et al. // Bloomsbury Academic. – Feb 8, 2006. – 232 p.

16. Sgall P. The Meaning of the Sentence in Its Semantic and Pragmatic Aspects / P. Sgall, E. Hajičová, J. Panevová // Dordrecht: D. Reidel Publishing Company. ISBN 90-277-1838-5. – 1986. – 353 p.

17. Vijay-Shanker K. The equivalence of four extensions of context-free grammars / K. Vijay-Shanker, D.J. Weir // Springer-Verlag: Mathematical systems theory. – vol. 27. – no. 6. – doi 10.1007/BF01191624. – issn 0025-5661. – 1994. – pp. 511-546.

18. Борисова Н.В. Модели и методы синтаксического анализа / Н.В. Борисова, О.В. Канищева // Бионика интеллекта, 2012. – № 2 (79). – С. 89-94.

19. 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. – ом 51,N N 11.-С.47-56

20. Лингвистический энциклопедический словарь / М.: Советская энциклопедия. – гл. ред. В. Н. Ярцева – ISBN: 5-85270-031-2 – 1989. – 688 с.

21. Open Office Spell Dictionary / ел. ресурс реж. доступу http://www.jamochamud.org/docs/org/dts/spell/dictionary/OpenOfficeSpellDictionary.html. – перевірено 08.10.2013.

22. Кудинов М.С. Частичный синтаксический разбор текста на русском языке с помощью условных случайных полей / М.С. Кудинов // Машинное обучение и анализ данных, 2013. – Т.1, №6. – С. 714-724.

23. Ponte, J. M. A Language Modeling Approach to Information Retrieval / J. M. Ponte, W. B. Croft // ACM SIGIR, 275-281, 1998.

24. Nallapati R. Sentence-Forest Language Model: A Graph-theoretic Framework to Capture Local Term Dependencies / R. Nallapati, J. Allan // CIIR Technical Report. – 2003.