УДК 004.912

Т.В. Шепелева, к.фіз.-мат.н., Г.С. Краліна

*Національний авіаційний університет*

ДОСЛІДЖЕННЯ І РОЗРОБКА ЕФЕКТИВНИХ
АЛГОРИТМІВ ТОКЕНІЗАЦІЇ

Сучасні технології попередньої обробки електронних текстів базуються на багаторівневому представленні природної мови. Компоненти, що складають мовну модель, – лінгвістичні процесори, які один за одним обробляють вхідний текст, – вхід одного процесора є виходом іншого. Як правило, виділяються наступні компоненти процесу попередньої обробки електронних текстів:

* токенізація – графематичний аналіз – виділення слів, цифрових комплексів, формул тощо;
* морфологічний аналіз – побудова морфологічної інтерпретації слів вхідного тексту;
* синтаксичний аналіз – побудова дерева залежностей речення в цілому;
* семантичний аналіз – побудова семантичного графа тексту.

Графематичний аналіз є початковим етапом обробки тексту, в ході якого визначаються елементи граматичної структури (слова, знаки пунктуації, числа, скорочення тощо). Результатом такого аналізу має бути інформація, що є необхідною для подальшої обробки морфологічним  та синтаксичним процесорами.

В роботі виконано огляд сучасних алгоритмів токенізації з точки зору їх ефективності, розглянута постановка задачі графематичного аналізу в термінах теорії формальних мов, обрана графематична модель російської та української мови, на підставі якої виконана алгоритмічна і програмна реалізація модуля для графематичного аналізу.

Запропоновано тести для аналізу ефективності алгоритмів графематичного аналізу електронних текстів.