УДК 004.934.2
Пйонтко Н. – ст. гр. СНм-51
*Тернопільський національний технічний університет імені Івана Пулюя*

# СИНТАКСИЧНИЙ АНАЛІЗ РЕЧЕНЬ УКРАЇНСЬКОЇ МОВИ

Науковий керівник: асистент Маєвський О.В.

Однією з найважливіших задач математичної лінгвістики є синтаксичний аналіз речень мови. Він є невід'ємною складовою таких задач, як: комп'ютерний переклад, синтез мови, перевірка правопису, розстановка розділових знаків, а також є одним із початкових етапів на шляху до розпізнавання мови.

Комп'ютерний синтаксичний аналіз є давно досліджуваною задачею, яка є до певної міри розв'язаною для різних мов. Однак незважаючи на тривале дослідження цієї проблеми спроектовані системи ще далекі від ідеалу, тим паче дуже мало хороших наробок (як теоретичних, так і, що найголовніше, практичних) для української мови і більшість з них є комерційними таємницями.

Задача синтаксичного аналізу розв'язується в декілька етапів: графемний аналіз, морфологічний аналіз і безпосередньо синтаксичний аналіз.

Графемний аналіз – це початковий етап аналізу, отримана інформація на якому використовується в морфологічному і синтаксичному аналізах. Задачею графемного аналізу є виділення у вхідному тексті: слів, розділювачів, збір слів написаних окремо, фразеологізмів, ПІП, електронних адресів, імен файлів, чисел та інше. Поставлені задачі вирішуються за допомогою використання регулярних виразів, словників фразеологізмів і власних назв, а також не складних алгоритмів перевірки приналежності слова або декількох слів певному шаблону.

Морфологічний аналіз – це етап на якому кожному слову приписується ряд атрибутів, які визначають, до якої частини мови воно належить, які ознаки має в середині цієї частини мови (наприклад, іменник чоловічого роду, знахідний відмінок, однина). Найбільш точним і надійним методом розв'язку задачі морфологічного аналізу є використання словника слів. Оскільки словники містять зазвичай велику кількість слів, використовуються спеціальні структури даних і алгоритми пошуку, щоб якомога зменшити час пошуку і об'єм затраченої пам'яті. Якщо організувати слова у вигляді масиву і відсортувавши їх в порядку зростання, можна використати інтерполяційний і бінарний пошуки, які мають складність $O(\log_2\log_2 N)$ і $O(\log_2 N)$ відповідно, де N – кількість слів в словнику. Важливою особливістю морфологічного аналізу є визначення атрибутів слів, яких немає в словнику. Оскільки основним методом творення слів в українській мові є флективний, то ми можемо не знайденому слову надати атрибути слова, яке максимально співпало з шуканим із сторони закінчення (кількість букв співпадіння не менше 4).

Синтаксичний аналіз останній і найскладніший етап. Для кожної мови створені свої підходи, які безперервно розвиваються і змінюються. Перш ніж здійснити синтаксичний аналіз, необхідно провести його фрагментацію – виділити прості речення, які входять в складні. Це здійснюється методом розбиття речення на сегменти, що містяться між розділовими знаками і застосуванням до отриманих сегментів спеціальних правил, що дозволяють отримати перелік простих речень.

При проведенні синтаксичного аналізу вводять поняття – синтаксична група – сукупність двох залежних слів, груп, або слова і групи, в якій виділений головний елемент. Застосовуючи спеціальні синтаксичні правила, здійснюється формування синтаксичних груп з подальшим їх укрупненням.