Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://doi.org/10.36930/40340213
Article received 26.02.2024
Article accepted 04.03.2024
UDC 004.8/.9:001.103.2]:81'33
Correspondence author: I. V. Kozak ivan.v.kozak@lpnu.ua
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

***І. В. Козак, Н. Е. Кунанець***
*Національний університет "Львівська політехніка", м. Львів, Україна*

# ПРОБЛЕМИ РОЗРОБЛЕННЯ ТЕКСТОВИХ КОРПУСІВ ЗАСОБАМИ ІНФОРМАЦІЙНИХ СИСТЕМ І ШЛЯХИ ЇХ ВИРІШЕННЯ

Відзначено, що актуальність побудови інформаційних систем для формування та підтримки текстових корпусів зумовлена зростанням кількості методів і засобів аналізу текстової інформації для конкретних рівнів лінгвістичного дослідження, а також обсягів текстових матеріалів для їх опрацювання. З'ясовано, що невпинно зростають вимоги до якості метатекстової інформації, її глибини та рівнів лінгвістичного опису, котрі зумовлені використанням таких корпусів з внесеною мета-інформацією для використання в подальших лінгвістичних дослідженнях та організації моделей машинного навчання. Спостережено тенденцію до використання алгоритмів машинного навчання для введення розмітки, а також під час аналізу "чистих" корпусів. Опрацьовано низку наукових праць стосовно створення текстових корпусів та практичних рекомендацій під час розроблення текстового корпусу. Виділено етапи побудови лінгвістичних текстових корпусів, з погляду розроблення інформаційної системи та проаналізовано процеси формації корпусу на кожному з етапів. На кожному з етапів проаналізовано виклики та проблеми, котрі постають перед корпусними лінгвістами під час створення текстового корпусу, можливості й обмеження індивідуальних розрізнених підходів до їх вирішення. Опрацьовано публікації, котрі описують розроблення архітектури, використання засобів та підходи до розроблення конкретних корпусів текстів. Виокремлено рішення, котрі володіють більшою кількістю переваг та успішно застосовують під час роботи з текстовими корпусами. На підставі детального аналізу процесів створення корпусу сформульовано вимоги на кожному з етапів розроблення корпусу, а також до інформаційної системи на високорівневому рівні. Запропоновано діаграму діяльності інформаційної системи для розроблення текстових корпусів. Результати дослідження доцільно використовувати для побудови інформаційних систем, які б давали змогу розробляти та підтримувати корпусі тексти. Подальші дослідження авторів будуть спрямовані на створення інформаційних моделей, аналіз новітніх індивідуальних рішень під час розроблення корпусів текстів і можливості їхньої інтеграції у інформаційну систему та проектування системи підтримки роботи з текстовими корпусами.

**Ключові слова:** корпусна лінгвістика; інформаційні технології; прикладна лінгвістика; лінгвістичний корпус; метадані; лінгвістичний аналіз.

## Вступ / Introduction

Корпусна лінгвістика (англ. *Corpus Linguistics*) – це розділ мовознавства, який вивчає мову шляхом аналізу великих колекцій текстів, відомих як корпуси. Зазвичай, під текстовим корпусом розуміють сукупність текстів природної мови, поданих у електронному вигляді, які якимось способом організовано і призначено для практичного і наукового вивчення цієї мови [10].

Синклер Джон [20] подає таке формулювання цього концепту: "корпус – це набір фрагментів мовного тексту в електронній формі, відібраних відповідно до зовнішніх критеріїв для представлення, наскільки це можливо, мови чи мовного різновиду як джерела даних для лінгвістичного дослідження".

Виходячи з попередніх визначень, а також розуміння того, що природні мови розвиваються надзвичайно динамічно, то і корпуси текстів мають оновлюватися паралельно з мінімальними затримками. Тому на сьогодні неможливо уявити корпус текстів, котрий укладено не електронно, а також використовуючи суто ручне опрацювання текстової інформації. Отож, можемо сформулювати твердження, що текстовий корпус – це база даних, що містить електронний масив текстів чи їх фрагментів, застосовану до нього систему розмітки, яка дає змогу вносити метатекстову інформацію, а також засоби для опрацювання та ефективного використання внесених знань для використання у подальших лінгвістичних дослідженнях.

Зі швидким розвитком моделей опрацювання текстової інформації та методів генераційного штучного інтелекту, потреба у правильно сформованих та анотованих корпусах стрімко зростає, особливо для тих мов, для котрих статична вибірка корпусів текстів є відносно малою, і для української мови.

*Об'єкт дослідження* – розроблення текстових корпусів засобами інформаційних систем.

*Предмет дослідження* – методи і засоби розроблення текстових корпусів українською мовою, що дасть

**Інформація про авторів:**
**Козак Іван Володимирович**, аспірант, кафедра інформаційних систем та мереж.
**Email:** ivan.kozak.lp@gmail.com; https://orcid.org/0009-0007-4953-2816
**Кунанець Наталія Едуардівна**, д-р наук із соц. комунікацій, професор, кафедра інформаційних систем та мереж.
**Email:** nek.lviv@gmail.com; https://orcid.org/0000-0003-3007-2462

**Цитування за ДСТУ:** Козак І. В., Кунанець Н. Е. Проблеми розроблення текстових корпусів засобами інформаційних систем і шляхи їх вирішення. Науковий вісник НЛТУ України. 2024, т. 34, № 2. С. 101–108.
**Citation APA:** Kozak, I. V., & Kunanets, N. E. (2024). Challenges in creating text corpora using information systems and ways to solve them. *Scientific Bulletin of UNFU*, 34(2), 101–108. https://doi.org/10.36930/40340213

змогу вирішити проблеми корпусних лінгвістів на кожному етапі їх створення.

*Мета роботи* – проаналізувати проблеми розроблення текстових корпусів засобами інформаційних систем, що дасть змогу з'ясувати шляхи їх вирішення.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

- розглянути особливості використання методів і засобів під час розроблення корпусів текстів, зокрема українською мовою;
- виокремити виклики та проблеми, котрі знаходяться перед корпусними лінгвістами на кожному з етапів створення корпусу;
- визначити вимоги до інформаційної системи для створення та підтримки корпусу текстів.

**Аналіз останніх досліджень та публікацій.** Автори у публікації [1] описують процес створення корпусу CCOHA (англ. *Clean Corpus of Historical American English*) шляхом алгоритмічної вивірки попередньо розміченого корпусу COHA (англ. *Corpus of Historical American English*), де особливу увагу звертають на використання нових тег-сетів, алгоритмів вивірки внесеної розмітки, обмеженостей тестового корпусу при внесенні змін до нього для дотримання авторського права та специфіки використання бази даних корпусу COHA. У роботі [5] розглянуто особливості та технологічні особливості створення корпусу української мови, зокрема використання нереляційних баз даних та існування паралельних версій текстового матеріалу. Увагу приділяють нормуванню процесів попереднього опрацювання текстів на докорпусному етапі [17]. Водночас вивірка розмітки залишається важливим, однак трудомістким процесом [12, 15], оптимізацію, якого намагаються вирішувати, зокрема, і автоматизацією. Джерелом інформації для корпусу дедалі частіше вказують веб, а способом отримання такої інформації – використання веб-скрепінгу (англ. *Web-Scrapping*) [2, 5, 13]. Важливою особливістю виокремлюють [7] уніфікацію та гармонізацію застосування мовних анотацій та токенізації текстів, котрі часто виникають через застосування різних підходів, а також трансформації лінгвістичних анотацій для отримання репрезентації іншого типу чи якості, котрі мають здійснюватися на підставі правил в автоматизованому чи напівавтоматизованому режимі, а не шляхом машинного навчання, адже такі анотації стануть золотими даними (англ. *Gold Data*) для подальшого навчання анотаторів чи NLP (англ. *Natural Language Processing*) [6]. З іншого боку, з розвитком штучного інтелекту, його можливостей та стійкою високою присутністю ШІ (англ. AI, *Artificial Intelligence*) в інформаційному полі, потрібно відзначити велику кількість праць щодо використання ШІ у корпусній лінгвістиці. Розглядають можливості використання великих мовних моделей (англ. *Large Language Models*, LLM) [3] та моделей глибинного навчання в синергії з традиційними лінгвістичними моделями [14]. Прикладні лінгвісти аналізують використання генеративного штучного інтелекту під час створення корпусу та введення розмітки [22], використанні його для аналізу "чистих" корпусів без введення розмітки [19] та обмежень таких підходів [9].

## Результати дослідження та їх обговорення / Research results and their discussion

Для того, щоб сформулювати вимоги до інформаційної системи побудови корпусів мови, розглянемо особливості їх формування. Створення корпусу тексту для аналізу мови має низку характерних особливостей, які забезпечують швидкий розвиток корпусної лінгвістики [23]:

1) використання комп'ютерних та інформаційних технологій для вивчення лінгвістичного матеріалу;
2) застосування методів аналізу тексту, заснованих на частотному та статистичному математичному підходах;
3) використання технологій, які забезпечують аналіз величезних обсягів лінгвістичної інформації в режимі реального часу;
4) впровадження засобів зберігання аналітичних даних, котрі дають змогу їх подальшого використання для попереднього навчання моделей (pre-trained models);
5) наявність ефективних інструментів для дослідження мовних моделей у реальній реалізації (вивчаються реальні тексти мовлення, уникаючи інтуїтивно-прогностичних висновків без будь-яких практичних доказів);
6) обґрунтованість конкретними фактами та статистичними розрахунками, що дають змогу позбутися суб'єктивності;
7) переоцінка традиційних підходів і використання інноваційних для вивчення мови та формування теорії лінгвістичної науки.

До створення текстового корпусу певної предметної галузі дослідник або група дослідників визначають об'єкт дослідження та мовний матеріал, на якому буде ґрунтуватися таке дослідження.

Перш ніж почати роботу з інформаційною системою задля створення нового чи роботи з наявним корпусом текстів, користувач має пройти реєстрацію (якщо така не була проведена попередньо) або авторизуватися у системі. Такий процес сприятиме безпеці введених даних, зменшить неконтрольоване використання серверів, забезпечить можливості для здійснення оплати за користування системою (за потреби) та поширення корпусу серед певного кола спеціалістів.

Логінація, авторизація та автентифікація користувачів загалом не є проблемним питанням під час створення корпусів, а отже і об'єктом цього дослідження, тому ми не будемо звертати увагу на виклики цього процесу, однак відзначимо низку функціональних вимог до нашої інформаційної системи.

Вимоги до інформаційної системи підтримки корпусів на етапі реєстрації користувачів повинні враховувати такі особливості:

1. *Збір інформації.* Система має забезпечувати можливість збирання необхідних даних від користувачів, наприклад ім'я, прізвище, адреса електронної пошти, контактний номер телефону, організація тощо.
2. *Автентифікація та авторизація.* Для забезпечення безпеки інформаційної системи, необхідно створити можливість перевірки і підтвердження ідентифікації користувачів.
3. *Перевірка введених даних.* Система має забезпечувати перевірку введених даних користувачів на правильність та достовірність (форматів, наявності спецсимволів у паролі та інше).
4. *Унікальний ідентифікатор.* Кожен користувач повинен мати унікальний ідентифікатор в системі для подальшого доступу і отримання послуг.
5. *Управління доступом.* Система має мати можливість керувати рівнем доступу користувачів до інформації. Це може містити різні ролі користувачів з різними правами доступу до функціональності системи чи/та конкретних корпусів.

Враховуючи ці вимоги, інформаційна система підтримки корпусів на етапі реєстрації користувачів має

сприяти безпечній та зручній реєстрації та управлінню даними користувачів.

Наступним етапом використання системи у межах роботи з корпусом є власне створення сутності корпусу. На цьому етапі одним з найскладніших завдань для корпусних лінгвістів є додавання правильної та обширної інформації про сам корпус, адже саме завдяки ній користувачі зможуть знайти та використовувати цей корпус для подальшого аналізу.

Серед вимог до інформаційної системи на етапі створення сутності корпусу потрібно виокремити такі:

1. *Надання доступу.* Система має забезпечувати можливість надання доступу до корпусу відповідним користувачам з урахуванням їхній ролей.
2. *Опис корпусу.* Інформаційна система має мати можливість додавання інформації про корпус, такої як назва, опис, мова, категорії або тематики, до яких він належить. Це допоможе користувачам зорієнтуватися у наявних корпусах та знайти той, який відповідає їхнім потребам.
3. *Присвоєння тегів.* Важливою функціональністю інформаційної системи є можливість присвоєння тегів для подальшого пошуку корпусу. Теги можна використати для класифікації корпусу за певними характеристиками, такими як тематика, автор, рік створення і т.ін.
4. *Попередній вибір системи тегування.* Система має надати користувачам можливість вибору системи тегування, яка відповідає їхнім потребам і завданням. Це можуть бути готові системи тегування, які використовують в схожих проектах, або можливість створення власних систем тегування з урахуванням специфічних вимог корпусу.


**Рис. 1.** Алгоритм створення текстового корпусу / Algorithm for creating a text corpus

Серед усіх завдань, виконання яких необхідне для створення текстового корпусу та збирання матеріалу для нього, В. В. Жуковська [24] називає дев'ять основних. Беручи за основу зазначені завдання, сформуємо алгоритм (див. рис. 1). Під час докладного аналізу кожного кроку алгоритму зосередимося на потенційних проблемах, питаннях і труднощах.

*Крок 1. Обрання типу вибірки та пошук лінгвістичного матеріалу.* Відбираючи матеріал для текстового корпусу, перед мовознавцями постає низка потенційних проблем.

По-перше, теоретичні основи корпусної лінгвістики, а також практичні міркування зобов'язують дослідників відбирати таку кількість лінгвістичних текстів, яка дає змогу охарактеризувати цей текстовий корпус за критерієм повноти. Саме повнота текстового корпусу є такою характеристикою, яка зумовлює потребу створення достатньо обґрунтованого щодо вичерпності та вибірковості зібрання матеріалів мови.

По-друге, текстовий корпус також потребує репрезентативності. Це означає, що вибірка має бути орієнтована на здатність обраної сукупності відтворювати основні характеристики цілісної сукупності і містити багатотипові тексти, які відповідають необхідним параметрам, заданими дослідниками.

Репрезентативності текстового корпусу (для великих корпусів) можна досягти за допомогою використання певних технік відбору текстового матеріалу. Серед останніх потрібно виокремити такі [11]:

- *пропорційна* – матеріал вводять до корпусу відповідно до ваги того чи іншого типу тексту у мові;
- *стратифікована* – відбір полягає у внесенні однакових за обсягом фрагментів різних типів, незалежно від їхньої ваги в мові, класифікації текстів за певними критеріями та їх поділу на "страти" або групи;
- *суцільного відбору з подальшою оцінкою репрезентативності* – до корпусу поступово вводять тексти, які репрезентують ті чи інші явища, релевантні для цього корпусу;
- *випадкової вибірки* – дослідники попередньо формулюють вимоги до об'єму корпусу, його структури та параметрів таких як обсяг, структура, тема, автор тощо.

Незважаючи на обраний підхід до формування текстової вибірки корпусу, на цьому етапі для дослідників важливим є мати розуміння того, як в текстовому корпусі репрезентовані ті чи інші явища, релевантні для цього корпусу.

Отже, інформаційна система для ефективної організації роботи з корпусом має підтримувати тегування чи збереження будь-яким іншим чином метаданих загальнотекстового рівня та забезпечувати аналітику на підставі цих даних відповідно до параметрів корпусу.

*Крок 2. Введення даних.* На етапі введення тексту, вибрані тексти перетворюють в електронний формат, оскільки однією з основних характеристик сучасного текстового корпусу є подання у такому форматі, щоб проводити автоматизоване опрацювання таких текстів, для формування корпусу.

Серед способів введення даних можна виокремити такі [24]:

- адаптація текстів, котрі уже репрезентовані в електронному форматі. Такий спосіб є найпростішим та, зазвичай, породжує меншу кількість дій на подальших етапах;
- перетворення текстів у паперовому варіанті за допомогою їх сканування. Цей спосіб використовують у тих випадках, коли необхідні текстові матеріали доступні тільки у паперовому вигляді, одначе їх репрезентація дає змогу ефективно провести цифрове розпізнавання;
- ручне введення текстового матеріалу є найбільш ресурсомістким способом. Одначе це єдиний спосіб переведення

текстової інформації в електронну форму, якщо його поточна репрезентація не дає змоги провести сканування (рукописи, пошкоджені сторінки і т.ін.).

Два останні способи зазвичай породжують більшу кількість помилок, а отже потребуватимуть ще більших витрат часу на наступних етапах.

Одним з найбільш поширених способів наповнення корпусу текстами певного жанру чи типу, котрі уже існують в електронному форматі, є використання так званих кроулерів (англ. *Web Crawler*) – пошукових роботів, котрі аналізують web-сторінки певного ресурсу чи групи ресурсів і зберігають знайдену інформацію [5]. Такий спосіб є одним з відносно швидких та дешевих способів наповнення корпусу, однак, водночас, володіє низкою обмежень, таких як обмеження текстів опублікованих у веб-мережі за довжиною, їх форматування та індексації.

Окрім цього, в мережі Інтернет є багато документів і текстів, які поширюються месенджерами та електронною поштою у вигляді як повідомлень, так і вкладень. Підписуючись на відповідні розсилки чи канали, колекцію тестових матеріалів певних стилів можна швидко збільшувати.

Окремим способом отримання текстового матеріалу, наявного в електронному форматі, варто назвати використання великих архівів текстів або дампів (англ. *Dump*) даних [20]. За отримання таких масивів даних, іноді вимагають оплату, особливо якщо це популярний і захищений авторським правом матеріал. У такому випадку розробникам корпусу потрібно ретельно розглянути вартість таких даних, їхню виправданість, можливість діалогу з власниками ресурсу. При цьому, потрібно зазначити, що використання таких даних особливо варто розглядати під час створення корпусів, матеріали яких плановано будуть захищені авторським правом та поширюватимуться на підставі платних ліцензій.

Виходячи зі способів отримання текстового матеріалу, можемо визначити такі вимоги до інформаційної системи на етапі внесення текстів:

- забезпечити завантаження текстів у різних форматах, модифікування та стандартизацію їх назв;
- дозволити вводити користувачу тексти вручну в затребуваних форматах;
- використовувати модульну структуру для потенційного підключення зовнішніх кроулерів.

*Крок 3-4. Попереднє опрацювання, конвертування і графематичний аналіз текстових даних.* На цьому етапі проводять перевірку всіх поданих природномовних текстів на предмет філологічних невідповідностей та їх виправлення. Також текстові дані зазнають ітераційних змін у форматуванні та піддаються конвертуванню за потреби.

Серед недоліків, отриманих під час переведення тексту в електронне його подання, потрібно виокремити такі:

- дрібні механічні помилки та одруки, зокрема і помилки, отримані під час розпізнавання відсканованого тексту;
- наявність зайвого форматування, табулювання, знаків переносу, розривів сторінок і т.ін. [24];
- наявність веб- та іншої нерелевантної розмітки;
- присутність зайвої для веб-сторінок текстової інформації, помилково скопійованих кроулерами, такої як "завантажити тут" чи кодів HTTP помилок з їх текстовим поясненням (до прикладу, "404 Not Found");
- дрібні текстові фрагменти, на зразок анотацій, бібліографічних описів чи метадані, котрі були помилково завантажені кроулерами [5];
- різнотипність форматування та кодування текстів, котрі були отримані в електронному форматі.

На цьому етапі текст також повинен бути проаналізованим на наявність у ньому рисунків, таблиць та інших нетекстових елементів. Усі такі елементи необхідно видалити з тексту, а текстова інформація має бути синтезованою з таких елементів за потреби.

Отже, інформаційна система на цьому етапі має забезпечити такий функціонал:

- сувору послідовність дій для забезпечення коректної та оптимальної роботи алгоритмів;
- опрацювання візуальних елементів і переведення їх до текстового формату;
- збереження тезаурусів певної мови для подальшого використання під час філологічних перевірок.

*Крок 5. Розмітка тексту.* Наступним кроком під час створення текстового корпусу є визначення та конструювання архітектури системи введення метаінформації.

Деякі дослідники вважають, що анотація для корпусу є зайвою та правильно є працювати з неанотованим корпусом, який не має додаткової внесеної інформації, котра може бути суб'єктивною [18]. Інші ж вважають, що внесення метаданих корпусу текстів є однією з невід'ємних його характеристик та дає змогу отримати низку переваг, таких як пришвидшений пошук, отримання точних аналітичних даних, багаторазове та мультицільове використання внесеної мета-інформації для подальших досліджень на підставі корпусу [8].

У випадку, якщо внесення метаданих до певного корпусу, лінгвісти вважають доцільним, цей процес розбивають на декілька етапів.

На першому етапі відбувається перехід від текстової інформації до формування метатекстових даних. Цей процес потребує особливої точності та уваги лінгвістів, адже неправильно розроблена архітектура системи ставить під питання подальшу доцільність розроблення, існування та експлуатації корпусу.

Розмітку документа на високорівневому рівні можна поділити на такі підвиди (див. рис. 2) [16]:

1. Розмітка рівня документа (заголовки, бібліографія, кодування тощо).
2. Структурна розмітка верхнього рівня (томи, нетекстові структури).
3. Структурна розмітка нижнього рівня (речення, слова і т.ін.).
4. Лінгвістична анотація тексту та його елементів (частин мови, лем, систематизація, граматична анотація і т.ін.).

Серед основних властивостей якісного підходу до визначення архітектури розмітки, науковці [4, 18] виокремлюють такі:

- металінгвістична інформація має бути виокремлюваною;
- наявність документації про використану архітектуру розмітки;
- підходи до анотації мають ґрунтуватися на лінгвістичному консенсусі;
- використання наявних стандартів і підходів до введення анотації.

На другому етапі виконують структурне визначення тексту – іншими словами, його поділ на розділи, сторінки, абзаци, речення, слова чи інші частини, котрі будуть піддаватися подальшій розмітці та аналізу. На цьому етапі потрібне чітке визначення структурних частин та способів їх ідентифікації.

Після визначення типів розмітки та створення відповідної системи тегів метатекстова інформація вводиться в текст. Основною складністю на цьому кроці є визна-

чити такий спосіб присвоєння метатекстової інформації, котрий буде водночас і репрезентативним для подальшого аналізу, так і достатньо легким та швидким для подальшого аналізу матеріалів. Серед таких способів можемо виокремити такі: просте додавання (кодування слідує за спецсимволом) [18], табличний спосіб (структурні елементи вводяться в таблицю бази даних), використання гіперпосилань та внесення метаданих за допомогою мов розмітки як в основний, так і в додаткові файли [16].

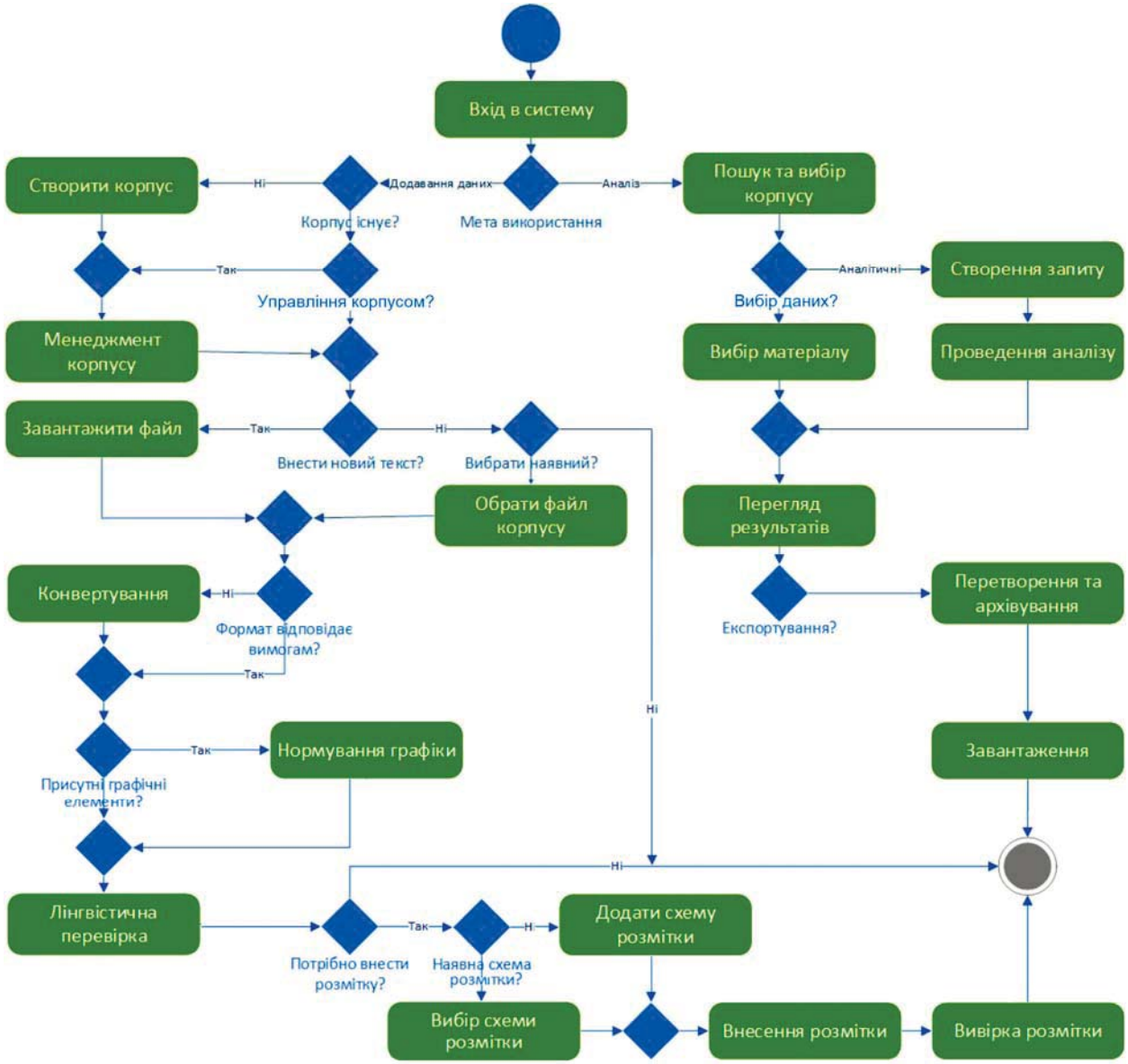


**Рис. 2.** Діаграма діяльності інформаційної системи підтримки текстових корпусів / Activity diagram of the informational system for text corpora management

Отже, інформаційна система на етапі внесення метаданих має задовольняти такі вимоги:

- забезпечити використання одного з альтернативних способів введення розмітки на вибір лінгвіста;
- надати можливість вибрати та задокументувати використовувану архітектуру розмітки, включно з DTD;
- автоматично вносити ту частину розмітки, котра є можливою для тієї чи іншої мови чи типу розмітки;
- створити можливості для роботи над розміткою у синхронному форматі для команди мовознавців;
- забезпечити версійність введеної розмітки.

*Крок 6. Коректування результатів розмітки.* Для більшості цілей, особливо якщо корпус великий і його потрібно зробити доступним для загального користування, важливо перевірити анотацію. Засоби автоматичної розмітки можуть досягти певного рівня правильності міток. Однак такий результат часто не є достатнім, тому автоматичне додавання тегів часто супроводжується етапом пост-редагуванням внесеної розмітки, на якому люди-аналітики виправляють помилки, отримані під час автоматичного додавання тегів і усувають будь-які неоднозначності [18]. Для забезпечення послідовності під час введення та коригування анотації часто запрошують корекцію анотації у двох лінгвістів і визначають, якою мірою вони погоджуються один з одним.

Отже, інформаційна система на етапі коректування анотації має:

- забезпечити паралельну роботу двох чи більше лінгвістів над цим самим фрагментом тексту;
- проводити аналітику за результатами корекції для прийняття узгодженого рішення;
- надати можливість документування прийнятих чи неузгоджених рішень;
- вести статистику помилок для подальшої аналітики, коректування коду та пріоритезації завдань коригування;
- за можливості, пріоритезувати коригування фрагментів, де точність введення розмітки є нижчою (якщо така інформація існує).

*Крок 7. Введення анотованих текстів у структуру корпусного менеджера.* На етапі підготовки текстового файла для опрацювання та використання виникає потребу використання корпусного менеджера – програми, яка здатна прочитати розмітку чи метадані і представити їх у зручній формі. Текстовий файл з тегами зазвичай не має додаткового значення для звичайного користувача у візуальному сприйнятті, натомість розмітка, що вставлена в текст, ускладнює його читання. Зазвичай корпусний менеджер розробляють спеціально для конкретного текстового корпусу, оскільки він повинен мати функціонал для аналізу розмітки і здатність виводити необхідну інформацію.

На цьому етапі також виконують усе доцільне попереднє опрацювання текстів, яке потребуватиме пошукового програмного забезпечення. Цей крок можна пропустити для невеликих корпусів, однак для великих корпусів, наприклад, національних, він є життєво необхідним.

Уся проаналізована інформація має бути збережена у спеціалізованих базах даних на кожному з рівнів (індивідуального тексту, групи текстів, жанру, підкорпусу, корпусу чи рівнів інших визначених структурних елементів корпусу). Під час введення нових текстів у корпус така інформація має бути оновлена. Отже, кінцевий користувач зможе отримувати інформацію з мінімальними затримками на аналіз, а вартість аналітики знизиться.

Для реалізації функціоналу із формування корпусу на цьому етапі інформаційна система має забезпечити ефективне функціонування конкордансера відповідно до заданої архітектури розмітки, а також проводити попередні аналітичні операції для внесених змін.

*Крок 8. Дистрибуція та підтримка корпусу.* У всіх зазначених вище етапах роботи з лінгвістичним корпусом також необхідна розподіленість текстового корпусу та можливість паралельної роботи над його елементами. Оскільки головною метою створення текстового корпусу є вивчення мови та її функціонування, а також використання в різних предметних галузях, які вимагають аналітичного опрацювання даних, корпус повинен мати змогу розширюватись та забезпечувати можливість спільного використання для аналітиків, які працюють над великим обсягом лінгвістичних даних. З цього приводу виникають питання щодо онлайн-внесення змін та врегулювання конфліктів, що виникають під час внесення таких змін.

Окрім цього, доступ до вихідних корпусних файлів (текстів із розміткою, файлів архітектури розмітки, документації) має бути доступним тільки для зареєстрованих та авторизованих користувачів, адміністраторів корпусу. В іншому випадку виникає ризик незаконного копіювання опрацьованих даних [21]. Відповідно, система має забезпечувати можливості додавання користувачів, визначення їхніх ролей, ієрархічну систему підпорядкування та визначення прав користувачів, обмеження та надання доступів до певних корпусів чи їх структурних елементів, забезпечувати захист від неавторизованого доступу.

По-друге, потрібно забезпечити процедури резервного копіювання під час збирання даних і етапу розроблення проекту побудови корпусу. Після того, як корпус буде завершено, його необхідно архівувати. Тут важливо розрізняти між резервним копіюванням та архівуванням. Резервне копіювання означає створення періодичної копії сховища файлів. Засоби архівування забезпечують перенесення інформації, що має важливе значення (особливо для дистрибуції), в окреме сховище, де вона буде зберігатися на невизначений термін або на погоджений період часу.

Окрім цього, важливо зберігати вихідні файли в тому стані, у якому їх було отримано для створення так званого "single source of truth". Такі тексти система має зберігати у форматі звичайного тексту, адже у крайніх випадках текст може бути відхилений, якщо його форматування не можна стандартизувати. Такий підхід рекомендовано, навіть якщо запланований пакет опрацьовуватиме мову розмітки, як-от HTML, XML, SGML, або вихідний файл текстового процесора (до прикладу MS Word).

***Функціональні вимоги до інформаційної системи.*** Функціональні вимоги до інформаційної системи для створення корпусів текстів і їх розмітки:

- *Створення корпусів текстів*. Можливість завантаження текстових даних у різних форматах (текстові файли, документи PDF, веб-сторінки тощо). Можливість автоматичного виділення мовних корпусів з текстів різних мов.
- *Функція опрацювання текстів*, в т.ч. й очищення від зайвих символів, токенізацію та лематизацію. Можливість створення корпусів з різних документальних джерел і додавання метаданих до текстів.
- *Розмітка текстів.* Можливість створення або використання наявних схем розмітки для позначення різних типів сутностей або атрибутів у тексті (наприклад, іменованих сутностей, синтаксичних структур тощо).
- *Створення інтерфейсу для ручного або півавтоматичного маркування текстів* за допомогою схем розмітки. Можливість перегляду та редагування розмічених текстів для перевірки правильності розмітки.
- *Функція збереження та експорту розмічених текстів і усього корпусу* у різних форматах, наприклад, XML, JSON, CSV, BZ2, LZMA та ZIP, RAR відповідно.
- *Управління корпусами та розміченими даними.* Можливість створення та управління різними корпусами текстів і їх розміченими версіями.
- *Функція пошуку та фільтрації корпусів за різними параметрами* (наприклад, мова, джерело, статус розмітки). Можливість створення та управління користувацькими ролями та доступом до корпусів і розмічених даних.
- *Аналіз та візуалізація даних*. Функції аналізу статистичних характеристик корпусів текстів (наприклад, розподіл частот слів, довжина текстів тощо). Можливість візуалізації розмічених даних у вигляді графіків, діаграм або хмари тегів для отримання інсайтів.
- *Функція статистичного порівняння різних корпусів текстів* або розмічених даних.

Інформаційна система, що відповідатиме цим функціональним вимогам, допоможе забезпечити ефективне створення, управління та аналіз текстових корпусів з розміченими даними (див. рис. 2).

**Обговорення результатів дослідження.** Проблематику створення корпусів текстів висвітлено у низках наукових праць, зокрема у публікаціях [1, 5] автори описують процес створення конкретного корпусу текстів, де зазначають специфіку використання систем тегування, алгоритмів вивірки внесеної розмітки та обмеженостей створених текстових корпусів. Відзначають потребу пошуку і застосування набору індивідуальних засобів для розроблення корпусу, а також – модифікації наявних алгоритмів для пристосування для роботи з корпусами українською мовою чи врахування особливостей певного корпусу. Значна кількість засобів поширюють-

ся на умовах відкритого доступу (англ. *open source*) з можливістю доступу до вихідного коду, котрий часто лежить в відкритих гіт-репозиторіях, а отже – можливостями адаптації коду та застосування у інтегрованих системах. Таку особливість пов'язуємо з численними зусиллями прикладних лінгвістів розробляти спеціалізовані некомерційні корпуси текстів для подальших досліджень.

У дослідженнях автори приділяють увагу уніфікації та нормуванню процесів розроблення текстових корпусів на різних етапах [17, 20, 24]. Алгоритм створення корпусу, запропоновано у цій роботі корелюється із алгоритмами, які можна синтезувати із наукових праць, присвячених загальній методології створення лінгвістичних корпусів [4, 18, 20], проте містить поправки, котрі дають змогу інформаційній системі оптимізувати залежні процеси під час розроблення корпусів, та будуть враховувати тенденції під час отримання текстового матеріалу зі застосуванням веб-кроулерів [2, 13], потреби у вивірці введеної розмітки із застосуванням підходу паралельної вивірки з подальшою синхронізацією результатів [12] для отримання більш якісного матеріалу текстового корпусу.

Запропонований метод розпаралеленого зберігання роззначених текстових масивів [16] з використанням баз даних різних типів враховує обмеження та досвід застосування як реляційних [1], так і нереляційних баз даних [5] та матиме переваги під час аналізу метатекстової інформації одного типу. Водночас для комбінованої аналітики такий підхід буде мати певні обмеження, над вирішенням котрих автори планують працювати надалі.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

*Наукова новизна отриманих результатів дослідження* – уніфіковано набір вимог до інформаційних систем підтримки корпусів текстів, розроблено алгоритм створення корпусів текстів, адаптований під використання інформаційної системи.

*Практична значущість результатів дослідження* – структурований набір вимог, можна буде використовувати як для розроблення, так і для оцінювання якості створених інформаційних систем для підтримки роботи з текстовими корпусами. Застосування запропонованого алгоритму допоможе оптимізувати процеси створення текстового корпусу з використанням інформаційних систем.

## Висновки / Conclusions

Унаслідок виконання роботи проаналізовано проблеми розроблення текстових корпусів засобами інформаційних систем. За результатами дослідження можна зробити такі основні висновки:

1. Побудова таких інформаційних систем дає змогу централізовано керувати колекцією текстів, даними та метаданими, полегшує організацію, пошук і управління корпусом, а також спрощує спільну роботу багатьох користувачів, забезпечує швидкий доступ до різних частин корпусу та можливість ефективного пошуку текстів за допомогою різних критеріїв, таких як слова, фрази, метадані тощо. Це дає змогу швидко знаходити потрібну інформацію та проводити точні лінгвістичні аналізи. Така інформаційна система дає змогу збільшувати обсяг корпусу, додавати нові тексти та оновлювати наявні дані, використовуючи для цих операцій один із наявних засобів, забезпечуючи можливість їх параметризації чи додавання нових. Інформаційна система дасть змогу для кожного з етапів створення текстового корпусу використовувати ті самі підходи, що і для текстів, введених у корпус раніше для забезпечення уніфікації метаданих усього корпусу чи підкорпусу. Такий параметр важливий для довготривалих лінгвістичних досліджень, які потребують великого обсягу текстових даних. Основною особливістю є те, що інформаційна система може містити автоматизовані процеси для опрацювання корпусу, такі як анотація, токенізація, лематизація тощо.
2. Виокремлено етапи створення лінгвістичного текстового корпусу та виклики під час його реалізації. Адаптовано та оптимізовано алгоритм створення текстових корпусів з врахуванням застосування інформаційних систем до процесу, що дасть змогу отримати якісні зміни під час створення корпусу.
3. Синтезовано вимоги до інформаційної системи підтримки текстових корпусів. Така інформаційна система має забезпечувати централізоване управління, швидкий доступ, модульність, масштабованість та автоматизацію, що є важливими факторами для ефективного створення та управління текстовими корпусами.

Результати цього дослідження доцільно використовувати для побудови інформаційних систем для роботи з корпусами. Подальші дослідження авторів будуть спрямовані на створення інформаційних моделей та проектування системи підтримки роботи з текстовими корпусами.

***Подяка.*** Світлої пам'яті кандидату технічних наук, доценту Ігорю Кульчицькому за внесок у становлення комп'ютерної лінгвістики.

## References

1. Alatrash, R., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2020). CCOHA: Clean Corpus of Historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6958–6966. Marseille, France: European Language Resources Association. URL: https://aclanthology.org/2020.lrec-1.859/
2. Alves, D., Thakkar, G., & Tadić, M. (2022). Building and Evaluating Universal Named-Entity Recognition English corpus, 1–15. https://doi.org/10.48550/arXiv.2212.07162
3. Anthony, L. (2023). Corpus AI: Integrating Large Language Models (LLMs) into a Corpus Analysis Toolkit. Presentation given at the *49th Annual Conference of the Japan Association for English Corpus Studies*, Kansai University, Osaka, Japan. URL: https://osf.io/srtyd/
4. Burnard, L. (2004). Metadata for corpus work. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 40–57). Oxford: Oxbow Books. URL: https://users.ox.ac.uk/~martinw/dlc/chapter3.htm
5. Chaplynskyi, D. (2023). Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale. *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, 1–10, Dubrovnik. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.unlp-1.1
6. Chiarcos, C., & Fäth, C. (2019). Graph-Based Annotation Engineering: Towards a Gold Corpus for Role and Reference Grammar. *Open Access Series in Informatics*, 70(9), 1–9. https://doi.org/10.4230/OASIcs.LDK.2019.9
7. Chiarcos, C., & Schenk, N. (2019). CoNLL-Merge: Efficient Harmonization of Concurrent Tokenization and Textual Variation. *Open Access Series in Informatics* (OASIcs), 70(7), 1–7. https://doi.org/10.4230/OASIcs.LDK.2019.7
8. Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast!. *Applied Corpus Linguistics*. 3(3), 100066, 1–5. https://doi.org/10.1016/j.acorp.2023.100066

9. Curry, N., Baker, P., & Brookes, G. (2023). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics,* 4(1), 100082, 1–9. https://doi.org/10.1016/j.acorp.2023.100082
10. Darchuk, N. (2013). Corpus linguistics: problems, methods, perspectives: educational program. Kyiv: Publishing house of KNU. [In Ukrainian].
11. Demska-Kulchytska, O. (2005). Representativeness as a feature of the text corpus. *Ukrayinska mova*. 3, 100–107. [In Ukrainian]. URL: https://core.ac.uk/download/pdf/149237952.pdf
12. Dobrić, N. (2022). Identifying errors in a learner corpus – the two stages of error location vs. error description and consequences for measuring and reporting inter-annotator agreement. *Applied Corpus Linguistics*, 3(1), 100039, 1–11. https://doi.org/10.1016/j.acorp.2022.100039
13. Egbert, J., & Wood, M. (2023). The corpus of United States state statutes – design, construction and use. *Applied Corpus Linguistics*, 3(2), 100047, 1–15. https://doi.org/10.1016/j.acorp.2023.100047
14. Ganpat, S. C., et al. (2020). A two-step hybrid unsupervised model with attention mechanism for aspect extraction. *Expert Systems with Applications,* 161, 113673, 1–13. https://doi.org/10.1016/j.eswa.2020.113673
15. Hill, M., & Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities,* 34. 825–843. https://doi.org/10.1093/llc/fqz024
16. Ide, N. (2002). Encoding Linguistic Corpora., 9 p. URL: https://aclanthology.org/W98-1102.pdf
17. Kulchytskyy, I. (2020). Text normalization during pre-corpus preparation: experience of application. *Journal of Lviv Polytechnic National University. Ser. Information Systems and Networks*, 7, 51–58. URL: https://doi.org/10.23939/sisn2020.07.051
18. Leech, G. (2005). Adding linguistic annotation. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 17–29). Oxford: Oxbow Books. URL: https://users.ox.ac.uk/~martinw/dlc/chapter2.htm
19. Lin, P. (2023). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, 3(3), 100065, 1–10. https://doi.org/10.1016/j.acorp.2023.100065
20. Sinclair, J. (2004). How to build a corpus. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 96–101). Oxford: Oxbow Books. URL: https://users.ox.ac.uk/~martinw/dlc/appendix.htm
21. Wynne, M. (2004). Archiving, distribution and preservation. In M. Wynne (Ed.), *Developing linguistic corpora: a guide to good practice* (pp. 87–96). Oxford: Oxbow Books. URL: https://users.ox.ac.uk/~martinw/dlc/chapter6.htm
22. Zappavigna, M. (2023). Hack your corpus analysis: How AI can assist corpus linguists deal with messy social media data. *Applied Corpus Linguistics*, 3(3), 100067, 1–5. https://doi.org/10.1016/j.acorp.2023.100067
23. Zhukovska V. (2015). Corpus Linguistics: History and Current Status. In *Modern linguistic studies. Tutorial* (pp. 168–203). Zhytomyr: Publishing house of Ivan Franko ZhDU. [In Ukrainian]. URL: https://www.academia.edu/22835661/Корпусна_лінгвістика_історія_становлення_та_сучасний_стан
24. Zhukovska, V. (2013). Introduction to corpus linguistics: a study guide. Zhytomyr: Publishing house of Ivan Franko ZhDU. [In Ukrainian]. URL: http://eprints.zu.edu.ua/18909/1/korpusna_lingv.pdf

***I. V. Kozak, N. E. Kunanets***

*Lviv Polytechnic National University, Lviv, Ukraine*

## CHALLENGES IN CREATING TEXT CORPORA USING INFORMATION SYSTEMS AND WAYS TO SOLVE THEM

The relevance of constructing information systems for the creation and maintenance of text corpora is noted to be determined by the increasing number of methods and tools for analysing textual information at specific levels of linguistic research, as well as the volumes of the textual materials taken for the processing. It has been revealed that there is a growing demand for the quality of the metadata, its depth, and levels of linguistic description, which is driven by the usage of corpora with incorporated meta-information for further linguistic research and organization of machine learning models. The trend towards using machine learning algorithms for annotation and analysis of "clean" corpora is highlighted. Several scientific works dedicated to the development of textual corpora and practical recommendations for corpus construction have been examined. The stages of constructing linguistic text corpora from the perspective of information system development have been identified, and the corpus formation processes at each stage have been analysed as well. The challenges and issues encountered by corpus linguists in creating text corpora, as well as the possibilities and limitations of individual differentiated approaches to their resolution, have been analysed at every stage. Publications describing the development of the architecture, tools, and approaches for specific text corpora have been reviewed. Specific solutions that have more advantages and are successfully applied in working with text corpora have been identified. Based on a detailed analysis of corpus construction processes, requirements for each stage of corpus development, as well as for the information system at a high level, have been formulated. The activity diagram of the information system for text corpus development has been proposed. The research results can be used for development of the information systems to construct and maintain the text corpora. In future researches, authors aim to focus on creating information models, analysing innovative individual solutions in corpus development, exploring the possibilities of their integration into an information system, and designing a system to support work with a text corpus.

**Keywords:** corpus linguistics; information technology; applied linguistics; linguistic corpus; metadata; linguistic analysis.