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-

формаційних систем і технологій обробки текстів є побудова знання-орієнтованих систем, функціонування яких ґрунтується на автоматизації процесу формалізації змісту природно-мовних текстів [3]. Наступною ланкою цього процесу є обробка формалізованого відображення змісту логіко-семантичними методами з метою рішення задач користувачів, орієнтованих та інтелектуальних аналізів.

Таким чином, особливістю лінгвістичної обробки є підпорядкування всіх її етапів формуванню елементів формалізованого представлення знань. Традиційно, задача розуміння природно-мовних текстів поділяється на три етапи : аналіз, інтерпретацію і синтез. В роботі розглядається тільки перший етап – етап аналізу і частково – етап інтерпретації. На етапі аналізу виокремлюється опис сутностей, відбитих у вхідному тексті, виявляються властивості цих сутностей і відношення між ними, які представляються у вигляді формальних моделей.

Існує декілька алгоритмів аналізу природно-мовних текстів. Класичними є підходи на основі продукційних та декларативних моделей. У основі першого підходу лежить поняття морфа та його застосування при формуванні відповідних лінгвістичних характеристик моделювання, при іншому підході для формування формалізованого опису використовується повна множина лексем і їх словоформ певної природної мови.

Аналіз військово-технічних текстів тісно пов'язаний з аналізами звичайних текстів, але і має свої нюанси, які пов'язані з особливостями графічного оформлення тексту.

Метою даної статті є дослідження особливостей графічного оформлення текстів з військово-технічної тематики та їх формалізованого представлення в інтересах подальшої автоматичної обробки.

Викладення основних результатів дослідження. На основі вище описаних підходів далі буде розглянуто особливості доморфемного аналізу. До доморфемного аналізу текстів відносять графемний те лексемний етапи аналізу. Результат роботи лінгвістичного процесору доморфемного аналізу представлено на рисунку 1.

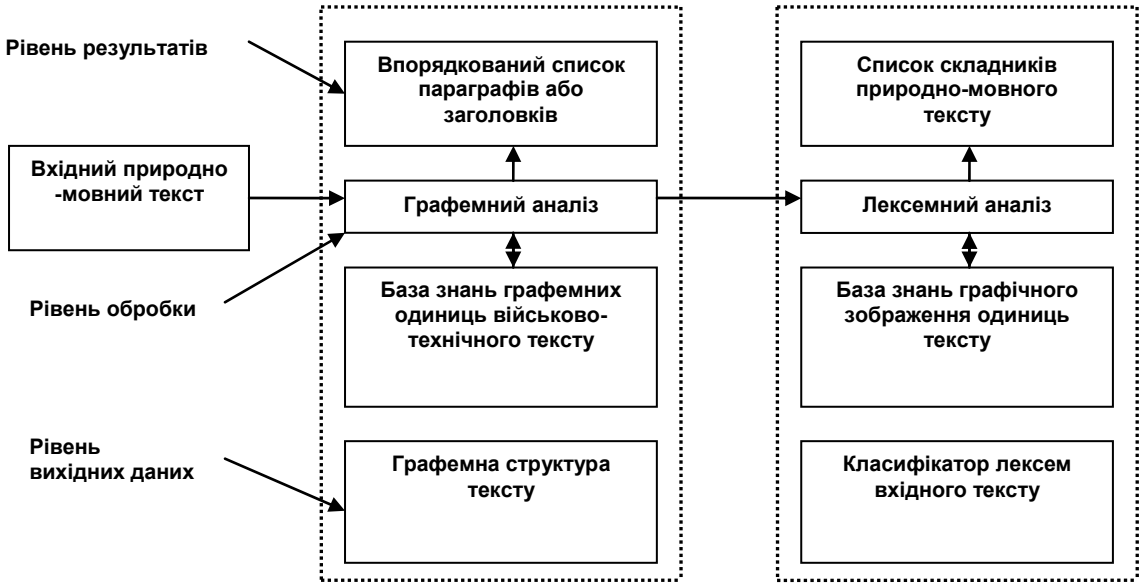

Рис.1. Структурно-логічна схема доморфемного аналізу тексту

В якості початкового етапу лінгвістичного аналізу використовується графемний аналіз. При цьому текст відповідно до знання-орієнтованого підходу розглядається як різновидність знакової системи [3]. Графемний аналіз – це знаковий рівень аналізу, основною задачею якого є визначення таких ознак тексту як заголовки та параграфи. В результаті дослідження ознак для графемного аналізу було виявлено що для визначення заголовків в тексті недостатньо використовувати лише довжину речення та формат відступу від інших речень. Тому до аналізу були додані й інші ознаки, зокрема розмір шрифту, позиціонування (по центру чи ні), також було враховано що заголовки не обов'язково закінчуються крапкою. В результаті графемного аналізу текст перетворюється у впорядкований список параграфів, або заголовків. Такий список у спрощеному вигляді представлено нижче:

ТЕКСТ = ЗАГОЛОВОК | {ПАРАГРАФ_1.1, …, ПАРАГРАФ_1.N};

Результат роботи графемного аналізу подають на вхід лексемному. Лексемний аналіз – це рівень аналізу що дозволяє виділяти в тексті слова, словарні групи, синтагми та речення. Слова, словарні групи та синтагми об'єднуються одною назвою – лексеми. Призначення цього етапу – побудова моделі лексемної структури, в якій виділені і зв'язані відношеннями (де це можливо) такі змістовні одиниці тексту, як фрагмент, речення та лексема. Для лексем виділяють класи лексем, які відрізняються своєю структурою і виконують різні функції в тексті. Крім того, аналізуються закономірності сполучуваності деяких лексичний одиниць, які уже на цьому етапі дозволяють об'єднувати декілька лексем в одну на тій основі, що вони виконують в тексті єдину функцію. В результаті отримуємо впорядкований список складників військово – технічного тексту :

ПАРАГРАФ_1.1 = {РЕЧЕННЯ_1.1, …, РЕЧЕННЯ_1.M};

…….

ПАРАГРАФ_1.N = {РЕЧЕННЯ_1,…, РЕЧЕННЯ_1.P};

РЕЧЕННЯ_1.N = {(СЛІСФ)_1,…,(СЛІСФ)_i};

……

РЕЧЕННЯ_N.P = {(СЛІСФ)_1,…,(СЛІСФ)_j}, де i = 1,I, j = 1,J – кількість спеціальних лексем чи словоформ відповідно у реченнях 1.1…N.P

Для прикладу можна зобразити доморфемний аналіз наступного військово-технічного тексту:

U.S. Navy

Since World War II the U.S. Navy has designated fleets in the Atlantic Ocean and Mediterranean Sea with even numbers, and those in the Pacific Ocean with odd numbers. The ***operating forces of the Navy*** *are divided into five numbered* ***fleets*** *(the Second Fleet, at Norfolk, Virginia; the Third Fleet at San Diego, California; the Fifth Fleet, at Manama, Bahrain; the Sixth Fleet, at Gaeta, Italy; and the Seventh Fleet at Yokusuka, Japan).*

The Navy maintains ***bases****. They ensure that American forces are located in important regions continuously. The Navy also maintains a* ***forward presence*** *(a deployment of substantial military force) outside of the United States to demonstrate its commitment to a region or to allied countries. This forward presence also enables the United States to act quickly in response to threats against American interests.*

Результатом графемного аналізу є список заголовків, або параграфів, а тому схематично вихідні дані можуть виглядати наступним чином:

U.S. Navy => ***HEADER***

Since World War II the U.S. Navy has designated fleets in the Atlantic Ocean and Mediterranean Sea with even numbers, and those in the Pacific Ocean with odd numbers. The operating forces of the Navy are divided into five numbered fleets (the Second Fleet, at Norfolk, Virginia; the Third Fleet at San Diego, California; the Fifth Fleet, at Manama, Bahrain; the Sixth Fleet, at Gaeta, Italy; and the Seventh Fleet at Yokusuka, Japan). ***= > PARAGRAPH***

The Navy maintains bases. They ensure that American forces are located in important regions continuously. The Navy also maintains a forward presence (a deployment of substantial military force) outside of the United States to demonstrate its commitment to a region or to allied countries. This forward presence also enables the United States to act quickly in response to threats against American interests. ***= > PARAGRAPH***

Вихідні дані графемного аналізу подаються на вхід лексемному. Задачею лексемного аналізу є класифікація лексем відповідно до їх графемного оформлення. Результат визначення типів лексем першого речення представлений нижче.

РЧ
СН
Since [лексема_типу1]
World [лексема_типу2] *War* [лексема_типу2] *II* [лексема_типу4]
the [лексема_типу1]
U.S. [лексема_типу3] *Navy* [лексема_типу1]
has [лексема_типу1]
designated [лексема_типу1]
fleets [лексема_типу1]
in [лексема_типу1]
the[лексема_типу1]
Atlantic [лексема_типу2] *Ocean* [лексема_типу2]
and [лексема_типу1]
Mediterranean [лексема_типу2] *Sea*[лексема_типу2]
with [лексема_типу1]
even [лексема_типу1]
numbers [лексема_типу1]
[лексема_типу5]

СН
and [лексема_типу1]
those [лексема_типу1]
in [лексема_типу1]
the [лексема_типу1]
Pacific [лексема_типу2] *Ocean* [лексема_типу2]
with [лексема_типу1]
odd [лексема_типу1]
numbers [лексема_типу1]

В даному випадку прийняті такі позначення: **РЧ** – речення, **СН** – синтагма. Розпарсення лексем відбувається за регулярними виразами. Після розпарсення кожна лексема відноситься до певного типу лексем. Наприклад, для **лексеми_типу1** регулярний вираз може виглядати наступним чином:

([a-z"]+) – тобто, слово, яке починається з маленької букви, та містить символи від **a** до **z**, та символи **"'"**. З наведеного прикладу видно, що це слова: *the , numbers, and, even* і т.д.

Для **лексеми_типу2** регулярний вираз буде мати такий вигляд:

([A-Z][a-z"]*) – слово, яке починається з великої букви та містить тіж самі символи, що і для лексеми типу 1. Наприклад, для наведеного вище тексту це будуть такі лексеми, як: *Atlantic*, *Ocean*, *Mediterranean* тощо.

Також повинні існувати регулярні вирази для розпарсення розділових та інших знаків: **(\\.{3}|[!?.:,;-]) , ([()]), ([[]]) , ([№@#$%&*])** і т.ін.

Завершальним етапом доморфемного аналізу є побудова графемної структури тексту, яка включає класифікацію лексем, що починаються з великої букви та визначення семантичних меж речення [4]. Для визначення типу лексеми, що починається з великої букви, передбачено словник власних назв. Якщо лексему з великої букви не знайдено у словнику власних назв, то її присвоюється значення: [лексема_типу1].

Отримана таким чином графемна структура тексту є вхідною для автоматичного морфологічного аналізу тексту.

Висновки. Провівши аналіз військово-технічних текстів об'ємом понад 100 сторінок виявлено велику кількість лексемних одиниць, що рідко зустрічаються в звичайних текстах. До лексем характерних для більшості текстів можна віднести: звичайне слово, слово з великої літери, слово через дефіс, слово на іншій мові. До специфічних лексем можна віднести такі: слово через дефіс де перша друга або обидві частини складаються з великих літер, цифрові комбінації, різні формати дат, скорочення що складаються лише з приголосних, спеціальні скорочення, характерні лише для текстів військової і військово-технічної тематики (наприклад: ртбр – радіотехнічна бригада), і потребують інших програмних засобів їх інтерпретації.

1. Петренко М.Г. Особливості розробки знання–орієнтованого лінгвістичного процесора. – Комп'ютерні засоби, мережі та системи. – 2006. – №5. 2. Апресян Ю.Д. Лингвистический процесор для сложных информационных систем. – М.: МГУ, 2005. – 287 с. 3. Балабін В.В., Замарує-ва І.В. Побудова систем машинного перекладу на основі знання-орієнтованого підходу // К.: ВІ КНУ. – Збірник наукових праць ВІ КНУ. – 2006. – №2. – С.68-74. 4. Замарує-ва І.В., Балабін В.В. Доморфемна обробка текстів в системах машинного перекладу // К.: ВІ КНУ. – Збірник наукових праць ВІ КНУ. – 2008. – №11. – С.78-85.

Надійшла до редколегії 18.05.12