УДК 802.0 - 561.32

І.Д. Карамишева[1], О.І. Заневський[2], С. Ю. Юриш[2]
Національний університет "Львівська політехніка"
[1]кафедра "Прикладна лінгвістика",
[2] кафедра "Інформаційні системи та мережі"

ІНТЕЛЕКТУАЛЬНА ІНФОРМАЦІЙНА СИСТЕМА РОЗПІЗНАВАННЯ СТРУКТУР ВТОРИННОЇ ПРЕДИКАЦІЇ В АНГЛІЙСЬКІЙ МОВІ

The following article presents the developed software for the automatic recognition of structures of secondary predication that are complex syntactic structures in modern English. The absence of equivalents of these structures in the Ukrainian language makes their perception difficult and therefore their translation as well. That is why the development of the program for their automatic recognition in the text is an actual problem, solved at the intersection of linguistics and informational technologies. The result of this development is the software written on the Visual Basic for Application programming language and integrated in MS Word. The article contains examples of the formalized representation of the mentioned structures and the main features of the algorithm application.

Представлено розроблену програму для автоматичного розпізнавання структур вторинної предикації, що є складними синтаксичними структурами сучасної англійської мови. Відсутність еквівалентних структур в українській мові утруднює їхнє сприйняття, а відтак і переклад, тому створення програми їх автоматичного розпізнавання у тексті є досить актуальною проблемою, що вирішується на перетині лінгвістики та інформаційних технологій. Результатом даної розробки є програма, написана на мові Visual Basic for Application та інтегрована у середовище MS Word. Стаття містить зразки формалізованого представлення вказаних структур та основні особливості реалізації алгоритму.

1 ВСТУП

У сучасному мовознавстві усе чіткіше можна прослідкувати тенденцію щодо використання методів точних наук для розв'язку різних лінгвістичних задач. Це свідчить про нові стандарти підходу до матеріалу дослідження. Одним із чинників впливу, що формує дану тенденцію, є розвиток сучасних інформаційних технологій, що призводить до чіткої структуризації усіх сфер знання та представлення понять у відповідних сферах. Перш за все, серед напрямків лінгвістичних досліджень у сфері "штучного інтелекту" можна

виділити *моделювання лінгвістичних об'єктів, автоматичну обробку текстів, машинний переклад, створення автоматизованого робочого місця лінгвіста/ перекладача з опорою на* ***взаємний діалог людини та машини***. Поєднання обробки лінгвістичних даних з використанням сучасних найрізноманітніших інформаційних та комп'ютерних технологій утворюють важливу та перспективну сферу досліджень у відносно молодій галузі лінгвістики – прикладній лінгвістиці. Незважаючи на те, що сфера прикладної лінгвістики включає такі вагомі напрямки дослідження як, наприклад, методику навчання іноземним мовам, питання теоретичного та практичного перекладу, лексикографію і т.д., перш за все, поєднання слів "прикладний" + "лінгвістика" асоціюється з автоматизованими/ автоматичними системами обробки мовної інформації.

Одними з найпростіших чинників, що свідчать про користь таких автоматизованих/ автоматичних систем, є економія часу та позбавлення від рутинної роботи дослідника-фахівця, що працює з мовними даними. Тут, скажімо, йдеться про певний підрахунок даних, чи визначення частоти компонентів, що входять до складу тих чи інших синтаксичних структур тексту. Слід зауважити, що при створенні автоматизованих систем для обробки мовної інформації, невід'ємними є такі поняття як "формалізація лінгвістичного об'єкта" та "модель", як результат виконання даної дії. Автоматизація розпізнавання структур вторинної предикації, результати розроблення програми якої і є предметом даного викладу, теж, беззаперечно, пов`язана з формалізацією та побудовою моделі певного лінгвістичного об'єкту [1-12].

2 ЛІНГВІСТИЧНЕ ЗАБЕЗПЕЧЕННЯ

Структури вторинної предикації (СВП), як об'єкт даного дослідження, є складним утворенням англійської мови. Якщо брати до уваги традиційні назви даних структур, у сферу дослідження потрапляють Complex Object (with the Infinitive, Participle I, Participle II, Gerund, non-Verbal), Complex Subject (with the Infinitive, Participle I, non-Verbal), For-to-Infinitive Construction, Absolute Constructions (with Participle I, Participle II, Infinitive, non-Verbal). За своїм структурним оформленням усі наведені структури є досить різнорідні, проте характеризуються такою спільною рисою як порушення узгодження суб'єктно-предикатних відносин у їх поверховій структурі. СВП – це структури, що складаються з вторинного підмета та присудка; вторинний присудок у СВП може виражатися як неособовою формою дієслова (тобто інфінітивом, дієприкметником *I* та *II*, герундієм) так і невербальною частиною мови [6].

Складною є розробка формалізованого представлення СВП. Дані структури є досить своєрідними та не знаходять прямого аналогу в українській мові. Вони становлять труднощі як для граматичного аналізу, при вивченні англійської мови як іноземної, так і у сфері перекладу, а, отже, є цікавими з точки зору моделювання для оптимізації їх виявлення та аналізу автоматизованими системами обробки мовної інформації.

При моделюванні СВП враховувались такі моменти: 1) компоненти, які складають ту чи іншу СВП; 2) форма цих компонентів (особливо це стосується форми вираження вторинного присудка); 3) розташування компонентів у межах самої СВП; 4) дистрибутивні характеристики, тобто позиція СВП стосовно інших компонентів речення; 5) графічне вираження СВП, тобто чи вводиться дана структура, скажімо, певним прийменником, чи виділяється вона комою і т. ін.

Будь-яка структура вторинної предикації (СВП) функціонує лише в межах речення, тобто при наявності структури первинної предикації (СПП). Найзагальніша модель речення з СВП матиме такий вигляд:

$S^1 + P^1 + S^2 + P^2$ (+ complements),

де:

S^1 – підмет (первинний підмет) речення, або суб'єкт;

P^1 – присудок (первинний присудок) речення, або предикат;

СВП, в свою чергу складається з **$S^2 + P^2$ (+ complements),** де

S^2 – вторинний підмет, який може виражатися **займенником** (найчастіше особовим займенником в об'єктивному відмінку), **іменником** (загальним іменником чи значно рідше власною назвою), **іменниковою групою**;

P^2 – вторинний присудок, який виражається найчастіше неособовими формами дієслова: **інфінітивом** (з маркером "*to*" та без нього), **дієприкметником I** (інфінітив + *ing*), **герундієм** (інфінітив + *ing*), **дієприкметником II** (інфінітив + *ed* для правильних дієслів, для неправильних дієслів – власна форма) або **невербальною частиною мови** (e.g., іменником, прикметником);

complement(s) – досить часто після вторинного присудка може стояти комплемент або додаток до цього дієслова, який вимагається семантикою даного дієслова і допомагає розкрити його зміст.

Наведемо приклади розроблених моделей для певного типу СВП, а саме: **Complex Object (Об'єктний комплекс).**

СВП, яка у традиційній граматиці має назву Complex Object, у реченні вводиться певним дієсловом (первинним присудком), що є характерною ознакою при виявленні даної структури. Ми будемо називати дане дієслово **ключовим**.

Для розпізнавання Complex Object важливими є такі характеристики:

- ключове дієслово, яке вводить дану структуру, а саме список даних дієслів та їх граматичні форми;
- список займенників (pronouns), якими виражено вторинний підмет;
- якщо вторинний підмет виражено іменником або іменниковою групою (nouns, noun groups), слід визначити скільки позицій буде займати дана іменникова група перед вторинним присудком;
- форми вираження вторинного присудка (Infinitive with and without "*to*", Participle I, Participle II, Gerund, non-Verbal).

2.1 COMPLEX OBJECT WITH THE INFINITIVE (ОБ'ЄКТНИЙ КОМПЛЕКС З ІНФІНІТИВОМ)

Перелік усіх типів моделей даного підтипу СВП:

1 модель: **S^2** – виражено займенником, **P^2** – виражено інфінітивом з маркером "*to*"
<key verb> <pronoun> <to> <infinitive> (<complement>)

e.g., Now, I want **you to take us up the Wando River**. -
Тепер я хочу, щоб ти взяв нас уверх по річці Вандо.

2 модель: **S^2** – виражено займенником, **P^2** – виражено інфінітивом без маркера

<key verb> <pronoun> <infinitive> (<complement>)

e.g., I never knew **her go out with anyone**. -
Я ніколи не знала, що вона зустрічається з кимось.

3 модель: S^2 – виражено іменником, або іменниковою групою, P^2 – виражено інфінітивом з маркером "*to*"
<key verb> <noun> <to> <infinitive> (<complement>)

…
<position 1> (<position 2>) (<position 3>)

e.g., … and he persuaded **the family to sue quickly**. -
… і він переконав, щоб сім'я швидко подала позов.

4 модель: S^2 – виражено іменником, або іменниковою групою, P^2 – виражено інфінітивом без маркера
<key verb> <noun> <infinitive> (<complement>)

…
<position 1> (<position 2>) (<position 3>)

She could feel **the blood pour from her cheeks**.
Вона відчувала, як (що) кров текла по її щоках.

Коментар. Підрахунок частоти компонентів, що є суттєвими для побудови моделі того чи іншого типу СВП, дає наступні результати:

- **вторинний підмет** у СВП підтипу Complex Object with the Infinitive (1751 прикладів) виражається:

- займенниками – у 1054 прикладах даного підтипу СВП (60,19 %),

з них 7 основних займенників (маємо на увазі pronoun in the objective case: *me, you, him, her, them, us, its*) зустрічаються у 1000 випадках (57,1 %)

інших 18 займенників – у 54 випадках (3,08 %);

- іменниками/ іменною групою – у 553 випадках (31,58 %). Найчастіше іменна група займає 3 позиції перед вторинним присудком (у 452 випадках з 553);

- власні назви – у 144 випадках (8, 22 %).

Можливо зробити висновок, що для даного типу СВП, а саме Complex Object with the Infinitive, характерним є вираження вторинного підмета S^2 особовими займенниками у об'єктивному відмінку (які у нас йдуть під рубрикою – основні займенники).

Даний тип СВП вводиться у речення певним **ключовим дієсловом**; зокрема, виявлено 48 таких ключових дієслів. Частотними у даній групі є СВП, що вводяться такими ключовими дієсловами (+ "*to*" – позначає, що після даного дієслова інфінітив, у ролі вторинного присудка СВП, вживається з часткою *to*, "-"- позначає, що інфінітив вживається без частки *to*), таблиця 1.

Зауважимо, що частотними ми вважаємо такі СВП, які становитимуть 1 % і більше від загальної кількості вибірки (3760 прикладів). Вказані частотні СВП налічують 1603

одиниць і становлять 91,55 % від загальної кількості даного підтипу СВП (Complex Object with the Infinitive – 1751 прикладів усього).

Таблиця 1.

Частота зустрічання ключових слів СВП.

want + to	295
feel -	47
see -	149
watch -	161
make -	448
expect + to	68
let -	265
hear -	170

3 АЛГОРИТМИ ТА ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ

Основним завданням у побудові комп'ютерної програми для автоматизації розпізнавання структур вторинної предикації є побудова та практичне застосування алгоритму, на базі формалізованого представлення даних лінгвістичних об'єктів.

Алгоритм функціонування даної програми складається з кількох етапів. Розпізнавання починається з того, що ведеться пошук ключових дієслів, які відрізняються для кожного виду структур (але у випадку такого типу СВП як Absolute Construction спочатку шукається кома). Далі у напрямку справа від ключового дієслова через два, три або чотири слова (позиції, що розмежовуються пробілами) йде пошук слова, яке має певні ознаки (тут йдеться про вторинний присудок, що виражається найчастіше неособовими формами дієслова, які маркуються певними формальними ознаками). Для Complex Object with the Infinitive дане слово є інфінітивом, тому актуальним виявилось створення списку найпоширеніших інфінітивів, для Complex Object with Participle I – це є дієприкметник теперішнього часу, що характеризується суфіксом *-ing* і т. д. Зауважимо, що дане слово (тобто вторинний присудок) стоїть в кінцевій позиції формалізованої частини конструкції. Якщо таке слово віднайдено, то залишається визначити чи слова/ слово, що знаходяться між ключовим дієсловом та вторинним присудком, відносяться до тих, що використовуються у нашому типі конструкцій. У випадку, якщо це є одне слово, це може бути займенник, список яких для кожної з конструкцій визначений. Коли наштовхуємося на групу слів, то тут йдеться про іменникову групу, за допомогою якої досить часто виражається вторинний підмет у досліджуваних конструкціях, і працювати з якою набагато складніше, тому що однозначно визначити її у формальному плані досить важко.

Для розпізнавання групи слів використовується велика кількість формальних ознак, на основі об`єднання яких приймається рішення, про те чи дана група є іменниковою, чи це є певний інший зворот. Серед таких формальних ознак для іменникової групи/ іменника найхарактернішими є велика буква, яка говорить про те, що це власна назва. Важливо, коли група починається з деякої множини ввідних слів (наприклад, в англійській мові маркерами, що вказують на присутність іменника є: *the*, *a*, *a little* і т.д.). Для того, щоб комп`ютер відрізняв іменникову групу серед ланцюжка слів, які за формальними ознаками видаються

програмі аналогічними, було введено кілька правил, що допомагають відмежовувати іменник від інших частин мови, які можуть мати схожі формальні ознаки.

Наприклад, суфікс прикметника у найвищому ступені порівняння - *est* та суфікс прикметника у вищому ступені порівняння *-er* збігаються з відповідними суфіксами у деяких іменників. Тому, слід було виявити і задати для програми так звані "винятки", тут мається на увазі список іменників з даним суфіксом, наприклад, *forest, manifest, test* і т.д. Іншою з ознака, яка вказує на присутність іменника є форма вираження множини: у англійській мові в стандартних випадках це закінчення *–s*.

До даного обмеження, знову ж таки, з'являється ряд винятків (тут йдеться про слова, що не є іменниками, але закінчуються на "*s*", тому сприймаються програмою розпізнавання як іменник у множині, наприклад, *was, as, whereas* і т.д.). З емпіричних досліджень, а саме, за підрахунком частоти компонентів досліджуваних структур, стало відомо, що найчастотнішими є іменникові групи, які містять одне, два, або три слова, тому було вирішено розглядати та вводити у програму лише три позиції. Таке рішення має свою перевагу, оскільки чим більше число слів треба розпізнати/ виявити, тим значніше сповільнюється швидкодія розпізнавання.

В процесі роботи алгоритму виявилось декілька ускладнень:

- по-перше, це вищезгадане обмеження на іменникову групу до трьох слів;
- по-друге, проблема зовнішньої схожості іменників і інфінітивів (*sag, start, bag* і т.д.), іменників і дієприкметників теперішнього часу (*moving, dancing, fighting* і т.д.), яка на даному етапі все ще перебуває в стані розв'язку.

Проте процент таких слів незначний. Крім того, трапляються, але досить рідко, випадки, в яких неправильне рішення приймається програмою по причині якогось некоректного форматування тексту.

На основі розробленого алгоритму написано програму на мові VBA, інтегровану у середовище MS Word, яка розпізнає структури вторинної предикації в тексті і виділяє їх, що є досить зручним при візуальному контакті з текстом. Програма, також, надає можливість вибрати при запуску зі всіх типів конструкцій лише певні, потрібні. Діалогове вікно вибору типів конструкції показано на рис.1.

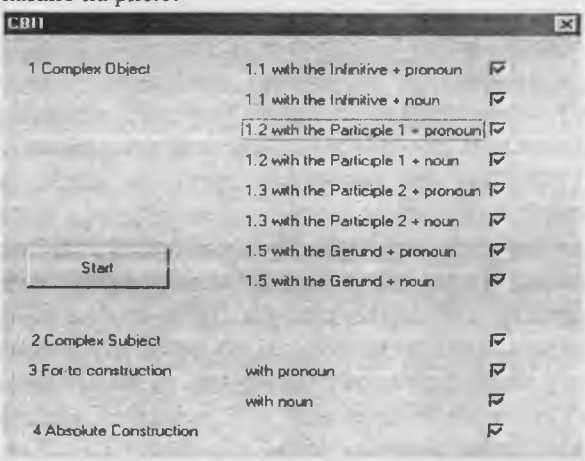

Рис. 1. *Діалогове вікно вибору типів конструкції*

Після натиснення кнопки `**Start**` з`являється шкала **Statusbar**, що відображає ступінь виконання програми у відсотках, доки розпізнавання не закінчиться. Вікно результату роботи програми показано на рис.2.

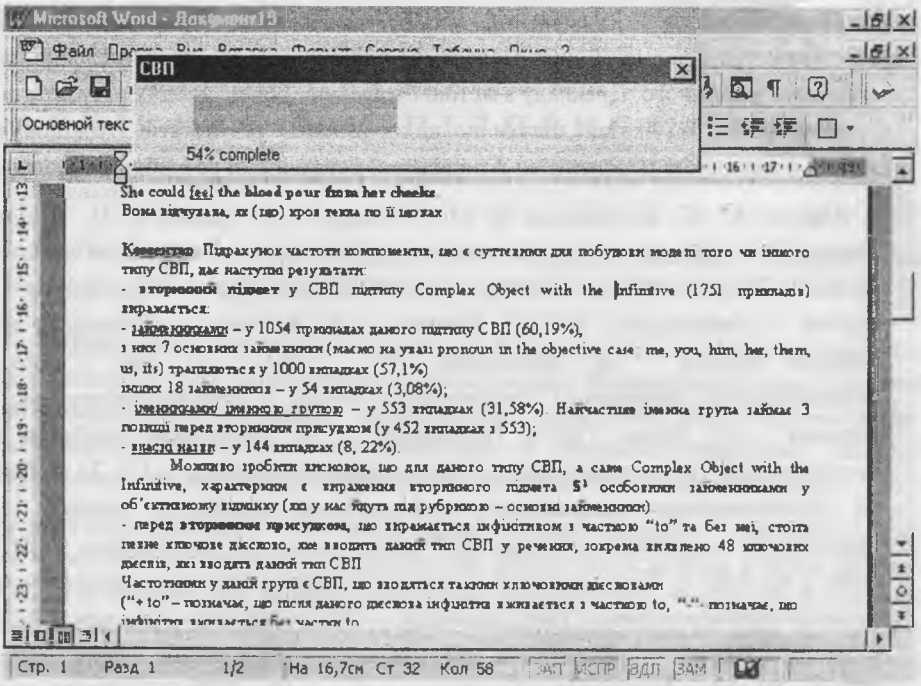

Рис. 2. *Вікно результату роботи програми*

4 ВИСНОВОК

Підсумовуючи, зазначимо, що представлена у даному викладі програма є експериментальною і знаходиться на шляху до вдосконалення, як з точки зору лінгвістичного забезпечення так і з точки зору певних технічних аспектів. Проте, беручи до уваги той факт, що для певного автоматичного оперування з лінгвістичними об'єктами, в першу чергу, слід виявити дані об'єкти серед множини інших, можна сказати, що поставлене важливе завдання у представленій програмі виконано: програма автоматично розпізнає структури вторинної предикації у тексті. Результати опрацювання текстового матеріалу цією програмою, а саме розпізнані в тексті речення, що містять СВП, можна використовувати різними шляхами:

1) Нагромадження бази даних речень, що містять СВП, та використання їх у навчально-методичних цілях. Наприклад, як матеріал для розробки вправ і тестів, що можуть використовуватись при вивченні студентами даних структур. Це є значний позитивний момент, оскільки речення, що містять СВП є складними в плані розуміння, а отже, і в плані їх перекладу;

2) Використання програми по розпізнаванню речень, що містить СВП, полягає у тому, що створена програма може бути покладена в основу "препроцесора" у програмі машинного перекладу. Ідея полягає в тому, що виявлені за допомогою програми речення, можна переробляти а) вручну, в режимі інтерактивної праці з текстом при його перекладі, тобто у попередньому редагуванні, або б) автоматично (при вдосконалені програми додатковими трансформаційними моделями) для спрощеного представлення даних речень для програм машинного перекладу з метою отримання більш якісних результатів.

Вважаємо, що вказані шляхи по використанню результатів функціонування даної програми окреслюють перспективи її подальшої детальнішої розробки та вдосконалення.

1. Апресян Ю. Д., Богславский И. М., Иомдин Л. Л., Крысин Л. П., Лазурский А. В., Перцов Н. В., Санников В. З. Лингвистическое обеспечение в системе автоматического перевода ЭТАП-1 // Разработка формальной модели естественного языка. Сб. научных трудов. – Новосибирск, 1981. 2. Баранов А. Н. Введение в прикладную лингвистику: Учебное пособие. – М.: Эдиториал УРСС, 2001. – С. 360. 3. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения. – М.: Наука, 1985.4. Городецкий Б. Ю. Актуальные проблемы прикладной лингвистики // Новое в зарубежной лингвистике. Выпуск XII. – С.5-22. 5. Городецкий Б. Ю. Компьютерная лингвистика: моделирование языкового общения // Новое в зарубежной лингвистике. Выпуск XXIV. Компьютерная лингвистика. – М.: Прогресс, 1989. – С. 5-32. 6. Карамишева І. Д. Основні ознаки структур вторинної предикації у сучасній англійській мові // Нова філологія. – 2002. - №3(14). – Запоріжжя: ЗДУ, 2002. – С. 65-73. 7. Марчук Ю. Н. Методы моделирования перевода. – М.: Наука, 1985. 8. Марчук Ю. Н. Основы компьютерной лингвистики. – М., 2000. 9. Откупщикова М. И. Моделирование языка // Прикладное языкознание. – С.- Петербург, 1996. – С. 100-112. 10. Персональное бюро переводов // Chip.# 9, 2002, - С. 56-62. 11.Соколов В. И. Особенности програмирования на VBA. (Московський форум http://www.vbasic.ru*). 12. Шалютин С. М. Искусственный интеллект. – М.: Мысль, 1985.*