УДК 004.09

ФОРМАЛЬНИЙ АНАЛІЗ МОРФЕМНИХ СКЛАДОВИХ ТЕЗАУРУСА МОВНИХ ОБРАЗІВ

Кравчук І.А.
Науковий керівник – проф., д.т.н. Квєтний Р.Н.

Лінгвістична обробка природномовних (ПМ) текстів є однією з центральних проблем інтелектуалізації інформаційних технологій. Традиційно аналіз ПМ текстів представлявся як послідовність процесів – морфологічний аналіз, синтаксичний аналіз, семантичний аналіз.

Морфологічний аналіз вхідних словоформ є першим етапом аналізу ПМ тексту. Результати морфологічного аналізу є базовою основою для всіх наступних етапів автоматичного аналізу та синтезу тексту. Морфеми поділяються на два основних типи – кореневі та афіксальні. Корінь є основною значущою частиною слова. Афікс є допоміжною частиною слова, що приєднується до кореня і служить для словотвору та вираження граматичних значень. Найбільш поширені два типи афіксів – префікси, що знаходяться перед коренем, та постфікси, що знаходяться після кореня. Постфікси розділяються на суфікси (мають словотвірне значення) і флексії або закінчення (вказують на зв'язок з іншими членами речення). В українській мові, яка належить до флексійних мов, використовуються як префікси, так і постфікси (суфікси та закінчення).

Аналізуючи морфологічну структуру слова флексійних мов, визначимо базові правила з P для слів w та y:

а) x є префіксом w ($x \triangleright w$), якщо $w = xy$,

б) z є суфіксом w ($z \triangleleft w$), якщо $w = yz$,

в) ν є закінченням w ($\nu \triangleleft w$), якщо $w = y\nu$ або $w = yz\nu$.

В загальному випадку слово з одним префіксом, суфіксом та закінченням має вигляд:

$$w = xyz\nu .$$

Формалізуємо правила визначення множин коренів, префіксів та суфіксів:

$Root(L) = \{y \mid (\exists y \in L)\}$,

$Pref(L) = \{x \mid (\exists y \in L)x \triangleright w\}$,

$Suf(L) = \{z \mid (\exists y \in L)z \triangleleft w\}$,

$End(L) = \{\nu \mid (\exists y \in L)\nu \triangleleft w\}$.

За допомогою цих правил забезпечується можливість перевірки належності довільного слова w до множини $L(G)$ всіх правильних слів у граматиці, що розглядається.

Отже, задача зводиться до обґрунтування конструктивних алгоритмів попереднього відокремлення морфем з складових тезауруса мовних образів.

УДК 004.932.4