ЛІНГВІСТИЧНЕ ЗАБЕЗПЕЧЕННЯ ВІЙСЬК (СИЛ) ТА ПЕРЕКЛАДОЗНАВСТВО

УДК 81'322.2

І.В. Замаруєва, д-р техн. наук, проф.,
В.В. Балабін, канд. філол. наук, проф.

АВТОМАТИЧНИЙ СИНТАКСИЧНИЙ АНАЛІЗ УКРАЇНОМОВНОГО ТЕКСТУ В ЗНАННЯ-ОРІЄНТОВАНІЙ СИСТЕМІ МАШИННОГО ПЕРЕКЛАДУ

У статті пропонуються принципи розробки процедури автоматичного синтаксичного аналізу україномовного речення відповідно до положень знання-орієнтованої системи машинного перекладу. Розглядаються три етапи аналізу цілісного тексту.

Ключові слова: автоматичний синтаксичний аналіз, знання-орієнтована система машинного перекладу, словник синтаксичних правил, словник інтерпретацій, просте речення, складне речення.

The paper deals with the principles of the procedure development of the automatic syntactical analysis of Ukrainian according to the knowledge-based machine translation system. Three analysis steps of integral text are described.

Key words: automatic syntactical analysis, knowledge-based machine translation system, the syntactical rules dictionary, the dictionary of interpretation, simple sentence, complex sentence.

Автоматичний синтаксичний аналіз є невід'ємною складовою будь-якої системи автоматичної обробки текстової інформації. Запропонований дослідниками модульний підхід вирішення цієї задачі значною мірою визначив шляхи рішення проблеми [1-3]. Проте на практиці цей компонент систем автоматичної обробки текстів повністю все ще залишається не реалізованим, а актуальність його розробки дедалі зростає, свідченням чого є дослідження останніх років [4-6]. У даному випадку специфіка лінгвістичних правил та їх алгоритмічного втілення залежить від поставлених перед дослідниками практичних потреб. Зокрема виникає необхідність напрацювання відповідної лінгвістичної бази у разі розробки системи машинного перекладу (СМП).

Мета поданої статті – узагальнити досвід розробки поверхнево-синтаксичного аналізу україномовного речення відповідно до запропонованих принципів знання-орієнтованої системи машинного перекладу [7]. Завдання і можливості підходу розглядаються з урахуванням їх практичної реалізації.

Синтаксичний аналіз в СМП має багатофункціональне призначення, а саме:

- усунення лексико-граматичної омонімії, отриманої на етапі морфологічного аналізу;
- розпізнавання термінів і понять, що є словосполученнями;
- побудова синтаксичної структури речення.

Кінцевою метою синтаксичного аналізу є представлення синтаксичної структури речення, яке є придатним для семантичного аналізу. Загалом, між синтаксичною і семантичною структурою є однозначний зв'язок. Так, синтаксичні відношення не існують без семантичних, які в свою чергу реалізуються в заданій предметній галузі.

Знання-орієнтований підхід до розробки СМП базується на принциповому положенні, що предметом аналізу виступають наявні в текстовій інформації знання з предметної галузі.

Реалізація знання-орієнтованого підходу до розробки систем машинного перекладу передбачає такі етапи: аналіз (розпізнавання і вилучення знань про світ з вхідного тексту), інтерпретацію результатів аналізу (вилучених знань) в термінах вирішуваної прикладної задачі та синтез знань засобами іншої мови. На етапі розпізнавання і вилучення знань про світ (предметну галузь) текст на вході розглядається як об'єкт різних рівнів аналізу: як знакова система, як граматична система і як система знань про світ (предметну галузь). Кожний рівень має свої закономірності і особливості і, отже, припускає наявність специфічних, притаманних тільки заданому рівню, методів обробки.

Синтаксичний аналіз загалом передбачає 3 етапи:

1) контекстно-синтаксичний аналіз;
2) синтаксичний аналіз простого речення;
3) міжфразовий синтаксичний аналіз (складні речення аналізуються як частковий випадок міжфразового синтаксису).

Об'єктом аналізу синтаксичного рівня мовної системи є синтаксичні закономірності взаємодії лексем у межах речення і речень у межах цілісного тексту. Вихідними даними є результати роботи попередніх модулів (доморфемного і морфологічного аналізу) лінгвістичного процесора [8, 9], та апріорно задані словники синтаксичних правил, які визначають ознаки синтаксичного поєднання лексем у словосполучення. Інформаційна база даних на цьому етапі зумовлюється особливостями певної предметної галузі. Вона включає словник лексико-семантичних валентностей дієслів, який обумовлює ознаки найбільш вірогідного оточення, та словник семантичних інтерпретацій, який на основі розпізнаних синтаксичних правил визначає стійкі словосполучення і поняття в заданій предметній галузі.

Синтаксичний аналіз здійснюється за декілька проходжень. Важливим для знання-орієнтованого підходу є збереження змістової цілісності тексту, тому розглянемо детальніше на прикладі цілісного тексту, представленого на рис.1.

Роль державних органів в системі національної безпеки

Державні органи відіграють головну роль у забезпеченні національної безпеки України.

До них належать: Верховна Рада України як орган законодавчого регулювання відносин національної безпеки; Президент України як глава держави, гарант державного суверенітету, територіальної цілісності України, дотримання Конституції України, прав і свобод людини і громадянина та Верховний Головнокомандувач Збройних Сил України; Рада національної безпеки і оборони України як координаційний орган з питань національної безпеки і оборони при Президенті України; Кабінет Міністрів України як вищий орган у системі органів виконавчої влади, що вживає заходів до забезпечення обороноздатності, національної безпеки України та громад.

Важливі функції у забезпеченні національної безпеки виконують також Конституційний Суд України, Прокуратура України, Національний банк України, міністерства і відомства.

Рис. 1. Приклад автентичного перекладу

Як вже зазначалося, на модуль синтаксичного аналізу подається результат попередніх етапів обробки (доморфемного і морфологічного аналізу). Так, текст розбивається на речення, визначається клас кожного речення (заголовок, підзаголовок, службове речення, мовне речення тощо), речення розбивається на синтагми, в межах кожної синтагми визначаються лексико-граматичні характеристики до кожної лексеми, що входять до синтагми [8, 9].

Наприклад, перше речення, що представлене на рис.1, після роботи попередніх етапів буде мати такий вигляд (див. рис. 2).

```
ЗР =>     МС =>
                    Роль[1*211000000/1*214000000/]->
                    державних[2*922000000/2*926000000/]->
                    органів[1*122000000/]->
                    в[23*004000000/23*006000000/]->
                    системі[1*213000000/1*216000000/]->
                    національної[2*212000000/]->
                  безпеки[1*212000000/1*221000000/1*224000000/]
        [КР]
```

Рис. 2. Приклад результату доморфемного і морфологічного аналізів заголовка

З рис. 2 видно, що, незважаючи на відсутність крапки, стоїть маркер кінця речення ([КР]), крім того для цього речення визначений клас ЗР – речення-заголовок. Граматична інформація до кожної лексеми представляється позиційно цифровим кодом, знаком "*" відокремлюється код лексико-граматичного класу від відповідних йому граматичних характеристик, через "/" подається альтернативна інформація.

Як вже говорилося, синтаксичний аналіз здійснюється в декілька етапів. Перший етап відповідає контекстно-синтаксичному аналізу. Задачею цього етапу є визначення іменникових груп в межах однієї синтагми, що можуть позначати терміни (цілісні поняття) в заданій предметній галузі. Визначення синтаксично поєднаних слів здійснюється завдяки застосування словника синтаксичних правил, який містить контекстно-синтаксичні правила узгодження, керування і прилягання. Формат подання таких умов представлений в таблиці 1. В таблиці визначені: тип синтаксичного правила, лексико-граматичні ознаки їх прояву в контексті, головне слово для сполуки, що проявилася та умови переходу на ступне правило. Декларативне подання правил контекстно-синтаксичного аналізу дає можливість звести цей етап синтаксичного аналізу в плані програмування до обробки таблиць.

На цьому етапі аналізу речення прочитується з кінця, тобто від маркера, що позначає кінець речення. Результат першого проходу для останнього речення з рис.1 наведений на рис.3.

Таблиця 1

Форма представлення правил контекстного поєднання лексем у межах синтагми

Який клас	Через який клас	З яким класом	За якими ознаками			Тип СП	Гол. слово	Яку дію вик.
			рід	числ.	відм.			
1*		2*	+	+	+	У	1*	М1
1*	24*	1*			+	У	1*/2	М1
…		…	…	…	…	…	…	…
1*		23*			2	К	23*	М1
1*		1*			2	К	1*/2	М2
…		…	…	…	…	…	…	…
14*		9*				П	9*	М3
…		…	…	…	…	…	…	…

```
МР =>     МС =>
(У) Важливі[2*221000000/2*224000000/] (ГС)функції[1*221000000/1*224000000/]
(К) (ГС)у[23*006000000/] (ГС)забезпеченні[1*316000000/]
(У) національної[2*212000000/] (ГС)безпеки[1*212000000/]
(П) (ГС)виконують[9*024329012/9*026329012/9*022329012/] також[14*000000000/]
(У) Конституційний[69*111000000/69*114000000/]
(К) (ГС)Суд[69*111000000/69*114000000/]
України[62/1*232000000/]
,[L16]->   МС =>
(К) Прокуратура[69*211000000/] України[62*232000000/]
,[L16]->        МС =>
(У) Національний [69*111000000/69*114000000/]
(К) (ГС)банк [1*111000000/1*114000000/] України[62/1*232000000/]
,[L16]->        МС =>
(У) (ГС)міністерства[1*312000000/1*321000000/1*324000000/] і[24*000000000/]
відомства[1*312000000/1*321000000/1*324000000/]          .[КР]
```

Рис. 3. Результат контекстно-синтаксичного аналізу речення

Результат контекстно-синтаксичного аналізу подається на модуль інтерпретації. Задачею даного модуля на етапі контекстно-синтаксичного аналізу є виявлення стійких словосполучень, а також термінів і понять в заданій предметній галузі. Для цього синтаксичні конструкції речення в межах однієї синтагми приводяться до початкової форми (лематизуються) і перевіряються за словником семантичних інтерпретацій. Ідентифіковані за словником синтаксичні конструкції визначаються далі як одна лексема і граматичні характеристики приписуються тільки для головного слова в синтаксичній конструкції. Стійкі словосполучення, до складу яких входять дієслова, перевіряються за словником лексико-семантичних валентностей, але інформація до кожного слова подається окремо. Для словників модуля інтерпретації вхідними даними є лексико-граматична інформація, а вихідними – семантична, яка подається у вихідних даних до знаку "*". Крім того, терміни і поняття можуть формуватися ситуативно, тобто на підставі класів, які входять до синтаксичної конструкції. Наприклад, для словосполучень: "*Конституційний Суд", "Національний банк",* якщо у словнику інтерпретацій і буде відсутній відповідник, вони все одно сприймаються системою як одне поняття, оскільки до лексико-граматичної інформації цих синтаксичних конструкцій входить лексико-граматичний клас "69*", який призначається ще на етапі доморфемного аналізу тексту.

У разі якщо для синтаксичної конструкції відсутній відповідник у словнику семантичних інтерпретацій, то кожне слово, що входить до відповідної синтаксичної конструкції, приводиться до початкової форми, але зберігається лексико-граматична інформація. Це стосується і понять ситуативного типу. Результат обробки модулем інтерпретації речення з рисунка 3 представлений на рис. 4.

```
MP =>     MC =>
(У) Важливий[2*221000000/2*224000000/] (ГС)функція[1*224000000/1*221000000/]
(К) (ГС)у[23*006000000/] (ГС)забезпечення[1*316000000/]
національна безпека[5/1*212000000/]
(П) (ГС)виконувати[1/9*024329012/9*026329012/9*022329012/] також[14*000000000/]
Конституційний[69/2*111000000/69*114000000/] Суд[69/1*111000000/69*114000000/]
 Україна [62/1*232000000/]
,[L16]->   MC =>
(К) Прокуратура[69/1*211000000/] Україна[62/1*232000000/]
,[L16]->        MC =>
(У) (ГС)Національний[69/2*111000000/1*114000000/]
(К) банк[1*111000000/1*114000000/] Україна[62/1*232000000/]
,[L16]->      MC =>
(У) (ГС)міністерство[1*312000000/1*321000000/1*324000000/] і[24*000000000/]
відомство[1*312000000/1*321000000/1*324000000/]          .[KP]
```

Рис.4. Результат контекстно-синтаксичного аналізу речення

На другому етапі синтаксичного аналізу будується синтаксична структура простого речення. Його призначенням є визначення підмета, присудка і другорядних членів речення, визначених в категоріях синтаксису. Якщо текст складається тільки з одного простого речення, то алгоритм синтаксичного аналізу завершує свою роботу і передає результати обробки алгоритму семантичного аналізу тексту. Результат роботи другого етапу синтаксичного аналізу для останнього речення з наведеного прикладу (рис.1) представлений на рис. 5.

На цьому етапі роботи алгоритму обробка речення починається з пошуку головних членів речення, причому перевага надається присудку. На рисунку 5 використані позначення: ***(DrO)*** – прямий додаток; ***(Atr)*** – означення; ***(VPrO)*** – прийменниковий додаток дієслова; ***(NPrO)*** – прийменниковий додаток іменника; ***(G)*** – генетивна група іменника; ***(Pr)*** – присудок; ***(AdM)*** – обставина способу дії; ***(S)*** – підмет; (∧) – логічна операція "і". Оскільки інформація про синтаксичні зв'язки також може бути багатозначною, то вона подається через знак "/". Синтаксична багатозначність розв'язується на етапі семантичного аналізу.

```
MP => MC =>
=> (DrO) (Atr) Важливий[2*] (ГС)функція[1*929/]
=> (VPrO/NPrO) у[23*] забезпечення[1*919/] (G) національна безпека[5/1*919/]
=> (Pr) (AdM) виконувати[1/9*029329012/] також[14*]
=> (S) (Atr)Конституційний[69/2*] Суд[69/1*911/] (G) Україна [62/1*239/](∧), [L16]
=> (S) Прокуратура[69/1*911/] (G)Україна[62/1*239/](∧), [L16]
=> (S) (Atr) Національний[69/2*] банк[1*911/] (G) Україна[62/1*239/](∧),[L16]
=> (S) міністерство[1*921/] і[24*]          відомство[1*921/]   .[KP]
```

Рис. 5. Результат другого етапу синтаксичного аналізу речення

Призначення третього етапу синтаксичного аналізу – встановити синтаксичні зв'язки між реченнями в тексті. При цьому складні речення розглядаються як частковий випадок міжфразового синтаксису. Результат третього етапу є кінцевим для синтаксичного аналізу і є вихідними даними для семантичного аналізу тексту. Кінцевий результат синтаксичного аналізу прикладу-тексту (див. рис.1) представлений на рис. 6. З рис. 6 видно, що в третьому реченні, яке є складнопідрядним, для підрядної частини визначений синтаксичний тип зв'язку ***(SnAtr/SnDrO)***, де ***SnAtr –*** означальне речення, ***SnPrO*** – з'ясувальне речення, через знак "/" подається альтернативна інформація, яка усувається на етапі семантичного аналізу. Крім того, дане речення через анафоричну зв'язку "*них*" синтаксично залежне від попереднього речення і також має альтернативну інформацію. Позначення ***SnPrO*** означає, що це з'ясувальне речення з прийменниковою конструкцією.

ЗР => МС => ***(S)*** Роль[1*911/] ***(G) (Atr)*** державний[2*] орган[1*922/]
=> ***(NPrO)*** в[23*] система[1*919] ***(G) (Atr)*** **національна безпека** [1*919/] [KP]

МР => МС => ***(S) (Atr)*** Державний[2*] орган[1*921/] => ***(Pr)*** відігравати[9*029329012/]
=> ***(DrO) (Atr)*** головний[2*] роль[1*919/] => ***(VPrO/NPrO)*** у[23*]забезпечення[1*919/] ***(G)*** **національна безпека**[1*919/] ***(G)*** **Україна** [62*239/] .[KP]

(SnPrO/SnAtr) МР => МС => ***(VPrO)*** До[23*] {*Державний[2*]орган[1*921/]*}він[5*9293/]
=> ***(Pr)*** належати[9*029329012/] :[L16]
=> ***(GrS)(G)*** **Верховна Рада**[66*331/] **Україна**[62/1*239/] ***(Atr)*** як[24*] (ГС) орган[1*919/] ***(G) (Atr)***законодавчий[2*] регулювання[1*919/] ***(G)*** відносини[1*929/] ***(G)*** **національна безпека**[5/1*919/] (∧);[L16]
=> ***(GrS) (G)*****Президент**[69/1*911/] ***(G)*** **Україна**[62/1*239/] ***(Atr){*** як[24*] ***(G)*** глава[1*919/] держава[1*919/] (∧),[L16] ***(G)*** гарант[1*919/] ***{(Atr)*** державний[2*] суверенітет[1*919/] (∧),[L16] ***(Atr)***територіальний[2*] ***(G)***цілісність[1*931/] **Україна**[62/1*239/] (∧) ,[L16] ***(G)*** дотримання [1*919/] ***(G)*****Конституція**[69/1*919/] **Україна**[62/1*239/] (∧),[L16] права [1*929/] і[24*] ***(G)***свобода[1*929] людина[1*919/] і[24*] громадянин[1*919/]}та[24*] ***(Atr)*** **Верховний** [69/2*] **Головнокомандувач**[69/1*919] ***(G)*** **Збройні сили України**[65*249/]}} (∧);[L16]
=> ***(GrS)*****Рада**[1*211/] **національної безпеки і оборони України** ***(AdM/Atr)*** як[24*] ***(Atr)*** координаційний[2*] орган[1*919/] ***(NPrO)*** з[23*] питання[1*929/] ***(G) (Atr)*** **національна безпека**[5/1*919/] і[24*] оборона [1*919/] ***(NPrO)*** при[23*] **Президент**[69/1*116000000/] ***(G)*** **Україна**[62/1*239/] (∧);[L16]
=> ***(GrS)*** **Кабінет Міністрів України**[65*931/] ***(Atr)*** як[24*] ***(Atr)*** високий[15*9990001] орган[1*919/] ***(NPrO)*** у[23*] система[1*919] ***(G)*** орган[1*929/] ***(Atr)*** виконавча[2*] влада[1*919] ,[L16]
=> ***(SnAtr/SnDrO)*** МС => ***(S) Atr*** {*Кабінет Міністрів України*[65*931/]} що[20*]
=> ***(Pr)*** вживати[9*019329012/] ***(DrO)*** захід[1*929/] ***(NPrO)*** до[23*] забезпечення[1*919] ***(G)*** обороноздатність[1*939/] (∧) ,[L16] ***(NPrO)*** **національна безпека** [1*919/] ***(G)*** **Україна**[62/1*239/] та[24*] громада[1*929/] .[KP]

МР => МС => ***(DrO) (Atr)*** Важливий[2*] (ГС)функція[1*929/] ***(VPrO/NPrO)*** у[23*] забезпечення[1*919/] ***(G)*** **національна безпека**[5/1*919/]
=> ***(Pr) (AdM)*** виконувати[1/9*029329012/] також[14*]
=> ***(S) (Atr)*****Конституційний**[69/2*] **Суд**[69/1*911/] ***(G)*** **Україна** [62/1*239/](∧), [L16]
=> ***(S)*** **Прокуратура**[69/1*911/] ***(G)*****Україна**[62/1*239/](∧), [L16]
=> ***(S) (Atr)*** **Національний**[69/2*] **банк**[1*911/] ***(G)*** **Україна**[62/1*239/](∧),[L16]
=> ***(S)*** міністерство[1*921/] і[24*] відомство[1*921/] .[KP]

Рис. 6. Результат третього етапу синтаксичного аналізу

Висновки. Таким чином, синтаксичний аналіз є невід'ємною складовою аналізу тексту як лінгвістичної системи і спрямований на розпізнавання, вилучення і формалізацію знань про фрагменти навколишнього світу (предметну галузь), що містяться в тексті. Такий підхід дає можливість розв'язувати різноманітні задачі штучного інтелекту в тому числі й автоматичний переклад.

Запропонований трьохетапний синтаксичний аналіз дозволяє не тільки будувати синтаксичну структуру речень, але й зберігати (через міжфразовий синтаксис) цілісність тексту, що є дуже важливим для адекватності перекладу. Занурення проміжних результатів (після кожного етапу) синтаксичного аналізу у предметну галузь дозволяє сформулювати вимоги до розподіленої структури і обсягу бази знань з предметної галузі. Переважно декларативне подання (у вигляді таблиць) на кожному етапі синтаксичних правил дозволяє реалізувати відомий принцип програмування: відокремлення даних від алгоритму їх обробки, що робить його відкритим як щодо нових мов, так і щодо "нових" прикладних задач з обробки текстової інформації.

1. Апресян Ю.Д., Богуславский И.М. и др. Лингвистическое обеспечение в системе автоматического перевода третьего поколения. – М., 1978. – 74 с. 2. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения. – М.: Наука, 1985. – 140 с. 3. Синтаксический анализ научного текста на ЭВМ. – К.: Наукова думка, 1999. – 272 с. 4. Баталина А. М., Епифанов М. Е., Кобзарева Т. Ю., Кушнарёва Е. В., Лахути Д. Г. Опыт экспериментальной реализации алгоритмов поверхностно-синтаксического анализа // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. "Диалог'2006" // www.dialog-21.ru/Archive/2006. 5. Кобзарева Т. Ю. Иерархия задач поверхностно-синтаксического анализа русского предложения // НТИ. – Сер. 2., № 1, 2007. – С. 23-35. 6. Кулагина О.С. О современном состоянии машинного перевода / Математические вопросы кибернетики. – М.: Наука, 1991.– Вып.3. – С. 5-51. 7. Замарує́ва І.В. Комп'ютерна модель розуміння природно-мовної текстової інформації // Проблемы программирования. – 1999. – №2. С.96–102. 8. Замарує́ва І.В., Балабін В.В. Доморфемна обробка текстів в системах машинного перекладу// Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К., 2008. – № 11. – С.78–84. 9. Замарує́ва І.В., Шипнівська О.О. Морфемна обробка текстів в системах машинного перекладу // Вісник Київського національного університету імені Тараса Шевченка. Військово-спеціальні науки. – К., 2008. – №20. – С.61–63.

Надійшла до редколегії 23.05.12

УДК:81'322.4

Л.О. Литвиненко, здобувач

ОСОБЛИВОСТІ ПОБУДОВИ ЛІНГВІСТИЧНОГО ПРОЦЕСОРА ДОМОРФЕМНОГО АНАЛІЗУ АНГЛІЙСЬКИХ ВІЙСЬКОВО-ТЕХНІЧНИХ ТЕКСТІВ

У статті розглянуто особливості побудови лінгвістичного процесора доморфемного аналізу англійських військово-технічних текстів. Проведено аналіз завдань та проблем графемного та лексемного аналізів текстів. Представлено реалізаційні аспекти доморфемного аналізу англійських військово-технічних текстів.

Ключові слова: лінгвістичний процесор, доморфемний аналіз, автоматична обробка природно-мовного тексту.

The features of construction of linguistic processor of preliminary morphological analysis of English military-technical texts are considered in the article. The analysis of tasks and problems of graphic and lexical analyses of texts is conducted. The realization aspects of preliminary morphological analysis of English military-technical texts are presented.

Keywords: linguistic processor, preliminary morphological analysis, automatic natural language text processing.

Вступна частина. Лінгвістична обробка природномовних текстів є однією з центральних проблем інтелектуалізації інформаційних технологій. Цій проблемі приділяється значна увага в розвинутих країнах Європи та США, свідченням чого є виділення величезних коштів на розробку лінгвістичного програмного забезпечення [1,2]. Велика кількість науково-дослідних програм спрямовані на розвиток лінгвістичних інформаційних систем. На сучасному етапі одним із перспективних напрямків вдосконалення інтелектуальних ін-