

DOI <https://doi.org/10.15407/usim.2020.03.060>
УДК 004.4

Г.В. ХОДЯКОВА, канд. пед. наук, доцент кафедри інформаційних технологій,
Миколаївський національний ун-т ім. В.А. Сухомлинського,
м. Миколаїв, вул. Шнеєрсона, 11, кв. 7,
khodiakovagalina@gmail.com

ВИКОРИСТАННЯ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ У РОЗВ'ЯЗАННІ ЗАДАЧ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

Представлено огляд популярних онлайн-сервісів з обробки та аналізу тексту та опис технології рішення деяких лінгвістичних задач: семантичного аналізу тексту; обчислення частотних характеристик тексту; знаходження типологічних індексів Грінберга для даної мови, розбору слова за складом, граматичного аналізу тексту, побудови синтаксичних графів. Описано перспективи використання вирішуваних завдань в подальших дослідженнях.

Ключові слова: комп'ютерна лінгвістика, комп'ютерна обробка тексту, аналіз тексту, завдання лінгвістичного аналізу тексту.

Вступ

Спроби оцінити лінгвістичні явища математичними методами робилися протягом багатьох десятиріч. Але останнім часом істотно зріс інтерес до комп'ютерної лінгвістики у зв'язку з переконливими результатами вирішення її завдань на основі використання персональних комп'ютерів і комп'ютерних мереж. Спільні зусилля вчених різних спеціальностей: лінгвістів, математиків, фізиків, фахівців в області інформаційних технологій дали реальні практичні результати і зараз вже можна навести приклади вирішення завдань: ідентифікація мови тексту, визначення автора, автоматична обробка текстів у текстових редакторах; інформаційний пошук у мережі Інтернет; розпізнавання мови, голосове введення; синтез мови, читання текстів; переклад з мови на мову в режимі реального часу; програми навчання іноземним мовам, голосові помічники *Siri, Amazon Alexa, Google Assistant, Яндекс, Аліса* тощо.

Огляд досліджень

Автоматичну обробку тексту природною мовою та проблем комп'ютерної лінгвістики досліджують науковці: Арапов М.В., Баранов А.Н., Ланде Д.В., Большакова Є.І., Клишинській Е.С., Носков А.А., Андрусенко Т.Б., Волошин В.К., Городецький Б.Ю., Пирогова Ю.К., Марчук Ю.М., Хроленко А.Т., Широков В.А. та ін. Наприклад, Арапов М.В. описує значення кількісних даних для вивчення мови, однорідність і регулярність зв'язків між одиницями словника [2]. Хроленко А.Т. розкриває найважливіші напрямки в розробці лінгво-культурологічної методології [9]. Баранов А.Н. описує деякі завдання комп'ютерної лінгвістики: машинний переклад, інформаційно-пошукові системи, лексикографію, термінознавство татермінографію, методику викладання мови, теорію перекладу, корпусну лінгвістику, лінгвістичні аспекти нейро-лінгвістичного програмування, теорію впливу [3]. Марчук Ю.М. присвятив

свою роботу лінгвістичним основам обробки текстів природною мовою за допомогою комп'ютера [7]. Таким чином, на цей час є велика кількість наукових робіт, присвячених комп'ютерній лінгвістиці. У них наявні проблеми обговорюються на рівні загальнонаукової методології. Ми ж вирішили описати процес розв'язання деяких прикладних лінгвістичних задач на технологічному рівні [11].

Постановка задачі

Спочатку аналіз та обробка лінгвістичної інформації здійснювалися з використанням комп'ютерних програм загального та спеціального призначення, які створювалися різними авторськими колективами. На початку 2000-х років застосовувалася, наприклад, програма ВААЛ [13], деякі інші засоби. Проте це були лише перші спроби використання подібних програм. Нині спеціалізовані онлайн-сервіси, які розробляються великими корпораціями, значно перевершують відомі раніше програмні засоби з обробки та аналізу текстів. Вони мають вищу якість, надійність і доступність. При цьому перелік розв'язуваних лінгвістичних задач істотно розширився.

Мета статті — дати опис нової технології вирішення деяких лінгвістичних задач шляхом використання популярних онлайн-ресурсів і на базі порівняльного методу. У статті наводиться вирішення завдань семантичного аналізу тексту; обчислення частотних характеристик тексту та використання цих даних для подальших досліджень, знаходження типологічних індексів Грінберга для даної мови, розбору слова за складом, граматичного аналізу тексту, побудови семантичних графів.

Огляд популярних онлайн-ресурсів для розв'язання задач комп'ютерної лінгвістики

Серед великої кількості онлайн-програм з обробки тексту ми обрали найвідоміші та надійні в плані результату. Якість програм оцінювалося через вивчення думки користувачів

та порівняння результатів роботи програм під час особистого експерименту. У такий спосіб було обрано програми, орієнтовані переважно на роботу з російськомовним текстом. Деякі з них дають змогу обробляти текст також інших мов, зокрема, української, англійської та німецької.

1. *Text.ru* (<http://text.ru/>) — це онлайн-сервіс перевірки тексту на унікальність, відображає відсоток унікальності тексту, знаходить дублікати тарерайт; виконує перевірку орфографії, характеристики *SEO*-аналізу тексту російською та англійською мовами.

2. *Istio.com* (<http://istio.com/rus/text/analyz/>) виконує семантичний аналіз тексту російською мовою. Показує такі параметри, як довжина тексту, водність, нудотність, найживаніші у тексті слова та інші параметри.

3. *Lenartools.ru* (<http://lenartools.ru/tools/lemmatop/>) знаходить слова (леми, уніграми) та словосполучення (біграми), які часто використовуються в певному списку адрес, приводячи їх до основної форми слова. Наприклад, ви можете завантажити в список Топ-100 сторінок із пошукової видачі за будь-яким запитом і сервіс покаже слова та словосполучення, які часто використовуються в текстах ваших конкурентів. Або можете завантажити список *URL* на відгуки про продукт і сервіс покаже слова та словосполучення, якими користувачі висловлюють своє схвалення, несхвалення, сумніви та тривоги стосовно певного об'єкта чи продукту. Використання *lemmatop* збільшує якість й ефективність текстів на сайтах.

4. Сайт Главред (<https://glvrd.ru/>) допомагає очистити текст від словесного сміття, перевіряє на відповідність інформаційному стилю. Він є відповідним для реклами, новин, статей, сайтів, інструкцій, листів і комерційних пропозицій, але не надається для редагування віршів, художньої прози.

5. *Etxt.ru* (<https://www.etxt.ru/antiplagiat/>) виконує перевірку тексту на унікальність. Безкоштовна версія програми може перевіряти текст обсягом до 3000 слів. Зберігає результати перевірки на сервері та за потреби надає їх.

6. Яндекс. Спеллер (<https://tech.yandex.ru/speller/>) допомагає знаходити та виправляти орфографічні помилки в російському, українському або англійському тексті. Мовні моделі Спеллер включають сотні мільйонів слів і словосполучень. Щоб виявляти помилки та підшукувати заміни, Спеллер використовує бібліотеку машинного навчання *CatBoost*.

7. Система «Антиплагіат» визначає також автора тексту (<https://content-watch.ru/text/>) є ефективним безкоштовним інструментом перевірки українських і російських текстів на унікальність та якість.

8. Ресурс Адвего (<https://advego.com/text/seo/>) — один із найпопулярніших засобів аналізу текстів 90 мовами. Він виконує перевірку тексту на унікальність, семантичний аналіз, обчислює індекс якості сайту.

9. *Psi-technology* (<https://psi-technology.net/servisfonosemantika.php>) виконує фоносемантичний аналіз слова російською мовою, оцінюючи його за 25 характеристиками (фоносемантичними шкалами).

10. Аналіз листів (<http://www.analizpisem.ru/index.html>) — ще одна програма фоносемантичного аналізу російської мови. За допомогою цієї програми визначається настрій автора та дається його характеристика за десятьма параметрами.

11. Автоматична обробка текстів (*aot.ru*) виконує кілька завдань, а саме: морфологічний аналіз, синтаксичний аналіз, побудову синтаксичних дерев у вигляді графів, переклад з російської мови на англійську, лінгвістичний пошук у розміченому морфологічним аналізатором масиві.

12. Сайт Карта слів (*kartaslov.ru*). Створює карту для заданого слова, яка показує: значення слова, асоціації із заданим словом, синоніми, робить морфологічний і морфемний розбір, показує речення та цитати із заданим словом.

Вивчаючи роботу зазначених сайтів, ми виявили, що аналіз текстів за деякими параметрами збігається. Наприклад, статистичні дані, перевірка орфографії виконується на всіх сайтах однаково. Інші види окремих показників тексту можуть значно відрізнятися. Наприк-

лад, аналіз на водність та нудотність тексту. Це відбувається тому, що онлайн-ресурси використовують різні алгоритми. Окрім того такий показник, як «водність тексту», залежить від наявної бази «стоп-слів», яка періодично оновлюється.

За допомогою *Text.ru* та *Advego.com* нами було проаналізовано на «водність тексту» ті самі чотири фрагменти тексту російською та українською мовами. Оскільки результати значно різнилися, ми додатково виконали «ручну перевірку». Результати обчислень наведено в табл. 1.

Таким чином, ми прийшли до висновку про необхідність використання порівняльного методу.

Використання онлайн-ресурсів для розв'язання задач комп'ютерної лінгвістики

Задача 1. Обчислення частотних характеристик тексту. Закон Ціпфа.

Складання частотного словника тексту є дуже поширеним завданням. За його допомогою можна встановлювати зв'язок між інформаційною цінністю одиниць мови та кількісними характеристиками слів і використовувати ці дані для подальшої обробки під час вирішення конкретних завдань, наприклад, у задачі атрибуції тексту або для *SEO*-аналізу тексту. Складаючи частотний словник, ми ставили за мету перевірити якість роботи комп'ютерної онлайн-програми та порівняти результати.

Таблиця 1. Порівняльний аналіз фрагментів тексту на «водність»

| Номер фрагмента тексту | <i>Text.ru</i> | <i>Advego.com</i> | <i>Advego.com</i> | Перевірка «вручну» |
|------------------------|----------------|-------------------|-------------------|--------------------|
| | Російською, % | Російською, % | Українською, % | Російською, % |
| 1 | 3 | 53,5 | 21,1 | 14 |
| 2 | 9 | 23 | 29,5 | 23 |
| 3 | 7 | 53,4 | 18,3 | 19 |
| 4 | 8 | 60 | 28,8 | 31 |

Тому попередньо в редакторі *MS Word* підготували частотний словник для тексту, наприклад, із 100 слів: завантажений текст виділили і перетворили на таблицю з однією колонкою; відсортували слова за абеткою в порядку зростання частоти вживання. Після сортування ми бачимо повторювані слова та їхню кількість, тому додаємо другий стовпчик, де записується частота слів, а рядки із зайвими словами видаляємо. Далі сортуємо цифри у другій колонці за порядком спадання і додаємо перший стовпчик, де записуємо ранг слова. Частотний словник готовий. Робота займає близько 15 хвилин.

В табл. 2 наведено фрагмент таблиці з підсумками частотного аналізу.

За допомогою одного з онлайн-ресурсів — *text.ru*, *istio.com* або Адвего можна також скласти частотний словник і проаналізувати текст за певними параметрами за 1–2 хвилини. Далі порівнюємо результати роботи та робимо висновки про якість результатів обробки тексту тією чи іншою програмою.

На підставі отриманих даних можна здійснювати подальші дослідження, наприклад, знаходити відносну частоту слів, перевіряти закон Ціпфа про розподіл частоти слів у тексті. Для будь-якого слова добуток його рангу та частоти появи буде величиною постійною.

Таблиця 2. Фрагмент таблиці з підсумками частотного аналізу

| <i>r</i> | <i>X</i> | <i>W</i> |
|----------|-----------|----------|
| 1 | В | 5 |
| 1 | Не | 5 |
| 2 | И | 3 |
| 2 | Слишком | 3 |
| 3 | Гостиница | 2 |
| 3 | Город | 2 |
| 3 | Чтобы | 2 |
| 3 | Бричка | 2 |
| 3 | Сидеть | 2 |
| 4 | Въехала | 1 |
| 4 | но | 1 |
| 4 | Ворота | 1 |

$$r * w = c,$$

де *w* — частота використання слова в тексті;

r — ранг слова у списку;

c — емпірична постійна величина (коефіцієнт Ціпфа).

Задача 2. Семантичний аналіз тексту.

Семантичний аналіз є показником якості *SEO*-тексту, який складається та розміщується на сторінках сайту з метою його просування. Написання тексту для оптимізації вимагає наявності в ньому низки показників, які не повинні перевищувати відповідні норми. Іншими словами, аналіз дає можливість зрозуміти, чи не переспамлено текст, наскільки великою є його нудотність, чи не завеликою є його водність, які слова повторюються занадто часто. За результатами семантичного аналізу можна здійснити необхідне коригування тексту, зробивши його привабливішим і для пошукової машини, і для людей.

У процесі семантичного аналізу виявляються такі показники:

- статистика тексту (кількість символів, кількість символів без пробілів, кількість слів, кількість унікальних слів);
- ключові слова;
- семантичне ядро;
- стоп-слова;
- класична та академічна нудотність тексту.

Виконуючи семантичний аналіз, деякі параметри тексту можна підрахувати самостійно. Для цього визначмо відносну частоту слова. У частотному словнику додаємо третій стовпчик, потім число повторів слова ділимо на загальну кількість слів у тексті. Пошук семантичного ядра виконується за допомогою частотного словника — це, як правило, сукупність ключових слів. Стоп-слова — це слова, що не мають змістового навантаження (прийменники, сполучники, займенники, іменники, що найчастіше вживані в Інтернеті, дієслова тощо).

Далі за відповідальними формулами обчислюється класична нудотність тексту:

$$K = \sqrt{w},$$

та академічна нудотність тексту

$$A = \left| \frac{w}{W} \right| \cdot 100\%,$$

Таблиця 3. Фрагмент таблиці з результатами аналізу морфемної структури слів

| Слово | Морфемная структура | | | | | | Кількість морфем | Кількість швів | |
|----------------|---------------------|---------|------------------|--------|------------|----------|------------------|----------------|---|
| | Приставка | Корінь | З'єднув. голосна | Суфікс | Закінчення | Постфікс | | | |
| по аллеям | | алле | | | ям | | 2 | 1 | |
| среди | | | | | | | 1 | | |
| цветущих | | цвет | | у | их | | 3 | 2 | |
| кустов | | куст | | | ов | | 2 | 1 | |
| сирени | | сирен | | | и | | 2 | 1 | |
| и пылающих | | пыл | | а | ющ | их | 4 | 3 | |
| альми | | ал | | | ыми | | 2 | 1 | |
| тюльпанами | | тюльпан | | | ами | | 2 | 1 | |
| тлумб | | клумб | | | | | 1 | | |
| прогуливалась | про | гул | | ива | ла | сь | 5 | 4 | |
| нарядная | | наряд | | н | ая | | 3 | 2 | |
| публика | | публик | | | а | | 2 | 1 | |
| дамы | | дам | | | ы | | 2 | 1 | |
| под кружевными | | кружев | | н | ы | ми | 4 | 3 | |
| зонтиками | | зонт | | ик | ами | | 3 | 2 | |
| скачущего | | скач | | а | ющ | его | 4 | 3 | |
| вида | | вид | | | а | | 2 | 1 | |
| молодые | | молод | | | ые | | 2 | 1 | |
| люди | | люд | | | и | | 2 | 1 | |
| взаимодействие | | взаим | дей | о | ств | и | е | 6 | 5 |
| ничто | ни | что | | | | | 2 | 1 | |
| не предвещало | пред | вещ | | | а | ло | 4 | 3 | |
| неприятностей | не | приятн | | | ость | ей | 4 | 3 | |

де w — кількість входжень ключового слова, W — загальна кількість слів у тексті.

Статистику тексту, як і інші значущі показники семантичного аналізу, можна отримати, використовуючи можливості редактора *MS Word* через меню «Статистика текста». Контрольна перевірка результатів здійснюється за допомогою онлайн-програм *text.ru*, *istio.com* або Адвего.

Задача 3. Типологічні індекси Грінберга.

Основні критерії типологічної характеристики мови припускають обчислення індексів син-

тетичності, аглютинації, деривації, показника словотворчої спроможності мови, індексів префіксальності, суфіксальності, словозміни.

Значення кожного з названих критеріїв залежить від числа морфем в досліджуваному фрагменті тексту. Тому попередньо робиться морфемний аналіз тексту, наприклад, зі 100 слів, і результати заносяться в таблицю. Морфемну структуру слова допомагає перевірити інтернет-ресурс *kartaslov.ru* (див. табл. 3).

На рис. 1 продемонстровано морфемний аналіз слова «лінгвістика».

На підставі отриманих даних підраховуються типологічні індекси Грінберга.

1. Індекс синтетичності $I = M/W$, де M — кількість морфем, W — кількість слів.

2. Індекс аглютинації $= A/J$, де A — кількість аглютинативних конструкцій, J — кількість швів між морфемами.

3. Рівень деривації $= R/W$, де показник словотворчої спроможності мови $= D/W$, R — кількість кореневих морфем у тексті, D — кількість дериваційних морфем (суфікси і префікси), W — кількість слів у тексті

4. Місце афіксу по відношенню до кореня
індекс префіксальності $= P/W$,
індекс суфіксальності $= S/W$,

де P — префікси,

S — суфікси, W — кількість слів у тексті;

індекс словозміни $= Pi/N$,

де Pi — чисті словозміни (флексії),

N — кількість слів, що не повторюються.

Задача 4. Аналіз тексту та побудова синтаксичних дерев у вигляді графів.

Діалог з програмою *aot.ru* є дуже простим. Необхідно ввести речення до 150 символів і програма побудує синтаксичний граф. Приклади наведено на рис. 2 та 3.

У програмі *aot.ru* є список відносин між вершинами графа:

SUB — суб'єкт / підмет;

CONTEN — зміст;

OBJ — об'єкт;

BELNG — належність чогось / когось;

PROPERT — ознака чогось / когось;

AGENT — агент;

LOK — локація;

QUANTIT — кількість;

MUA: B — обидві позиції висловлюють однакові відносини, а також інші.

На даний час широко застосовується аналіз текстів на антиплагіат. Ці ж програми можуть використовуватися для атрибуції тексту. Ідентифікація автора встановлюється за деякими загальними та приватними ознаками тексту, що складають авторський стиль. Функціонують системи визначення авторства тексту *text.ru*, *etxt.ru*, система «Антиплагіат» тощо. У них використовуються статистичні методи, за-



| | |
|-------|-----------|
| лингв | корень |
| ист | суффикс |
| ик | суффикс |
| а | окончание |

Рис. 1. Розбір за складом слова «лінгвістика» на сайті *kartaslov.ru*

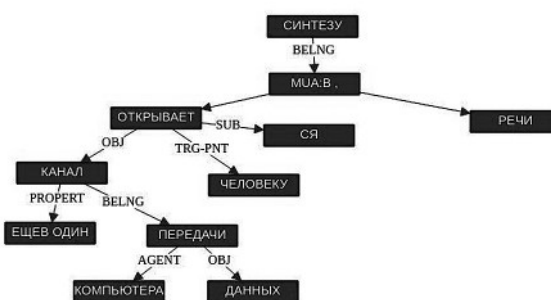


Рис.2. Побудова синтаксичного графа для речення: «Благодаря синтезу речи, открывается еще один канал передачи данных от компьютера к человеку»



Рис.3. Побудова синтаксичного графа для речення: «Невидимая нить коснулась его лба и неслышно лопнула»

сновані на пошуку «авторського інваріанта». Він характеризує мовну особливість тексту. Як інваріант можуть виступати: частка голосних або приголосних, частота вживання певної частини мови, ймовірність переходів від однієї частини мови до іншої, «улюблені» слова тощо. Комп'ютерна обробка текстів виконується також під час фоносемантичного аналізу слів та тексту (програми *Psi-technology*, *analizpisem*) та в інших задачах.

Висновки

У статті описано процес вирішення деяких прикладних лінгвістичних завдань на технологічному рівні. А саме: завдання семантичного аналізу тексту; обчислення частотних характеристик тексту; визначення типологічних індексів Грінберга для певної мови, розбір слова за складом, синтаксичний аналіз тексту, побудова синтаксичних дерев у вигляді графів. Особливістю запропонованого підходу є використання сучасних онлайн-ресурсів, які належать до вільного програмного забезпечення, і якісність яких було перевірено в процесі їхнього використання.

Результати досліджень можна застосувати для розв'язання інших завдань. При визначенні автора тексту попередньо здійснюється частотний аналіз слів, словосполучень і зворотів мови.

Семантичний аналіз тексту використовується *SEO*-фахівцями, робота яких полягає в оптимізації сайтів під пошукові запити. Результати обчислення типологічних індексів потрібні фахівцям із теоретичної лінгвістики.

Грамматичний аналіз текстів і побудова синтаксичних графів використовуються під час створення програм-перекладачів та програм із розпізнавання мови й автоматичного вводу тексту.

Стаття може бути корисною лінгвістам та фахівцям в області *SEO*-аналізу, які розв'язують задачі аналізу та обробки текстів.

Перспективу подальших досліджень ми вбачаємо в розширенні списку завдань комп'ютерної лінгвістики та пошуку конкретних технологічних рішень цих завдань на основі використання інформаційних технологій.

ЛІТЕРАТУРА

1. Андрусенко Т. Б. Лингвистические структуры в компьютерных средах. Киев, 1994. 160 с.
2. Арапов М. В. Квантитативная лингвистика. М.: Наука, 1988. 184 с.
3. Баранов А. Н. Введение в прикладную лингвистику. М., 2003. 347 с.
4. Большакова Е. И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учебное пособие. М.: МИЭМ, 2011. 272 с.
5. Волошин В. К. Комп'ютерна лінгвістика : навчальний посібник. Суми : Університетська книга, 2004. 381 с.
6. Городецкий Б. Ю. Компьютерная лингвистика : моделирование языкового общения. Новое в зарубежной лингвистике. М., 1989. Вып. 24.
7. Марчук Ю. Н. Компьютерная лингвистика. М., 2007. 317 с.
8. Пирогова Ю. К. Рекламный текст : Семиотика и лингвистика. М., 2000. 263 с.
9. Хроленко А. Т. Основы лингвокультурологии : учебное пособие. М.: Наука, 2008. 184 с.
10. Широков В. А. Феноменология лексикографических систем. Київ: Наукова думка, 2004. 327 с.
11. Юдин Э. Г. Системный подход и принцип деятельности. М. : Наука, 1978.
12. Проект ВААЛ. - VAAL. URL: <http://www.vaal.ru/index.php>. (дата звернення: 25.11.2019).

Надійшла 05.01.2020

REFERENCES

1. Andrusenko T. B., 1994. Lingvisticheskie struktury v kompyuternykh sredakh [Linguistic structures in computer environments]. Kiev, 160 p. (in Russian).
2. Arapov M. V., 1988. Kvantitativnaya lingvistika [Quantitative linguistics]. Moscow:Nauka, 184 p. (in Russian).
3. Baranov A. N., 2003. Vvedenie v prikladnuyu lingvistiku [Introduction to Applied Linguistics], Moscow, 347 p. (in Russian).
4. Bolshakova E. I., 2011. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i kompyuternaya lingvistika, uchebnoe posobie [Automatic Natural Language Processing and Computational Linguistics: Tutorial], MIEM, Moscow, 272 p. (in Russian).
5. Voloshin V. K., 2004. Kompyuterna lingvistika : navchalnyi posibnik [Computer linguistics: tutorial], UnIversitetska kniga, Sumy, 381 p. (in Ukrainian).
6. Gorodetskiy B. Yu., 1989. Kompyuternaya lingvistika: modelirovanie yazykovogo obscheniya [Computational linguistics: modeling language communication], Novoe v zarubezhnoy lingvistike, Moscow, 24. (in Russian).

7. Marchuk Yu. N., 2007. Kompyuternaya lingvistika [Computer linguistics], Moscow, 317 p. (in Russian).
8. Pirogova Yu. K., 2000. Reklamnyy tekst: Semiotika i lingvistika [Promotional Text: Semiotics and Linguistics], Moscow, 263 p. (in Russian).
9. Hrolenko A. T. Osnovy lingvokulturologii: uchebnoe posobie [Basics of linguoculturology: study guide], Flinta, Nauka, Moscow, 184 p. (in Russian).
10. Shirokov V. A., 2004. Fenomenolohiya leksykohtrafichnykh system [Phenomenology of lexicographic systems], Naukova dumka, Kyiv, 327 p. (in Ukrainian).
11. Yudin E. G., 1978. Sistemnyy podkhod i printsip deyatelnosti, Nauka, Moscow. (in Russian).
12. Proekt VAAL. – VAAL. [online] Available at: <<http://www.vaal.ru/index.php>>. [Accessed 25 Nov. 2019]. (in Russian).

Received 05.01.2020

H.V. Khodiakova, Doctor of Philosophy in Education, Associate Professor at the Department of Information Technology of the V. A. Sukhomlinsky Nikolayev National University, Mykolaiv, 11, Shneerson Str., Apt. 7, khodiakovagalina@gmail.com

INFORMATION TECHNOLOGIES IN SOLVING COMPUTER LINGUISTICS PROBLEMS

Introduction. At present there are many scientific researches devoted to the problem of computer linguistics which describe it at the level of general scientific methodology. In our article, we study the possibilities of modern computer means for text information analysis in computer linguistics tasks and the process of solving some applied linguistic tasks at the technological level.

The purpose of this article is to describe a new technology for solving some linguistic problems by using popular online resources and using a comparative method.

Results. The article gives a review of popular online text processing and analysis services. The technology of solving the following linguistic tasks is described: semantic text analysis; calculation of text frequency characteristics; finding typological Greenberg indexes for a given language, parsing a word by composition, grammatical text analysis, building syntactic graphs.

Conclusion. The results of the considered problems solution can be used for other studies in the field of linguistics. In particular, a preliminary frequency analysis of words, word combinations and turns of speech is made when defining an author of the text. Semantic text analysis is used by SEO-specialists, whose work is to optimize sites for search queries. The results of calculating typological indices are needed by specialists in theoretical linguistics. Grammatical text analysis and the construction of semantic graphs are needed when creating translator programs, as well as programs for speech recognition and automatic text input. The article is intended for linguists and specialists in the field of SEO-analysis, solving problems on analysis and text processing.

Keywords: *computer linguistics, computer text processing, text analysis, tasks of linguistic text analysis.*

Г.В. Ходякова, канд. пед. наук, доцент кафедры информационных технологий,
Николаевский национальный ун-т им. В.А. Сухомлинского,
г. Николаев, ул. Шнеерсона, 11, кв. 7,
khodiakovagalina@gmail.com

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В РЕШЕНИИ ЗАДАЧ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Введение. В настоящее время существует большое количество научных работ, посвященных компьютерной лингвистике. Описанные в них проблемы обсуждаются на уровне общенаучной методологии. Рассматриваются возможности современных компьютерных средств для анализа текстовой информации в задачах компьютерной лингвистики и процесс решения некоторых прикладных лингвистических задач на технологическом уровне.

Цель статьи — описать новую технологию решения некоторых лингвистических задач путем использования популярных онлайн-ресурсов и на основе сравнительного метода.

Результаты. В статье приводится обзор популярных онлайн-сервисов по обработке и анализу текста. Описана технология решения следующих лингвистических задач: семантического анализа текста; вычисления частотных характеристик текста; нахождения типологических индексов Гринберга для данного языка, разбор слова по составу, грамматического анализа текста, построение синтаксических графов.

Выводы. Результаты решения рассмотренных задач могут использоваться для других исследований в области лингвистики. В частности, при определении автора текста предварительно проводится частотный анализ слов, словосочетаний и оборотов речи. Семантический анализ текста используется *SEO*-специалистами, работа которых заключается в оптимизации сайтов под поисковые запросы. Результаты вычисления типологических индексов необходимы специалистам по теоретической лингвистике. Грамматический анализ текстов и построение семантических графов нужны при создании программ-переводчиков, а также программ по распознаванию речи и автоматическому вводу текста. Статья предназначена для лингвистов и специалистов в области *SEO*-анализа, решающих задачи по анализу и обработке текстов.

Ключевые слова: компьютерная лингвистика, компьютерная обработка текста, анализ текста, задачи лингвистического анализа текста.