УДК 004.852

М.С. Ярощук
В.С. Озеранський

ОРФОГРАФІЧНИЙ ТА СЕМАНТИЧНИЙ АНАЛІЗИ ТЕКСТУ

Вінницький національний технічний університет

Анотація

Доповідач намагається пояснити значущість орфографічного та семантичного аналізів тексту, розглянувши причини появи різних видів помилок. Акцентує увагу на тому, що машинний аналіз тексту почав здійснюватися від початку становлення кібернетики, тому є важливим і необхідним для суспільства. Автор статті розповідає, що орфографічний аналіз тексту зобов'язує наявність еталонного словника, семантичний аналіз тексту сприяє доцільному використанню ключових слів у тексті.

Ключові слова: аналіз, текст, помилка, слово, дослідження

Abstract

The speaker tries to explain the importance of spelling and semantic analyzes of the text by examining the causes of different types of errors. Emphasizes that machine text analysis has begun since the beginning of cybernetics, and is therefore important and necessary for society. The author of the article says that spelling analysis of the text requires the presence of a reference dictionary, semantic analysis of the text contributes to the expedient use of keywords in the text.

Keywords: analysis, text, error, word, research

Слово «аналіз» походить із грецької мови, що означає «розкладання». Як науковий термін аналіз – це методичне наукове дослідження предметів, явищ та дій шляхом розкладу, розчленування їх у думці на складові частини. [1]

Оскільки суспільство використовує аналітичні методи, тому термін «аналіз» застосовують як синонім до понять «дослідження», «розв'язання різнопланових завдань». Отже, аналіз є одним із етапів будь-якого наукового дослідження, в результаті якого дослідник виявляє основні ознаки об'єкту, його призначення, відповідність ідентичним. [2]

Щодо аналізу тексту – то це процес, який полягає в тому, що дослідник отримує інформацію відповідно до запиту дослідження. У науці відомі різні види аналізу тексту, ми пропонуємо вашій увазі орфографічний та семантичний, оскільки в повсякденному житті робота з текстом у будь-якій галузі потребує великої відповідальності.

Набираючи текст, можна припуститися помилок, які за своєю специфікою поділяються на дві групи: орфографічні та друкарські огріхи.

Причини появи орфографічних помилок:

1) морфологічний принцип українського правопису, за яким написання слів, морфем не відповідає вимові, тому слід знати багато правил, щоб уникнути орфографічних помилок
2) низький рівень грамотності серед населення
3) впровадження нової редакції Українського правопису
4) уведення до словника загальновживаних слів діалектизмів
5) поява нових слів (неологізмів), правопис яких ще остаточно не визначений

Другий тип помилок, які можна помітити, працюючи із текстом, - це друкарські огріхи. Вони виникають у результаті різних причин. По-перше, клавіатура може бути незвичною, по-друге, товсті пальці, по-третє, набір тексту відбувається поспіхом.

У зв'язку з тим, що обидва види помилок різні, систему автокорекції можна побудувати як для кожного типу окремо, так і в цілому як універсальний механізм. Щодо друкарських огріхів, то їх дуже швидко в тексті можна помітити, а от з приводу орфографічних помилок необхідно вдатися до більш серйозного дослідження. [3]

Із появою кібернетики та обчислювальної техніки питання аналізу тексту щодо орфографічних помилок одразу було розв'язано завдяки автоматизованій перевірці грамотності. [4, 7]

Суб'єкт, який набирає текст, повинен бути переконаний у тому, що матеріал оформлено без орфографічних помилок. Роботу над мовними недоліками допомагає виконувати заздалегідь

внесений словник. Слова, що входять до його складу, є своєрідним еталоном, з яким порівнюють слова, взяті з тексту. Причому до такого еталонного словника можна вводити нові слова.

Якщо ж в словнику немає відповідного слова, то відбувається пошук-дослідження. Для слова, що перевіряється, створюється набір таких слів, які можуть утворитися шляхом операцій: вилучення, вставка, заміна й перестановка. [5]

Хоча є ряд недоліків, але в цілому робота над орфографічними помилками відбувається швидко і відносно якісно, що сприяє виникненню у людей почуття довіри до системи автокорекції.

Семантичний аналіз тексту дозволяє вичленувати в реченні ключові слова, визначити їхній зв'язок з іншими словами в реченні, з'ясувати залежність значення слова від контексту.

Семантичний аналіз – надзвичайно важкий процес, оскільки людина закріплює за кожним словом свій образ, який машина не завжди може розкрити.

Виконуючи семантичний аналіз тексту, дослідник дізнається таку інформацію: кількість слів, які визначають зміст тексту, тобто його семантичне ядро, а також частоту їхнього вживання. Якщо правильно сформоване семантичне ядро, то аналіз тексту відбувається дуже швидко. Підбір необхідних слів, грамотно побудовані фрази ефективно впливатимуть на читача, спонукаючи до певних дій, у яких зацікавлені замовники тексту. Отже, семантичний аналіз тексту необхідний, в першу чергу, задля інформативності та пробудження зацікавленості.

Щоб текст відповідав запитам дослідників, обов'язково необхідно враховувати статистичні показники. Проводячи семантичний аналіз, науковці визначають такі статистичні показники: кількість символів з пробілами і без них, загальна кількість слів, зокрема унікальних та значущих, кількість стоп-слів, кількість води, граматичних помилок, процент класичної та академічної нудоти, семантичне ядро. Відбувається порівняльний аналіз кількості унікальних слів (без повторів), значущих слів (іменників) та стоп-слів (неповнозначних). Наприклад, якщо в тексті кількість води вища за 70%, то необхідно зменшити кількість стоп-слів. Якщо ж у тексті дуже часто повторюється одне й те ж слово, то показник класичної нудоти буде вищим за норму (7). Підвищення коефіцієнта академічної нудоти залежить від повторення в тексті великої кількості слів. [6, 8]

Орфографічний та семантичний аналізи текстів сприяють підвищенню грамотності, рівня культури мовлення, вихованню в громадян почуття відповідальності за написане, надруковане.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Яременко В., Сліпушко О. Новий тлумачний словник української мови у трьох томах (1 том А-К). – К.:Видавництво «Аконіт», 2001. – 928 с.
2. inforaz Глава 2. Аналіз документів (рос) [Електронний ресурс] / inforaz. – Режим доступу: http://inforaz.narod.ru/analiz-2.html
3. Хабр Робимо спелчекер на фонетичних алгоритмах своїми руками (рос) [Електронний ресурс] / Хабр. – Режим доступу: https://m.habr.com/ru/post/325364/
4. Воронько В., Костинський О. Комп'ютерний аналіз текстів (рос) [Електронний ресурс] / Воронько В., Костинський О. – Режим доступу: https://archive.svoboda.org/programs/sc/2001/sc.062601.asp
5. Хабр Penisland, або як визначати помилку (рос) [Електронний ресурс] / Хабр. – Режим доступу: https://m.habr.com/ru/post/105450/
6. Cropas Семантичний аналіз (рос) [Електронний ресурс] / Cropas. – Режим доступу: https://cropas.by/seo-slovar/semanticheskij-analiz/
7. Колодний В.В. Метод некритеріального структурування множини альтернатив за допомогою аналізу тернарних трирівневих ранжувань / В.В. Колодний, В.В. Зубко // «ІНТЕРНЕТ-ОСВІТА-НАУКА-2014» : Збірник матеріалів конференції. – Вінниця : ВНТУ, 2014. – С. 13-14.
8. Колодний В.В. Застосування гештальт-ранжувань для виявлення переваг ОПР / В.В. Колоний, В.В. Зубко // «ІНТЕРНЕТ-ОСВІТА-НАУКА-2016» : Збірник матеріалів конференції. – Вінниця : ВНТУ, 2016. – С. 43-44.

Ярощук Мирослав Сергійович - студент групи 2КН - 16б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: krystallsilver@gmail.com

Науковий керівник: ***Озеранський Володимир Сергійович*** - кандидат технічних наук, старший викладач, Вінницький національний технічний університет, м. Вінниця, e-mail: ozeransky@ukr.net

Yaroshchuk Myroslav - Department of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email: krystallsilver@gmail.com

Supervisor: **Ozeransky Volodumur** - Ph.D., senior lecturer, Vinnytsia National Technical University, Vinnytsia, e-mail: ozeransky@ukr.net