Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

СТВОРЕННЯ ФОНЕТИЧНО РЕПРЕЗЕНТАТИВНОГО ТЕКСТУ ДЛЯ УКРАЇНСЬКОЇ МОВИ

Кваліфікаційна робота
на здобуття ОС «бакалавр»
студентки IV курсу
галузі знань 03 «Гуманітарні науки»,
спеціальність 035 «Філологія»
спеціалізації 035.10 «Прикладна лінгвістика», ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»
Анастасії Дмитрівни ПЛАСКАЛЬНОЇ

Науковий керівник:
асистент кафедри української мови та прикладної лінгвістики
Валентина РОБЕЙКО

«Допущено до захисту»
Протокол засідання кафедри
української мови та прикладної лінгвістики
від «5» 06 2026 року № 16
завідувач кафедри _____
к.філол.н., доц. Сергій РІЗНИК

КИЇВ – 2026

Зміст

Анотація

У бакалаврській роботі досліджено проблему створення фонетично репрезентативного тексту для української мови. Актуальність роботи зумовлена потребою розвитку сучасних мовних технологій, зокрема систем автоматичного розпізнавання мовлення, синтезу мовлення, створення мовленнєвих корпусів та фонетичних баз даних для української мови. Якість роботи таких систем значною мірою залежить від наявності мовного матеріалу, який забезпечує достатнє представлення фонетичних одиниць та їхніх комбінацій у природному мовленні.

Метою роботи є розроблення фонетично репрезентативного тексту українською мовою на основі статистичного аналізу трифонів. Для досягнення поставленої мети було проаналізовано теоретичні засади фонетичної репрезентативності, особливості використання фонетично збалансованих текстів у мовних технологіях, а також сучасні підходи до створення мовленнєвих корпусів.

Матеріалом дослідження став український переклад твору Льюїса Керролла «Аліса в Країні чудес». Для обробки тексту було розроблено програмний комплекс мовою Python, який здійснює автоматичне транскрибування тексту у фонетичний запис та статистичний аналіз трифонних послідовностей. У результаті було сформовано базу трифонів, визначено їхню частотність та проаналізовано структуру трифонного інвентарю корпусу.

На основі отриманих статистичних даних створено фонетично репрезентативний текст, що містить найважливіші трифонні поєднання української мови. Запропонований текст є зв’язним, граматично коректним і придатним для використання під час запису мовленнєвих корпусів, тестування систем автоматичного розпізнавання мовлення та синтезу мовлення.

Практичне значення роботи полягає у можливості використання розробленого тексту та програмних засобів у фонетичних дослідженнях, мовленнєвих технологіях і подальших роботах зі створення українських мовленнєвих ресурсів.

Ключові слова: фонетично репрезентативний текст, трифон, транскрибування, українська мова, корпусна лінгвістика, автоматичне розпізнавання мовлення, синтез мовлення, мовленнєвий корпус.

Abstract

The bachelor's thesis addresses the problem of creating a phonetically representative text for the Ukrainian language. The relevance of the study is determined by the growing demand for speech technologies, including automatic speech recognition systems, speech synthesis systems, speech corpora, and phonetic databases for Ukrainian. The quality of such systems largely depends on the availability of linguistic material that adequately represents phonetic units and their combinations in natural speech.

The aim of the study is to develop a phonetically representative Ukrainian text based on triphone frequency analysis. To achieve this goal, the theoretical foundations of phonetic representativeness, the role of phonetically balanced texts in speech technologies, and modern approaches to speech corpus creation were examined.

The research material was the Ukrainian translation of Lewis Carroll's *Alice's Adventures in Wonderland*. A software toolkit implemented in Python was developed to automatically convert text into phonetic transcription and perform statistical analysis of triphone sequences. As a result, a triphone database was compiled, triphone frequencies were calculated, and the structure of the triphone inventory was analyzed.

Based on the obtained statistical data, a phonetically representative text was created. The proposed text contains the most significant triphone combinations of Ukrainian and preserves grammatical correctness, coherence, and naturalness. It can therefore be used for speech corpus recording, speech recognition evaluation, and speech synthesis research.

The practical significance of the thesis lies in the possibility of applying both the developed text and the software tools in phonetic research, speech technologies, and future projects related to Ukrainian language resources.

Keywords: phonetically representative text, triphone, phonetic transcription, Ukrainian language, corpus linguistics, automatic speech recognition, speech synthesis, speech corpus.

ВСТУП

Із розвитком науково-технічної галузі розпочався новий етап суспільного прогресу, у якому цифрові технології посідають провідне місце в організації повсякденного життя людини. Комп'ютерні системи та інтелектуальні пристрої стали невід'ємною складовою комунікативного середовища, що зумовило потребу в ефективних засобах взаємодії людини з технікою. Одним із найбільш перспективних напрямів такої взаємодії є використання природної мови, зокрема створення систем автоматичного синтезу та розпізнавання мовлення. Дослідження у цій галузі активно розвиваються ще з середини XX століття і не втрачають актуальності донині.

Попри значні досягнення, ефективність сучасних систем автоматичного синтезу й розпізнавання українського мовлення у природних умовах залишається недостатньою. Основною причиною цього є висока варіативність мовного сигналу, яка зумовлена індивідуальними особливостями дикторів, впливом акустичного середовища, технічними характеристиками каналів передавання, а також складністю фонетичної та акцентуаційної системи української мови. Це зумовлює необхідність пошуку об'єктивних засобів оцінювання якості мовних технологій, які б дозволяли адекватно визначати рівень їхньої ефективності.

У цьому контексті особливого значення набуває використання фонетично репрезентативних текстів — компактних за обсягом, але інформаційно насичених текстових матеріалів, що забезпечують повне або максимально повне покриття фонетичних одиниць мови. Такі тексти дають змогу оцінити якість функціонування систем синтезу й розпізнавання мовлення, виявити їхні недоліки, а також здійснювати адаптацію систем до мовлення конкретного диктора. Крім того, фонетично насичені тексти застосовуються у фундаментальних і прикладних дослідженнях, зокрема при створенні мовних корпусів, розробці експертних систем та вивченні особливостей усного мовлення.

Наукова проблема дослідження полягає у відсутності чітко визначеної та обґрунтованої моделі створення фонетично репрезентативного тексту української мови, який би забезпечував системне й повне відображення її фонемного складу та типових фонетичних контекстів. Ця проблема становить структурований комплекс взаємопов'язаних питань, що охоплюють визначення параметрів фонетичної репрезентативності, принципів добору мовного матеріалу, способів його статистичного аналізу та критеріїв оцінки отриманого результату.

Актуальність дослідження зумовлена як теоретичними, так і практичними потребами сучасної лінгвістики та мовних технологій. З одного боку, вона пов'язана з необхідністю поглибленого вивчення фонетичної організації української мови та розширення наукових уявлень про її звукову систему. З іншого боку, актуальність визначається потребою у створенні ефективних інструментів для розробки й тестування автоматичних систем синтезу та розпізнавання мовлення, а також відсутністю завершених досліджень у цьому напрямі в українській науковій традиції.

Мета дослідження полягає у розробленні теоретично обґрунтованої та практично реалізованої моделі фонетично репрезентативного тексту української мови, яка забезпечує повне покриття її фонетичних одиниць і може бути використана як універсальний інструмент для аналізу та оцінки мовлення.

Відповідно до поставленої мети визначено такі **завдання дослідження**:

1. проаналізувати наукову літературу, присвячену проблемі створення фонетично репрезентативних текстів і мовних корпусів;
2. зібрати та опрацювати текстовий матеріал, що відображає фонетичне та інтонаційне розмаїття українського мовлення;
3. розробити програмний інструмент для автоматичної фонетичної транскрипції текстів із урахуванням наголосу та позиційних характеристик голосних звуків з метою отримання статистичних даних;

4. здійснити аналіз отриманих фонетичних комбінацій (зокрема трифонів) і сформувати на їх основі фонетично репрезентативний текст.

Об’єкт дослідження — фонетична система сучасної української мови.

Предмет дослідження — принципи і методи побудови фонетично репрезентативного тексту як моделі цієї системи.

Матеріал дослідження – текст роману Льюїса Керрола «Аліса в Країні Чудес» у перекладі Валентина Корнієнка, який було використано як базовий корпус для автоматизованого фонетичного аналізу. Вибір саме цього твору зумовлений достатнім обсягом текстового матеріалу, стилістичною різноманітністю, наявністю широкого спектра фонетичних поєднань, а також природністю мовлення персонажів і авторського викладу. Загальний обсяг проаналізованого тексту становив понад 87000 трифонних реалізацій, що дозволило отримати статистично репрезентативні результати для подальшого формування фонетично репрезентативного тексту української мови.

У процесі дослідження текст було автоматично наголошено, транскрибовано та проаналізовано за допомогою розробленого програмного комплексу. Основну увагу приділено статистичному аналізу трифонних поєднань, а також виявленню найбільш частотних і специфічних фонетичних моделей українського мовлення.

Методи дослідження визначалися метою, завданнями та специфікою практичної реалізації роботи. Під час дослідження було застосовано комплекс теоретичних, статистичних і програмно-алгоритмічних методів.

Для опрацювання наукових джерел, присвячених проблемам створення фонетично репрезентативних текстів, систем автоматичного синтезу та розпізнавання мовлення, використано метод аналізу, зіставлення та узагальнення наукової літератури. Це дало змогу визначити основні принципи фонетичної репрезентативності, сучасні підходи до формування фонетично збалансованих корпусів та критерії оцінювання фонетичного покриття тексту.

Для автоматизації обробки текстового матеріалу застосовано алгоритмічні методи та засоби комп'ютерної обробки даних. У межах дослідження було розроблено програми розстановки наголосів й автоматичної транскрипції українського тексту з урахуванням наголосу та позиційних характеристик голосних звуків. Програма також здійснювала сортування унікальних трифонних комбінацій і пошук слів, у яких вони реалізуються.

Для отримання статистичних даних про частотність фонетичних одиниць використано метод статистичного аналізу. У процесі роботи було здійснено підрахунок уживання окремих фонем, а також унікальних комбінацій із трьох фонем (трифонів), що дало можливість визначити найбільш репрезентативні фонетичні структури української мови.

На завершальному етапі дослідження використано метод моделювання, який полягав у побудові фонетично репрезентативного тексту на основі отриманих статистичних даних та відібраних фонетичних комбінацій. Під час укладання тексту враховувалися критерії фонетичної повноти, природності мовлення, граматичної зв'язності та компактності текстового матеріалу.

Наукова новизна дослідження полягає у розробці підходу до створення фонетично репрезентативного тексту української мови на основі автоматизованого аналізу фонетичних комбінацій, зокрема трифонів, із урахуванням наголосу та позиційних характеристик звуків. Запропонований підхід поєднує традиційні лінгвістичні принципи з методами комп'ютерної обробки мовних даних.

Практичне значення отриманих результатів полягає у можливості використання створеного фонетично репрезентативного тексту як універсального інструменту для оцінювання та порівняння систем автоматичного синтезу й розпізнавання мовлення, а також у дослідженнях усного мовлення. Отримані результати можуть бути застосовані у формуванні мовних корпусів, розробці мовних технологій, а також у навчальному процесі

під час викладання дисциплін, пов’язаних із фонетикою та комп’ютерною лінгвістикою.

Структура роботи: бакалаврська робота складається зі вступу, трьох розділів, чотирнадцяти підрозділів, висновків, списку використаних джерел та чотирнадцяти додатків. Загальний обсяг роботи становить 70 сторінок. Список використаних джерел налічує 40 позицій. У додатках подано допоміжні матеріали, програмні коди, результати транскрибування текстів та статистичні дані щодо трифонного складу досліджуваних корпусів.

РОЗДІЛ 1

1.1. Теоретичні засади створення фонетично репрезентативного тексту

Проблема створення фонетично репрезентативних текстів є важливим напрямом сучасної прикладної лінгвістики, експериментальної фонетики та мовних технологій. Формування компактного тексту, здатного відображати основні закономірності функціонування фонетичної системи мови, потребує врахування не лише окремих фонем, але й їхніх комбінаторних та позиційних варіантів, частотності вживання та особливостей реалізації у зв'язному мовленні. [Кочерган 2010, с. 312–315; Селіванова 2006]

У цьому розділі розглядаються теоретичні засади створення фонетично репрезентативних текстів, специфіка фонетичної системи української мови як об'єкта дослідження, а також сучасні підходи до використання корпусів у фонетичних і мовнотехнологічних дослідженнях. Особливу увагу приділено аналізу міжнародного досвіду створення фонетично збалансованих корпусів і текстів для систем синтезу та розпізнавання мовлення. [Загнітко 2007, с. 108–112; McEnery, Hardie 2012]

Фонетично репрезентативні тексти залишаються важливим інструментом сучасної прикладної фонетики, корпусної лінгвістики та мовних технологій. Вони використовуються для дослідження особливостей звукової системи мови, тестування систем синтезу й автоматичного розпізнавання мовлення, а також для створення стандартизованого мовного матеріалу. Особливу цінність такі тексти мають у тих випадках, коли необхідно отримати компактний, але фонетично насичений матеріал, здатний максимально повно відобразити основні фонетичні явища мови. [Селіванова 2006; Crystal 2008]

У сучасних системах автоматичної обробки мовлення значну роль справді відіграють статистичні методи машинного навчання, що ґрунтуються на використанні великих масивів текстових і мовленнєвих даних. Такі масиви отримали назву текстових або мовленнєвих корпусів. Зазвичай корпусом вважається електронно організований, структурований та анотований масив

текстів або аудіозаписів, підготовлений для виконання мовознавчих або прикладних завдань. Водночас навіть великі корпуси не усувають потреби у фонетично репрезентативних текстах, оскільки саме вони дозволяють цілеспрямовано контролювати покриття фонетичних одиниць і забезпечують зручний матеріал для експериментальних фонетичних досліджень та тестування мовних систем. [Бацевич 2004, с. 58; Загнітко 2007, с. 112; Jurafsky, Martin 2025]

Опорний текстовий корпус використовується як база для отримання статистичних характеристик мови, зокрема частотного розподілу фонем, алофонів, складів і типових фонетичних послідовностей. [Biber, Conrad, Reppen 1998; McEnery, Hardie 2012]

Однак поряд із розвитком великих корпусних систем не менш актуальним залишається завдання вироблення об'єктивних критеріїв оцінки їх якості, і в цьому випадку тестовим матеріалом для оцінки та порівняння систем автоматичного синтезу та розпізнавання мовлення можуть стати невеликі фонетично представницькі тексти (ФПТ). Створення компактних текстових матеріалів, здатних репрезентувати звукову систему мови, дозволить оцінити повноту покриття системою автоматичного синтезу й розпізнавання мовлення фонетичних одиниць цільової мови та виявити можливі недоліки її роботи. ФПТ використовуються насамперед для тестування систем, проведення акустико-фонетичних досліджень, аналізу мовлення носіїв мови. Крім того, на таких текстах зручно проводити швидку адаптацію мовної системи до нового диктора. [Rabiner, Juang 1993; Huang, Acero, Hon 2001]

Фонетично представницьким (репрезентативним) називають такий текстовий матеріал, у якому частотний розподіл фонетичних одиниць (фонем, алофонів, складів) відповідає загальномовному розподілу, встановленому на основі аналізу великого опорного корпусу. [Вінцюк 1987, с. 37–42]

Поняття фонетичної репрезентативності тісно пов’язане з поняттям фонетичної збалансованості. Фонетична репрезентація природним чином

передбачає присутність у тексті всіх фонем цільової мови в їхній основній дистрибуції, тобто в тому основному позиційному розподілі. Тобто у таких текстах повинні бути представлені всі фонеми мови, бажано в типових позиційних і комбінаторних умовах. Крім окремих фонем, враховується також розподіл дифонів, трифонів, складових структур і просодичних характеристик. [Тоцька 1995, с. 24–31; Тоцька 1981, с. 28–34]

Фонетично збалансовані й фонетично репрезентативні тексти традиційно використовуються як матеріал для вивчення фонетичних характеристик мови. Перевага використання фонетично представницьких текстів полягає, насамперед, у їхній компактності поряд з інформаційною насиченістю. З одного боку, такі тексти зазвичай невеликі за обсягом, з другого — відбивають фонетичне різноманіття мовної системи не гірше довільно взятих текстових масивів значного обсягу. [Clark, Yallop, Fletcher 2007; Ladefoged, Johnson 2014]

Досягнення такої репрезентативності потребує ретельного добору лексичних одиниць і граматичних конструкцій, що містять необхідні фонетичні елементи, а також вилучення малозначущих фрагментів тексту. Це досягається шляхом копіткої роботи з конструювання тексту – наповненням його словами, що містять потрібні фонетичні одиниці, а також скороченням його обсягу шляхом видалення елементів з низькою інформативністю. У результаті отримуємо зручний для прочитання матеріал (зазвичай не більше 600 слів), що дозволяє досліджувати характер реалізації та варіювання в мовленні носіїв певної мови значущих фонетичних характеристик та сформувати повноцінний мовний портрет того, хто говорить. У результаті формується зв’язний і природний текст, придатний для читання носіями мови та проведення фонетичного аналізу. [Young et al. 2006; Taylor, Black, Caley 1998]

На сьогодні в українській лінгвістичній науці не існує такого фонетично вивіреного тексту, у якому зустрічалися б фонеми в усіх можливих контекстах. Саме тому проблема створення фонетичного репрезентативного тексту української мови є актуальною як у теоретичному, так і в прикладному

аспектах. Результати подібної роботи можуть бути використані під час дослідження дитячого мовлення, мовлення білінгвів, інтонаційних особливостей української мови, а також у розробці систем синтезу й автоматичного розпізнавання українського мовлення. [Сучасна українська лінгвістика 2006; ГРАК 2026]

Оскільки ця тема не достатньо поширена в просторі вітчизняної мовознавчої науки, при виконанні роботи важливим є звернення до зарубіжного досвіду створення фонетично репрезентативних текстів і корпусів. Одним із найвідоміших прикладів фонетично репрезентативного тексту англійською мовою є уривок «The North Wind and the Sun» («Північний вітер і Сонце»), який широко використовується Міжнародною фонетичною асоціацією (International Phonetic Association, IPA) як еталонний матеріал для фонетичної транскрипції різними мовами світу. Основною перевагою цього тексту є наявність більшості фонем англійської мови у природному контексті. Такі фонетично збалансовані тексти традиційно використовуються як матеріал для вивчення фонетичних характеристик англійської мови. Крім того, текст характеризується нейтральним змістом, логічною структурою та природністю синтаксису, що робить його зручним для порівняльних фонетичних досліджень. Цей текст слугує основою для об'єктивного вивчення мовлення і сприяє вдосконаленню мовних технологій таких як: аналіз звукової системи мови, перевірка навичок транскрибування у студентів, тестування синтезаторів мовлення та систем ASR (автоматичного розпізнавання мовлення), діагностика та корекція мовленнєвих порушень. [International Phonetic Association 2026; Journal of the International Phonetic Association 2026]

Основною метою створення «The North Wind and the Sun» було забезпечення всебічного покриття всіх її фонем (як голосних, так і приголосних), а також основних алофонів, інтонаційних моделей і просодичних особливостей. Такий текст також мав би дозволити отримати достовірні акустичні та артикуляційні дані для аналізу мовлення носіїв англійської мови.

Тому для забезпечення вирішення поставлених завдань було прийнято основні принципи відбору матеріалу. Процес створення фонетично репрезентативного тексту включав такі етапи: 1) аналіз фонологічної системи англійської мови, тобто визначення всіх фонем та їх варіантів (алофонів), типових для певного діалекту (наприклад, General American або Received Pronunciation); 2) частотний аналіз – урахування частоти вживання звуків, слів і синтаксичних структур у повсякденному мовленні; 3) розробка корпусу-джерела, що містив добір слів і фраз, які разом охоплюють всі цільові фонетичні одиниці; 4) урахування просодії – включення інтонаційних моделей, темпу, ритму, наголосу та паузування, характерних для живого мовлення; 5) перевірка збалансованості – за допомогою комп'ютерного аналізу перевіряється, наскільки повно і рівномірно представлені фонетичні одиниці. При цьому текст має бути природним і граматично зв'язним. [Handbook of the International Phonetic Association 1999; International Phonetic Association 2026]

Сюжет тексту взято з байки Езопа «Північний вітер і Сонце», яка є короткою алегорією про суперечку між двома природними силами, які змагаються за вплив. Цей простий сюжет ідеально підходить для фонетичного аналізу, оскільки містить широкий діапазон звуків, типових для англійської мови та є достатньо коротким, щоб легко запам'ятатися або бути продекламованим. У тексті «The North Wind and the Sun» представлені майже всі фонеми англійської мови, зокрема: 1) голосні як короткі (/ɪ/, /ʌ/), так і довгі (/iː/, /ɑː/), а також дифтонги (/aɪ/, /əʊ/); 2) приголосні (дзвінкі і глухі, проривні, фрикативні, носові, латеральні та напівголосні); 3) різні позиції фонем (на початку, у середині й наприкінці слів); 4) просодичні характеристики (інтонація, паузи, фразовий наголос). Також текст має нейтральний зміст, що не містить емоційно забарвленої лексики, що значно допомогло науковцям. До того ж цей ФПТ для англійської мови може бути адаптований до будь-якої мови, зберігаючи ту ж саму структуру і зміст. [Moran, McCloy 2019; Journal of the International Phonetic Association 2026]

Міжнародна фонетична асоціація (IPA) використала цей текст як основу для демонстрації транскрипції мовлення різними мовами світу. Це дозволило порівнювати фонологічні системи різних мов, застосовуючи єдину модель тексту. У випадку англійської мови існують різні варіанти цього тексту — для різних стандартів вимови, зокрема: 1) Received Pronunciation (RP) — британський варіант; 2) General American (GA) — американський стандарт; 3) а також варіанти для інших діалектів, як-от австралійський чи канадський англійські варіанти. Сьогодні «The North Wind and the Sun» використовується для різноманітних лінгвістичних досліджень: навчання фонетичної транскрипції, зіставного фонетичного аналізу між різними мовами, оцінки точності синтезу або розпізнавання (у мовних технологіях), лінгвістичного картографування діалектів. [International Phonetic Association 2026; Handbook of the International Phonetic Association 1999]

Отже, ФПТ «The North Wind and the Sun» є прикладом успішного поєднання природної мови та лінгвістичного планування. Завдяки ретельному добору лексики, граматичних структур та звукових одиниць, текст став універсальним інструментом для вивчення і викладання фонетики. Його простота та ефективність зробили його еталонним у фонетичних дослідженнях. [Moran, McCloy 2019]

Однак сучасні дослідження у сфері мовних технологій демонструють принципово інший підхід до створення фонетично репрезентативного матеріалу. Якщо традиційні тексти на кшталт «The North Wind and the Sun» створювалися переважно вручну, з опорою на інтуїтивно-лінгвістичний добір мовного матеріалу, то сучасні корпуси формуються із застосуванням автоматизованих алгоритмів, що базуються на статистичному аналізі великих масивів даних. Такий перехід від інтуїтивного до формалізованого підходу є закономірним результатом розвитку корпусної лінгвістики та комп’ютерної обробки мовлення. [Jurafsky, Martin 2025; Huang, Acero, Hon 2001]

Одним із найвідоміших прикладів реалізації цього підходу є корпус TIMIT (Texas Instruments and MIT Acoustic-Phonetic Corpus), створений для досліджень англійського мовлення. Як зазначають розробники корпусу, його метою було забезпечення «акустично-фонетично збалансованої бази даних для досліджень мовлення» (Garofolo et al.). У межах цього проєкту було сформовано набір так званих *phonetically rich sentences* (наприклад: *"She had your dark suit in greasy wash water all year"*) — фонетично насичених речень, які підбиралися таким чином, щоб максимально охопити всі фонеми англійської мови та їхні типові комбінації. [Handbook of the International Phonetic Association 1999; Moran, McCloy 2019; Garofolo et al. 1993; Linguistic Data Consortium 2026]

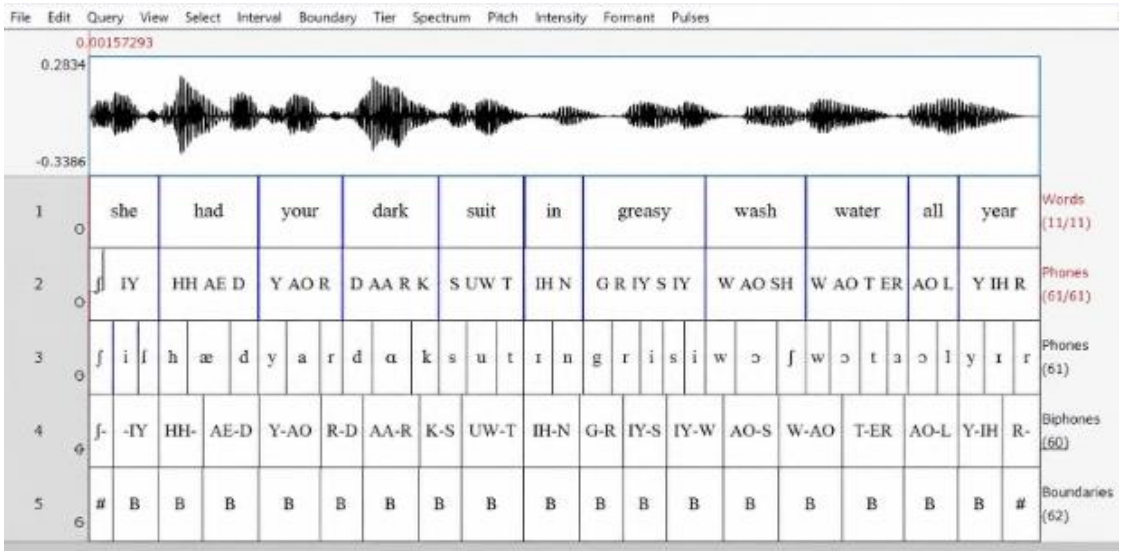

Рис. 1.1.1. Приклад фонетично насиченого речення та його фонемної розмітки в корпусі TIMIT

Особливістю TIMIT є те, що він поєднує дві, на перший погляд, суперечливі характеристики: статистичну репрезентативність і природність мовлення. Корпус містить записи 630 носіїв американської англійської мови з різних діалектних регіонів, що дозволяє врахувати варіативність вимови. Крім того, кожен аудіозапис супроводжується детальною фонетичною розміткою на рівні фонем. Це робить TIMIT еталонним ресурсом для тестування систем автоматичного розпізнавання мовлення (ASR). [Garofolo et al. 1993; Rabiner, Juang 1993]

Подібний підхід використано й у проєкті ATR English Speech Corpus for Speech Synthesis, створеному для систем синтезу мовлення. На відміну від TIMIT, де значну роль відігравав експертний відбір речень, у цьому корпусі активно застосовувалися алгоритмічні методи. Зокрема, використовувався так званий *greedy algorithm*, який дозволяє ітеративно відбирати речення, що максимально збільшують покриття фонетичних одиниць. [Sagisaka 2026; ATR English Speech Corpus 2026]

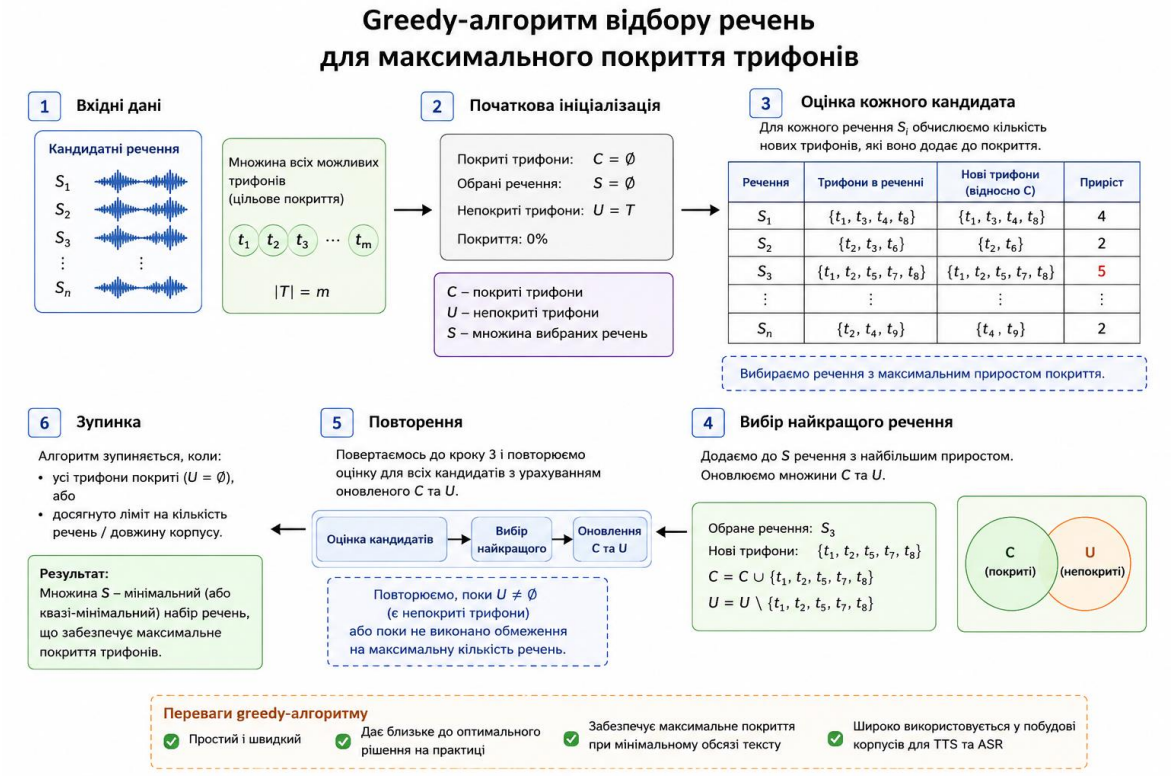

Рис. 1.1.2. Принцип відбору речень для максимального покриття трифонів (за моделлю greedy-алгоритму)

Як зазначають автори дослідження, «sentences were selected to maximize phonetic coverage while minimizing corpus size», тобто добір матеріалу здійснювався з метою досягнення оптимального співвідношення між повнотою фонетичного покриття та компактністю корпусу. Особливу увагу було приділено не лише окремим фонемам, а й їхнім комбінаціям — дифонам і трифонам, які є критично важливими для якості синтезу мовлення. [Sagisaka 2026; ATR English Speech Corpus 2026]

Таким чином, у сучасних дослідженнях відбувається зміщення акценту з окремих фонем на фонетичні контексти. Це пояснюється тим, що у природному мовленні реалізація звука значною мірою залежить від його оточення. Саме тому покриття дифонів і трифонів є більш релевантним критерієм для оцінки якості мовних технологій, ніж просте включення всіх фонем. [Young et al. 2006; Jurafsky, Martin 2025]

Ще одним важливим прикладом є британський корпус EUSTACE, у якому було запропоновано розширене розуміння фонетичної репрезентативності. На відміну від попередніх підходів, тут увага приділялася не лише сегментному рівню (фонеми), а й супрасегментним характеристикам мовлення. Як зазначають автори проєкту, речення «were controlled for phonetic and prosodic content», тобто враховувалися інтонація, ритм, наголос, темп мовлення та довжина фраз. [Centre for Speech Technology Research 2026; Taylor, Black, Caley 1998]

Keyword Series A example sentences: utterance-medial; fixed utterance length; variable word length.

Right-headed keywords		
Sentence type	**Accented condition sentence**	**Unaccented condition sentence**
‖. . . σσ♯σ♯. . . ‖	John saw Jessica MEND it again	JOHN saw Jessica mend it AGAIN
‖. . . σ♯σσ♯. . . ‖	John saw Jessie COMMEND it again	JOHN saw Jessie commend it AGAIN
‖. . . ♯σσσ♯. . . ‖	John saw Jess RECOMMEND it again	JOHN saw Jess recommend it AGAIN
Left-headed keywords		
Sentence type	**Accented condition sentence**	**Unaccented condition sentence**
‖. . . ♯σ♯σσ. . . ‖	I saw the MACE unreclaimed again	I SAW the mace unreclaimed AGAIN
‖. . . ♯σσ♯σ. . . ‖	I saw the MASON reclaimed it all	I SAW the mason reclaimed it ALL
‖. . . ♯σσσ♯. . . ‖	I saw the MASONRY cleaned again	I SAW the masonry cleaned AGAIN

Full discussion of the design of the sentences in Keyword Series A is provided on pages 147 to 149 in Chapter 4 of *English speech timing: a domain and locus approach* (see DISSERTATION page).

Рис. 1.1.3. Приклад урахування просодичних характеристик у корпусі EUSTACE.

Такий підхід є надзвичайно важливим, оскільки саме просодичні характеристики значною мірою визначають природність синтезованого мовлення та точність його розпізнавання. Відтак поняття фонетичної репрезентативності у сучасній лінгвістиці розширюється і включає не лише фонемний склад, але й інтонаційні та ритмічні особливості мовлення. [Clark, Yallop, Fletcher 2007; Ladefoged, Johnson 2014]

Схожі тенденції простежуються й у дослідженнях інших мов. Зокрема, у роботі «A High Quality and Phonetic Balanced Speech Corpus for Vietnamese» (2019) описано створення фонетично збалансованого корпусу для в'єтнамської мови. Автори зазначають, що корпус був сформований на основі «phonetic balance and coverage criteria», із використанням автоматизованого відбору речень із великих текстових баз. [Nguyen, Do, Pham 2019]

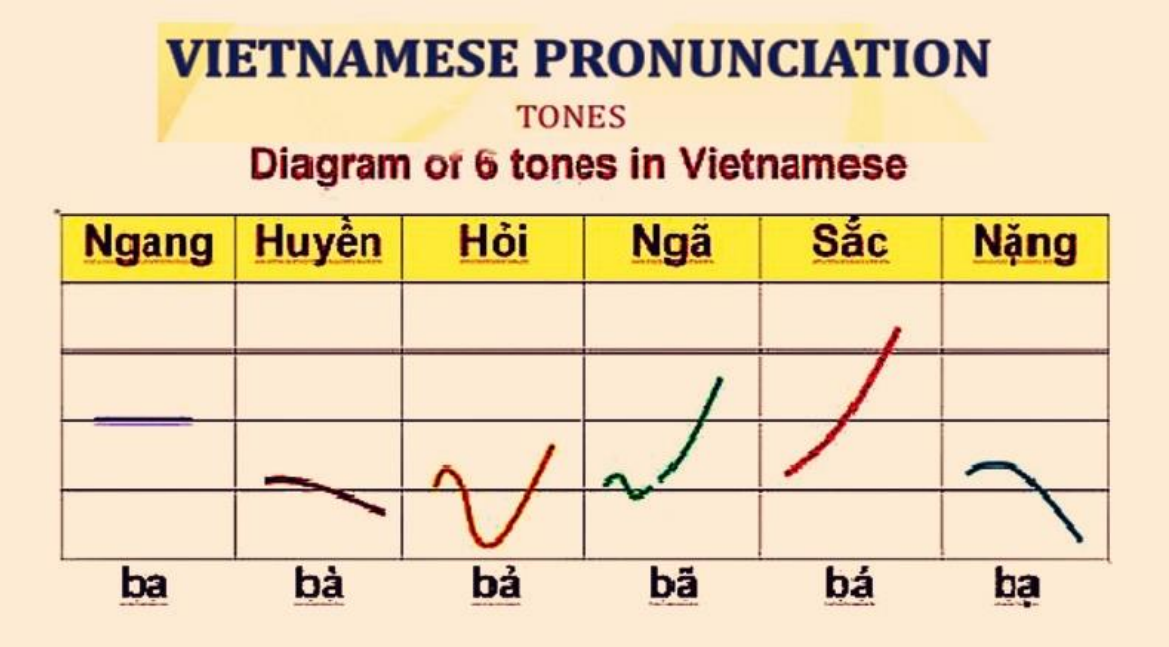

Рис. 1.1.4. Урахування тональних характеристик у фонетично збалансованому корпусі в'єтнамської мови. [Nguyen, Do, Pham 2019]

Особливістю цього дослідження є врахування тональної природи в'єтнамської мови. Оскільки тон у таких мовах виконує смислорозрізнювальну функцію, автори включили до критеріїв репрезентативності не лише фонеми, але й тональні варіації. Це ще раз підтверджує, що модель фонетично

репрезентативного тексту має враховувати специфіку фонологічної системи конкретної мови. [Nguyen, Do, Pham 2019]

Отже, аналіз теоретичних засад створення фонетично репрезентативних текстів засвідчує, що сучасні підходи до дослідження мовлення поєднують методи традиційної фонетики, корпусної лінгвістики та комп'ютерної обробки мовних даних. Незважаючи на активний розвиток великих мовленнєвих корпусів і статистичних моделей машинного навчання, компактні фонетично репрезентативні тексти не втратили своєї актуальності. Вони залишаються важливим інструментом для тестування систем автоматичного синтезу та розпізнавання мовлення, проведення акустико-фонетичних досліджень, а також аналізу особливостей мовлення носіїв мови. [McEnery, Hardie 2012; Jurafsky, Martin 2025]

У процесі опрацювання наукової літератури було встановлено, що поняття фонетичної репрезентативності охоплює не лише наявність усіх фонем мови, а й урахування їхніх позиційних і комбінаторних варіантів, частотності вживання, просодичних характеристик та природності мовлення. Таким чином, сучасне розуміння фонетично репрезентативного тексту значно ширше за традиційне уявлення про фонетично збалансований набір слів або речень. [Crystal 2008; Селіванова 2006]

Аналіз зарубіжного досвіду дозволив виокремити два основні підходи до створення фонетично репрезентативного матеріалу. Перший ґрунтується на ручному укладанні зв'язного тексту з урахуванням фонетичного покриття, прикладом чого є текст «The North Wind and the Sun». Його перевагами є природність, логічна зв'язність і зручність для фонетичних досліджень. Другий підхід базується на автоматизованому статистичному аналізі великих текстових корпусів і використанні алгоритмів добору речень, що реалізовано в сучасних корпусах TIMIT, ATR та EUSTACE. Такий підхід дозволяє забезпечити точніше покриття фонетичних контекстів, зокрема дифонів і трифонів, які мають ключове значення для систем синтезу й автоматичного розпізнавання

мовлення. [Garofolo et al. 1993; Sagisaka 2026; Centre for Speech Technology Research 2026]

Дослідження також показало, що критерії фонетичної репрезентативності залежать від специфіки фонологічної системи конкретної мови. Якщо для англійської мови особливе значення мають позиційні варіанти фонем і трифонні комбінації, то для тональних мов, зокрема в'єтнамської, важливим компонентом стає врахування тональних характеристик мовлення. Це свідчить про необхідність адаптації принципів створення фонетично репрезентативного тексту до особливостей української фонологічної системи. [Nguyen, Do, Pham 2019; Shosted, Andrusenko 2017]

Отже, проведений теоретичний аналіз підтверджує актуальність створення фонетично репрезентативного тексту української мови та дозволяє сформулювати основні принципи його побудови: повнота фонемного покриття, урахування фонетичних контекстів і просодичних характеристик, статистична збалансованість, природність мовлення та компактність тексту. Саме ці принципи становитимуть основу подальшої практичної частини дослідження, присвяченої розробці власної моделі фонетично репрезентативного тексту української мови. [Загнітко 2007, с. 201–204; McEnery, Hardie 2012]

1.2. Специфіка фонетичної системи української мови

Фонетична система української мови є складною багаторівневою структурою, що охоплює сукупність голосних і приголосних фонем, їхніх позиційних варіантів, комбінаторних змін та інтонаційних характеристик. Саме фонетична система стала основним об’єктом дослідження під час створення фонетично репрезентативного тексту, оскільки ефективність такого тексту безпосередньо залежить від повноти відображення звукових одиниць мови та типових умов їхнього функціонування. [Тоцька 1995, с. 5–8; Мойсієнко 2010, с. 9–14]

Сучасна українська літературна мова традиційно налічує 38 основних фонем: 6 голосних і 32 приголосні. Голосні фонеми української мови представлені звуками /а/, /е/, /и/, /і/, /о/, /у/. На відміну від багатьох інших мов, українська вокалічна система характеризується відносно стабільною артикуляцією голосних, однак у природному мовленні ненаголошені голосні можуть зазнавати часткової редукції та позиційних змін. Особливо це стосується голосних [е], [и] та [о], які в ненаголошеній позиції реалізуються зі змінами артикуляційних характеристик. [Жовтобрюх, Кулик 1972, с. 34–42; Тоцька 1981, с. 17–25]

Рис. 1.2.1. Система голосних фонем української мови [Мойсієнко 2010, с. 20–23]

Особливу складність становить система приголосних фонем української мови. Значна кількість приголосних має пари за твердістю та м’якістю, а також бере участь у процесах асиміляції, спрощення та подовження. Для української мови характерними є:

- протиставлення твердих і м’яких приголосних;
- наявність африкат [дз͡], [дж͡], [ц], [ч];
- широке поширення асимілятивних процесів;
- позиційні зміни звуків у мовленнєвому потоці [Мойсієнко 2010, с. 20–23]

М’якість приголосних є однією з найважливіших особливостей української фонетики. Наприклад, фонеми /л/ і /л'/ або /н/ і /н'/ можуть розрізняти значення слів, що підвищує функціональне навантаження ознаки м’якості. Крім того, значна кількість фонетичних процесів виникає саме на межі твердих і м’яких приголосних. [Прокопова 1958, с. 61–74; Тоцька 1995, с. 48–57]

Групи приголосних (усього 32 звуки)				Приголосні звуки
Губні (лабіальні)	губно-губні (білабіальні)			[б], [п], [м], [в])
	губно-зубні (лабіо-дентальні)			[ф]
Язикові (26 звуків)	Передньоязикові (22 звуки)	Апікальні (11 звуків)	тверді	[д], [т], [л], [н], [дз͡], [дж͡]
			м’які	[д'], [т'], [л'], [н'], [дз͡'], –
		Какумінальні, або церебральні (5 звуків)	тверді	[ж], [ч], [ш], [р]
			м’які	–, –, –, [р']
		Дорсальні (6 звуків)	тверді	[з], [ц], [с]
			м’які	[з'], [ц'], [с']
	Середньоязиковий (1 звук – постійно м’який)			[й]
	Задньоязикові (3 звуки – тверді)			[г], [к], [х]
Глотковий (фарингальний) (1 звук – твердий)				[ґ]

Рис. 1.2.2. Класифікація приголосних фонем української мови. [Плющ 2005, с. 45–49]

Важливою особливістю української фонетичної системи є значна кількість комбінаторних змін звуків у мовленнєвому потоці. У процесі мовлення звуки впливають один на одного, що призводить до виникнення асиміляції за

дзвінкістю, глухістю, м’якістю, місцем і способом творення. Наприклад, у слові «боротьба» глухий приголосний [т'] під впливом наступного дзвінкого [б] реалізується як дзвінкий звук. Подібні процеси значно ускладнюють автоматичну фонетичну обробку тексту та потребують урахування під час створення програм транскрибування. [Сербенська 2014, с. 96–104; Мойсієнко 2010, с. 121–136]

Окрему роль у фонетичній системі української мови відіграють просодичні характеристики мовлення. Наголос в українській мові є вільним і рухомим, а його позиція впливає на реалізацію голосних звуків та інтонаційну структуру висловлювання. Саме тому під час створення фонетично репрезентативного тексту було необхідно враховувати не лише фонемний склад, але й особливості наголошування. [Іщенко 2009, с. 8–13; Тоцька 1995, с. 93–101]

Під час створення фонетично репрезентативного тексту особливого значення набуває використання текстових і мовленнєвих корпусів. Корпусом називають електронно організований, структурований та анотований масив текстів або аудіозаписів, призначений для мовознавчих досліджень і прикладних завдань. У сучасній комп’ютерній лінгвістиці корпуси є основним джерелом статистичних даних про частотність мовних одиниць, типові фонетичні поєднання та особливості функціонування мови у природному мовленні. [McEnery, Hardie 2012, p. 1–12; Biber, Conrad, Reppen 1998, p. 4–10]

У світовій практиці для досліджень мовлення активно використовуються спеціалізовані фонетичні корпуси, зокрема TIMIT Acoustic-Phonetic Corpus, ATR English Speech Corpus та EUSTACE Speech Corpus. Такі корпуси містять аудіозаписи мовлення, фонетичну транскрипцію та різні типи мовної анотації, що дозволяє використовувати їх для синтезу мовлення, автоматичного розпізнавання мовлення та статистичного аналізу фонетичних одиниць. [Garofolo et al. 1993; ATR English Speech Corpus 2026; EUSTACE Speech Synthesis Project 2026]

Окремі дослідження також демонструють важливість використання фонетично збалансованих текстів для покриття фонемного складу мови. Наприклад, текст «The North Wind and the Sun» традиційно використовується Міжнародною фонетичною асоціацією для демонстрації особливостей вимови різних мов світу. [Handbook of the International Phonetic Association 1999, p. 191–194; Moran, McCloy 2019, p. 305–309]

Для української мови проблема створення фонетично збалансованих корпусів залишається актуальною. Значна частина наявних корпусів орієнтована переважно на текстовий рівень і не містить достатньої кількості фонетично анотованого матеріалу. Саме тому створення фонетично репрезентативного тексту є важливим етапом формування якісних українських мовленнєвих ресурсів. [ГРАК 2026; Скляренко 2006, с. 201–208]

У межах дослідження як текстову основу було використано переклад твору Льюїса Керрола «Аліса в Країні Чудес» Валентина Корнієнка. Цей текст характеризується значним лексичним різноманіттям, широким спектром синтаксичних конструкцій та великою кількістю фонетичних комбінацій, що робить його придатним для побудови фонетично репрезентативного матеріалу.

Таким чином, фонетична система української мови характеризується значною кількістю фонетичних одиниць, складною системою їхніх комбінаторних змін та важливою роллю просодичних характеристик. Саме ці особливості визначають складність створення фонетично репрезентативного тексту та обумовлюють необхідність використання автоматизованих методів аналізу мовного матеріалу. Використання корпусного підходу та статистичного аналізу фонетичних одиниць дозволяє забезпечити більш повне покриття фонетичної системи української мови та створити текстовий матеріал, придатний для сучасних мовних технологій. [Jurafsky, Martin 2025; Huang, Acero, Hon 2001, p. 35–49]

1.3. Узагальнення та зіставлення підходів

Аналіз опрацьованої літератури дозволяє виокремити дві основні тенденції у створенні фонетично репрезентативних текстів. [Crystal 2008; Jurafsky, Martin 2025]

1. Традиційний (лінгвістичний) підхід. Він ґрунтується на ручному укладанні зв'язного тексту з урахуванням фонетичного покриття. Його класичним прикладом є «The North Wind and the Sun». Перевагами цього підходу є:

- природність мовлення;
- логічна зв'язність тексту;
- зручність використання в навчанні та експериментах.

Однак недоліком є обмежена точність фонетичного покриття, оскільки такий текст не завжди охоплює всі можливі комбінації фонем. [Handbook of the International Phonetic Association 1999, p. 191–194; Moran, McCloy 2019, p. 305–310]

2. Сучасний (корпусно-алгоритмічний) підхід. Він передбачає автоматичне формування набору речень на основі статистичного аналізу корпусів. Його переваги:

- висока точність фонетичної збалансованості;
- можливість контролю покриття дифонів і трифонів;
- масштабованість і відтворюваність результатів.

Недоліком може бути зниження природності тексту та стилістичної цілісності. [McEnery, Hardie 2012, p. 151–165; Garofolo et al. 1993; Taylor, Black, Caley 1998, p. 147–151]

Отже, зіставлення традиційного та сучасного підходів до створення фонетично репрезентативних текстів свідчить про те, що кожен із них має як переваги, так і певні обмеження. Традиційний лінгвістичний підхід забезпечує природність, логічну цілісність і зручність сприйняття тексту, що є важливим для фонетичних експериментів за участю носіїв мови. Водночас ручне

укладання тексту не завжди дозволяє досягти повного й статистично точного покриття всіх фонетичних комбінацій. Clark, Yallop, Fletcher 2007, p. 15–21; Ladefoged, Johnson 2014, p. 1–10]

Сучасний корпусно-алгоритмічний підхід, навпаки, орієнтований насамперед на точність фонетичної репрезентації. Використання статистичного аналізу та автоматизованого відбору речень дозволяє враховувати не лише окремі фонеми, а й їхні позиційні та комбінаторні варіанти, що є особливо важливим для систем синтезу й автоматичного розпізнавання мовлення. Однак надмірна орієнтація на статистичне покриття може негативно впливати на природність і стилістичну зв'язність тексту. [Rabiner, Juang 1993, p. 17–26; Huang, Acero, Hon 2001, p. 35–49; Young et al. 2006, p. 1–12]

Таким чином, найбільш ефективним видається поєднання обох підходів: використання алгоритмічних методів для досягнення максимальної фонетичної збалансованості та лінгвістичного редагування для забезпечення природності, граматичної коректності й зручності сприйняття тексту. Саме такий комбінований принцип покладено в основу цього дослідження під час створення фонетично репрезентативного тексту української мови. [Nguyen, Do, Pham 2019, p. 295–301; McEnery, Hardie 2012, p. 245–252]

1.4. Висновки до розділу 1

Таким чином, сучасні підходи до створення фонетично репрезентативних текстів демонструють тенденцію до інтеграції методів корпусної лінгвістики, статистичного аналізу та комп'ютерного моделювання. На відміну від традиційних текстів, сучасні моделі орієнтуються не лише на покриття фонем, але й на врахування фонетичних контекстів і просодичних характеристик. [McEnery, Hardie 2012, p. 245–252; Jurafsky, Martin 2025]

Узагальнення розглянутих підходів дозволяє сформулювати основні принципи створення фонетично репрезентативного тексту української мови:

- повне покриття фонемного складу;
- урахування позиційних і комбінаторних варіантів;
- орієнтація на частотність мовних одиниць;
- включення просодичних характеристик;
- забезпечення природності та зв'язності тексту;
- оптимізація обсягу тексту. [Handbook of the International Phonetic Association 1999, p. 191–194; Moran, McCloy 2019, p. 320–323; Nguyen, Do, Pham 2019, p. 298–300]

Отже, створення фонетично репрезентативного тексту є складним багаторівневим процесом, що поєднує теоретичні та прикладні аспекти мовознавства. Відсутність такого тексту для української мови підтверджує актуальність дослідження та необхідність розробки власної моделі, яка поєднуватиме точність статистичних методів із природністю мовлення. [Селіванова 2006; Скляренко 2006, с. 201–208; Huang, Acero, Hon 2001, p. 35–49]

Таким чином, сучасні підходи до створення фонетично репрезентативних текстів демонструють тенденцію до інтеграції методів корпусної лінгвістики, статистичного аналізу та комп'ютерного моделювання. На відміну від традиційних текстів, сучасні моделі орієнтуються не лише на покриття фонем,

але й на врахування фонетичних контекстів і просодичних характеристик. [McEnery, Hardie 2012, p. 245–252; Jurafsky, Martin 2025]

Узагальнення розглянутих підходів дозволяє сформулювати основні принципи створення фонетично репрезентативного тексту української мови:

- повне покриття фонемного складу;
- урахування позиційних і комбінаторних варіантів;
- орієнтація на частотність мовних одиниць;
- включення просодичних характеристик;
- забезпечення природності та зв’язності тексту;
- оптимізація обсягу тексту. [Handbook of the International Phonetic Association 1999, p. 191–194; Moran, McCloy 2019, p. 320–323; Nguyen, Do, Pham 2019, p. 298–300]

Отже, створення фонетично репрезентативного тексту є складним багаторівневим процесом, що поєднує теоретичні та прикладні аспекти мовознавства. Відсутність такого тексту для української мови підтверджує актуальність дослідження та необхідність розробки власної моделі, яка поєднуватиме точність статистичних методів із природністю мовлення. [Селіванова 2006; Скляренко 2006, с. 201–208; Huang, Acero, Hon 2001, p. 35–49]

РОЗДІЛ 2

2.1. Загальна організація дослідження та матеріал аналізу

Практична реалізація дослідження потребувала створення програмних засобів для автоматизованої обробки великих текстових масивів. Ручне наголошування, транскрибування та статистичний аналіз фонетичних одиниць потребували б значних часових витрат і не забезпечували б достатнього рівня точності та повторюваності результатів.

У другому розділі описано структуру та принципи роботи розробленого програмного комплексу, призначеного для автоматичного розставлення наголосів, фонетичної транскрипції українського тексту та статистичного аналізу фонетичних одиниць. Також розглянуто особливості реалізації файлу налаштувань, організації файлової структури проєкту та використання автоматизованих алгоритмів для формування фонетично репрезентативного тексту.

Фонетично репрезентативний (представницький) текст є одним із ключових інструментів сучасної прикладної фонетики та мовних технологій. Під таким текстом розуміють компактний, фонетично насичений текстовий матеріал, у якому максимально повно представлені фонетичні одиниці певної мови та їхні типові комбінації. Основне призначення такого тексту полягає у забезпеченні можливості об’єктивного аналізу мовлення, тестування систем автоматичного синтезу й розпізнавання мовлення, а також формування фонетично збалансованих мовних корпусів.

У сучасних умовах розвитку комп’ютерної лінгвістики фонетично репрезентативні тексти набувають особливого значення, оскільки використовуються як універсальний тестовий матеріал для дослідження мовлення носіїв мови та оцінювання якості мовних технологій. Для української мови проблема створення такого тексту є особливо актуальною через відсутність стандартизованого фонетично збалансованого матеріалу, який би забезпечував повне покриття фонетичної системи мови.

Якщо розглядати можливість використання фонетично репрезентативного тексту для формування мовних корпусів та тестування експертних і автоматичних мовних систем, необхідно насамперед забезпечити максимальну повноту покриття фонетичних одиниць української мови. Це зумовлено тим, що якість роботи систем автоматичного синтезу й розпізнавання мовлення безпосередньо залежить від того, наскільки повно в навчальному або тестовому матеріалі представлені звукові одиниці та їхні контексти.

Фонетична система сучасної української мови є складною багаторівневою структурою. Традиційно в українській мові виділяють 38 основних фонем: 6 голосних і 32 приголосні. Однак реальна кількість можливих фонетичних реалізацій значно більша, оскільки кожна фонема може функціонувати в різних позиційних і комбінаторних умовах. Фонеми здатні поєднуватися між собою в межах складів, слів і фраз, утворюючи значну кількість варіантів звукових послідовностей.

Наприклад, для голосної фонеми /а/ та м’яких приголосних /т'/, /с'/, /ц'/ можливі такі комбінації: /ат'с'/, /ат'ц'/, /ас'т'/, /ас'ц'/, /ац'т'/, /ац'с'/, /т'ас'/, /т'ац'/, /с'ат'/, /с'ац'/, /ц'ат'/, /ц'ас'/, /т'с'а/, /т'ц'а/, /с'т'а/, /с'ц'а/, /ц'т'а/, /ц'с'а/ тощо. Подібні поєднання демонструють, що навіть невелика кількість фонем здатна утворювати надзвичайно велику кількість фонетичних контекстів.

Особливої складності додає необхідність урахування просодичних характеристик мовлення. Інтонація істотно впливає на реалізацію звуків у мовленнєвому потоці, тому під час побудови фонетично репрезентативного тексту важливо враховувати не лише сегментний, а й супрасегментний рівень мовлення. У межах дослідження було враховано такі основні інтонаційні типи:

- інтонація завершеності;
- інтонація незавершеності;
- питальна інтонація;
- оклична інтонація.

Урахування інтонаційних варіантів значно розширює кількість можливих фонетичних реалізацій. Якщо умовно розглядати трифонні комбінації як мінімальну модель фонетичного контексту, то для 38 фонем кількість потенційних комбінацій становить:

$38 \times 38 \times 38 = 54\,872$ варіанти.

Якщо ж враховувати інтонаційні реалізації, кількість умовних фонетичних одиниць збільшується до 55, а кількість потенційних трифонних комбінацій становить:

$55 \times 55 \times 55 = 166\,375$ варіантів.

Наведені розрахунки демонструють надзвичайну складність завдання створення фонетично репрезентативного тексту української мови. Очевидно, що практично неможливо створити невеликий текст, який би містив абсолютно всі можливі комбінації фонем. Саме тому під час розробки ФПТ важливим є не механічне охоплення всіх варіантів, а відбір найбільш частотних і функціонально значущих фонетичних структур.

Для забезпечення максимальної репрезентативності тексту необхідно використовувати матеріал різних функціональних стилів і типів мовлення. Саме тому під час дослідження було важливо сформувати текстову базу, здатну відобразити звукове та інтонаційне різноманіття сучасної української мови.

Одним із текстів, який традиційно вважається достатньо фонетично насиченим для української мови, є переклад повісті Льюїса Керрола «Аліса в Країні Чудес», виконаний Валентином Корнієнком. Цей твір активно використовується під час запису дикторів та формування мовних матеріалів завдяки багатству лексики, різноманітності синтаксичних конструкцій і широкому спектру фонетичних поєднань. Саме тому уривки з цього твору стали основою текстового матеріалу в межах цього дослідження.

Після формування текстової бази виникла необхідність у проведенні детальної фонетичної транскрипції. Особливу увагу було приділено позиції

голосних звуків щодо наголосу, оскільки наголос істотно впливає на акустичну реалізацію голосних та їхню редукцію у мовленні.

Очевидно, що ручне транскрибування значного обсягу текстового матеріалу потребувало б надмірних часових витрат і могло б призвести до появи суб'єктивних помилок. Саме тому в межах дослідження було прийнято рішення автоматизувати цей процес шляхом створення спеціалізованого програмного забезпечення.

Було розроблено дві програми:

1. програму автоматичного розставлення наголосів у тексті;
2. програму автоматичної фонетичної транскрипції українського тексту.

Крім безпосередньої транскрипції, програма також виконувала автоматичне сортування унікальних трифонних комбінацій та пошук слів, у яких вони реалізуються. Це дозволило значно пришвидшити процес аналізу текстового матеріалу та мінімізувати вплив людського фактора.

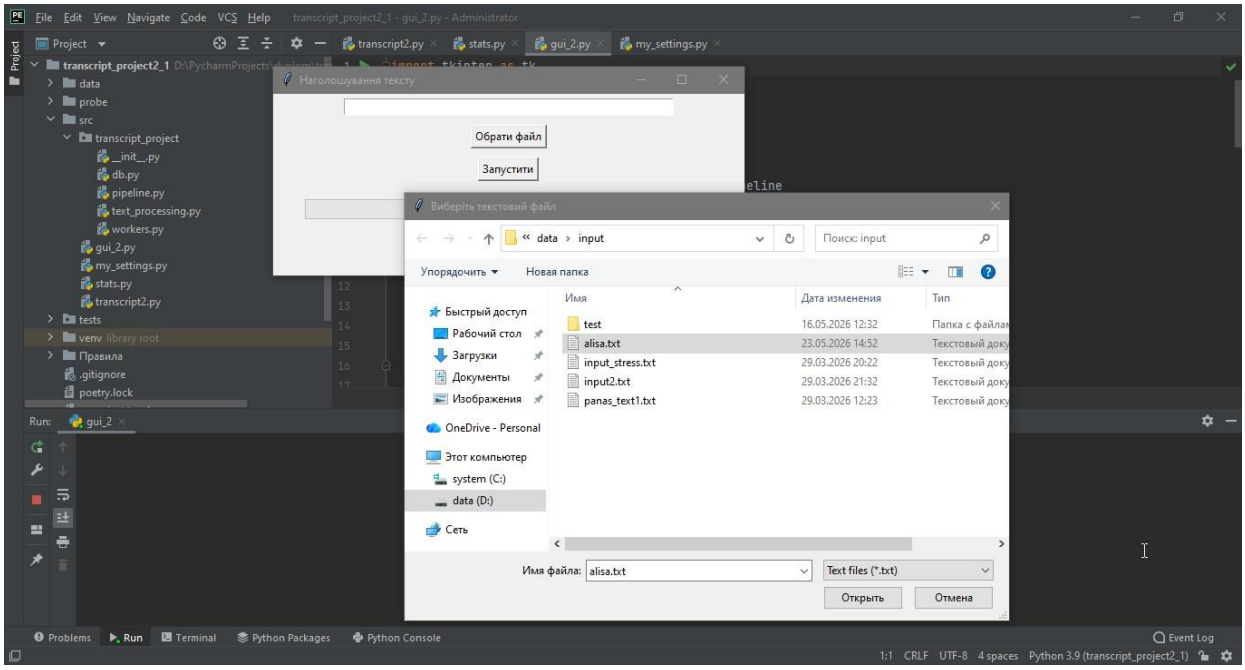

Рис. 2.1.1. Приклад роботи програми автоматичної транскрипції українського тексту.

Під час розробки алгоритмів транскрипції особливо важливим було визначення нормативної теоретичної бази. Спочатку було розглянуто декілька сучасних праць із фонетики української мови, зокрема:

• «Сучасна українська літературна мова» за редакцією М. Я. Плющ;

• «Голос і звуки рідної мови» Олександри Сербенської.

Однак основним джерелом для формування правил транскрипції було обрано підручник «Сучасна українська літературна мова. Лексикологія. Фонетика», створений на базі Інституту філології Київського національного університету імені Тараса Шевченка та затверджений Міністерством освіти і науки України.

Вибір саме цього джерела був зумовлений декількома чинниками:

• ґрунтовністю опису фонетичних явищ;

• системністю викладу матеріалу;

• науковою обґрунтованістю правил транскрипції;

• авторитетністю авторського колективу.

Особливу роль відіграв і той факт, що авторами підручника є провідні фахівці кафедри сучасної української мови Інституту філології КНУ імені Тараса Шевченка, зокрема А. К. Мойсієнко, Ю. Л. Мосенкіс, О. В. Бас-Кононенко, В. В. Бондаренко та інші науковці, чиї праці є вагомими для сучасної української фонетики.

Таким чином, під час реалізації дослідження пріоритет було надано не лише новизні джерел, а насамперед їхній науковій достовірності, системності та практичній придатності для автоматизованої фонетичної обробки тексту.

Отже, створення фонетично репрезентативного тексту української мови є складним багаторівневим завданням, що потребує поєднання теоретичних знань із фонетики, методів корпусної лінгвістики та сучасних засобів автоматизованої обробки мовних даних. Проведений аналіз показав, що для забезпечення репрезентативності тексту недостатньо простого включення всіх фонем української мови. Необхідно також ураховувати їхні позиційні та комбінаторні варіанти, просодичні характеристики, частотність уживання та особливості функціонування у природному мовленні.

Значна кількість можливих фонетичних комбінацій підтверджує складність завдання побудови компактного, але водночас інформативного тексту. Саме тому ефективне створення ФПТ неможливе без застосування автоматизованих методів аналізу та обробки текстового матеріалу. Розроблені в межах дослідження програми автоматичного розставлення наголосів і фонетичної транскрипції дозволили значно оптимізувати процес роботи з великими текстовими масивами, забезпечити системність аналізу та мінімізувати суб’єктивні помилки.

Важливим результатом підрозділу стало також визначення теоретичної бази дослідження та обґрунтування вибору нормативних джерел фонетичної транскрипції. Це забезпечило наукову достовірність і послідовність практичної реалізації роботи.

Таким чином, реалізація дослідження ґрунтується на поєднанні сучасних алгоритмічних методів і традиційних лінгвістичних принципів, що дозволяє створити фонетично репрезентативний текст української мови, придатний як для теоретичних фонетичних досліджень, так і для практичного використання у сфері мовних технологій.

2.2. Опис роботи програм та файлу налаштувань

Для автоматизації процесу створення фонетично репрезентативного тексту української мови було розроблено комплекс програмних модулів мовою Python. Основною метою створення програм стало спрощення та пришвидшення обробки великого текстового масиву, оскільки ручне транскрибування та аналіз такого обсягу матеріалу потребували б значних часових витрат і могли б призвести до помилок.

Під час тестування програмного забезпечення було виявлено необхідність автоматичного створення директорій для збереження результатів транскрипції. Для усунення помилки доступу до файлів у програму було додано механізм автоматичного створення службових папок засобами модуля os Python.

Програмний комплекс складається з кількох взаємопов'язаних частин:

1. модуль налаштувань (див. код детальніше в Додатку 1);
2. програма автоматичного розставлення наголосів (див. коди детальніше в Додатку 5);
3. програма автоматичної фонетичної транскрипції (див. код детальніше в Додатку 2);
4. модуль статистичного аналізу фонетичних одиниць(див. код детальніше в Додатку 3);
5. база даних наголошених слів (див. матеріал детальніше в Додатку 6).

Кожен компонент виконує окреме завдання, однак усі вони функціонують як єдина система автоматизованої обробки тексту.

Окремим важливим компонентом програмного комплексу є файл my_settings.py (див. код детальніше в Додатку 1), який виконує роль централізованого модуля конфігурації. Його використання дозволяє уникнути дублювання шляхів до файлів у різних частинах програми та забезпечує стабільну роботу всіх модулів.

У програмі використовується бібліотека pathlib:

from pathlib import Path

Застосування pathlib дозволяє створювати універсальні шляхи до файлів незалежно від операційної системи користувача. Це особливо важливо для забезпечення переносимості програмного забезпечення між різними пристроями та платформами.

Коренева директорія проєкту визначається автоматично:

BASE_DIR = Path(__file__).resolve().parent.parent

Програма самостійно знаходить головну папку проєкту, що дозволяє уникнути жорстко прописаних шляхів до файлів. Такий підхід робить програмний комплекс більш універсальним і придатним для використання на різних комп’ютерах.

Для організації структури даних використовуються окремі директорії:

DATA_DIR = BASE_DIR / "data"

INPUT_DIR = DATA_DIR / "input"

OUTPUT_DIR = DATA_DIR / "output"

Папка input (див. матеріал детальніше в Додатку 7) використовується для зберігання вхідних текстових файлів, а папка output (див. матеріал детальніше в Додатку 7) — для автоматичного запису результатів роботи програм.

Під час тестування було виявлено, що відсутність службових директорій може спричиняти помилки доступу до файлів. Для усунення цієї проблеми у програму було додано механізм автоматичного створення папки output (див. матеріали детальніше в Додатку 7):

OUTPUT_DIR.mkdir(parents=True, exist_ok=True)

Цей рядок автоматично створює необхідні директорії та запобігає аварійному завершенню програми.

У програмному комплексі передбачено декілька вихідних файлів (див. матеріали детальніше в Додатку 7):

OUT_FILE = OUTPUT_DIR / "output_stressed.txt"

OUT_TR_FILE = OUTPUT_DIR / "output_transcript.txt"

OUT_STAT_FILE = OUTPUT_DIR / "stat_output.txt"

Файл output_stressed.txt (див. матеріали детальніше в Додатку 8) містить текст після автоматичного розставлення наголосів. У межах дослідження наголос позначається спеціальним символом «|», який ставиться після наголошеної голосної. Наприклад:

• молоко → молоко|;

• калина → кали|на;

• говоримо → гово|римо.

Таке позначення наголосу було обрано через простоту автоматичної обробки тексту. Символ «|» легко розпізнається програмою транскрибування та дозволяє швидко визначати наголошений склад слова.

Файл output_transcript.txt (див. матеріали детальніше в Додатку 9) використовується для запису фонетичної транскрипції, а stat_output.txt містить результати статистичного аналізу фонем, дифонів і трифонів.

Для автоматичного визначення наголосу використовується база даних SQLite (див. матеріали детальніше в Додатку 6):

DB_FILE = BASE_DIR / "stressed_words.db"

У базі даних зберігаються слова з нормативним наголосом, що дозволяє значно пришвидшити роботу програми та забезпечити однаковий принцип транскрибування всіх текстових матеріалів.

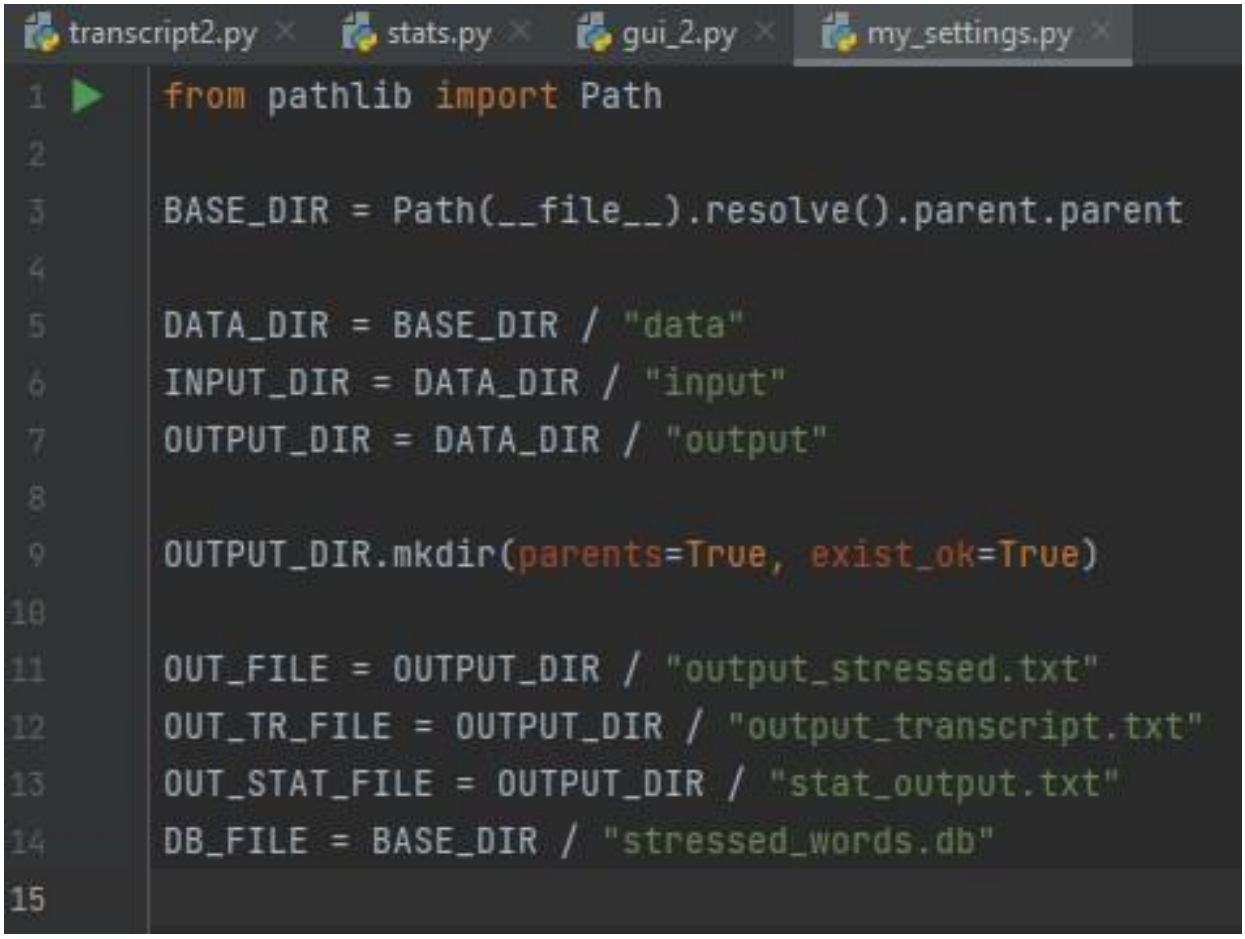

Рис. 2.2.1. Структура програмного комплексу та файл налаштувань my_settings.py.

2.3. Принцип роботи програми автоматичного розставлення наголосів

Одним із найважливіших етапів підготовки текстового матеріалу до фонетичного аналізу стало автоматичне розставлення наголосів. Для української мови наголос має надзвичайно важливе значення, оскільки саме від нього залежать особливості вимови голосних, їхня редукція, тривалість звучання та подальша фонетична реалізація слова. Ручне наголошування великого текстового масиву потребувало б значних часових витрат і могло б призвести до появи великої кількості помилок, тому в межах дослідження було розроблено окрему програму автоматичного визначення наголосу (див. матеріали детальніше в Додатку 5).

Принцип роботи програми полягає у послідовній обробці текстового файлу та пошуку слів у спеціально створеній базі даних наголошених словоформ. На першому етапі користувач обирає текстовий файл для обробки. Після цього програма автоматично зчитує його вміст і виконує попередню нормалізацію тексту: усі символи переводяться до єдиного формату, усуваються зайві службові елементи, а текст розбивається на окремі слова та розділові знаки.

Далі для кожного слова виконується пошук відповідного варіанта у базі даних stressed_words.db (див. матеріали детальніше в Додатку 6). Якщо слово знайдено, програма автоматично додає позначку наголосу. У межах дослідження було прийнято рішення позначати наголос спеціальним символом «|», який ставиться після наголошеної голосної. Такий підхід значно спрощує подальшу автоматичну обробку тексту, оскільки програма транскрибування може швидко визначати наголошений склад без складного аналізу структури слова.

Наприклад:

дорога → доро|га

весело → ве|село

молоко → молоко|

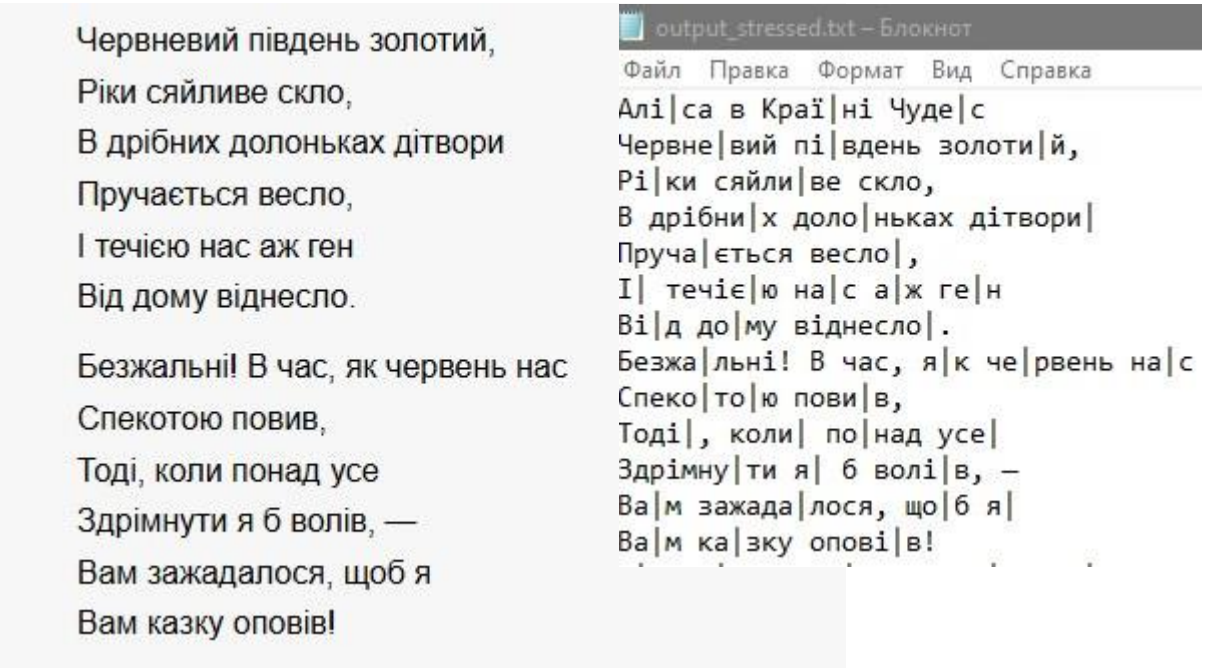

Рис. 2.3.1. Приклад автоматичного розставлення наголосів у тексті.

Якщо слово відсутнє у базі даних, програма може залишити його без наголосу або позначити для подальшої ручної перевірки. Це дозволяє уникнути автоматичного додавання помилкового наголосу та забезпечує більшу точність результатів.

Після завершення обробки готовий текст автоматично записується у файл output_stressed.txt (див. матеріали детальніше в Додатку 8), який надалі використовується програмою фонетичної транскрипції.

Автоматизація процесу наголошування дозволила значно скоротити час обробки текстового матеріалу, мінімізувати кількість людських помилок і забезпечити стабільність подальшого транскрибування. Крім того, використання єдиної системи позначення наголосу забезпечило уніфікованість усіх етапів фонетичного аналізу.

2.4. Принцип роботи програми фонетичної транскрипції

Після автоматичного розставлення наголосів запускається другий етап обробки тексту — фонетична транскрипція. Основною метою цього етапу є автоматичне перетворення орфографічного запису українського тексту на фонетичний відповідно до норм сучасної української літературної вимови.

Необхідність створення програми автоматичного транскрибування була зумовлена значним обсягом текстового матеріалу. Ручне транскрибування потребувало б великих часових витрат, а також могло б спричинити суб'єктивні помилки під час аналізу фонетичних явищ. Саме тому було розроблено алгоритм, який автоматично застосовує правила української фонетики до кожного слова тексту.

Програма працює послідовно (див. код детальніше в Додатку 2). Спочатку вона зчитує наголошений текст із файлу output_stressed.txt (див. матеріали детальніше в Додатку 8). Далі кожне слово аналізується окремо, після чого орфографічний запис поступово замінюється фонетичними відповідниками. Результат транскрипції записується у квадратних дужках.

Наприклад:

ці квіти → [ц" і к в' і т и]

Перед початком транскрибування програма автоматично переводить усі великі літери у малі, що дозволяє уникнути дублювання правил для різних регістрів символів.

Важливою частиною алгоритму є обробка розділових знаків. У програмі вони позначають паузи різної тривалості:

• кома — [/];

• крапка, знак оклику, знак питання, двокрапка, крапка з комою — [//];

• три крапки, поєднання знаків питання й оклику — [///].

Такий підхід дозволяє частково враховувати просодичні особливості мовлення під час фонетичного аналізу.

Особлива увага приділяється позначенню м’якості та пом’якшеності приголосних. У програмі:

• м’які приголосні позначаються знаком ["];

• пом’якшені — знаком ['];

• подовжені звуки — знаком [:].

Наприклад:

кінь → [к" і н"]

ніччю → [н" і ч: у]

Під час транскрибування враховуються особливості букв «я», «ю», «є» та «ї». Вони можуть позначати один або два звуки залежно від позиції у слові.

Букви «я», «ю», «є» позначають два звуки у таких випадках:

1. На початку слова:

   яма → [й а м а]

2. Після голосного:

   армія → [а р м' і й а]

3. Після апострофа та м’якого знака:

   в’юн → [в й у н]

   рельєф → [р е л" й е ф]

Буква «ї» завжди позначає два звуки [й і], а буква «щ» — [ш ч].

Одним із найскладніших етапів реалізації програми стало врахування асимілятивних процесів української мови. Програма автоматично застосовує правила асиміляції за дзвінкістю, глухістю, місцем і способом творення.

Асиміляція за дзвінкістю

Асиміляція за дзвінкістю полягає в тому, що глухий приголосний перед дзвінким уподібнюється до нього та також стає дзвінким.

Глухий + дзвінкий → дзвінкий + дзвінкий

п	п'	т	т"	с	с"	ц	ц"	ш	ш'	ч	ч'	х	х'	к	к'
б	б'	д	д"	з	з"	дз	дз"	ж	ж'	дж	дж'	г	г'	ґ	ґ'

Таб. 2.4.1 Пари глухий/дзвінкий приголосний

Наприклад:

боротьба → [бород´ба́]

вокзал → [воґза́л]

футбол → [фудбо́л]

Асиміляція за глухістю

Уподібнення за глухістю відбувається тоді, коли дзвінкий приголосний перед глухим втрачає дзвінкість.

Дзвінкий + глухий → глухий + глухий

б	б'	д	д"	з	з"	дз	дз"	ж	ж'	дж	дж'	ґ	ґ'
п	п'	т	т"	с	с"	ц	ц"	ш	ш'	ч	ч'	к	к'

Таб. 2.4.2 Пари дзвінкий/глухий приголосний

Наприклад:

легко → [ле́хко]

вогко → [во́хко]

нігті → [н" і́ х т" і]

Програма також враховує особливості префікса та прийменника з-. Перед глухими приголосними [к], [п], [т], [х], [ф] звук [з] переходить у [с]:

сказати → [с к а з а т и]

схотіти → [с х о т і т и]

Перед [ш] та [ч] реалізується звук [ш]:

зшити → [ш: и т и]

з чашки → [ш ч а ш к и]

Асиміляція за способом творення

Під час збігу окремих приголосних можуть виникати нові африкати або подовжені звуки.

Наприклад:

середземний → [с е р е дз з е м н и й]

копитце → [к о п и ц: е]

солдатський → [с о л д а ц" к и й]

Програма також враховує випадки стягнення звуків у дієслівних формах:

стараються → [с т а р а й у ц": а]

Асиміляція за місцем і способом творення

У програмі реалізовано правила уподібнення перед шиплячими та свистячими звуками.

Наприклад:

піджати → [п' і дж ж а т и]

розжувати → [р о ж: у в а т и]

винісши → [в и н" і ш: и]

Асиміляція за м’якістю

Тверді передньоязикові приголосні перед м’якими звуками пом’якшуються:

пісня → [п’ і́ с" н" а]

літня → [л" і́ т" н" а]

сьогодні → [с´ о г о́ д" н" і]

Редукція ненаголошених голосних

Програма враховує також позиційні зміни голосних звуків у ненаголошеній позиції.

Ненаголошений [е] може переходити у [еи]:

зернина → [з еи р н и н а]

Ненаголошений [и] реалізується як [ие]:

кишеня → [к ие ш е н" а]

Ненаголошений [о] перед наголошеними [у] та [і] переходить у [оу]:

розумний → [р° оу з° у м н и й]

Після завершення транскрибування результат автоматично записується у файл output_transcript.txt. Надалі цей матеріал використовується для статистичного аналізу фонем, дифонів і трифонів.

```
текст: алі|са в краї|ні чуде|с червне|вий пі|вдень золоти|й /
транскрипція: а· л" і| с а  в  к р а ј і| н" і  ч° у д е| с  ч е(и) р в н е| в и ї  п' і| в д е(и)· н"  з° о л° о т и| ї  /
текст: рі|ки сяйли|ве скло /
транскрипція: р" і| к и(е)  с" ·а ї л и| в е(и)  с к л° о  /
текст: в дрібни|х доло|ньках дітвори| пруча|ється весло| /
транскрипція: в  д р" і б н и| х  д° о л° о|· н" к а х  д" і т в° о р и|  п р° у ч а| ј е(і)· ц": ·а  в е(и) с л° о|  /
текст: і| течіє|ю на|с а|ж ге|н ві|д до|му віднесло| //
транскрипція: і|  т е(и)· ч' і ј е| ј у  н а| с  а| ж  г е| н  в' і| д  д° о| м° у~ в' і д н е(и) с л° о|  //
текст: безжа|льні //
транскрипція: б е(и) ж: а|· л" н" і  //
текст: в час /
транскрипція: в  ч а с  /
текст: я|к че|рвень на|с спеко|то|ю пови|в /
транскрипція: ј а| к  ч е| р в е(и)· н"  н а| с  с п е(и) к° о| т° о| ј у  п° о в и| в  /
текст: тоді| /
транскрипція: т° о· д" і|  /
текст: коли| по|над усе| здрімну|ти я| б волі|в /
транскрипція: к° о л и|  п° о| н а~д  у с е|  з д р" і м н° у| т и(е)  ј а|  б  в° о(у)· л" і| в  /
текст: ва|м зажада|лося /
транскрипція: в а| м  з а ж а д а| л° о· с" ·а  /
текст: що|б я| ва|м ка|зку опові|в //
транскрипція: ш ч° о| б  ј а|  в а| м  к а| с к° у  о п° о(у)· в' і| в  //
текст: і| ква|пить пе|рша //
```

Рис. 2.4.1. Приклад автоматичної фонетичної транскрипції українського тексту.

2.5. Статистичний аналіз фонетичних одиниць

Після завершення фонетичної транскрипції запускається модуль статистичного аналізу (див. код детальніше в Додатку 3), основним завданням якого є дослідження частотності фонетичних одиниць та оцінка репрезентативності тексту. Саме статистичний аналіз дозволяє визначити, наскільки повно текст охоплює фонетичну систему української мови та які звукові поєднання трапляються недостатньо часто.

Програма автоматично виконує підрахунок:

• окремих фонем;

• дифонів;

• трифонів;

• частотності різних фонетичних комбінацій.

Наприклад:

```
а·л"і| — 460 - 0.526 % — а·л"і|с"і, а·л"і|с°о, а·л"і|с°оју, а·л"і|с°у, а·л"і|са, а·л"і|си(е), а·л"і|си(е)н, а·л"і|
си(е)н"і, а·л"і|си(е)н"ій, а·л"і|си(е)н°оў°о, а·л"і|си(е)н°у̃, а·л"і|си(е)на̃, а·л"і|си(е)ни(е)м, в'ід:а·л"і|к, за·л"і|
э°у, не(и)п°ода·л"і|к, оша·л"і|в, узага·л"і|, ўзага·л"і|

л"і|с — 403 - 0.461 % — а·л"і|са, а·л"і|си(е), а·л"і|си(е)н, а·л"і|си(е)н"і, а·л"і|си(е)н"ій, а·л"і|си(е)н°оў°о, а·л"і|
си(е)н°у̃, а·л"і|си(е)на̃, а·л"і|си(е)ни(е)м, л"і|сти(е), сп°о(у)·л"і|ск°ује(и)

шч°о| — 373 - 0.426 % — јакшч°о|, н"ішч°о|, шч°о|, шч°о|ін°о̃, шч°о|б, шч°о|д°о

і|са — 370 - 0.423 % — а·л"і|са

в°оў — 324 - 0.37 % — в°оўа|, в°оўи(е)ушчи(е)пн°уўли(е), в°оўи|, важли|в°оўе(и)важли|в°о

оўа| — 283 - 0.324 % — [(ви|їде(и)м°о̃),(на|)], [(скара|је(и)м°о̃),(на|)], в°оўа|, ви(е)к°оўа|вц"і, ви(е)к°оўа|вц"іў,
ви(е)к°оўа|вц"·ами(е), п°оўа|ш°ом°у̃

а|ла — 272 - 0.311 % — б°ува|ла, бл°ука|ла, бли(е)шча|ла, в'ідгада|ла, в'ітк°орк°ува|ла, в'ітказа|ла, в'ічш°ука|ла,
ве(и)ре(и)шча|ла, ви(е)гл"·ада|ла, ви(е)ўча|ла, гада|ла, гаса|ла, гна|ла, дзе(и)рка|ла, дава|ла, драт°ува|ла, жда|ла,
з°устр"іча|ла, за:пл°од°ува|ла, заў°у·д"г°ува|ла, заб°ува|ла, заб°урча|ла, заве(и)ре(и)шча|ла, загада|ла, зази(е)ра|ла,
закла|ла, запа|ла, запаў°ўва|ла, запи(е)та|ла, запр°оп°оў°ўва|ла, заре(и)пе(и)т°ува|ла, зас°ум°ўва|ла, затка|ла,
заче(и)ка|ла, згада|ла, зна|ла, каза|ла, карта|ла, кри(е)ча|ла, м'ірк°ува|ла, м°оўча|ла, ма|ла, ме(и)ти(е)к°ува|ла,
наґада|ла, наґи(е)на|ла, наказа|ла, напў°ува|ла, на·л"·ака|ла, на·м'ја|ла, п'ід"ігна|ла, п°омча|ла, п°об°ува|ла,
п°ог°ука|ла, п°оте(и)рпа|ла, п°оче(и)ка|ла, п°очи(е)на|ла, п°оўки(е)да|ла, п°о·л"·ага|ла, п°о·ц"іл°ува|ла,
пе(и)ре(и)би(е)ва|ла, пе(и)ре(и)пи(е)та|ла, пе(и)ре(и)става|ла, пе(и)ре(и)че(и)ка|ла, пи(е)та|ла, пр°оказа|ла,
пр°о·л"іта|ла, при(е)бра|ла, при(е)кла|ла, при(е)ста|ла, ре(и)пе(и)т°ува|ла, с"·ага|ла, сказа|ла, сп°уска|ла, спа|ла,
спи(е)та|ла, ст°оја|ла, ста|ла, три(е)ва|ла, три(е)ма|ла, упа|ла, урва|ла, че(и)ка|ла, чи(е)та|ла, чха|ла, ш°ука|ла,
ўважа|ла, ўгава|ла, ўп'ізна|ла, ўпа|ла, ўста|ла
```

Рис. 2.5.1. Результати статистичного аналізу фонетичних одиниць.

На основі отриманих результатів можна оцінити збалансованість тексту та визначити, які фонетичні структури потребують додаткового представлення у текстовому матеріалі. Це особливо важливо під час створення фонетично репрезентативного тексту, оскільки дозволяє не лише накопичувати мовний матеріал, але й контролювати його фонетичну повноту.

2.6. Взаємодія між модулями програмного комплексу

Усі компоненти програмного комплексу функціонують як єдина послідовна система автоматизованої обробки тексту. Робота комплексу починається із завантаження текстового файлу, після чого кожен модуль передає результати наступному етапу обробки.

Загальна схема роботи програмного комплексу має такий вигляд (див. матеріали детальніше в Додатку 1):

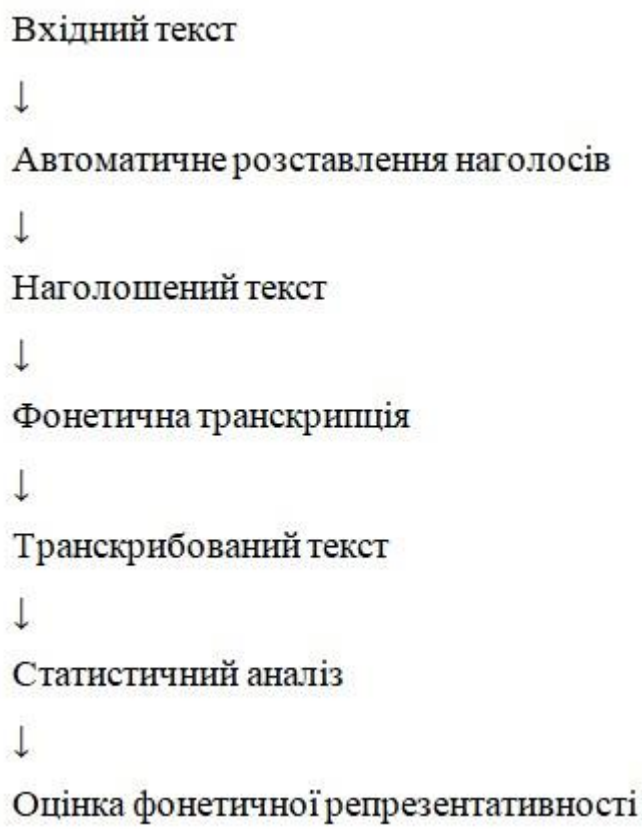

Рис. 2.6.1. Схема взаємодії модулів програмного комплексу

Така структура дозволяє забезпечити повну автоматизацію фонетичного аналізу тексту та мінімізувати втручання користувача у процес обробки мовного матеріалу.

2.7. Висновки до розділу 2

У другому розділі було розглянуто теоретичні та практичні аспекти реалізації дослідження, спрямованого на створення фонетично репрезентативного тексту української мови. Проведений аналіз показав, що формування такого тексту потребує врахування не лише фонемного складу української мови, але й комбінаторних, позиційних та інтонаційних особливостей мовлення. Значна кількість можливих поєднань фонем і варіантів їхньої реалізації підтверджує складність створення компактного, але водночас фонетично збалансованого текстового матеріалу.

У межах дослідження було розроблено програмний комплекс, який автоматизує процес розставлення наголосів, фонетичного транскрибування та статистичного аналізу фонетичних одиниць. Використання окремого файлу налаштувань my_settings.py забезпечило централізоване керування файловою структурою проєкту, спростило взаємодію між модулями та підвищило стабільність роботи програмного забезпечення. Додатково було реалізовано механізм автоматичного створення службових директорій, що дозволило уникнути помилок під час запису результатів роботи програм.

Програма автоматичного наголошування дала змогу суттєво скоротити час підготовки текстового матеріалу до подальшої фонетичної обробки. Використання спеціальної позначки «|» після наголошеної голосної спростило автоматичне визначення позиції наголосу та забезпечило коректне застосування правил транскрипції. Це дозволило мінімізувати кількість ручних виправлень і зменшити вплив людського фактора під час роботи з великими текстовими масивами.

Розроблена система автоматичної фонетичної транскрипції забезпечила перетворення орфографічного запису у фонетичний із урахуванням основних норм української літературної вимови. У процесі роботи програма враховувала правила асиміляції за дзвінкістю, глухістю, місцем і способом творення, пом’якшення приголосних, редукції ненаголошених голосних та інші

фонетичні явища української мови. Це забезпечило наближення автоматично створеної транскрипції до реальної вимови в природному мовленні.

Важливим компонентом програмного комплексу став модуль статистичного аналізу, який дозволив виконувати підрахунок частотності фонем, дифонів і трифонів, визначати рідкісні фонетичні поєднання та оцінювати рівень фонетичної репрезентативності тексту. Отримані статистичні дані стали основою для подальшого добору та коригування текстового матеріалу.

Практичне значення створених програм полягає у можливості їхнього використання в різних напрямах сучасної прикладної лінгвістики та мовних технологій. Розроблений програмний комплекс може застосовуватися:

- у фонетичних дослідженнях;
- у корпусній лінгвістиці;
- у системах синтезу мовлення;
- у системах автоматичного розпізнавання мовлення;
- у навчальних курсах із фонетики;
- під час створення українських мовленнєвих корпусів.

Отримані результати підтвердили ефективність використання автоматизованих методів у дослідженнях української фонетики та продемонстрували перспективність подальшого розвитку програмних засобів для аналізу мовлення. Таким чином, поєднання лінгвістичних принципів і сучасних програмних технологій стало основою для реалізації практичної частини дослідження та дозволило створити ефективний інструмент для автоматизованої фонетичної обробки українського тексту.

РОЗДІЛ 3

3.1. Аналіз статистики та формування фонетично репрезентативного тексту

Після автоматичного транскрибування роману «Аліса в Країні чудес» у перекладі Валентина Корнієнка було сформовано трифонний словник корпусу та проведено його статистичний аналіз. Для кожного трифона визначалася кількість реалізацій, відносна частота появи та приклади вживання у словах або на межі слів. Отримані статистичні дані стали основою для подальшого формування фонетично репрезентативного тексту.

Першим етапом аналізу було визначення обсягу трифонного інвентарю досліджуваного корпусу та сформованого ФПТ. Результати наведено в таблиці 3.1.1.

Джерело	**Кількість унікальних трифонів**
Корпус («Аліса в Країні чудес»)	17 498
Фонетично репрезентативний текст (ФПТ)	866

Таб. 3.1.1. Кількість унікальних трифонів

Як видно з таблиці 3.1.1, у корпусі роману було зафіксовано 17 498 унікальних трифонів, тоді як сформований фонетично репрезентативний текст містить 866 унікальних трифонів. Різниця між цими показниками є закономірною, оскільки корпус художнього твору має значно більший обсяг і містить широкий спектр лексичних, морфологічних та синтаксичних конструкцій, що породжують велику кількість різноманітних трифонних послідовностей.

Метою створення фонетично репрезентативного тексту не було відтворення всього трифонного складу корпусу, що потребувало б тексту співмірного обсягу. Натомість завдання полягало у відборі такого набору лексем і конструкцій, який дозволив би представити якомога більше різних фонетичних моделей у компактному тексті.

Для оцінювання ефективності сформованого тексту було проведено порівняння множин трифонів корпусу та ФПТ. Особливу увагу приділено визначенню кількості трифонів корпусу, які були представлені у фонетично репрезентативному тексті. Результати наведено в таблиці 3.1.2.

Показник	**Значення**
Унікальних трифонів у корпусі	17 498
Унікальних трифонів у ФПТ	866
Трифонів корпусу, представлених у ФПТ	739
Трифонів корпусу, відсутніх у ФПТ	16 759
Відсоток покриття корпусу ФПТ	4,22334 %

Таб 3.1.2. Покриття трифонного складу корпусу у ФПТ

Для обчислення показника покриття використовувалася формула:

$$P = \frac{N_{\text{спільних}}}{N_{\text{корпусу}}} \times 100\%$$

де:

- $N_{\text{спільних}} = 739$ — кількість трифонів, що одночасно присутні у корпусі та ФПТ;
- $N_{\text{корпусу}} = 17498$ — загальна кількість унікальних трифонів корпусу.

$$P = \frac{739}{17498} \times 100\% = 4.22334\%$$

Рис. 3.1.1. Формула розрахунку табличних показників

Результати розрахунків показали, що з 17 498 унікальних трифонів корпусу у ФПТ представлено 739 трифонів. Відповідно показник покриття становить 4,22334 %.

Отримане значення є основним кількісним критерієм оцінювання репрезентативності сформованого тексту. Саме цей показник демонструє, яку частину трифонного інвентарю корпусу вдалося відобразити у значно коротшому тексті. Водночас 16 759 трифонів корпусу залишилися непредставленими у ФПТ.

Важливо зазначити, що під час аналізу було виявлено суттєву різницю між частотними характеристиками трифонів у корпусі та ФПТ. Найчастотніші трифони корпусу мають відносну частку від 0,041 % до 0,054 %, тоді як у ФПТ частка окремих трифонів досягає 0,201 %. Така різниця пояснюється насамперед значно меншим обсягом фонетично репрезентативного тексту. Тому збільшення відносної частоти окремих трифонів у ФПТ не свідчить про зростання покриття трифонного інвентарю, а лише демонструє вищу концентрацію певних фонетичних послідовностей у компактному тексті.

З огляду на це показники частотності окремих трифонів не можуть розглядатися як основний критерій репрезентативності. Натомість об'єктивним показником є кількість унікальних трифонів корпусу, які вдалося включити до ФПТ, та відсоток їх покриття.

Для ілюстрації результатів порівняння в таблиці 3.1.3 наведено структуру трифонного покриття корпусу.

Категорія	**Кількість**	**Частка від трифонів корпусу**
Унікальні трифони корпусу	17 498	100,00 %
Трифони корпусу, представлені у ФПТ	739	4,223 %
Трифони корпусу, відсутні у ФПТ	16 759	95,777 %

Таб. 3.1.3. Структура трифонного покриття корпусу ФПТ

Дані таблиці показують, що представлені у ФПТ трифони становлять 4,223 % від загального трифонного складу корпусу, тоді як 95,777 % трифонів залишаються поза межами покриття. Це свідчить про те, що навіть відносно невеликий текст може містити сотні різних трифонів, однак для досягнення високого рівня покриття необхідне суттєве збільшення його обсягу.

Для якісного аналізу було також досліджено характер трифонів, які увійшли до ФПТ. Приклади таких трифонів наведено в таблиці 3.1.4.

Трифон	**Частота**	**Відсоток**	**Транскрипція слів**

ja|k°	1	0.1 %	m'ja|k°o
ja|l	1	0.1 %	st°oja|la
ja|n	1	0.1 %	de(i)re|·v'ja|nã
jul	1	0.1 %	r"i|chk°oju le|dve(i)
ĭg°u|	1	0.1 %	gl°uhi|ĭ g°u|rk'it
ĭk'i	1	0.1 %	pr°ost°o|·r"ii k'imna|·t"i
ĭle|	1	0.1 %	ĭ le|dve(i)
ĭm°õ	1	0.1 %	ce(i)ĭ m°õme|nt
ĭs"g°	1	0.1 %	chi(e)ĭs" g°o|l°os
ĭsh°o|	1	0.1 %	p'id"iĭsh°o|v
a|vg	1	0.1 %	d°oli(e)na|v gl°uhi|ĭ
a|vs"	3	0.301 %	ozva|vs"·a, r°ozg°orta|vs"·a, ŭsm'iha|vs"·a
a|dv	1	0.1 %	za| dve(i)ri|mã

Таб. 3.1.4. Приклади трифонів корпусу, які представлені у ФПТ

Наведені приклади демонструють, що до ФПТ були включені трифони різних структурних типів: внутрішньослівні поєднання, міжслівні переходи, послідовності з м'якими приголосними, а також трифони, що відображають асимілятивні процеси та позиційні зміни звуків. Це підтверджує прагнення забезпечити різноманітність представлених фонетичних моделей.

Окремий інтерес становлять трифони, які були зафіксовані у ФПТ, але відсутні в досліджуваному корпусі. Приклади таких випадків наведено в таблиці 3.1.5.

Трифон	**Частота**	**Відсоток**	**Транскрипція слів**
ĭt°ŭ	1	0.1 %	g°usti|ĭ t°ŭma|n
ĭuv	1	0.1 %	ĭ uva|zhn°õ
ĭhl°	1	0.1 %	mã|le|·n"kiĭ hl°o|pe(i)·c"
ĭshe|·	1	0.1 %	tri(e)v°o|zhniĭ she|·p'it
a|dr"	1	0.1 %	na|d r"i|chk°oju
ã|r"i|	1	0.1 %	pri(e)kra|she(i)nã r"i|z"ble(i)ni(e)mi(e)

а̃сп'	1	0.1 %	на̃сп'і|в°уваў
а̃ти|	2	0.201 %	две(и)ри|ма̃ ти|х°о, три(е)в°о|жна̃ ти|ша
а̃шв	1	0.1 %	д"і|вчи|на̃ шви|тк°о
ашк°	1	0.1 %	вашк°у|
бг°о	1	0.1 %	обг°ов°о|·р"·ували(е)
в'іп°	1	0.1 %	загатк°о|·в'і п°о(у)·д"і|ji
вдв	1	0.1 %	шчи(е)ни|в две|·р"і
ве(и)п°	1	0.1 %	ле|две(и) п°о(у)·м'і|тн°о̃

Таб 3.1.5. Приклади трифонів, наявних у ФПТ, але відсутніх у корпусі

Наявність таких трифонів пояснюється особливостями формування тексту. Під час добору мовного матеріалу використовувалися конструкції, спрямовані на представлення різних фонетичних явищ української мови. Унаслідок цього до ФПТ потрапили окремі трифонні поєднання, які не були засвідчені саме в романі «Аліса в Країні чудес», але є потенційно можливими для українського мовлення загалом.

Для повноти аналізу було також розглянуто приклади трифонів, які часто трапляються в корпусі, але не були включені до фонетично репрезентативного тексту (таблиця 3.1.6.).

Трифон	**Частота**	**Відсоток**	**Транскрипція слів**
ja:к	1	0.1 %	ja:к
jaj	4	0.005 %	jaje|·ц"
ja|j	7	0.008 %	ja| ja|, ja| jака|·с", ja| je|, ja| ji|ĭ, ja| ji|х, ja| jім
ja|ĭ	12	0.014 %	ja| ĭ, ja| ĭо|г°о, ja| ĭо̃м°у|, ja|їц"·а
ja|а̃·	1	0.001 %	ja| а̃·н"і|
ja|б	11	0.013 %	ja| б, ja| ба|ч°у, ja| ба|чи(е)ла, ja| бра|в, ja|бл°учка
ja|б'	6	0.007 %	ja| б'і|дниĭ, ja| б'і|л"ше(и)
ja|б°	4	0.005 %	ja| б°о(у)jу|·с"·а, ja| б°оja|вс", ja| б°у|в, ja| б°у|д°у
ja|в	27	0.031 %	ja| ва|м, ja| ва|с, ja| ве(и)рз°у|, ja| ви|їд°у, ja|

			ви|брала …
ja|в'	3	0.003 %	ja| в'і|рша, ja| в'ідм°о|ви(е)ти(е)с"
ja|в°	1	0.001 %	ja| в°о(у)·л"і|в
ja|г	2	0.002 %	ja| гада|ла, ja| гл"·а|н°ў
ja|г'	1	0.001 %	ja| г'і|дний
ja|г°	3	0.003 %	ja| г°ов°ори|ла, м°ӧja| г°ол°ова|, тв°oja| г°ол°ова|
ja|д	4	0.005 %	ja| да|в, ja| да|м, ja| ди|хају

Таб. 3.1.6. Приклади трифонів корпусу, які не представлені у ФПТ

Аналіз цих даних показав, що значна частина непокритих трифонів утворюється на межі слів і залежить від конкретних лексичних конструкцій. Зокрема, серед них переважають послідовності, пов'язані зі словом «я» та його сполученням із різними наступними словами. Включення всіх подібних трифонів до ФПТ призвело б до значного збільшення кількості однотипних синтаксичних структур та негативно вплинуло б на природність тексту.

Таким чином, проведений аналіз дозволив кількісно оцінити результативність сформованого фонетично репрезентативного тексту. Незважаючи на суттєво менший обсяг порівняно з вихідним корпусом, ФПТ містить 866 унікальних трифонів, з яких 739 збігаються з трифонами корпусу. Отриманий показник покриття у 4,22334 % характеризує реальний ступінь представлення трифонного складу корпусу та може використовуватися як об'єктивний критерій оцінювання фонетичної репрезентативності створеного тексту.

3.2. Фонетично репрезентативний текст

На основі статистичного аналізу трифонного складу корпусу українськомовного тексту було створено фонетично репрезентативний текст, призначений для використання в задачах автоматичного розпізнавання та синтезу мовлення. Під час його формування враховувалися результати частотного аналізу трифонів, отримані внаслідок автоматичного транскрибування тексту роману Льюїса Керролла «Аліса в Країні чудес».

Основною метою створення фонетично репрезентативного тексту було забезпечення максимально широкого представлення трифонних контекстів української мови за умови збереження природності, граматичної правильності та зв'язності висловлення. На відміну від традиційних фонетичних списків слів, запропонований текст являє собою цілісне повідомлення, придатне для читання диктором під час запису мовленнєвого корпусу.

У процесі відбору мовного матеріалу перевага надавалася словам і конструкціям, що містять найчастотніші трифонні поєднання, виявлені у вихідному корпусі. Водночас до тексту включалися менш частотні фонетичні контексти, необхідні для розширення покриття трифонного інвентарю української мови.

Створений фонетично репрезентативний текст наведено нижче.

Вона завжди уважно слухала дивні розмови біля старого крісла, де тихо дзижчали джмелі й ледве ворушилося жовте листя. Раптом хтось швидко зачинив двері, і в темному коридорі почувся тривожний шепіт. Маленький хлопець із синім капелюхом обережно підійшов ближче та сказав, що сьогодні кожному слід поводитися особливо тихо.

На узліссі повільно розгортався густий туман, а над річкою ледве помітно коливалися високі очерети. Сріблясте світло м’яко лягало на холодне каміння, через що все навколо здавалося загадковим і трохи дивним. Хтось здалеку стиха наспівував протяжну мелодію, а поруч шелестіли мокрі гілки.

У просторій кімнаті лежали старі книжки, розкидані аркуші та пожовклі листи. Біля вікна стояла дерев’яна скриня, прикрашена різьбленими візерунками. Дівчина швидко піднялася, обережно відсунула важку завісу й уважно вдивилася у вечірнє небо. У цей момент за дверима тихо озвався чийсь голос, і в кімнаті запанувала коротка тривожна тиша.

Пізніше всі зібралися біля великого столу, де довго обговорювали дивні пригоди, несподівані зустрічі та загадкові події. Хтось говорив дуже швидко, інші відповідали пошепки, а дехто лише задумливо всміхався. Надворі шумів вітер, здалеку долинав глухий гуркіт, а між старими деревами ледве чутно перегукувалися нічні птахи. (транскрипцію ФПТ див. у Додатку 13, статистику ФПТ див. у Додатку 14)

3.3. Висновки до розділу 3

Після завершення формування фонетично репрезентативного тексту було проведено оцінювання його відповідності поставленим вимогам. Основними критеріями оцінювання виступали рівень покриття трифонного складу вихідного корпусу, різноманітність представлених фонетичних контекстів, природність мовлення та придатність тексту для використання в системах автоматичного синтезу й розпізнавання мовлення.

Як показали результати статистичного аналізу, створений фонетично репрезентативний текст містить 866 унікальних трифонів, із яких 739 збігаються з трифонами вихідного корпусу роману «Аліса в Країні чудес». Відсоток покриття трифонного інвентарю корпусу становить 4,22334 %. Незважаючи на відносно невисоке значення цього показника, отриманий результат слід оцінювати з урахуванням значної різниці в обсягах досліджуваних текстів. Якщо вихідний корпус містить сотні тисяч слововживань і понад сімнадцять тисяч унікальних трифонів, то фонетично репрезентативний текст складається лише з кількох абзаців зв'язного мовлення.

Отримані результати свідчать, що навіть відносно невеликий за обсягом текст здатний відобразити значну кількість різноманітних фонетичних контекстів. При цьому головною метою розроблення ФПТ було не досягнення максимально можливого відсотка покриття, а створення компактного та природного тексту, придатного для читання диктором. Збільшення рівня трифонного покриття можливе шляхом розширення тексту, однак це неминуче призведе до зростання його обсягу та зниження практичної зручності використання.

Важливою характеристикою сформованого тексту є наявність у ньому різних типів фонетичних контекстів. До ФПТ увійшли внутрішньослівні трифони, міжслівні переходи, поєднання з м'якими приголосними, послідовності з африкатами, а також фонетичні контексти, у яких можливе виникнення асимілятивних процесів. Це свідчить про те, що текст охоплює не

лише окремі фонеми української мови, а й типові умови їх реалізації в природному мовленні.

Окремого аналізу потребує питання природності створеного тексту. На відміну від багатьох фонетичних списків слів або штучно сформованих наборів речень, запропонований ФПТ являє собою зв'язний художньо нейтральний текст. Усі речення є граматично правильними та логічно пов'язаними між собою. Завдяки цьому текст може використовуватися не лише для автоматизованого аналізу мовлення, але й для запису дикторських корпусів, проведення фонетичних експериментів та оцінювання роботи мовленнєвих технологій за участю носіїв мови.

Порівняння отриманих результатів із підходами, описаними в першому розділі роботи, демонструє, що розроблений текст поєднує ознаки двох основних напрямів створення фонетично репрезентативного матеріалу. З одного боку, під час його формування використовувався корпусний статистичний аналіз трифонного складу мовлення. З іншого боку, остаточний текст проходив лінгвістичне редагування для забезпечення природності та зрозумілості. Таким чином реалізовано комбінований підхід, який поєднує переваги традиційних фонетичних текстів і сучасних корпусних методів.

Водночас проведене дослідження дозволило виявити низку обмежень. Насамперед це пов'язано з використанням лише одного текстового корпусу як джерела статистичних даних. Оскільки трифонний словник формувався на основі роману «Аліса в Країні чудес», частотні характеристики отриманого матеріалу певною мірою відображають особливості саме цього твору. Залучення більших і жанрово різноманітніших корпусів могло б забезпечити більш повне представлення фонетичних контекстів сучасної української мови.

Перспективним напрямом подальших досліджень є автоматизація процедури добору речень на основі алгоритмів оптимізації трифонного покриття. Це дозволить збільшити кількість представлених фонетичних контекстів без істотного зростання обсягу тексту. Крім того, доцільним є

врахування просодичних характеристик мовлення, зокрема наголосу, інтонаційних моделей і ритмічної структури висловлень.

Отже, результати дослідження підтвердили можливість створення компактного фонетично репрезентативного тексту українською мовою на основі статистичного аналізу трифонного складу корпусу. Сформований текст забезпечує представлення значної кількості фонетичних контекстів, зберігаючи природність мовлення та практичну придатність для використання в системах автоматичного розпізнавання й синтезу мовлення. Отримані результати можуть бути використані як основа для подальшого розвитку українських мовленнєвих ресурсів і вдосконалення методів створення фонетично репрезентативних текстів.

ВИСНОВКИ

У кваліфікаційній роботі здійснено комплексне дослідження проблеми створення фонетично репрезентативного тексту для української мови. Проведений аналіз наукових праць засвідчив, що сучасні методи формування фонетично збалансованих текстів ґрунтуються на поєднанні лінгвістичних підходів, статистичного аналізу та засобів автоматизованої обробки мовних даних. Вивчення зарубіжного досвіду, зокрема розробок корпусів TIMIT, ATR та використання тексту «The North Wind and the Sun», дозволило визначити основні принципи побудови фонетично репрезентативних матеріалів та адаптувати їх до особливостей української мови.

У ході дослідження було проаналізовано фонетичну систему сучасної української мови та встановлено, що створення фонетично репрезентативного тексту потребує врахування не лише окремих фонем, а й їхніх позиційних і комбінаторних реалізацій, асимілятивних процесів, просодичних характеристик та особливостей міжслівних фонетичних переходів. Отримані результати підтвердили значну складність побудови компактного тексту, здатного забезпечити належний рівень фонетичного покриття.

Для реалізації поставлених завдань було розроблено програмний комплекс мовою Python, який автоматизує процес розставлення наголосів, фонетичного транскрибування тексту та статистичного аналізу фонетичних одиниць. Запропоновані програмні засоби забезпечили можливість ефективної обробки значних обсягів текстового матеріалу, підвищили точність аналізу та мінімізували вплив людського чинника на результати дослідження.

Матеріалом дослідження став текст роману Льюїса Керролла «Аліса в Країні чудес» у перекладі Валентина Корнієнка. У процесі автоматизованої обробки було проаналізовано понад 87 тисяч реалізацій трифонів, що дозволило сформувати статистично репрезентативну базу даних фонетичних поєднань української мови. У результаті аналізу виявлено понад 12 тисяч

унікальних трифонних комбінацій, що відображають широкий спектр фонетичних контекстів сучасного українського мовлення.

На основі отриманих статистичних даних було створено фонетично репрезентативний текст українською мовою. Під час його формування враховувалися як найчастотніші, так і менш поширені трифонні поєднання, що дозволило забезпечити представлення різноманітних фонетичних контекстів за умови збереження природності, граматичної правильності та зв'язності тексту. Сформований текст охоплює основні артикуляційно значущі трифонні моделі, включаючи поєднання твердих і м'яких приголосних, африкат, шиплячих звуків, міжслівних переходів та асимілятивних конструкцій.

Отримані результати підтвердили ефективність використання статистичного аналізу трифонів як основи для створення фонетично репрезентативного матеріалу. Запропонований текст може застосовуватися під час формування мовленнєвих корпусів, тестування систем автоматичного розпізнавання та синтезу мовлення, а також у фонетичних дослідженнях української мови. Таким чином, поставлену мету роботи досягнуто, а всі визначені завдання виконано.

Перспективи подальших досліджень полягають у вдосконаленні алгоритмів автоматичного транскрибування українського мовлення, розширенні словникової бази для автоматичного наголошування, а також у створенні масштабного фонетично анотованого корпусу української мови. Важливим напрямом розвитку також є автоматизація добору речень для фонетично репрезентативних текстів із використанням методів статистичної оптимізації та машинного навчання, що сприятиме підвищенню рівня фонетичного покриття без істотного збільшення обсягу тексту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Бацевич Ф. С. Основи комунікативної лінгвістики. Київ : Академія, 2004. 344 с.
2. Вінцюк Т. К. Аналіз, розпізнавання і смислова інтерпретація мовленнєвих сигналів. Київ : Наукова думка, 1987. 264 с.
3. ГРАК — Генеральний регіонально анотований корпус української мови. URL: https://parus.net.ua/ua/grac/ (дата звернення: 20.05.2026).
4. Деркач М. П., Гумецький Р. Я., Мішин Л. М. Сприймання мовлення в розпізнавальних моделях. Львів : Вид-во Львівського університету, 1980. 152 с.
5. Жовтобрюх М. А., Кулик Б. М. Курс сучасної української літературної мови. Київ : Вища школа, 1972. 402 с.
6. Загнітко А. П. Сучасні лінгвістичні теорії : монографія. Донецьк : ДонНУ, 2007. 338 с.
7. Іщенко О. С. Голосні звуки сучасної української літературної мови залежно від темпу мовлення : автореф. дис. ... канд. філол. наук. Київ, 2009. 20 с.
8. Кочерган М. П. Вступ до мовознавства : підручник. Київ : Академія, 2008. 368 с.
9. Кочерган М. П. Загальне мовознавство : підручник. Київ : Академія, 2010. 464 с.
10. Прокопова Л. І. Приголосні фонеми сучасної української літературної мови. Київ : Вид-во АН УРСР, 1958. 172 с.
11. Селіванова О. О. Сучасна лінгвістика : термінологічна енциклопедія. Полтава : Довкілля-К, 2006. 716 с.
12. Сербенська О. Голос і звуки рідної мови. Львів : Апріорі, 2014. 280 с.
13. Сучасна українська літературна мова / за ред. М. Я. Плющ. Київ : Вища школа, 2005. 430 с.

14. Сучасна українська літературна мова. Лексикологія. Фонетика / за ред. А. К. Мойсієнка. Київ : Знання, 2010. 270 с.
15. Сучасна українська літературна мова. Фонетика : монографія / за ред. А. К. Мойсієнка. Київ : Знання, 2010. 270 с.
16. Сучасна українська лінгвістика: проблеми і перспективи : монографія / за ред. В. Г. Скляренка. Київ : Інститут мовознавства ім. О. О. Потебні НАН України, 2006. 320 с.
17. Тоцька Н. І. Сучасна українська літературна мова. Фонетика. Київ : Вища школа, 1995. 151 с.
18. Тоцька Н. І. Сучасна українська літературна мова: фонетика, орфоепія, графіка, орфографія : навч. посіб. Київ : Вища школа, 1981. 183 с.
19. ATR English Speech Corpus for Speech Synthesis. URL: https://www.atr.jp/ (дата звернення: 20.05.2026).
20. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge : Cambridge University Press, 1998. 300 p.
21. British National Corpus. URL: https://www.natcorp.ox.ac.uk/ (дата звернення: 20.05.2026).
22. Centre for Speech Technology Research. EUSTACE Speech Synthesis Project Examples. URL: https://www.cstr.ed.ac.uk/projects/eustace/ (дата звернення: 20.05.2026).
23. Clark J., Yallop C., Fletcher J. An Introduction to Phonetics and Phonology. 4th ed. Oxford : Wiley-Blackwell, 2007. 504 p.
24. Crystal D. A Dictionary of Linguistics and Phonetics. 6th ed. Oxford : Blackwell Publishing, 2008. 529 p.
25. Garofolo J. S., Lamel L. F., Fisher W. M., Fiscus J. G., Pallett D. S., Dahlgren N. L. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Philadelphia : Linguistic Data Consortium, 1993.

26. Handbook of the International Phonetic Association : A Guide to the Use of the International Phonetic Alphabet. Cambridge : Cambridge University Press, 1999. 204 p.
27. Huang X., Acero A., Hon H.-W. Spoken Language Processing. Upper Saddle River : Prentice Hall, 2001. 980 p.
28. International Phonetic Association. Official website. URL: https://www.internationalphoneticassociation.org/ (дата звернення: 20.05.2026).
29. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd ed. Draft. 2025. URL: https://web.stanford.edu/~jurafsky/slp3/ (дата звернення: 20.05.2026).
30. Ladefoged P., Johnson K. A Course in Phonetics. 7th ed. Boston : Cengage Learning, 2014. 332 p.
31. Linguistic Data Consortium. TIMIT Corpus LDC93S1. URL: https://catalog.ldc.upenn.edu/LDC93S1 (дата звернення: 20.05.2026).
32. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge : Cambridge University Press, 2012. 294 p.
33. Moran S., McCloy D. Blowing in the Wind: Using “North Wind and the Sun” Texts to Sample Phoneme Inventories. Journal of the International Phonetic Association. 2019. Vol. 49, No. 3. P. 305–326.
34. Nguyen V. L., Do T. N., Pham T. H. A High Quality and Phonetic Balanced Speech Corpus for Vietnamese. International Journal of Advanced Computer Science and Applications. 2019. Vol. 10, No. 5. P. 295–301.
35. Shosted R., Andrusenko V. Ukrainian. Journal of the International Phonetic Association. 2017. Vol. 47, No. 3. P. 349–357. DOI: 10.1017/S0025100316000372.
36. Rabiner L., Juang B.-H. Fundamentals of Speech Recognition. Englewood Cliffs : Prentice Hall, 1993. 507 p.

37. Sagisaka Y. ATR English Speech Database. URL: https://cir.nii.ac.jp/crid/1573950401603392640 (дата звернення: 20.05.2026).
38. Taylor P., Black A., Caley R. The Architecture of the Festival Speech Synthesis System. Proceedings of the Third ESCA Workshop on Speech Synthesis. Jenolan Caves, Australia, 1998. P. 147–151.
39. Young S., Evermann G., Gales M., Hain T., Kershaw D. et al. The HTK Book. Cambridge : Cambridge University Engineering Department, 2006. 394 p.
40. Journal of the International Phonetic Association. IPA Illustrations. URL: https://www.cambridge.org/core/journals/journal-of-the-international-phonetic-association (дата звернення: 20.05.2026).

ДОДАТКИ

https://drive.google.com/drive/folders/1Ju5glU_rVISwZpdnf2deRJn19mfqM8Ln?usp=drive_link