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

СТВОРЕННЯ ПОЛІФУНКЦІОНАЛЬНОГО ПАРАЛЕЛЬНОГО АНГЛО-УКРАЇНСЬКОГО КОРПУСУ ТЕКСТІВ НА ОСНОВІ КОНТЕНТУ ВІДЕОІГОР

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

Наукові керівники:
к.філол.н., доцент кафедри української мови та прикладної лінгвістики
Оксана ЗУБАНЬ
к.т.н., доцент кафедри інформаційних систем та технологій
Микола КОСТІКОВ

«Допущено до захисту»
Протокол засідання кафедри
української мови та прикладної лінгвістики
від «05» червня 2026 року № 16

завідувач кафедри _____
к.філол.н., доц. Сергій РІЗНИК

КИЇВ – 2026

ЗМІСТ

АНОТАЦІЯ

Дипломну роботу присвячено створенню паралельного корпусу для машинного перекладу на основі текстів відеоігор та оцінюванню якості згенерованого перекладу, отриманого через велику мовну модель Gemini 3.1 Flash-Lite. Актуальність зумовлена стрімким зростанням індустрії відеоігор, посиленням попиту на українську локалізацію в контексті дерусифікації культурного простору та браком структурованих мовних ресурсів ігрового домену для української мови.

Об'єктом дослідження є переклад текстів відеоігор у парі англійська – українська; предметом – методологія побудови паралельного корпусу таких текстів і якість згенерованого перекладу відносно людського еталону. Мета – розробити методологію створення корпусу, сформувати його й оцінити придатність сучасних великих мовних моделей для перекладу ігрових текстів.

Матеріалом слугували локалізаційні файли відеогри Baldur's Gate 3 (Larian Studios). Методологія охоплює витягування текстів із бінарних форматів гри, вирівнювання пар за спільними ідентифікаторами, евристичну класифікацію за функціональними типами, фільтрацію, дедуплікацію та збагачення метаданими; систему реалізовано мовою Python. Корпус містить 186 311 унікальних паралельних пар за сімома функціональними класами; для 4 992 записів стратифікованої вибірки додано згенерований переклад моделі Gemini 3.1 Flash-Lite.

Оцінювання за п'ятьма автоматичними метриками виявило центральний результат: разючу розбіжність між лексичними метриками (BLEU 23,16) і семантичною оцінкою (BERTScore 91,93). Це підтвердило неадекватність метрик буквального збігу, що занижують якість художнього й діалогічного перекладу через варіативність формулювань. Сліпе експертне оцінювання та якісний аналіз за класами окреслили межі застосовності великих мовних моделей середнього класу й потребу постредагування складних фрагментів.

Ключові слова:
паралельний корпус, машинний переклад, локалізація відеоігор, велика мовна модель, метрики оцінки якості, українська локалізація, Baldur's Gate 3

ANNOTATION

The thesis addresses the construction of a parallel corpus for machine translation based on video game texts and the evaluation of the quality of generated translation produced by the Gemini 3.1 Flash-Lite large language model. Its relevance stems from the rapid growth of the video game industry, the rising demand for Ukrainian localization amid the derussification of the cultural space, and the lack of structured language resources for the gaming domain in Ukrainian.

The object of the study is the translation of video game texts in the English–Ukrainian pair; the subject is the methodology of building a parallel corpus of such texts and the quality of AI-generated translation relative to a human reference. The aim is to develop a corpus-construction methodology, build the corpus, and assess the suitability of modern large language models for translating game texts.

The material comprises the localization files of Baldur's Gate 3 (Larian Studios). The methodology covers extracting texts from the game's binary formats, aligning pairs by shared identifiers, heuristic classification by functional type, filtering, deduplication, and metadata enrichment; the system is implemented in Python. The corpus contains 186,311 unique parallel pairs across seven functional classes; for 4,992 records of a stratified sample, a synthetic translation by Gemini 3.1 Flash-Lite was added.

Evaluation with five automatic metrics revealed the central finding: a striking divergence between lexical metrics (BLEU 23.16) and a semantic one (BERTScore 91.93). This confirmed the inadequacy of surface-overlap metrics, which underestimate the quality of literary and dialogic translation due to wording variation. Blind expert evaluation and qualitative analysis by class outlined the limits of mid-tier large language models and the need for post-editing of complex fragments.

Keywords:

parallel corpus, machine translation, video game localization, large language model, translation quality metrics, Ukrainian localization, Baldur's Gate 3

СПИСОК УМОВНИХ СКОРОЧЕНЬ

МП – машинний переклад

ОПП – освітньо-професійна програма

ALPAC – Automatic Language Processing Advisory Committee

APE – Automatic Post-Editing – автоматичне постредагування машинного перекладу

API – Application Programming Interface – програмний інтерфейс додатка

BERT – Bidirectional Encoder Representations from Transformers

BG3 – Baldur's Gate 3

BLEU – Bilingual Evaluation Understudy – метрика автоматичної оцінки якості перекладу

CAT – Computer-Assisted Translation – комп'ютерно-асистований переклад

ChrF++ – Character n-gram F-score (розширений варіант ChrF++) – метрика автоматичної оцінки на основі символьних і словних n-грам

CSV – Comma-Separated Values

D&D – Dungeons & Dragons – настільна рольова система

DLT – Distributed Language Translation

EBMT – Example-Based Machine Translation – приклад-орієнтований машинний переклад

EN – англійська мова (English) – у позначеннях паралельних пар

FAMT – Fully-Automated Machine Translation – повністю автоматизований машинний переклад

GNMT – Google Neural Machine Translation

GPT – Generative Pre-trained Transformer

HAMT – Human-Assisted Machine Translation – машинний переклад з допомогою людини

HTML – HyperText Markup Language

JSON – JavaScript Object Notation

JSONL – JSON Lines – формат збереження множинних JSON-об'єктів

LLM – Large Language Model – велика мовна модель

MAHT – Machine-Assisted Human Translation – людський переклад з машинною допомогою

METEOR – Metric for Evaluation of Translation with Explicit ORdering – метрика автоматичної оцінки якості перекладу

NMT – Neural Machine Translation – нейронний машинний переклад

OPUS – Open Parallel Corpus – відкритий репозиторій паралельних корпусів

RBMT – Rule-Based Machine Translation – машинний переклад на основі правил

RPG – Role-Playing Game – рольова гра

SMT – Statistical Machine Translation – статистичний машинний переклад

TER – Translation Edit Rate – метрика автоматичної оцінки якості перекладу через підрахунок редагувань

TM – Translation Memory – пам'ять перекладів

TMX – Translation Memory eXchange – формат обміну пам'яттю перекладів

UI – User Interface – інтерфейс користувача

UK – українська мова (Ukrainian) – у позначеннях паралельних пар

XML – Extensible Markup Language

ВСТУП

Локалізація відеоігор як сегмент перекладознавчої практики переживає сьогодні безпрецедентне піднесення. Світовий ринок відеоігор уже понад десятиліття перевищує за обсягом сукупні доходи кіно- та музичної індустрій разом узятих, а українська спільнота гравців невпинно зростає, особливо у контексті активної дерусифікації культурного простору в останні роки. Зростання попиту на якісну українську локалізацію збігається в часі з технологічною революцією у машинному перекладі: великі мовні моделі (LLM) на кшталт GPT-4, Claude, Gemini демонструють якість, яка ще десять років тому здавалася науковою фантастикою. Водночас застосування цих систем до спеціалізованого ігрового домену стикається з фундаментальними обмеженнями, що випливають зі стилістичної гетерогенності, культурної насиченості й технологічної специфіки відеоігрового тексту.

Актуальність дослідження.

Прикладна лінгвістика й перекладознавство останніх років все виразніше орієнтуються на корпусну методологію – побудову й використання великих структурованих наборів мовного матеріалу як емпіричної основи теоретичних узагальнень і практичних інструментів. Для успішної роботи сучасних систем машинного перекладу – як статистичних, так і нейронних – паралельні корпуси є фундаментальним ресурсом, без якого ні тренування моделей, ні їх адекватне оцінювання неможливі. Однак для англо-української мовної пари в галузі відеоігрового перекладу спеціалізовані доменні корпуси на сьогодні фактично відсутні у відкритому доступі. Ця методологічна прогалина має кілька конкретних наслідків: оцінити якість сучасних систем машинного перекладу (МП) у застосуванні до ігрових текстів неможливо в стандартизований спосіб; для адаптації загальних моделей до ігрового домену через донавчання бракує доменних даних; систематичні дослідження якості людських перекладів ігор українською мовою позбавлені емпіричної основи. Створення спеціалізованого корпусу, що містить оригінальний англомовний текст, його людський

професійний переклад і згенерований переклад, отриманий через велику мовну модель Gemini 3.1 Flash-Lite, дозволяє зробити крок до закриття цієї прогалини й одночасно перевірити, наскільки такі моделі придатні до автоматизованої локалізації відеоігор.

Стан дослідження проблеми.

Теоретичною основою дослідження слугують роботи з кількох суміжних галузей. Загальні питання перекладу відеоігор та їхньої локалізації як перекладознавчої категорії розглянуто у фундаментальних працях М. А. Бернала-Меріно (Bernal-Merino, 2014), К. Манджирон і М. О'Гаган (Mangiron & O'Hagan, 2006; O'Hagan & Mangiron, 2013), А. Ф. Косталес (Costales, 2012). Стилістика й функціональні особливості мови відеоігор кількісно проаналізовані у вже згаданій монографії Бернала-Меріно. Теоретичні засади перекладознавчих стратегій (форенізація, доместикація, скопос-теорія) висвітлено в роботах Л. Венуті (Venuti, 1995), В. Шяучюнє та В. Любінєне (Šiaučiūnė & Liubinienė, 2011), українських дослідників – І. В. Корунця. Загальнотеоретичні рамки перекладознавства як дисципліни систематизовано в працях Дж. Манді (Munday, 2016), Е. Пайма (Pym, 2010) та Б. Гатіма і Дж. Манді (Hatim & Munday, 2004). Загальні питання когезії й когерентності тексту як лінгвістичних категорій ґрунтовно розроблені М. Геллідеєм та Р. Гасан (Halliday & Hasan, 1976). Історія машинного перекладу подана в класичних оглядах Дж. Гатчінса (Hutchins, 2005) і Д. Арнольда зі співавторами (Arnold et al., 1994). Проблематику сучасного машинного перекладу й обмеження автоматичних систем розглянуто в монографії Дж. Друган (Drugan, 2013); архітектурні засади нейронного машинного перекладу й трансформерів – у роботах Д. Богданова, К. Чо, Й. Бенджио (Bahdanau et al., 2014) та А. Васвані (Vaswani et al., 2017). Метрики автоматичної оцінки якості МП – BLEU, METEOR, TER, ChrF++, BERTScore – розроблено й описано в роботах К. Папінені (Papineni et al., 2002), С. Банерджі та А. Лаві (Banerjee & Lavie, 2005), М. Сновера (Snover et al., 2006), М. Поповіч (Popović, 2015), Т. Чжан (Zhang et al., 2019). Корпусна лінгвістика як методологічний підхід обґрунтована в роботах Дж. Сінклера (Sinclair, 1991), Т.

Маккенері й Е. Гарді (McEnery & Hardie, 2012), а паралельні корпуси для МП – у працях Ф. Кьона (Koehn, 2005, 2009) і Й. Тідемана (Tiedemann, 2012).

Попри значну розробленість окремих компонентів – теорії локалізації відеоігор, технологій машинного перекладу, методології корпусної лінгвістики – комплексне дослідження, що поєднало б ці компоненти на матеріалі англо-українських ігрових текстів, на момент проведення цієї роботи, наскільки нам відомо, не виконувалося. Це визначає теоретичну обґрунтованість обраної теми.

Об'єкт дослідження –

функціонально-стилістичні та лінгвістичні характеристики англомовних та українськомовних перекладних текстів відеоігор (традиційних і згенерованих).

Предмет дослідження –

методологія, технологічна реалізація і структура паралельного корпусу EN–UK на матеріалі текстів відеогри Baldur's Gate 3, що поєднує оригінал, офіційний український переклад і згенерований переклад, отриманий за допомогою моделі Gemini 3.1 Flash-Lite.

Мета роботи –

створити триколонковий паралельний англо-український корпус текстів відеогри Baldur's Gate 3 і на його основі оцінити якість перекладу, згенерованого моделлю Gemini.

Завдання дослідження:

1) проаналізувати теоретичні засади перекладу відеоігор, сучасний стан машинного перекладу та метрик оцінки його якості;

2) розробити методологію побудови паралельного корпусу і реалізувати програмну систему збору й обробки текстів;

3) сформувати корпус, що поєднує оригінал, офіційний український переклад і згенерований переклад, отриманий через LLM Gemini;

4) оцінити якість згенерованих перекладів у порівнянні із людським еталоном за допомогою автоматичних метрик і якісного аналізу.

Матеріал дослідження.

Матеріалом дослідження є оригінальний англомовний текст із локалізаційних файлів відеогри Baldur's Gate 3, розробленої компанією Larian Studios і випущеної у 2023 році, а також офіційний український переклад цієї гри й переклад, згенерований за допомогою LLM Gemini через офіційний API Google AI Studio. Загальний обсяг первинного текстового матеріалу становить понад два мільйони слововживань англомовного тексту; кінцевий обсяг сформованого корпусу становить 186 311 унікальних паралельних пар - українсько-англійських перекладних відповідників.

Методологічна основа і методи дослідження.

Методологічну основу роботи становить поєднання теоретичних і емпіричних підходів. Теоретичну базу формують: перекладознавча теорія локалізації (функціональний підхід, скопос-теорія), методи загальної корпусної лінгвістики, теорія машинного перекладу. У практичній частині застосовано низку методів: корпусний метод збору й упорядкування мовних даних — для формування власне корпусу; програмна реалізація пайплайну витягування й вирівнювання — як інженерний компонент дослідження; кількісне оцінювання за автоматичними метриками (BLEU, METEOR, TER, ChrF++, BERTScore) – для валідації якості згенерованих перекладів; якісний аналіз перекладацьких рішень на ілюстративних прикладах – для виявлення характерних патернів сильних і слабких сторін машинного перекладу в ігровому домені. Технологічну реалізацію дослідження забезпечує стек на основі мови програмування Python із розробкою модульного пайплайну витягування, вирівнювання й обробки текстів (із залученням спеціалізованих лінгвістичних та інженерних бібліотек lxml, google-genai, sacrebleu, bert-score та evaluate) та застосункового програмного інтерфейсу великої мовної моделі Gemini 3.1 Flash-Lite, яку залучено з порівняльною метою – для формування згенерованої української колонки, що зіставляється з офіційною локалізацією. Програмний код пайплайну розміщено у відкритому репозиторії: https://github.com/ZavarOvek/crispy_invention.

Наукова новизна.

Уперше для української мови сформовано паралельний корпус ігрового домену, що поєднує оригінал, офіційний людський переклад і згенерований переклад, отриманий через модель Gemini 3.1 Flash-Lite. Запропоновано методологію побудови такого ресурсу, придатну для перенесення на матеріал інших відеоігор. Виконано систематичну валідацію якості згенерованих перекладів, згенерованих LLM Gemini, на матеріалі стилістично гетерогенного RPG-проєкту, що дає емпіричні підстави для оцінки межі використання LLM у спеціалізованому перекладі.

Практичне значення.

Створений корпус слугує ресурсом для подальших досліджень якості МП у домені відеоігор, тренування доменно-адаптованих моделей, а також моделей постредагування. Запропонована методологія побудови переноситься на тексти інших ігор, що дозволяє в майбутньому розширювати корпус і будувати порівняльні дослідження між проєктами різних жанрів. Результати валідації показують межі використання сучасних LLM у спеціалізованому ігровому перекладі – інформація, корисна як для академічних досліджень, так і для практиків локалізаційної індустрії, що оцінюють доцільність впровадження автоматизованих рішень у свої робочі процеси.

Структура роботи.

Робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел і додатків. Перший розділ присвячено теоретичним засадам перекладу відеоігор: розглянуто жанрово-стилістичні особливості мови відеоігор, локалізацію відеоігор як перекладознавчу проблему, лінгвістичні явища, що ускладнюють переклад і МП. Другий розділ присвячено машинному перекладу – його еволюції від систем за правилами до сучасних великих мовних моделей – та метрикам автоматичної оцінки якості. У третьому розділі описано методологію побудови паралельного корпусу: огляд паралельних корпусів у машинному перекладі, обґрунтування вибору матеріалу, структуру корпусу, принципи фільтрації даних та етико-юридичні аспекти. Четвертий розділ містить опис практичної реалізації – архітектуру програмної системи, технологію витягування текстів і генерації третьої колонки, кількісну характеристику

отриманого корпусу й результати валідації якості згенерованих перекладів. У висновках узагальнено основні результати дослідження й окреслено перспективи його продовження. Загальний обсяг роботи – 86 сторінок основного тексту, список використаних джерел нараховує 41 позицію, з них 39 англомовних і 2 україномовні джерела.

РОЗДІЛ 1. ОСОБЛИВОСТІ ТЕКСТІВ ТА МЕТОДИКИ ПЕРЕКЛАДУ ВІДЕОІГР

1.1. Жанрово-стилістичні особливості мови відеоігор

Хоча відеоігри як вид мультимодального контенту увійшли у фокус перекладознавчих досліджень порівняно недавно, їхня мовна специфіка вже отримала у науковій літературі досить чітку характеристику. Відеогра не є монолітним текстом: у межах одного продукту співіснують репліки персонажів, інтерфейсні написи, описи предметів, наративні фрагменти, журнальні записи, листи, кінематичні монологи, технічні повідомлення системи. Кожен із цих типів тексту має власні стилістичні особливості, власну функцію у грі й власні обмеження, які перекладач – а в нашому випадку й машина, і дослідник корпусу – змушений враховувати.

Принципову відмінність ігрового перекладу від суміжних видів медіаперекладу – літературного й аудіовізуального – обґрунтували К. Манджирон та М. О'Гаган, увівши термін «локалізації відеоігор» (game localisation) як окрему перекладознавчу категорію (Mangiron & O'Hagan, 2006). На відміну від книжки, яку читач сприймає лінійно, або фільму, який глядач переглядає в наперед визначеній послідовності, відеогра є інтерактивним продуктом: гравець сам формує траєкторію взаємодії з текстом, і однакові репліки можуть прозвучати в різних емоційних обставинах залежно від попередніх рішень. Це робить переклад відеогри ближчим до перекладу гіпертексту, ніж до традиційного художнього перекладу. Дослідники також підкреслюють, що локалізатор має право – а часом і обов'язок – вдаватися до значно сміливіших трансформацій, ніж літературний перекладач: переписування жартів, заміна культурних реалій, навіть зміна імен персонажів задля збереження ігрового досвіду цільової аудиторії. Цей принцип, який К. Манджирон і М. О'Гаган позначають як «пріоритет ігрового досвіду над текстуальною вірністю», має прямий наслідок для побудови корпусу: пара (EN, UK_human) у такому корпусі часто буде не парою формальних еквівалентів, а парою функціональних

аналогів, що вимагає обережності при автоматичній оцінці якості будь-якого згенерованого перекладу порівняно з людським еталоном.

Відеоігрова продукція традиційно класифікується за кількома ознаками. Перша й найочевидніша – жанрова: стратегії, рольові ігри (RPG), шутери, пригодницькі ігри, симулятори, ігри-головоломки. Друга – за платформою: персональні комп'ютери, ігрові консолі, мобільні пристрої. Третя – за кількістю гравців: однокористувацькі, локальні багатокористувацькі та масові онлайн-ігри. Окремо виділяють розважальні, навчальні й спортивні ігри. З погляду перекладознавства релевантною є ще одна класифікація – за рівнем творчої свободи перекладача: ігри, що ґрунтуються на наявному всесвіті (книжкових серіях, коміксах, іншому медіа-матеріалі) і вимагають узгодження з раніше встановленою термінологією, та ігри з оригінальним всесвітом, де перекладач фактично співтворить термінологію мовою перекладу.

Матеріалом нашого дослідження є тексти гри Baldur's Gate 3 (Larian Studios, 2023) – рольової гри (RPG) у жанрі темного фентезі, що ґрунтується на правилах настільної рольової системи Dungeons & Dragons. BG3 належить до того самого класу проєктів, що й Baldur's Gate, Dragon Age, Pillars of Eternity, The Witcher: це ігри з розгалуженими діалогами, високим обсягом наративного тексту, варіативними сюжетними гілками й глибокою стилістичною неоднорідністю. Саме ця стилістична гетерогенність робить BG3 цінним матеріалом для побудови корпусу: одна гра вже сама по собі покриває кілька функціональних регістрів.

У процесі перекладу й локалізації відеоігор перекладачі стикаються з трьома базовими типами труднощів, що виходять за межі суто мовних. Перший – відсутність контексту. Переклад гри зазвичай починається задовго до її повного завершення, а через ризики копіювання та витоку розробники часто видають перекладачам лише таблицю вихідного тексту без візуального чи сюжетного контексту. Окремі репліки можуть бути фрагментами діалогу, розкиданими в межах різних файлів, тож перекладач не бачить, у якій ситуації, з якою інтонацією та кому саме персонаж промовляє ту чи ту фразу. Другий тип труднощів – фрагментація тексту, у тому числі технічна: ігровий інтерфейс часто

має жорсткі обмеження на довжину рядка (фіксована кількість символів у меню, кнопці, спливному вікні), і за перевищення тексту може зміщуватися за рамки інтерфейсу або обриватися. У таких випадках локалізатор працює радше як редактор, що шукає синонім або вдається до модуляції, ніж як перекладач у класичному значенні. Третій – синтез стилів: у межах однієї гри перекладач може потребувати знань із царин, далеких від філологічної компетенції, від механіки до окультизму, і змушений швидко орієнтуватися в незнайомій термінології.

Стилістичний аналіз показує, що тексти відеоігор охоплюють практично весь спектр функціональних стилів літературної мови. Сучасна перекладознавча класифікація виділяє розмовні, офіційно-ділові, суспільно-інформативні, наукові, художні та релігійні тексти. У відеогрі певного жанру з високою ймовірністю співіснуватиме більшість із них. Так, у RPG-іграх рівня BG3 присутні наукові й науково-популярні фрагменти (бестіарії, описи магічних шкіл, рукописи учених), офіційно-ділові тексти (укази, контракти, листи між владними особами), публіцистичні тексти (внутрішньоігрові газети, прокламації), художні тексти (вірші, легенди, фрагменти романів-у-романі), розмовно-побутові репліки (діалоги з другорядними персонажами, торгівля, флірт). Це принципово відрізняє ігровий текст від, наприклад, технічної документації чи художньої літератури – там стилістична однорідність задана жанром, тоді як у грі вона мусить узгоджуватися всередині одного продукту.

Відмінною рисою стилю наукових фрагментів у відеоіграх є їхній малий обсяг. Це здебільшого уривки з підручників або статей у 10–15 речень, що містять високу щільність когнітивної інформації – об'єктивних відомостей, оформлених через нейтральну лексику, термінологію, теперішній час дієслова, абревіатури й цифрові дані (Корунець, 2003). Художні тексти, навпаки, несуть переважно емоційно-естетичну інформацію: тропи, варіативність синтаксису, лексика різного рівня – від архаїзмів до жаргонізмів. Розмовно-побутовий стиль у RPG-іграх посідає провідну позицію саме тому, що через діалоги відбувається отримання завдань, розкриття світу й рух сюжету; його лексичні особливості – фразові дієслова, скорочені форми, колоквіалізми, слова-філери – створюють відчуття природного мовлення.

Кількісне підтвердження стилістичної неоднорідності відеоігрових текстів дає дослідження М. А. Бернала-Меріно (Bernal-Merino, 2014). Аналіз корпусу англомовних ігрових текстів показав, що неформальні письмові тексти (листи, щоденники, телеграми) демонструють високу частку розмовних елементів – скорочених форм дієслів (52,9 %), слів-філерів (5,5 %), колоквіалізмів (8,9 %), деривативної лексики (7,3 %). Натомість формальні письмові тексти (газетні статті, звіти, наукові довідки) тяжіють до літературної лексики (16,6 %), архаїзмів (16,9 %) та пасивного стану (10,6 %). Інтерфейсні тексти, незалежно від жанру гри, формують окремий полюс: у них переважає літературна лексика (36,6 %), архаїзми (34,1 %) та імперативна модальність (41,4 %), а часові форми зміщені до теперішнього (40,9 %), що пов'язано з функцією опису актуальних дій і станів. Фразові дієслова, типові для усного мовлення, в інтерфейсі вживаються рідше або замінюються однолексемними еквівалентами; навпаки, в усних діалогах вони переважають у три-чотири рази (Bernal-Merino, 2014, с. 111–119).

Жанр відеогри безпосередньо впливає на її стилістичний профіль. Шутери (як-от Call of Duty) характеризуються короткими динамічними репліками з імперативами та технічною лексикою, мінімалізмом описів. Стратегії (як-от Crusader Kings 3) використовують формалізовані повідомлення, статистичні дані й пасивний стан інструкцій (*«Територія захоплена»*, *«Населення: 1200»*). Рольові ігри – клас, до якого належить наш матеріал – містять найрозгорнутіші діалоги, художні описи світу, архаїзми у фентезійних сетингах і спеціалізовану термінологію у науково-фантастичних. Ігри-пригоди (як-от Red Dead Redemption 2) поєднують емоційно забарвлені діалоги з неформальними письмовими текстами – листами, щоденниковими записами, телеграмами.

Окремо варто зупинитися на віртуальних антропонімах – власних назвах персонажів, рас, локацій, артефактів, які становлять вагому частку лексики саме розмовних та неформальних письмових текстів (16,2 % і 10,1 % відповідно), тоді як в інтерфейсі їхня частка не перевищує 3,3 % (Bernal-Merino, 2014, с. 111–112). Ця асиметрія важлива для побудови корпусу: при автоматичному оцінюванні якості перекладу метрики, чутливі до точного відтворення власних назв (наприклад, BLEU), будуть давати різні показники для різних типів текстів навіть

за ідентичної якості перекладу решти лексики. Іншими словами, корпус, що репрезентує гру повноцінно, мусить містити метадані про тип кожного запису, інакше зведена оцінка якості перекладу буде систематично спотворена.

Особливим викликом у локалізації ігор є плейсхолдери – змінні шаблони на кшталт *«<ім'я гравця> отримав <кількість> очок»* або *«<назва_фракції> атакує Вас!»*. Такі конструкції критично залежать від морфологічних характеристик мови перекладу: рід, число, відмінок, узгодження з прикметником. Українська як флективна мова потребує тут значно складніших рішень, ніж англійська, де достатньо підставити лексичну одиницю в незмінному вигляді (Bernal-Merino, 2014, с. 138). Для побудови нашого корпусу це означає, що записи, які містять плейсхолдери, мають або зазнавати спеціальної обробки, або позначатися окремою міткою, бо вони фундаментально відрізняються від звичайних повних речень.

Підсумовуючи, можна стверджувати, що відеоігровий текст – не один тип тексту, а сукупність типів, що синтезуються у межах одного продукту. Будь-який паралельний корпус, побудований на матеріалі відеогри, буде стилістично гетерогенним за визначенням, і ця гетерогенність – не вада, а особливість, яку треба експліцитно описувати в метаданих. Саме тому подальші розділи нашої роботи приділяють окрему увагу типології записів і принципам їхньої сегментації за стилістично-функціональною ознакою.

1.2. Локалізація відеоігор як перекладознавча проблема

Коли йдеться про переклад комп'ютерних ігор, програм або вебсайтів, у фаховому дискурсі переважно вживають термін «локалізація». Локалізація охоплює щонайменше два складники: власне переклад текстового матеріалу та технічну адаптацію продукту до коректної роботи в цільовій культурі й мові – дотримання звичних форматів дати й часу, систем мір, мовного кодування, напрямку письма, специфіки введення тексту. Локалізацію загалом визначають як процес перетворення продукту так, щоб він лінгвістично, культурно, технічно та юридично відповідав вимогам цільової культури й мови (Costales, 2012).

Власне переклад є лише однією зі складових цього багатовимірного процесу, причому не завжди найскладнішою з технічного погляду.

Залежно від глибини охоплення розрізняють кілька рівнів локалізації. Паперова локалізація передбачає переклад зовнішніх атрибутів продукту – інструкції користувача, рекламних матеріалів, написів на пакованні; сам програмний продукт залишається незміненим. Економічна локалізація охоплює внутрішньоігровий текст – інтерфейс, субтитри, описи предметів і завдань; це найпоширеніший рівень для більшості ринків. Поглиблена локалізація додає переозвучення реплік мовою перекладу, що принципово підвищує занурення гравця. Надлишкова локалізація передбачає заміну графічних об'єктів – написів на ігрових моделях, текстур з мовним матеріалом, а часом і символіки. Найвищим рівнем є глибока локалізація – адаптація сценарію, а інколи й ігрової механіки до конкретної країни чи регіону, що включає роботу з культурно чутливим контентом, цензуру (наприклад, обов'язкове прибирання нацистської символіки в Німеччині) та зміну імен персонажів задля уникнення небажаних асоціацій (Bernal-Merino, 2014, с. 92–95). Вибір рівня локалізації для конкретного проєкту визначається не лише художніми міркуваннями, але й бюджетом, потенційним прибутком на ринку та юридичними вимогами.

Принципова відмінність локалізації ігор від локалізації звичайних програмних продуктів полягає у значно вищій частці творчої складової. Якщо локалізатор прикладного програмного забезпечення здебільшого працює з усталеними кліше та контрольованою термінологією, то локалізатор гри має балансувати між точністю передачі смислу та збереженням ігрового досвіду – атмосфери, гумору, темпу діалогів, культурних алюзій. Саме на цьому ґрунтується спостереження А. Ф. Косталес, що переклад комп'ютерної гри є передусім функціональним процесом, головна мета якого – відтворити для аудиторії всі ігрові можливості оригіналу так, щоб гравці у різних мовно-культурних спільнотах отримали порівнянний емоційний і ігровий досвід (Costales, 2012). Цей принцип, у свою чергу, тісно пов'язаний зі скопос-теорією Г. Фермеера: переклад вважається успішним тоді, коли реципієнт навіть не помічає, що текст від початку був написаний не його рідною мовою. В. Шяучюнє

та В. Любінєне обґрунтовують застосовність скопос-теорії до перекладу відеоігор, наголошуючи, що локалізатор має право вдаватися до стількох змін для забезпечення повного розуміння і занурення з боку користувача (Šiaučiūnė & Liubinienė, 2011). Водночас вони зазначають, що цей підхід дотепер сприймається в галузі неоднозначно – перекладознавчої спільноти розцінює сильні відхилення від оригіналу як вияв перекладацької самовпевненості.

У межах функціонального підходу А. Ф. Косталес виділяє три базові стратегії перекладу відеоігор – форенізацію, доместикацію та відмову від перекладу (no translation) (Costales, 2012). Терміни «форенізація» і «доместикація» введені Л. Венуті: за доместикації перекладений текст максимально наближається до норм цільової культури, що сприяє зближенню автора оригіналу й читача в цільовій мові; за форенізації, навпаки, читач здійснює крок у напрямку автора, сприймаючи особливості іншої культури часом ціною певного дискомфорту в рідній мові (Venuti, 1995). У відеоіграх обидві стратегії застосовуються вибірково – і часто в межах одного продукту: фракції чи раси, що мали б сприйматися гравцем як «чужі», локалізуються форенізаційно зі збереженням іншомовних або вигаданих лексичних одиниць, тоді як основні діалоги й інтерфейс адаптуються через доместикацію. Стратегія «no translation» застосовується найчастіше до власних назв самих ігор (Tekken, Assassin's Creed, World of Warcraft) і до базової термінології, що стала впізнаваним брендом.

Окремим викликом локалізації є мультимодальна природа ігрового контенту. На відміну від книжки чи навіть фільму, відеогра поєднує текст, зображення, звук, інтерактивність та системні обмеження інтерфейсу, що змушує перекладача враховувати фактори, далекі від філологічних. Зростання складності сюжетних ліній, частота культурно специфічних алюзій та жартів, потреба в синхронізації перекладу з обмеженнями інтерфейсу й анімації – усе це робить традиційні методи локалізації ресурсомісткими (Bernal-Merino, 2014, с. 2). Машинний переклад у цьому контексті виглядає привабливо як інструмент прискорення роботи з великими обсягами тексту – описами предметів, технічними повідомленнями, інструкціями. Однак специфіка ігрового контенту,

зокрема потреба у транскреації та адаптації культурних нюансів, фундаментально обмежує самостійне застосування МП без людського редагування (Bernal-Merino, 2014, с. 63–65, 172–173).

Економічний контекст локалізації для української мови додатково ускладнює ситуацію. Витрати на професійний переклад великого ігрового продукту лишаються високими, а обсяг ринку – порівняно невеликим, що знижує комерційну привабливість україномовних версій для розробників. Як наслідок, частина ігор виходить без офіційної української локалізації, а наявні переклади часто з'являються із затримкою або робляться силами фанатських спільнот. Інтеграція машинного перекладу в локалізаційний пайплайн потенційно може знизити витрати й зробити україномовні версії доступнішими, проте це вимагає спеціалізованих моделей і ресурсів, натренованих на ігровій термінології та реаліях. Саме нестача таких ресурсів – а серед них центральне місце посідають доменні паралельні корпуси – є однією з ключових перешкод для якісного автоматизованого перекладу ігор українською мовою.

Практичні випадки використання машинного перекладу в локалізації демонструють обидві сторони цього інструменту. Серія Call of Duty неодноразово фіксувала локалізаційні помилки, спричинені автоматизованим перекладом без належного редагування: дослівне передавання команди *«Reloading!»* як «*Перезавантаження!*» замість контекстуально коректного «Перезаряджаюсь!» стало хрестоматійним прикладом такого типу збоїв (Hansen et al., 2022, с. 5). Натомість локалізація The Witcher 3 студією CD Projekt Red є зразком успішного гібридного підходу: машинний переклад використовувався як допоміжний інструмент для прискорення роботи з повторюваними сегментами, тоді як остаточна редакція виконувалася вручну з урахуванням культурних особливостей і діалектних відтінків – особливо у гумористичних діалогах персонажа Лютика, де потрібно було зберегти ритм і стилістику (Hansen et al., 2022, с. 4–5). Ще одним показовим випадком є аматорська локалізація гри Bully на індонезійську мову, виконана командою ентузіастів: дослідження показало, що такий переклад забезпечує загальну зрозумілість сюжету, але демонструє суттєві проблеми з плавністю мовлення та обробкою культурно

чутливих діалогів (Nugroho, 2017, с. 40). Загальний висновок з цих випадків полягає в тому, що сучасний машинний переклад може ефективно адаптовувати повторювані й технічні сегменти, але креативні діалоги, метафори та варіативні наративи досі потребують ручної обробки. Це створює методологічну вимогу до будь-якого корпусу, призначеного для досліджень якості ігрового МП: він має містити збалансовану репрезентацію всіх типів тексту, а не лише тих, з якими МП справляється найкраще.

1.3. Лінгвістичні явища, що ускладнюють традиційний та машинний переклад

Якщо попередні підрозділи описували відеоігрову мову як цілісне явище – стилістично гетерогенне, мультимодальне, вписане в специфічну перекладацьку практику локалізації, – то цей підрозділ присвячено лінгвістичним явищам, які створюють найбільші труднощі для перекладу взагалі та для машинного перекладу зокрема. Виокремлення цих явищ важливе для нашого дослідження з двох причин. По-перше, тексти відеогри концентрують їх особливо щільно: фрагментованість діалогів, нелінійність наративу, висока частка референційних виразів, що розгортаються у часі гри, неминуче призводять до того, що ігровий корпус містить чимало одиниць, на яких автоматичні системи перекладу провалюються. По-друге, саме ці явища визначають межі застосовності будь-якої метрики автоматичної оцінки якості: одна й та сама метрика буде давати різні показники на сегменті з простим описовим реченням і на сегменті з еліптичним діалогом, навіть якщо згенерований переклад однаково якісний за обома.

Найзагальнішою категорією лінгвістичних явищ, релевантних для перекладу, є зв'язність тексту, яка традиційно поділяється на змістову (когерентність) і структурну (когезію) (Halliday & Hasan, 1976). Когерентність забезпечує цілісність тексту як комунікативної одиниці на рівні плану змісту, тоді як когезія працює на рівні плану вираження – через формальні мовні засоби, які пов'язують речення між собою. Когерентну послідовність речень становить така послідовність, у якій описано логічний і послідовний перебіг подій.

Класичний приклад: *«Вода перелилася через край склянки. Вона почала стікати на стіл»* – це когерентна послідовність, бо причина передує наслідку. Натомість послідовність *«Вона почала стікати на стіл. Вода перелилася через край склянки»* порушує логіку, а пара *«Вчитель перевірив домашнє завдання. Картопля вариться на кухні»* – приклад відсутньої когерентності, оскільки ці два речення не пов'язані змістовою лінією. Когерентність будь-якого тексту забезпечують три види когезії – логічна, стилістична та семантико-синтаксична, причому остання вважається основною (власне лінгвістична когезія).

Серед когезивних явищ, що особливо ускладнюють переклад, центральне місце посідають анафора, дейксис та еліпсис. Анафора (з давньогрецької ἀναφορά – «повторення») – це засіб уникнення повтору однакових слів через використання займенників, синонімів або інших субститутів, які відсилають до раніше згаданого об'єкта. Анафорична структура складається з антецедента (іменної групи) та анафора (анафоричного слова). У контекстуальному перекладі найбільший інтерес становить займенникова анафора, оскільки саме вона активно бере участь у вторинній номінації – необхідності повторно згадати певний об'єкт без буквального повтору його назви. Складність для машинного перекладу полягає в тому, що рід і число займенника в українській залежать від роду й числа антецедента, але система МП, працюючи з фрагментованим текстом, часто не має доступу до антецедента в попередньому реченні, особливо в умовах посегментної обробки.

Дейксис (з давньогрецької δεῖξις – «вказівка») охоплює клас слів і фраз, що вказують на час, місце або особу в межах конкретної комунікативної ситуації. Виділяють особовий дейксис (я, ти, він), просторовий (тут, там, ось), часовий (сьогодні, завтра, зараз) та дискурсивний (вищезгадані питання, наведений нижче приклад). Семантика дейктичних слів стійка, але їхня референція кожного разу залежить від контексту: «тут» у репліці одного персонажа і «тут» у репліці іншого можуть позначати різні локації. У відеоігровому тексті дейксис набуває додаткової складності через те, що референційне «зараз» гри розгортається в умовному ігровому часі, а просторовий дейксис часто прив'язаний до візуальної сцени, недоступної ні перекладачеві, ні автоматичній системі. Еліпсис (з

давньогрецької ἔλλειψις – «нестача, пропуск») є пропуском мовних елементів, які можна відновити з контексту. Класичні приклади – діалогічні репліки на кшталт *«– Хочеш чаю? – Зеленого»*, де відповідь редукована до прикметника, бо присудок легко відновлюється; або повторювана відповідь «Так» на різні запитання, де її конкретний зміст щоразу інший. Такі еліпсиси найбільше поширені в розмовному діалогічному мовленні, що становить основну тканину RPG-ігор. Машинний переклад еліптичних конструкцій особливо проблематичний: якщо цільова мова не дозволяє застосувати такий самий тип еліпсису або якщо пропущений матеріал впливає на синтаксис цільового речення, система змушена реконструювати пропущене на підставі ймовірнісних моделей, що часто веде до граматично коректних, але семантично неточних результатів.

Окремо від цих текстуальних явищ варто розглянути проблеми, специфічні для машинного перекладу як технологічного процесу. Першою з них є лексична багатозначність – здатність слова чи виразу мати кілька значень залежно від контексту. Класичний приклад – англійське слово bank, яке може позначати фінансову установу або берег річки, що вимагає від системи аналізу семантичного оточення для коректного вибору еквівалента (Drugan, 2013, с. 20). Системи на правилах (RBMT) намагалися вирішувати цю проблему через словники з контекстними правилами, проте ручне кодування таких правил виявилося неефективним через велику кількість винятків і динамічну природу мови (Drugan, 2013, с. 49). Статистичні системи (SMT) спираються на частотність співвідношень у паралельних корпусах, але їхня якість прямо залежить від обсягу та репрезентативності даних, що особливо ускладнює обробку рідкісних або неоднозначних випадків (Drugan, 2013, с. 111–112).

Другою фундаментальною проблемою є синтаксична складність мов, особливо при перекладі між мовами зі значно відмінною структурою. Німецькі складені слова, японські частки, що визначають відношення між іменниками, флективні системи слов'янських мов – усі ці явища потребують глибокого аналізу граматичних відношень, який не зводиться до пошуку лексичних відповідників. RBMT використовували трансферні правила для адаптації

синтаксису, але їхня жорстка архітектура часто призводила до помилок при нестандартних конструкціях (Drugan, 2013, с. 28–29). SMT, навпаки, автоматично виділяють сегменти тексту, але їхній підхід ігнорує лінгвістичну структуру, що ускладнює переклад довгих або складних речень (Drugan, 2013, с. 106–108).

Третій виклик – залежність якості МП від наявних даних. Статистичні й нейронні методи потребують мільйонів паралельних речень для адекватного навчання, що робить їх обмежено придатними для рідкісних або низькоресурсних мов. Українська мова, попри те що не належить до найрідкісніших, у багатьох галузях має серйозний дефіцит спеціалізованих корпусів – і ігровий домен є саме таким випадком. Переклад з рідкісної мови на іншу рідкісну часто змушує системи використовувати проміжну мову (зазвичай англійську), що неминуче призводить до накопичення помилок (Drugan, 2013, с. 112–113). RBMT, хоча менш залежні від корпусних даних, потребують трудомісткого створення словників і правил, що обмежує їх застосування до мов з обмеженими лінгвістичними ресурсами (Drugan, 2013, с. 115). Гібридні системи, що поєднують правила й статистику, пропонують компроміс, але їхня розробка вимагає значних інвестицій у спеціалізовані модулі (Drugan, 2013, с. 114–115).

Найскладнішими для машинного перекладу залишаються прагматичні аспекти мови – ідіоми, метафори, діалектизми, культурні алюзії. Англійський вираз kick the bucket має бути перекладений як «померти», але буквальний переклад утворює беззмістовний результат. RBMT-системи здатні опрацьовувати такі випадки через ручне додавання правил, але їхнє охоплення обмежене (Drugan, 2013, с. 85). SMT-системи ідентифікують ідіоми через частотність у корпусах, але їхня ефективність падає при роботі з художніми текстами або нестандартними виразами (Drugan, 2013, с. 119). Глибинне навчання пропонує нові можливості через автоматичне виявлення контекстуальних шаблонів, проте, як зазначає М. О'Гаган, нейромережеві моделі часто ігнорують культурні нюанси – а саме вони критично важливі для локалізації ігор (O'Hagan & Mangiron, 2013, с. 215). Дослідники також

підкреслюють, що навіть сучасні нейронні моделі часто нехтують семантичною цілісністю текстів і дискурсивною узгодженістю на дистанціях, більших за одне речення (Drugan, 2013, с. 124–125).

Сукупно ці явища – текстуальні (когезія, анафора, дейксис, еліпсис) та технологічні (багатозначність, синтаксична складність, залежність від даних, прагматика) – окреслюють контури того простору, у якому має функціонувати спеціалізований корпус відеоігрового перекладу. Корпус для МП має, з одного боку, репрезентувати всі ці явища у достатньому обсязі, щоб модель, навчена або оцінена на ньому, мала змогу зустрічатися з реальною складністю мовного матеріалу; з іншого – він має бути достатньо чистим і впорядкованим, щоб ці явища були прозорими для аналізу. Саме до питання про те, як паралельні корпуси загалом і ігрові корпуси зокрема влаштовані, як вони класифікуються та які вимоги до них висуває сучасна лінгвістика, ми звертаємося в підрозділі 3.1.

Висновки до розділу 1

У першому розділі ми розглянули теоретичні засади дослідження, рухаючись від загальної характеристики мови відеоігор до лінгвістичних явищ, що ускладнюють їх переклад. Здійснений огляд дозволяє сформулювати кілька важливих узагальнень.

По-перше, мова відеоігор є принципово гетерогенним мовним матеріалом: у межах одного продукту співіснують функціональні стилі, що в традиційному перекладознавстві аналізувалися окремо – від наукового до розмовно-побутового, від офіційно-ділового до художнього. Кількісні дослідження М. А. Бернала-Меріно показують, що різні типи ігрового тексту (інтерфейс, наратив, діалоги, неформальні письмові вставки) мають принципово різні стилістичні профілі – за часткою архаїзмів, скорочених форм, фразових дієслів, віртуальних антропонімів. Це зумовлює критичну вимогу до будь-якого корпусу ігрового домену: записи мають супроводжуватися метаданими щодо типу тексту, інакше зведена оцінка якості перекладу буде систематично спотвореною.

По-друге, локалізація відеоігор як перекладознавча практика суттєво відрізняється від суміжних видів медіаперекладу – літературного,

аудіовізуального, технічного. Інтерактивна природа ігор, нелінійність сюжету, потреба в транскреації культурних реалій, а також технічні обмеження інтерфейсу й системи плейсхолдерів змушують локалізатора часто діяти радше у функціональній, ніж у формально-еквівалентній парадигмі. Це означає, що пара (EN, UK_human) у корпусі ігрового перекладу часто є парою функціональних аналогів, а не лексично точних відповідників – обставина, яку треба враховувати при автоматичному оцінюванні якості будь-якого згенерованого перекладу у порівнянні з таким еталоном.

По-третє, лінгвістичні явища, що ускладнюють переклад загалом – анафора, дейксис, еліпсис, лексична багатозначність, синтаксична складність, прагматика, – концентруються в ігрових текстах особливо щільно. Фрагментованість діалогів, посегментна організація локалізаційних файлів, відсутність ширшого контексту під час перекладу – усе це загострює проблеми, з якими стикаються сучасні системи машинного перекладу. Навіть нейронні моделі, що значно перевершують попередні підходи в роботі з типовими реченнями, демонструють зниження якості саме на тих явищах, які складають серцевину ігрової мови.

Сформульовані в цьому розділі теоретичні позиції визначають конкретні методологічні рішення, що детально описуватимуться у наступних розділах: вибір матеріалу (Baldur's Gate 3 як стилістично гетерогенний RPG-проєкт із наявною офіційною українською локалізацією), формат корпусу (JSONL з полями для оригіналу, людського й згенерованого перекладу та метаданими типу тексту), методику валідації (набір автоматичних метрик у поєднанні з якісним аналізом). Перш ніж перейти до цих практичних питань, у наступному розділі ми завершуємо теоретичну частину розглядом еволюції систем машинного перекладу та метрик оцінки їхньої якості – компонентів, без яких неможливо коректно сформулювати ані задачу побудови корпусу, ані критерії його придатності.

РОЗДІЛ 2. МАШИННИЙ ПЕРЕКЛАД: ЕВОЛЮЦІЯ СИСТЕМ ТА МЕТРИКИ ОЦІНКИ ЯКОСТІ

2.1. Історія і типологія машинного перекладу

Машинний переклад як окрема науково-технічна дисципліна сформувався у першій половині XX століття, коли інженерна та лінгвістична традиції вперше об'єдналися навколо завдання автоматизації міжмовної комунікації. Перші концептуальні передумови сягають ще XVII століття з ідеями універсальних філософських мов, проте перші практичні спроби механізації перекладу датуються 1933 роком, коли Г. Арцруні у Франції та П. Троянський у СРСР отримали патенти на механічні пристрої для словникового перекладу (Hutchins, 2005, с. 5–6). Справжній імпульс розвиткові МП дала післявоєнна доба і початок холодної війни: потреба в оперативному аналізі та перекладі великих обсягів технічних текстів змусила американські й радянські уряди фінансувати перші масштабні дослідницькі проєкти.

Поворотним моментом стала публічна демонстрація системи Georgetown-IBM 1954 року, яка автоматично переклала 49 російських речень на англійську за допомогою словника на 250 слів і шести граматичних правил (Hutchins, 2005, с. 11). Попри явну примітивність, ця демонстрація стала каталізатором для створення дослідницьких груп у США, СРСР та Європі. Однак уже в 1960-х настала криза: підготовлений у 1966 році звіт ALPAC заявив про нерентабельність МП порівняно з людським перекладом, і фінансування галузі різко скоротилося (Hutchins, 2005, с. 12–13). Лише у 1970-х роках напрям відродився завдяки трансферному підходу – більш гнучкій архітектурі, що поділяла процес перекладу на етапи аналізу, перенесення структури і генерації; представниками цього підходу стали системи GETA-Ariane та METAL (Arnold et al., 1994, с. 14).

1980-ті роки принесли зміну парадигми – корпусні методи. Дослідницька група IBM розробила статистичний підхід до перекладу (SMT), за якого замість ручного кодування правил система виводила переклад на основі ймовірностей спільних входжень слів у паралельних текстах. Першою реалізацією цього підходу стала система Candide, побудована на корпусі канадських парламентських стенограм (Hansard) (Berger et al., 1994; Hutchins, 2005, с. 24). Паралельно в Японії розвинувся прикладний підхід EBMT (Example-Based Machine Translation), що шукав аналоги в базі вже перекладених фраз. У 1990-х роках МП пережив етап комерціалізації: системи Systran, Logos та інші впроваджувалися в інституціях ЄС, у НАТО та великих корпораціях (Hutchins, 2005, с. 15–16; Arnold et al., 1994, с. 14). Сучасний етап, що розпочався у 2010-х, ознаменувався переходом до нейронних мереж – спершу до рекурентних архітектур типу «кодувальник – декодувальник» (Cho et al., 2014; Sutskever et al., 2014) з механізмом уваги (Bahdanau et al., 2014, с. 1–2), а згодом до трансформерних моделей, що повністю переформували галузь.

Поряд із цією хронологічною типологією існує кілька інших, корисних для класифікації сучасних систем. За кількістю мов системи поділяють на бінарні (працюють в одній мовній парі) та багатомовні. За спрямованістю – на одно- та двонапрямлені, причому в останніх вихідна й цільова мови можуть мінятися місцями. Окрему типологію формує ступінь автоматизації, що визначає роль людини в процесі. Виділяють три рівні: FAMT (Fully-Automated Machine Translation) – повністю автоматичний переклад без людської участі; HAMT (Human-Assisted Machine Translation) – переклад зі штучно обмеженим словником і граматикою для контрольованих умов; MAHT (Machine-Assisted Human Translation) – переклад, виконуваний людиною з комп'ютерною підтримкою (термінологічні бази, перевірка узгодженості тощо). Останній тип, до якого належать сучасні CAT-інструменти, на практиці виявився найпродуктивнішим: повністю автоматичний переклад без редагування людиною залишається радше технологічним ідеалом, ніж робочою практикою – особливо у складних доменах на кшталт відеоігрової локалізації.

2.2. Машинний переклад за правилами (RBMT)

Машинний переклад за правилами (Rule-Based Machine Translation, RBMT) ґрунтується на формалізованих лінгвістичних описах вихідної та цільової мов – словниках, граматиках, правилах перетворення синтаксичних структур. У межах цього підходу традиційно виділяють три моделі: прямий переклад, трансферну модель і модель інтерлінгви (Arnold et al., 1994, с. 75). Найперші системи RBMT, починаючи з Georgetown-IBM 1954 року, реалізовували саме прямий переклад: кожне слово або фраза вихідного речення замінювалися відповідниками зі словника, а порядок слів зберігався максимально близьким до оригіналу. Цей метод, концептуально простий, виявився малопридатним для складних граматичних конструкцій і багатозначних слів.

Трансферний підхід, що з'явився у 1970-х роках, поділив процес перекладу на три послідовні етапи: аналіз вихідного тексту з побудовою синтаксичного дерева, трансфер – адаптація цієї структури до правил цільової мови, і генерація – формування тексту мовою перекладу. Цю архітектуру реалізувала система Systran, яка від 1970-х і до 2016 року використовувалася для перекладу технічних документів у Європейській Комісії та армії США (Hutchins, 2005, с. 15). Третій підхід RBMT – інтерлінгвальний – передбачав створення універсальної проміжної мови, незалежної від конкретних мов оригіналу й перекладу. Проєкти на кшталт DLT (Distributed Language Translation) використовували абстрактні семантичні представлення – наприклад, кодували поняття «сніданок» як MEAL-TYPE: morning, що теоретично дозволяло генерувати переклад на будь-яку мову з єдиного представлення (Arnold et al., 1994, с. 75–79). Однак побудова повноцінної інтерлінгви виявилася ресурсно непідйомним завданням, і цей підхід так і не вийшов за межі експериментальних реалізацій.

Серед сучасних і досі функціональних систем RBMT варто згадати: Systran – одну з найдовговічніших систем, яка пройшла еволюцію від чисто правил до гібридного підходу; PROMT – комерційну російськомовну систему, що активно використовувалася у бізнес-середовищі; Apertium – відкриту платформу,

орієнтовану на переклад між спорідненими мовами (романськими, тюркськими) і таку, що активно використовується в академічних та open-source-проєктах. Основними перевагами RBMT залишаються контрольованість і прозорість результатів: лінгвістичні правила дозволяють точно керувати поведінкою системи, що критично важливо для вузькоспеціалізованих доменів. Класичним прикладом є канадська система Météo, створена для перекладу прогнозів погоди й така, що сягнула близько 95% точності завдяки обмеженому словниковому запасу та шаблонним синтаксичним структурам (Nirenburg, 1993, с. 245–247). RBMT-системи менше залежать від обсягів даних, що робить їх привабливими для рідкісних мов, де відсутні великі паралельні корпуси, а кожен етап перекладу можна простежити, спрощуючи виправлення помилок. Головним недоліком є висока вартість підтримки: створення, оновлення й розширення лінгвістичних баз потребує постійної праці фахівців-лінгвістів, що погано масштабується на нові мовні пари й нові домени.

2.3. Статистичний машинний переклад (SMT)

Статистичний машинний переклад (Statistical Machine Translation, SMT): замість того щоб кодувати знання про мову у вигляді правил, написаних людьми, SMT-системи виводили переклад зі статистичних закономірностей, виявлених у великих паралельних корпусах. Поява SMT наприкінці 1980-х – на початку 1990-х років була зумовлена прагненням подолати головне обмеження RBMT – непомірну вартість створення й підтримки лінгвістичних описів. Першою комерційною реалізацією статистичного підходу стала вже згадана система Candide від IBM, що навчалася на корпусі канадських парламентських стенограм (Hansard) і досягла за тогочасними мірками вражаючих результатів (Berger et al., 1994).

Базова ідея SMT полягає у визначенні найімовірнішого перекладу на основі статистичних закономірностей, виявлених у паралельних даних. Класична архітектура SMT поєднує три компоненти. Перший – мовна модель (language model), що оцінює природність послідовностей слів у цільовій мові: фраза *«великий будинок»* отримує вищу ймовірність, ніж *«будинок великий»*,

оскільки перша частіше зустрічається в українському корпусі. Другий – модель перекладу (translation model), що встановлює відповідності між фразами вихідної й цільової мов: англійське heavy rain з високою ймовірністю перекладається як *«сильний дощ»*, а не *«важкий дощ»*. Третій компонент – декодер, що поєднує обидві моделі для пошуку оптимального перекладу серед мільйонів можливих варіантів.

Перший серйозний прорив у SMT відбувся з переходом від послівних до фразових моделей у 2000-х роках, реалізованих у відкритій платформі Moses. Аналіз цілих фраз замість окремих слів дозволив значно краще обробляти ідіоми (kick the bucket → «померти», а не буквальне *«копнути відро»*), а також складні синтаксичні конструкції. Іншим важливим напрямом стала робота з мультипаралельними корпусами: текст, перекладений на кілька мов, дозволяв використовувати додаткову інформацію для уточнення перекладу між тими парами мов, для яких прямих корпусних даних бракувало.

Серед відомих представників SMT-сімейства варто згадати IBM Models (моделі IBM 1–5), математичні засади яких закладено в основоположній праці (Brown et al., 1993), що стали класичним підходом до вирівнювання слів у паралельних текстах і досі вивчаються як методологічна основа машинного перекладу; ранню версію Google Translate (до 2016 року), що працювала саме на статистичному підході та використовувала близько 200 мільярдів слів з матеріалів ООН для тренування; та Moses – open-source-платформу, що залишається одним з найпоширеніших дослідницьких інструментів у галузі. Перевагою SMT була адаптивність: системи могли навчатися на будь-яких паралельних даних, що дозволяло швидко створювати перекладачі для нових мовних пар (Koehn, 2009, с. 20–23). Натомість головним недоліком стала залежність від обсягу й якості корпусних даних – без великих паралельних

текстів навчити SMT неможливо, а для більшості мовних пар такі корпуси були або відсутні, або обмежені одним доменом. Крім того, SMT мали суттєві проблеми з мовами, багатими на морфологію (зокрема, флективними слов'янськими), оскільки розглядали словоформи як окремі лексичні одиниці, не моделюючи їхні граматичні відношення. Хоча з 2010-х років SMT поступилися місцем нейронним системам, у деяких випадках – гібридних архітектурах, низькоресурсних мовах, доменах з обмеженими ресурсами – статистичний підхід зберігає актуальність.

2.4. Нейронний машинний переклад і великі мовні моделі

Нейронний машинний переклад (Neural Machine Translation, NMT) став наступним фундаментальним зсувом у парадигмі. Якщо SMT моделював переклад як комбінацію незалежних компонентів (мовна модель, модель перекладу, декодер), то NMT поєднує всі ці функції в єдиній нейромережевій архітектурі, навченій з кінця в кінець (end-to-end). Структурно NMT-моделі використовують векторні представлення слів – embeddings, неперервні представлення в багатовимірному просторі, де семантично близькі одиниці розташовані поруч – підхід, запропонований у праці Mikolov et al. (2013). Це дозволяє моделі узагальнювати: побачивши пару (king, король) на тренуванні, модель може коректно перекласти queen → «королева», навіть якщо ця конкретна пара в корпусі не зустрічалася часто.

Сучасний етап нейронного машинного перекладу розпочався з робіт Д. Богданова, К. Чо та Й. Бенджио, у яких було запропоновано механізм уваги (attention mechanism) – спосіб динамічно зосереджувати модель на релевантних частинах вхідного речення під час генерації кожного слова перекладу (Bahdanau

et al., 2014). Цей підхід дозволив значно покращити обробку довгих речень, де класичні рекурентні моделі втрачали інформацію про початок при генерації кінця – цю ваду частково долали архітектури довгої короткочасної пам'яті LSTM, запропоновані ще у праці Hochreiter & Schmidhuber (1997). У 2016 році Google перевів свій сервіс на систему GNMT (Google Neural Machine Translation), а починаючи з 2017 року в галузі домінують моделі архітектури Transformer (Vaswani et al., 2017), у якій механізм уваги повністю замінив рекурентні з'єднання.

На основі трансформерної архітектури згодом виникли великі мовні моделі (Large Language Models, LLM) – ще ширший клас систем, здатних не лише перекладати, а й виконувати широкий спектр завдань природномовної обробки. Моделі типу GPT-4, Claude, Gemini показують конкурентні результати в багатьох мовних парах і, на відміну від класичних NMT-систем, можуть адаптуватися до конкретного стилю, тону й контексту через інструкції в підказці (prompt). Водночас їхня генеративна природа створює нові методологічні виклики: нестабільність результатів (та сама вхідна послідовність може давати різні переклади), обмежена відтворюваність результатів, схильність до галюцинацій (вигадування неіснуючих фактів), а також нерівномірна якість на різних типах текстів. У нашому дослідженні саме одна з таких моделей – Gemini – використовується для генерації перекладу, що формує третю колонку корпусу.

Сучасні комерційні системи машинного перекладу – Google Translate, DeepL, DeepSeek, Microsoft Translator – спираються на трансформерні архітектури, тренуються на масштабних багатомовних корпусах і демонструють вражаючу якість на більшості мовних пар. Однак навіть найсучасніші рішення лишаються вразливими до доменних і жанрових відмінностей. Загальна модель, навчена переважно на новинних, технічних і веб-текстах, систематично гірше працює зі специфічними доменами – і відеоігри, з їхньою стилістичною гетерогенністю, ігровою термінологією та культурними алюзіями, є яскравим прикладом такого «незручного» домену. Саме це й мотивує створення спеціалізованих доменних корпусів, що дозволяють або тренувати, або хоча б адекватно оцінювати моделі в конкретній галузі.

2.5. CAT-інструменти та пам'ять перекладів

Окремий клас програмних рішень, тісно пов'язаний з машинним перекладом, але концептуально від нього відмінний, становлять CAT-інструменти (Computer-Assisted Translation tools) – автоматизовані помічники для перекладача-людини. Якщо МП-системи виконують переклад замість людини, то CAT-інструменти підтримують і прискорюють працю людини-перекладача через комплекс технологій: розбиття тексту на сегменти, ведення термінологічних баз, перевірку узгодженості, контроль термінів виконання, розподіл ролей у командних проєктах. Усі результати автоматично зберігаються в єдиній базі даних, доступній усім учасникам перекладацького процесу. У типології автоматизації CAT-інструменти належать до категорії МАНТ – переклад, виконуваний людиною з комп'ютерною підтримкою.

Центральним компонентом більшості CAT-інструментів є пам'ять перекладів (Translation Memory, TM) – база раніше перекладених фрагментів тексту, що дозволяє автоматично пропонувати готові варіанти для повторюваних або подібних сегментів. Концептуально TM виконує функцію того самого паралельного корпусу, з тією відмінністю, що вона будується інкрементально, у процесі поточної роботи перекладача, і має на меті безпосереднє практичне використання, а не дослідницький аналіз. Зв'язок між TM і паралельним корпусом методологічно важливий для нашої роботи: будь-яка добре впорядкована пам'ять перекладів може бути експортована як паралельний корпус, а будь-який паралельний корпус – імпортований у CAT-інструмент як стартова TM. Стандартним форматом обміну цими даними є TMX, про який йдеться в підрозділі 3.1.

Серед поширених CAT-інструментів варто згадати SDL Trados – комерційний стандарт індустрії, що використовується великими бюро перекладів і інституціями; MemoQ – конкурентоздатну альтернативу з активною спільнотою; Wordfast – більш доступний інструмент для індивідуальних перекладачів; OmegaT – open-source-рішення, популярне серед аматорських локалізаційних команд (зокрема й тих, що працюють з відеоіграми); SmartCAT і

Memsource – хмарні платформи, що набули поширення у 2010-х. Наявність розвиненої інфраструктури CAT-інструментів безпосередньо впливає на формат сировинних даних, з яких будуються паралельні корпуси: дуже часто наявний український переклад відеогри існує саме у формі робочих TM-файлів локалізаційної команди, які вже були використані в продакшені.

2.6. Метрики автоматичної оцінки якості машинного перекладу

Завершальним компонентом теоретичної рамки нашого дослідження є метрики автоматичної оцінки якості машинного перекладу. Ці метрики дозволяють здійснювати об'єктивне й відтворюване оцінювання перекладеного тексту без залучення експертів-людей, що критично важливо як для дослідницьких порівнянь, так і для практичної валідації корпусу. Залежно від підходу метрики можуть враховувати точність n-грамних збігів з еталоном, морфологічну подібність на рівні символів, або семантичну близькість у векторних просторах сучасних мовних моделей. У нашому дослідженні валідація згенерованої колонки корпусу – uk_gemini у порівнянні з людським еталоном uk_human – буде здійснюватися саме за цими метриками, тому розгляд кожної з них структурно орієнтовано на наступне практичне застосування.

2.6.1. BLEU

Метрика BLEU (Bilingual Evaluation Understudy), запропонована К. Папінені та співавторами у 2002 році, є найпоширенішим інструментом автоматичної оцінки якості МП і часто використовується як «золотий стандарт» галузі (Papineni et al., 2002). BLEU оцінює якість перекладу через порівняння його з одним або кількома еталонними перекладами, використовуючи статистику збігів n-грам (зазвичай від уніграм до 4-грам). Ключова ідея – визначити частоту, з якою фрази певної довжини з'являються одночасно в у згенерованому і еталонному перекладах. Для запобігання штучного завищення оцінки через повторення BLEU застосовує модифіковану точність (modified precision), а також штраф за надмірну стислість (brevity penalty), який активується, якщо довжина згенерованого тексту значно менша за еталонний.

Сильна сторона BLEU – простота й обчислювальна ефективність на великих корпусах, що зробило її стандартним інструментом порівняння МП-систем у науковій літературі; водночас непослідовність у параметрах її обчислення породжує проблему відтворюваності результатів, на що звертає увагу М. Пост у праці про стандартизацію звітності BLEU (Post, 2018). Однак метрика має істотні обмеження. Вона має слабку чутливість до синонімії – переклад «автомобіль» не отримає кредиту за відповідність еталонному «машина», навіть якщо обидва є коректними. BLEU погано враховує перестановки слів і семантичну близькість загалом: лексично точний переклад з порушеним порядком слів отримає вищу оцінку, ніж лексично відмінний, але стилістично й семантично кращий. Як показують самі автори метрики, BLEU добре працює для прямолінійних, формально-точних речень, але систематично недооцінює змістовно правильні переклади з варіативною лексикою (Papineni et al., 2002, с. 128–129). Це робить BLEU особливо ненадійною на художніх і розмовних текстах – тобто саме на тих типах, що становлять значну частину відеоігрового матеріалу.

2.6.2. METEOR

Метрика METEOR (Metric for Evaluation of Translation with Explicit ORdering), запропонована С. Банерджі та А. Лаві у 2004–2005 роках, стала прямою відповіддю на обмеження BLEU. METEOR зберігає ідею порівняння згенерованого перекладу з еталоном, але розширює спектр зіставлень за рахунок лінгвістичних характеристик: стемінгу (зведення слів до основи, щоб «біг» і «бігти» вважалися спорідненими), синонімії (через ресурс на кшталт WordNet) і парафраз (Banerjee & Lavie, 2005). Алгоритм METEOR складається з двох етапів: спершу встановлюються відповідності між словами згенерованого перекладу й еталона за кількома типами зіставлень, потім обчислюється інтегральна оцінка на основі гармонійного середнього точності й повноти зі штрафом за фрагментацію (chunk penalty), що знижує оцінку при порушенні порядку слів.

Як показують дослідження, METEOR демонструє кращу кореляцію з людськими оцінками, ніж BLEU, особливо у випадках нестандартного порядку

слів або синонімічних замін, що робить її доречною для аналізу художніх перекладів та діалогів (Banerjee & Lavie, 2005, с. 137–139). Це особливо цінно для оцінки креативного контенту, де дослівний збіг не є гарантією якості – типовий випадок ігрових діалогів. Однак METEOR має критичну залежність від доступності лінгвістичних ресурсів для конкретної мови. Для української мови, де ресурси на кшталт WordNet обмежені, повноцінне застосування METEOR потребує адаптації або заміни лексичних модулів. У нашому дослідженні ми використовуватимемо метрику в режимі без додаткових українських ресурсів, що обмежує її потенціал, але зберігає пряму порівнюваність із результатами інших робіт.

2.6.3. TER

Translation Edit Rate (TER), розроблена М. Сновером та співавторами у 2006 році, реалізує принципово інший підхід до оцінки якості (Snover et al., 2006). Замість підрахунку збігів TER підраховує мінімальну кількість редагувань – вставок, видалень, замін, перестановок – необхідних для приведення згенерованого перекладу у відповідність до еталона. Чим нижчий TER, тим ближчий переклад до еталонного. Інтуїтивно ця метрика моделює зусилля редактора-людини, який приводить машинний переклад до прийнятної якості, що особливо корисно для практичних задач локалізації, де редакційні витрати – це безпосередньо вартість роботи. TER добре працює з мовами, що мають гнучкий порядок слів (українська, німецька), оскільки явно враховує перестановки як одну з операцій.

Слабкою стороною TER є нечутливість до семантичної еквівалентності: навіть якщо згенерований переклад синонімічний еталону, але лексично відмінний, він отримає високу оцінку редагувань і, відповідно, низьку оцінку якості. Це обмежує застосовність TER для літературних або метафоричних текстів, але робить її цінною для технічних та інструктивних матеріалів. У поєднанні з BLEU й METEOR метрика TER забезпечує цілісну картину якості, дозволяючи оцінювати як лексичну точність, так і синтаксичну узгодженість у різних аспектах.

2.6.4. ChrF++

Метрика ChrF (Character n-gram F-score), запропонована М. Поповіч у 2015 році, виходить за межі лексичного рівня й обчислює подібність між перекладом і еталоном на рівні символьних n-грам – типово 6-грам (Popović, 2015). Це принципово важлива деталь для флективних мов, до яких належить українська: словоформи на кшталт «будинок», «будинку», «будинкові» при лексичному порівнянні розглядаються як зовсім різні одиниці, тоді як на рівні символів вони мають високу спільність, що відображає їхню морфологічну спорідненість. ChrF комбінує символьну точність і повноту через F-міру, не потребує токенізації (тобто менш чутлива до особливостей конкретного токенізатора) і дає значення в діапазоні від 0 до 100, де вищий бал означає кращу подібність до еталона. На практиці в цьому дослідженні (підрозділ 4.6) обчислюється розширений варіант метрики – ChrF++, який доповнює символьні n-грами словними уніграмами й біграмами, що дещо підвищує кореляцію з людською оцінкою; реалізація взята зі стандартного інструментарію sacrebleu. Для одноманітності далі в роботі вживається назва ChrF++ як позначення фактично обчисленої метрики.

ChrF++ активно використовується у дослідженнях МП для низькоресурсних мов, а також у випадках, коли BLEU дає необ'єктивні оцінки через незначні морфологічні розбіжності. Для англо-української пари ChrF++ потенційно є точнішим індикатором якості, ніж BLEU, оскільки враховує флективну природу української мови, у якій правильно вибраний корінь з помилково вибраним відмінком краще за повністю невдалий вибір кореня – і саме цю різницю ChrF++ здатна вловити, на відміну від BLEU.

2.6.5. BERTScore

BERTScore, запропонована Т. Чжан та співавторами у 2019 році, становить якісно новий клас метрик – семантичні метрики на основі попередньо навчених трансформерних моделей (Zhang et al., 2019). На відміну від попередніх метрик, що оперують поверхневими збігами на рівні слів або символів, BERTScore зіставляє векторні представлення слів згенерованого перекладу й еталона, отримані з нейронної мовної моделі (BERT, RoBERTa або інших). Кожне слово

згенерованого перекладу зіставляється з найсхожішим словом еталона за косинусною відстанню в багатовимірному просторі, після чого обчислюються precision, recall і F1-score. Зазвичай використовується саме F1-варіант метрики, значення якого варіюється від 0 до 1.

Принципова перевага BERTScore – здатність фіксувати семантичні еквіваленти навіть тоді, коли переклад містить стилістичні або синтаксичні відхилення. Переклад *«здається, дощ припиняється»* і еталон *«дощ, схоже, перестає лити»* матимуть низьку оцінку BLEU, але високу – BERTScore, бо їхні векторні представлення відображають подібний зміст. Це робить BERTScore найбільш «людиноподібною» з усіх автоматичних метрик, які ми використовуємо в роботі. Однак ця метрика має й недоліки: залежність від конкретної мовної моделі (різні моделі дають різні оцінки), вища обчислювальна вартість, нижча відтворюваність між дослідженнями, що використовують різні бекенди. Для української мови важливо обирати модель, яка адекватно представляє українську – наприклад, multilingual BERT або XLM-R.

2.6.6. Зведена характеристика і обґрунтування набору метрик

Для вибору ефективних метрик для оцінки перекладу текстів відеоігр, вважаємо за необхідне систематизувати базові характеристики проаналізованих метрик.

BLEU – n-грамна метрика на рівні слів; точкова оцінка лексичної точності; чутлива до перестановок; погано враховує синонімію; стандарт галузі для порівняльних досліджень.

METEOR – гармонійне середнє точності й повноти з урахуванням стемінгу та синонімії; краща кореляція з людськими оцінками; залежна від лінгвістичних ресурсів цільової мови.

TER – кількість редагувань для приведення до еталона; інтуїтивно моделює редакційні витрати; добре працює з гнучким порядком слів; нечутлива до семантики.

ChrF++ – F-міра на основі символьних і словних n-грам; не потребує токенізації; особливо корисна для флективних мов; добре відображає морфологічну подібність.

BERTScore – косинусна подібність векторних представлень з трансформерної моделі; найкраще відображає семантику; залежна від обраного бекенду; вища обчислювальна вартість.

Жодна окрема метрика не дає повної картини якості перекладу – кожна моделює свій аспект і має систематичні сліпі плями. Тому методологічно обґрунтованим рішенням є використання набору метрик, що компенсують слабкості одна одної. У нашому дослідженні ми оцінюватимемо згенеровану колонку корпусу у порівнянні з людською за всіма п'ятьма розглянутими метриками: BLEU дасть стандартний орієнтир для порівняння з іншими роботами; METEOR додасть лексичну гнучкість; TER покаже редакційну дистанцію; ChrF++ забезпечить морфологічну адекватність для української як флективної мови; BERTScore відобразить семантичну глибину. Тільки комбінація цих оцінок дозволяє сформувати обґрунтоване уявлення про якість і обмеження згенерованих перекладів, отриманих через LLM.

Висновки до розділу 2

У другому розділі ми розглянули машинний переклад як технологію в його історичному й концептуальному розвитку, а також метрики автоматичного оцінювання якості машинного перекладу – два компоненти, без яких не може бути коректно поставлене ані завдання побудови корпусу, ані критерії його придатності.

По-перше, еволюція МП пройшла три якісні етапи: системи на правилах (RBMT), статистичні системи (SMT) і нейронні системи (NMT) разом з LLM. Ключовою тенденцією цієї еволюції було поступове зменшення ролі ручного лінгвістичного програмування й зростання ролі даних. RBMT потребували трудомістко закодованих правил; SMT навчалися автоматично з паралельних корпусів; NMT і LLM здатні узагальнювати з нечітких прикладів і адаптуватися через підказки. Цей рух закономірно ставить паралельний корпус у центр

сучасної методології МП – навіть для оцінювання моделей, навчених на закритих даних, потрібні корпуси для верифікації.

По-друге, всупереч поширеному уявленню про «вирішену задачу машинного перекладу», сучасні системи зберігають серйозні обмеження саме у тих галузях, що відповідають характеристикам відеоігрового тексту: жанрова й стилістична варіативність, прагматична складність, культурні алюзії, специфічна термінологія. Загальні моделі, тренувані на новинах і технічних текстах, систематично гірше працюють у спеціалізованих доменах. Це робить актуальним створення доменних корпусів – як для тренування доменно-адаптованих моделей, так і просто для коректного оцінювання якості існуючих систем у конкретній галузі.

По-третє, CAT-інструменти й пам'ять перекладів демонструють тісний методологічний зв'язок між інженерією локалізаційної практики та академічними дослідженнями: TM-бази є за своєю природою паралельними корпусами, і будь-який добре впорядкований корпус може бути імпортований у CAT як стартова TM. Це підтверджує практичну цінність нашого корпусу не лише як дослідницького ресурсу, а і як потенційного матеріалу для подальшої локалізаційної роботи з відеоіграми.

По-четверте, набір з п'яти метрик автоматичної оцінки – BLEU, METEOR, TER, ChrF++, BERTScore – забезпечує комплексне покриття аспектів якості перекладу: від поверхневих лексичних збігів до глибинної семантичної подібності. Жодна з метрик окремо не дає повної картини, але в сукупності вони дозволяють обґрунтовано аналізувати якість згенерованої колонки корпусу у порівнянні з людським еталоном. Саме цей набір метрик буде застосовано у практичній частині роботи.

РОЗДІЛ 3. МЕТОДОЛОГІЯ ПОБУДОВИ ПАРАЛЕЛЬНОГО АНГЛО-УКРАЇНСЬКОГО КОРПУСУ ТЕКСТІВ ВІДЕОІГР

3.1. Корпуси у машинному перекладі: огляд

Поняття корпусу в сучасній лінгвістиці охоплює великий, систематично організований набір текстів або їхніх фрагментів, призначений для лінгвістичного аналізу й машинної обробки (Sinclair, 1991). На відміну від колекції випадково зібраних документів, корпус характеризується свідомо запроєктованою структурою: тексти у ньому добираються за визначеними критеріями (жанр, домен, період, мова), супроводжуються метаданими, а часто й лінгвістичною розміткою (морфологічною, синтаксичною, семантичною). Корпусний підхід зумовив фундаментальну зміну в багатьох галузях прикладної лінгвістики, перетворивши умоглядні припущення про мовну норму на емпірично перевірювані твердження, підкріплені статистикою реального вживання (McEnery & Hardie, 2012).

За структурою й призначенням корпуси поділяються на кілька типів. Одномовні корпуси – найбільш ранній і поширений тип – містять тексти однією мовою і використовуються для лексикографії, описової граматики, стилістичних досліджень, а також для тренування мовних моделей. Порівняльні (comparable) корпуси охоплюють тексти кількома мовами, відібрані за подібними критеріями (наприклад, новинні повідомлення з певного періоду англійською та українською), але без прямої перекладацької відповідності між ними. Паралельні корпуси, які становлять основний інтерес для машинного перекладу, містять тексти однією мовою разом з їхніми перекладами іншою мовою, причому між ними встановлено вирівнювання – на рівні документів, абзаців, речень або навіть слів. Окремий клас становлять мультипаралельні корпуси, у яких один вихідний текст представлено перекладами трьома й більше мовами; класичним прикладом

тут є корпуси протоколів європейських інституцій (Europarl), що містять матеріал двадцятьма та більше мовами одночасно (Koehn, 2005).

Для машинного перекладу паралельні корпуси є фундаментальним ресурсом – без них неможливі ані статистичні, ані сучасні нейронні системи. Якість і обсяг паралельних корпусів безпосередньо визначає якість моделей, навчених на них: модель не може коректно перекласти явище, прикладів якого вона не бачила під час навчання, і не може коректно оцінити переклад у домені, не представленому в оцінювальному корпусі. Це робить особливо важливими спеціалізовані доменні корпуси – паралельні корпуси, обмежені певною предметною областю (медицина, право, технічна документація, відеоігри). Доменний корпус значно менший за загальномовний, але концентрує специфічну термінологію, типові синтаксичні конструкції та стилістичні особливості області.

Серед загальнодоступних паралельних ресурсів, що містять українську мову, центральне місце посідає проєкт OPUS – найбільший відкритий репозиторій паралельних корпусів, який агрегує матеріали з десятків джерел (субтитри фільмів, документи Європарламенту, технічна документація, релігійні тексти, виправлення в Wikipedia) (Tiedemann, 2012). Попри значний загальний обсяг, специфічні домени представлені нерівномірно: переклади субтитрів і технічної документації покривають OPUS добре, тоді як ігрова сфера, юридичні тексти України, спеціалізована медична термінологія представлені фрагментарно або відсутні зовсім. На момент проведення нашого дослідження окремого спеціалізованого паралельного корпусу англо-українських ігрових текстів у відкритому доступі ми не виявили – що значною мірою й мотивує тему цієї роботи.

Якість паралельного корпусу визначається кількома взаємопов'язаними характеристиками. Першою з них є точність вирівнювання: пара EN-UK на рівні речення має містити саме переклад одне одного, а не випадковий збіг. Другою – чистота даних: відсутність технічних артефактів (HTML-тегів, керуючих символів, нерозкритих плейсхолдерів), послідовне кодування, нормалізація пунктуації. Третьою – репрезентативність: корпус має покривати різні стилі,

регістри, синтаксичні структури в межах свого домену. Четвертою – наявність метаданих: чим більше відомо про походження конкретного запису (тип тексту, автор перекладу, час створення), тим ширшим є спектр можливих застосувань корпусу.

Окремої уваги заслуговує формат зберігання й обміну паралельних корпусів. Класичним стандартом, розробленим в індустрії перекладу, є TMX (Translation Memory eXchange) – XML-формат, спроєктований для обміну пам'яттю перекладів між CAT-інструментами; він зручний для перекладацьких пайплайнів, але незручний для машинного навчання та обробки великих обсягів даних, оскільки накладні витрати на парсинг XML значні. У сучасних дослідженнях МП значно поширенішим є формат JSONL (JSON Lines), у якому кожен рядок файлу – самостійний JSON-об'єкт, що містить запис корпусу з усіма його полями. JSONL природно інтегрується з бібліотеками обробки даних (pandas, HuggingFace Datasets), легко зчитується потоково для великих корпусів, дозволяє довільну структуру метаданих і зручно версіонується системами контролю версій. Альтернативними простими форматами є TSV/CSV (зручні для невеликих корпусів і ручного перегляду) та паралельні plaintext-файли з вирівнюванням за номером рядка (часто використовується в OPUS). У нашому дослідженні ми обираємо саме JSONL як формат, що оптимально поєднує гнучкість, читабельність і сумісність зі стандартними інструментами обробки текстових даних.

Підсумовуючи, можна стверджувати, що паралельний корпус – це не просто двомовний текстовий ресурс, а інженерний продукт, побудований за чіткими методологічними принципами. Його якість визначає якість усього, що буде на ньому ґрунтуватися – від статистичних моделей до оцінок ефективності МП. Для відеоігрового домену англо-українська пара лишається на сьогодні мало забезпеченою такими ресурсами, що й створює методологічну прогалину, яку покликана заповнити ця робота.

3.2. Обґрунтування вибору текстового матеріалу: Baldur's Gate 3

Вибір конкретної відеогри як текстового матеріалу для побудови паралельного корпусу не є випадковим і ґрунтується на низці взаємопов'язаних критеріїв. Перш ніж описати, чому ми обрали саме Baldur's Gate 3, варто сформулювати самі критерії, яким має відповідати текстовий матеріал. Кандидатна гра мала: по-перше, належати до жанру з високою стилістичною гетерогенністю, аби корпус репрезентував різні типи ігрового тексту; по-друге, мати офіційну українську локалізацію – без неї ми не отримуємо людський еталон, без якого валідація згенерованих перекладів неможлива; по-третє, зберігати тексти у форматі, придатному для автоматизованого вилучення, що принципово знижує трудовитрати на побудову корпусу; по-четверте, мати достатній обсяг текстового матеріалу, щоб корпус був статистично репрезентативним; по-п'яте, бути відносно сучасною, аби лексика й стилістика відображали актуальний стан мови.

Baldur's Gate 3 (Larian Studios, 2023) задовольняє всі ці критерії. Як рольова гра в жанрі темного фентезі, побудована на правилах настільної системи Dungeons & Dragons (5-те видання), вона містить надзвичайно широкий спектр текстових типів: розгалужені діалоги між десятками персонажів зі своїми стилістичними особливостями, кінематичні наративні фрагменти, описи предметів і заклинань (близькі до науково-технічного стилю), внутрішньоігрові книги й рукописи (від поетичних до філософських), листи й щоденникові записи персонажів, інтерфейсні елементи (підказки, повідомлення системи, описи навичок), а також численні бестіарії та довідкові тексти. Стилістично BG3 є практично ідеальним зрізом усіх типів ігрового тексту, окреслених у підрозділі 1.1.

Окрему перевагу BG3 становить наявність повної офіційної української локалізації, виконаної професійною командою – це порівняно рідкісна ситуація для великих сюжетних RPG, оскільки історично українська мова часто не входила до пріоритетного переліку локалізацій великих видавців. Наявність якісного людського перекладу дає нам можливість використати його як еталон для оцінки згенерованої колонки корпусу. Текстові дані BG3 зберігаються у форматі XML з чіткою структурою, що містить унікальні ідентифікатори

(contentuid) і версійні позначки для кожного запису, що принципово спрощує вирівнювання EN-UK пар. Загальний обсяг гри становить понад два мільйони слів англомовного тексту, що дає достатній матеріалу для будь-яких подальших досліджень.

Додатковим аргументом є те, що BG3 належить до проєктів з оригінальним всесвітом (хоча й заснованим на правилах D&D, але з власним сюжетом, персонажами й локаціями). Це означає, що локалізатори мали значну творчу свободу при перекладі, і пара (EN, UK_human) у нашому корпусі демонструватиме всю варіативність перекладацьких рішень – від буквальних відповідників до сміливих транскреацій культурних реалій. Така варіативність робить корпус особливо цінним для досліджень якості МП: він не зводить переклад до простого пошуку лексичних еквівалентів, а ставить перед автоматичною системою справжні перекладацькі виклики.

3.3. Загальні методи вилучення текстів із сайтів ігор

Перш ніж зосередитися на конкретиці BG3, варто коротко окреслити загальний ландшафт методів вилучення текстових даних з відеоігор. Це необхідно з двох причин: по-перше, для розуміння місця нашого підходу в загальній методології; по-друге, тому що в перспективі описана нами методологія потенційно може бути перенесена на матеріал інших ігор, де технічні умови суттєво відрізняються.

У більшості відеоігор текстові дані (інтерфейс, субтитри, описи, діалоги) зберігаються в архівах або файлах із нетиповими розширеннями. Залежно від технологічної платформи й політики розробника, доступ до цих даних може коливатися від тривіального – звичайний текстовий редактор відкриває файл як простий текст – до надскладного, що потребує зворотної інженерії й написання спеціалізованих парсерів. Між цими крайнощами лежить більшість реальних випадків, де доступ потребує спеціалізованих утиліт, створених або самими розробниками, або модерською спільнотою.

Класичним прикладом проміжної складності є ігри Blizzard Entertainment (Starcraft, Warcraft III), у яких дані зберігаються у форматі MPQ (Mike O'Brien

Pack) – пропрієтарному архівному форматі, що містить ігрові дані, графіку, звуки й текст. Для роботи з такими архівами існують утиліти на кшталт Lazik's MPQ Editor, створені модерською спільнотою. Інший показовий випадок – серія ігор Dawn of War від Relic Entertainment, де дані зберігаються в архівах із розширенням .sga і можуть бути розпаковані за допомогою спеціалізованої утиліти Mod_Studio. Розпакований текстовий файл (наприклад, W40k.ucs у Dawn of War: Soulstorm) уже доступний для подальшої обробки звичайними текстовими інструментами і містить близько 40 тисяч слів локалізаційного тексту з індексацією рядків. Ці методологічні прецеденти демонструють принципову можливість систематичного вилучення ігрових текстів навіть за умов закритого формату – за умови наявності спільнотних інструментів.

Гнучкі ігрові рушії на кшталт Unreal Engine, Unity, або власних розробок (як у випадку Larian Studios), частіше зберігають локалізаційні дані у відкритих або напіввідкритих форматах: XML, JSON, CSV, або власних структурованих текстових форматах. Це принципово спрощує автоматичне вилучення текстів і робить такі ігри пріоритетним матеріалом для побудови дослідницьких корпусів. Саме до цієї категорії належить Baldur's Gate 3.

3.4. Структура паралельного корпусу: формат, поля, метадані

Кінцевий продукт нашого дослідження – паралельний корпус – реалізується у форматі JSON Lines (.jsonl), обґрунтуванню якого було приділено окрему увагу в підрозділі 3.1. Кожен рядок файлу становить самостійний JSON-об'єкт, що описує один запис корпусу – пару (точніше, трійку) текстів і пов'язані з ними метадані. Така архітектура дозволяє безпосередньо інтегрувати корпус у сучасні бібліотеки обробки текстових даних (HuggingFace Datasets, pandas, jq), потоково обробляти великі обсяги без завантаження всього файлу в пам'ять, легко розширювати схему додатковими полями без порушення сумісності з попередніми версіями.

Базова структура запису має такий вигляд: текстові поля en, uk_human, uk_gemini містять, відповідно, оригінальний англомовний текст з ігрових файлів BG3, офіційний український переклад тієї самої одиниці та згенерований

моделлю Gemini український переклад (поле uk_gemini заповнюється лише для записів, що увійшли до стратифікованої вибірки, для решти воно має значення null). Поле id містить унікальний ідентифікатор запису, що відповідає внутрішньому ідентифікатору гри (contentuid у термінології Larian), а поле version – версію відповідного рядка локалізації. Поле metadata містить структуровані додаткові відомості – функціональний тип тексту, клас довжини, кількість символів, набір ознак наявності внутрішньої розмітки та її симетричності між мовами, а також службові показники якості, повторюваності й належності запису до вибірки.

Блок metadata фіксує функціональні й технічні характеристики кожного запису: функціональний тип тексту (поле type зі значеннями dialogue, ui_short, narrative, mechanic_description, book_or_document, telepathic, ui_keybind); клас довжини (length_class: short / medium / long / extreme) і кількість символів в обох мовах (char_count_en, char_count_uk); набір булевих ознак наявності розмітки – тегів інтерфейсу Larian (has_lstag), курсиву (has_italic), напівжирного (has_bold), розривів рядка (has_br) і плейсхолдерів (has_placeholder); ознаку тегової асиметрії між мовами (tag_asymmetry) з її деталізацією (tag_asymmetry_details); якість людського перекладу (uk_human_quality зі значеннями на кшталт clean чи marker_mismatch); кількість повторень запису в корпусі (duplicate_count); а також належність запису до стратифікованої вибірки, для якої згенеровано переклад (in_sample). Саме такий набір метаданих уможливлює стратифікований аналіз якості перекладу за класами, довжиною й станом розмітки, а не лише усереднену оцінку по всьому корпусу.

Нижче наведено приклад одного запису корпусу у форматі JSON Lines (поля подано повністю, текст збережено без змін):


```
{
  "id": "h000006d4gcefbg4092gbb39gfeb27a3bb0a7",
  "version": 1,
  "en": "Dust. On.<i> </i>My. <i>Tongue!</i>",
  "uk_human": "<i>Повний рот</i>... піску!",
  "uk_gemini": "Пил. На. <i>Моєму.</i> <i>Язиці!</i>",
  "metadata": {
    "type": "dialogue",
    "length_class": "short",
    "char_count_en": 37,
    "char_count_uk": 22,
    "has_lstag": false,
    "has_italic": true,
    "has_bold": false,
```

```
      "has_br": false,
      "has_placeholder": false,
      "tag_asymmetry": true,
      "tag_asymmetry_details": {
        "i": [
          3,
          1
        ]
      },
      "uk_human_quality": "clean",
      "duplicate_count": 1,
      "in_sample": true
    }
  }
```


Принциповим методологічним рішенням є збереження всіх трьох колонок – оригіналу, людського й згенерованого перекладу – в одному файлі, а не в трьох окремих. Альтернативний варіант із роздільним зберіганням «золотого» (EN+UK_human) і «згенерованого» (EN+UK_gemini) підкорпусів формально був би чистішим у термінах призначення, але втрачав би ключову можливість, що випливає з нашої архітектури: тренування моделей постредагування на парах (uk_gemini, uk_human) – окрема цінна задача в галузі гібридного машинного перекладу. Один файл із трьома колонками гнучкіший у використанні: дослідник, що потребує суто паралельного EN-UK корпусу, фільтрує записи з непорожнім uk_human; дослідник, що працює з оцінкою якості LLM на ігрових текстах, бере uk_gemini як згенерований переклад і uk_human як еталон; дослідник, що тренує моделі постредагування, бере (uk_gemini, uk_human) як вхід і вихід відповідно.

Поле metadata з типом тексту особливо важливе для коректної інтерпретації результатів валідації. Як показав підрозділ 1.1, ігровий текст принципово гетерогенний: метрики якості перекладу (BLEU, METEOR, тощо.) даватимуть різні показники на діалогах, інтерфейсі та наративі навіть за однакової об'єктивної якості перекладу – просто через структурні особливості цих типів тексту. Зведена оцінка по корпусу без розбивки за типами тексту була б методологічно неінформативною. Тому записи в нашому корпусі мають супроводжуватися чіткою класифікацією за функціональним типом, що дозволяє в подальшому проводити аналіз як по корпусу загалом, так і по окремих його стратах.

3.5. Принципи фільтрації первинних даних і отримання чистоти текстових даних

Первинні дані, отримані безпосередньо з ігрових файлів, не придатні до прямого використання у текстовому корпусі. Вони містять значну кількість технічних артефактів – керуючих кодів, плейсхолдерів, форматувальних тегів, дублікатів, а також записів, що з різних причин непридатні для лінгвістичного аналізу (порожні рядки, технічні мітки, номери версій). Тому етап фільтрації й очищення даних є критичною складовою побудови корпусу, від якості якого безпосередньо залежить наукова цінність кінцевого продукту.

Перший рівень фільтрації – видалення технічних артефактів і нормалізація. До цього етапу належать: розгортання або видалення HTML- і XML-тегів, що не несуть лінгвістичного навантаження (теги стилізації <i>, <b>, переносів рядка); видалення керуючих символів (\n, \t, табуляції в недоречних позиціях); нормалізація пунктуації (різні типи лапок та апострофів зводяться до уніфікованих варіантів); приведення пробільних символів до канонічної форми (видалення подвійних пробілів, обрізка крайових пробілів). Окремо обробляються плейсхолдери на кшталт [1], [2] чи спеціалізовані шаблони з префіксами [IE_], [GLO_], [GEN_] – їх можна або зберігати в записах із позначкою у метаданих, або виносити такі записи в окремий підкорпус для спеціальної обробки.

Другий рівень – класифікація за довжиною. У межах цього етапу записи не відсіюються за обсягом, а отримують у метаданих відповідну мітку length_class (short / medium / long / extreme). Такий підхід обрано як методологічно гнучкіший: він дозволяє в подальшому проводити аналіз як по корпусу загалом, так і за окремими довжинними стратами, водночас не втрачаючи короткі UI-елементи й довгі книжкові тексти, які мають окрему лінгвістичну й перекладацьку цінність. Конкретні пороги класифікації визначаються на основі статистичного розподілу довжин у фактичних даних BG3.

Третій рівень – валідація вирівнювання EN ↔ UK_human. Хоча у випадку BG3 вирівнювання задається

самою структурою файлів (кожен запис UK-локалізації має той самий ідентифікатор, що й відповідний запис EN-оригіналу), на практиці можливі ситуації, коли запис присутній лише в одній мові (наприклад, оновлення гри додало новий текст англійською, який ще не локалізовано). Такі записи виключаються з основного корпусу, але можуть зберігатися в окремому підкорпусі лише з англомовною колонкою – потенційно корисному для досліджень обсягу неперекладеного матеріалу.

Четвертий рівень – обробка дублікатів. У великих ігрових проєктах часто трапляються однакові текстові записи з різними ідентифікаторами (наприклад, повторювані репліки другорядних персонажів, стандартні фрази системних повідомлень). Стратегія обробки залежить від мети: для тренувальних корпусів дублікати слід зберігати, бо вони відображають реальну частотність явищ у домені; для оцінювальних корпусів – навпаки, дедуплікація необхідна, щоб одна й та сама пара не впливала на агрегатну оцінку непропорційно. У нашому корпусі ми обираємо проміжний варіант: повні дублікати дедуплікуються, але інформація про частоту повторень зберігається в метаданих.

Окремою задачею є генерація колонки згенерованих текстів uk_gemini, що методологічно належить до етапу побудови, а не очищення, але має власні принципи контролю якості. Промпт для Gemini формулюється з урахуванням специфіки ігрового перекладу: модель має зберігати ігрову термінологію, дотримуватися стилю звертання (для діалогів – переважно «ти», для інтерфейсу – нейтрально-наказовий регістр), не вигадувати нових власних назв, обробляти плейсхолдери як незмінні токени. Параметр temperature встановлюється на низькому рівні (0,1–0,3), що зменшує варіативність і підвищує відтворюваність результатів. Усі помилки API обробляються з повтором запитів, а проміжні результати зберігаються інкрементально, щоб у разі переривання процесу можна було продовжити з останньої успішно опрацьованої позиції.

3.6. Етичні та юридичні аспекти укладання паралельного корпусу

Робота з матеріалом, що є інтелектуальною власністю комерційного розробника, потребує окремого етико-юридичного обґрунтування. Тексти Baldur's Gate 3 – як оригінальні англомовні, так і офіційний український переклад – захищені авторським правом Larian Studios. Створення похідних робіт, відтворення значних фрагментів чи публічне поширення таких матеріалів без дозволу правовласника становило б порушення авторського права.

Однак специфіка нашого дослідження дозволяє коректно вирішити це питання в межах принципу академічної доброчесності а використання первинних даних (fair use). Корпус, що формується в межах цієї роботи, призначений винятково для академічних цілей у рамках кваліфікаційної бакалаврської роботи. Він не передбачає публічного поширення – ні через відкриті репозиторії на кшталт HuggingFace Datasets, ні через GitHub, ні через будь-які інші канали публічного доступу. Корпус використовується тільки для проведення дослідження, формулювання його результатів і ілюстрації методологічних принципів у тексті роботи. Окремі приклади з корпусу, що цитуються в роботі, мають обмежений обсяг і слугують винятково ілюстративним цілям, що відповідає традиційному принципу академічного цитування.

Інший важливий аспект – доступ до матеріалу. Дослідник отримує тексти з ліцензійної копії гри, яка була законно придбана; жодних піратських копій, неавторизованих витоків чи вилучення текстів з обхідних каналів не використовується. Вилучення текстів виконується з власних файлів встановленої гри, що для приватного дослідницького використання загалом не суперечить умовам ліцензійної угоди (EULA), хоча конкретні обмеження варіюються від видавця до видавця і потребують уважного прочитання. Larian Studios, як показує політика компанії щодо модерських проєктів і дослідницьких ініціатив, традиційно не висуває обмежень на академічні дослідження своїх ігор за умови некомерційного використання й відсутності публічного поширення матеріалів.

Окремо треба згадати етичний аспект, пов'язаний з використанням великої мовної моделі Gemini для генерації перекладів. Доступ до моделі здійснюється через офіційний API Google AI Studio, відповідно до умов використання сервісу. Згенерований моделлю переклад використовується лише в межах дослідження, його якість критично оцінюється порівняно з людським еталоном, що відповідає стандартним академічним практикам у галузі досліджень МП. Жодних спроб видавання згенерованих перекладів за людські, ані будь-якого іншого використання, що могло б ввести в оману, не передбачається.

Висновки до розділу 3

У третьому розділі ми сформулювали методологічну основу побудови паралельного корпусу – від огляду наявних паралельних корпусів і обґрунтування вибору текстового матеріалу до структурних і етико-юридичних рішень. Кілька ключових позицій варто підсумувати.

По-перше, паралельний корпус є фундаментальним ресурсом для будь-якого підходу до машинного перекладу – від статистичного до нейронного, а якість і репрезентативність корпусу безпосередньо визначають якість моделей, навчених або оцінюваних на ньому. Для англо-українського ігрового домену загальнодоступних ресурсів такого типу на сьогодні бракує, і саме ця прогалина становить методологічне виправдання нашого дослідження.

По-друге, Baldur's Gate 3 обрано як матеріал дослідження не випадково: гра задовольняє всі поставлені критерії (стилістична гетерогенність, наявність офіційної української локалізації, придатний для автоматизованої обробки формат, достатній обсяг, актуальність). Її особливістю як матеріалу для корпусу є поєднання творчої свободи перекладачів з технічно передбачуваною структурою файлів – комбінація, що рідко зустрічається в одному проєкті.

По-третє, формат JSONL з трьома мовними колонками (en, uk_human, uk_gemini) і структурованими метаданими був обраний як оптимальне поєднання гнучкості, читабельності та сумісності зі стандартними інструментами обробки текстових даних. Архітектура корпусу свідомо

передбачає різнобічне використання: від простого паралельного аналізу до тренування моделей постредагування.

По-четверте, чотирирівнева система фільтрації – нормалізація, фільтрація за довжиною, валідація вирівнювання, обробка дублікатів – забезпечує необхідну якість даних, тоді як принципи генерації третьої колонки (контрольовані параметри, спеціалізований промпт, інкрементальне збереження) гарантують відтворюваність результатів.

По-п’яте, обмеження сфери використання корпусу академічними цілями в межах кваліфікаційної роботи знімає юридичні питання, пов'язані з інтелектуальною власністю Larian Studios, і відповідає прийнятим у науковій спільноті етичним нормам роботи із захищеними матеріалами.

РОЗДІЛ 4. АВТОМАТИЧНЕ УКЛАДАННЯ ТА ВАЛІДАЦІЯ ПАРАЛЕЛЬНОГО АНГЛО-УКРАЇНСЬКОГО КОРПУСУ ТЕКСТІВ ВІДЕОІГР

4.1. Архітектура програмної системи

Програмна система побудови корпусу реалізована мовою програмування Python (версія 3.11) у вигляді модульного пайплайну, де кожен модуль виконує одну логічно завершену операцію над даними і зберігає проміжний результат у власному файлі. Така архітектура має дві переваги: по-перше, кожен модуль можна запустити й протестувати окремо без потреби виконувати весь пайплайн; по-друге, у разі збою на пізніх етапах не потрібно повторювати раніші – достатньо завантажити останній проміжний результат і продовжити. Це особливо важливо для модуля генерації перекладу через зовнішній API, де виконання займає десятки годин і будь-яке переривання без інкрементального збереження означало б істотну втрату часу й ресурсів.

Загальна архітектура пайплайну складається з дев'яти основних модулів:

Перший модуль здійснює інвентаризацію HTML/XML-тегів у вихідних англомовних даних – розвідувальний крок, що дає підстави для подальших рішень про обробку розмітки.

Другий модуль вилучає текстові записи з XML-файлів локалізації для обох мов, формуючи проміжні JSONL-файли з ідентифікаторами, версіями та первинними текстами.

Третій модуль вирівнює англомовні й україномовні записи за спільним ключем contentuid, формуючи паралельні пари.

Четвертий модуль класифікує кожну пару за функціональним типом тексту на основі евристик, виявлених при аналізі вибірки.

П'ятий модуль виконує фільтрацію й нормалізацію: видалення технічних артефактів, фіксацію метаданих, дедуплікацію.

Шостий модуль формує стратифіковану вибірку для генерації перекладу.

Сьомий модуль звертається до API моделі Gemini для отримання згенерованих перекладів і записує результати у фінальну колонку корпусу.

Восьмий модуль обчислює статистичні характеристики корпусу (розподіли за класами, довжиною, наявністю тегів).

Дев'ятий модуль обчислює метрики автоматичної оцінки якості згенерованої колонки відносно людського еталону (BLEU, METEOR, TER, ChrF++, BERTScore).

Детальний технічний опис усіх модулів пайплайну, їхніх входів, виходів і ключових алгоритмів подано в Додатку А, а відомості про доступ до повного вихідного коду – у Додатку Г.

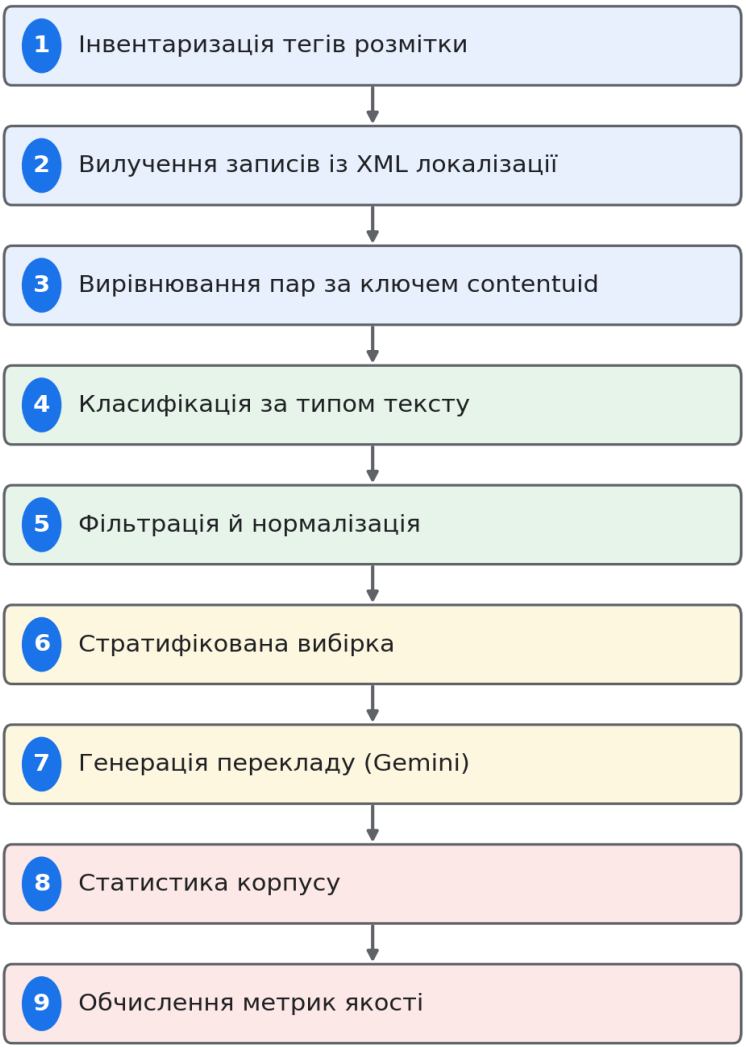

Рисунок 4.1. Загальна архітектура пайплайну: дев'ять модулів від інвентаризації розмітки до обчислення метрик якості.

Кожен модуль приймає на вхід один або кілька JSONL-файлів попереднього етапу й видає на виході новий JSONL-файл, збагачений додатковою інформацією або відфільтрований за певними критеріями. Усі файли проміжних результатів зберігаються в окремій директорії data/intermediate, що дозволяє в будь-який момент повернутися до первинних даних або до результату конкретного етапу для повторного аналізу. Фінальний корпус формується у файлі data/final/corpus.jsonl, до якого долучається файл описової статистики data/final/stats.md.

Технологічний стек системи включає кілька зовнішніх бібліотек, кожна з яких виконує специфічну функцію. Для парсингу XML-файлів локалізації використовується стандартна бібліотека xml.etree.ElementTree, що достатньо ефективна для обробки файлів розміром у десятки мегабайтів. Для роботи з LLM та обчислення семантичних метрик застосовуються бібліотеки google-genai (офіційний клієнт Google AI Studio API), bert-score (для метрики BERTScore), sacrebleu (для BLEU, ChrF++), evaluate від HuggingFace (для METEOR, TER). Для управління секретами (API-ключ Gemini) використовується python-dotenv. Стійкість до помилок зовнішніх API-запитів забезпечує бібліотека tenacity, що реалізує експоненційні повтори при тимчасових збоях. Усі залежності зафіксовано у файлі requirements.txt з конкретними версіями для забезпечення відтворюваності результатів.

Технічно процес запуску гри й вилучення з неї текстових файлів виконується поза основним Python-пайплайном – для цього використовуються спеціалізовані інструменти, призначені для розпакування ігрових архівів. У нашому випадку це BG3 Modder's Multitool (Shinyhobo, 2024) – графічна оболонка над бібліотекою LSLib (Norbyte, 2024), що дозволяє розпаковувати контейнерні архіви формату .pak та конвертувати бінарні файли локалізації .loca у формат XML. Розпаковані XML-файли потім поступають на вхід Python-пайплайну, що описаний вище. Деталі цього етапу розкриваються в наступних підрозділах.

4.2. Автоматичне вилучення англомовних текстів з Baldur's Gate 3

Текстові ресурси Baldur's Gate 3 зберігаються в Steam-копії гри в директорії Data, що містить серію контейнерних архівів з розширенням .pak. Архіви організовані за функціональним принципом: основні ресурси гри (Gustav.pak, Shared.pak, Engine.pak), мовні пакети (English.pak, Ukrainian.pak та інші), пакети з графічними й звуковими ресурсами. Для цілей нашого дослідження релевантні саме мовні пакети, що містять локалізаційні файли всіма мовами, на які гру було офіційно перекладено.

Розпакування .pak-архівів виконано за допомогою інструмента BG3 Modder's Multitool. Цей інструмент становить графічну оболонку над відкритою бібліотекою LSLib, розробленою спільнотою модерів і дослідників ігор Larian Studios. На відміну від суто командного інтерфейсу LSLib (виконуваного файлу divine.exe), Multitool автоматично виявляє встановлену копію гри, надає графічну можливість вибору архівів для розпакування й конвертації, а також може виконувати ці операції пакетно. У межах нашого дослідження було розпаковано пакет English.pak, що дав папку з локалізаційними ресурсами англійською мовою у внутрішній структурі директорій.

Усередині розпакованого пакета файли локалізації організовані як ієрархія директорій з гендерно-варіативними підпапками: основний файл english.loca знаходиться в директорії Localization/English, поряд з ним розташовані менші файли english_M_to_X.loca та english_to_F.loca, що містять гендерні варіанти реплік для випадків, коли репліка змінюється залежно від статі персонажа гравця або співрозмовника. У підпапці Gender/ дублюються ті самі дані з детальнішим розподілом. Для основного корпусу використано файл english.loca, що містить переважну більшість текстового матеріалу гри (понад 99% від загального обсягу); гендерні варіанти залишено поза рамками цього дослідження як спеціалізована страта, що потребує окремої методологічної розробки.

Формат .loca – це власний бінарний контейнер Larian Studios для локалізаційних даних. Для роботи з ним у Python-пайплайні файл попередньо конвертовано в текстовий XML-формат за допомогою тієї самої утиліти LSLib. Після конвертації отримано файл english.loca.xml обсягом приблизно 33 мегабайти. Цей файл має чітку структуру: кореневий елемент contentList містить

послідовність елементів content, кожен з яких має два атрибути – contentuid (37-символьний шістнадцятковий ідентифікатор з префіксом «h» і чотирма символами «g» як розділювачами між сегментами хешу) та version (цілочисельний номер версії запису, що збільшується з кожним патчем при зміні тексту), – а власне текстовий вміст розташовується між відкривальним і закривальним тегами content. Приклад типового запису: елемент content з атрибутом contentuid рівним «h000006d4gcefbg4092gbb39gfeb27a3bb0a7» та атрибутом version рівним одиниці містить текст *«Sorry, darling, I haven't got time for underlings.»*.

Парсинг XML-файлу виконано стандартним модулем xml.etree.ElementTree. Для кожного запису витягуються три поля: ідентифікатор contentuid, номер версії, текстовий вміст. Текст зберігається «як є», без жодних модифікацій: усі внутрішні теги (<LSTag>, <i>, <b>,
), плейсхолдери підстановок ([1], [2]), маркери наративу й телепатичного мовлення (*текст*, ((*текст*))) зберігаються в незмінному вигляді. На цьому етапі будь-яка очисна обробка вважалася передчасною – вона проводиться у наступних модулях за чітко сформульованими принципами. Усього з файлу english.loca.xml витягнуто 232 876 записів, що становить весь основний обсяг англомовного текстового матеріалу гри.

Розпакування .pak-архівів і конвертація .loca-файлів виконуються одноразово і їхні результати зберігаються в директорії data/raw нашого проєкту. Подальші запуски Python-пайплайну вже не потребують доступу до встановленої копії гри – вся робота йде з XML-файлами, що повністю автономні. Це методологічно важливо: воно дозволяє відтворити будь-який етап дослідження виключно на основі збережених даних, без потреби в комерційному програмному забезпеченні.

4.3. Інтеграція до паралельного корпусу офіційної української локалізації

Українська локалізація Baldur's Gate 3 з'явилася в офіційному оновленні гри під назвою Patch 7, випущеному в вересні 2024 року. Локалізація виконана

перекладацькою спілкою «Шлякбитраф» (SBT Localization) у партнерстві з Larian Studios. Це порівняно рідкісний випадок для великих рольових ігор: на момент виходу гри 2023 року українська не входила до офіційно підтримуваних мов, і саме спільнота, до якої Larian відкрила доступ через офіційну партнерську програму, забезпечила появу повноцінного україномовного перекладу понад рік потому. Тексти української локалізації покривають усі основні текстові ресурси гри: діалоги, інтерфейс, описи предметів, наративні фрагменти, внутрішньоігрові книги й документи.

Технічно українська локалізація постачається з грою у форматі окремого .pak-архіву – Ukrainian.pak – що розташовується в тій самій директорії Data, що й англомовний пакет. Структурно архів повністю аналогічний до English.pak: ієрархія Localization/Ukrainian/ містить основний файл ukrainian.loca, гендерні варіанти ukrainian_M_to_X.loca та ukrainian_to_F.loca, а також підпапку Gender/. Розпакування та конвертація виконані тим самим інструментом BG3 Modder's Multitool, що й для англійської. У результаті отримано файл ukrainian.loca.xml зі структурою, ідентичною англомовному XML.

Ключовою методологічною перевагою такої організації даних є те, що ідентифікатори contentuid у файлах обох мов збігаються один до одного. Кожен запис, що має певний contentuid в англомовному файлі, має той самий contentuid в українському файлі – за умови, що цей запис було перекладено. Це означає, що вирівнювання паралельних пар не потребує жодних алгоритмів автоматичного зіставлення речень (sentence alignment), що зазвичай є нетривіальною задачею у побудові паралельних корпусів, – достатньо побудувати словник ідентифікатор-текст для кожної мови й об'єднати їх за ключем.

З українського XML витягнуто 218 232 записи – на 14 644 менше, ніж з англомовного. Ця різниця відповідає записам, що присутні у файлі English.pak, але відсутні у Ukrainian.pak (або мають порожній текстовий вміст у ньому, проте окрема перевірка показала, що порожніх записів у українському файлі немає – всі 14 644 розбіжності зумовлені саме відсутністю в українському XML). Це становить 6,3% від загального обсягу англомовного матеріалу – досить значущу частку, що показує неповноту української локалізації станом на досліджуваний

період. Зазначений підкорпус англомовних записів без українського відповідника зберігається окремо у файлі data/intermediate/en_only.jsonl і не включається в основний паралельний корпус, але становить потенційний інтерес для майбутніх досліджень – зокрема як матеріал для оцінки покриття автоматизованих перекладацьких систем або для тренування моделей машинного перекладу, що могли б закрити цю прогалину.

Вирівнювання здійснюється тривіальним алгоритмом: для англомовного й українського JSONL-файлів будуються словники, де ключем є contentuid, значенням – повний запис; потім будується третій словник, у якому для кожного contentuid, присутнього в обох вихідних словниках, формується паралельний запис з полями id, version, en, uk_human. Записи, присутні лише в англомовному словнику, виносяться в окремий файл. Загалом отримано 218 232 паралельну пару EN-UK (тобто покриття становить 93,7% від англомовного матеріалу) – це базовий обсяг корпусу до подальшої класифікації, фільтрації й дедуплікації, описаних у наступних підрозділах. Приклад такого вирівняного запису у форматі JSONL наведено в Додатку Б.

Окремої уваги заслуговує питання якості людського перекладу, що в нашому дослідженні виконує функцію еталона для оцінки колонки згенерованих текстів. Аналіз вибірки записів демонструє, що команда Шлякбитраф здебільшого дотримується принципу збереження ігрової стилістики й функціональної точності, а не суто буквальної відповідності оригіналу. Так, у прикладах з вибірки можна побачити, що англійська репліка *«Aye. Our arses.»* перекладена як *«Еге. Наших дуп.»* – зі збереженням просторічної стилістики, а звертання *«Sorry, darling»* – як *«Вибач, серденько»*, що передає природну українську пестливість, без буквалізму. Подекуди виявляються поодинокі дефекти, як-от один запис із групи телепатичного мовлення, де в українському варіанті втрачено відкривальний маркер «((*», або два записи, де після завершального маркера «*))» залишено технічний фрагмент «
». Такі випадки виокремлено в метаданих окремою позначкою uk_human_quality зі значенням marker_mismatch і виключено з обчислення метрик зіставлення для відповідних страт, що детально описано в підрозділі 4.6.

4.4. Генерація перекладу за допомогою штучного інтелекту – чат-боту Gemini

Колонка згенерованих текстів – третій складник кожного паралельного запису – формується шляхом автоматичного перекладу англомовного оригіналу за допомогою LLM.

Вибір конкретної моделі еволюціонував у ході роботи. Початково генерацію планувалося виконувати моделлю Gemini 2.5 Flash-Lite – на той момент це була найдоступніша щодо вартості модель родини 2.5, із прийнятним для дослідницьких цілей балансом якості та швидкості. Однак на практиці виявилося, що безкоштовний рівень доступу до цієї моделі обмежено двадцятьма запитами на добу, чого категорично недостатньо для генерації кількох тисяч перекладів у розумний термін. Тому модель було замінено на Gemini 3.1 Flash-Lite – представника наступного покоління, для якого на момент дослідження діяв значно щедріший безкоштовний ліміт: п'ятсот запитів на добу, п'ятнадцять запитів на хвилину, двісті п'ятдесят тисяч токенів на хвилину. Перехід на новіше покоління моделі не лише розв'язав проблему лімітів, а й, як показали порівняльні тести, покращив якість оброблення структурних елементів тексту (зокрема збереження маркера у квадратних дужках книжкових описів).

Остаточний вибір моделі – Gemini 3.1 Flash-Lite – методологічно обґрунтований не лише доступністю. Ця модель належить до класу «полегшених» (lite) моделей, орієнтованих на масове застосування з мінімальними витратами. Саме такі моделі найімовірніше використовуватимуть реальні споживачі автоматизованого перекладу з обмеженим бюджетом – незалежні локалізатори, дослідники, ентузіасти. Оцінка якості перекладу через модель цього класу, а не через найдорожчі флагманські системи, дає реалістичнішу картину того, чого можна очікувати від доступного машинного перекладу в ігровому домені на сучасному етапі. Це свідомий методологічний вибір на користь екологічної валідності дослідження.

Взаємодія з моделлю здійснюється через офіційний програмний інтерфейс Google AI Studio за допомогою клієнтської бібліотеки google-genai для Python.

Параметри генерації зафіксовано на значеннях, що забезпечують високу детермінованість виходу: температура (temperature) – 0,2, що знижує випадковість і робить переклад передбачуванішим; параметр ядрової вибірки (top_p) – 0,9; параметр обмеження кандидатів (top_k) – 40; максимальна кількість токенів виходу – 2048, із запасом для найдовших наративних і книжкових записів. Низька температура – принципове рішення: для дослідницької відтворюваності важливо, щоб повторний запуск на тих самих даних давав максимально близький результат, а не варіювався випадково.

Центральним елементом генерації є інструкція-промпт, що супроводжує кожен запит до моделі. Її проєктування виявилося ітеративним процесом: початкова версія промпта, хоча й давала прийнятні переклади загалом, демонструвала кілька систематичних відхилень, виявлених під час верифікаційного тестування на стратифікованій вибірці з десяти записів. Промпт пройшов три послідовні ітерації вдосконалення.

Перша редакція промпта містила загальні правила: перекладати з англійської українською, зберігати теги розмітки (<LSTag>, <i>, <b>,
) незмінними, зберігати рантайм-плейсхолдери ([1], [2]) і стилістичні маркери (*наратив*, ((*телепатія*))), не перекладати власні назви, використовувати неформальне звертання «ти» в діалогах. Тестування цієї редакції на моделі 3.1 Flash-Lite виявило два систематичні відхилення. По-перше, інструкція «не перекладати власні назви» інтерпретувалася моделлю буквально – імена персонажів і локацій залишалися латиницею (Halsin замість Гальсін), що суперечить практиці української локалізації, де власні назви транслітеруються кирилицею. По-друге, у телепатичному мовленні модель подекуди застосовувала формальне звертання «ви» замість інтимного «ти», характерного для цього регістру.

Друга редакція усунула обидва відхилення через уточнення формулювань. Замість розмитого «не перекладати власні назви» введено чітку вимогу транслітерувати власні назви українською кирилицею за

усталеними нормами фентезійного перекладу, з наведенням конкретних прикладів відповідностей (Halsin → Гальсін, Shadowheart → Тінесерда, Nightsong → Нічна Пісня). Окремо обумовлено виняток: іншомовні внутрішньоігрові терміни, позначені курсивом (як-от ґітьянська лексема zaith'isk), транслітеруються зі збереженням курсиву й апострофів. Для телепатичного мовлення додано явну вимогу використовувати звертання «ти». Повторне тестування на тих самих десяти записах підтвердило усунення обох відхилень без появи нових.

Прикметно, що одне відхилення свідомо залишено невиправленим – систематична семантична помилка в перекладі числових співвідношень. У механічних описах модель стабільно неправильно інтерпретувала конструкцію на кшталт *«increased by half»* чи *«halved»*, передаючи її як *«збільшується вдвічі»* замість *«зменшується наполовину»*. Ця помилка відтворювалася на обох поколіннях моделі (2.5 і 3.1 Flash-Lite), що вказує на її системний, а не випадковий характер – вона зумовлена обмеженою здатністю моделей цього класу до точного семантичного аналізу математичних відношень у мовному контексті. Виправлення цієї помилки на рівні промпта означало б штучне втручання в природну поведінку моделі й спотворило б валідаційну картину. Натомість збереження її дозволяє зафіксувати цей клас помилок як емпіричний результат дослідження (детальніше у підрозділі 4.6).

Генерація виконувалася в режимі «один запит – один запис», без об'єднання кількох текстів у пакетні запити (батчинг). Це рішення має методологічне підґрунтя. Батчинг прискорив би обробку й заощадив квоту, проте він порушив би незалежність окремих перекладів: модель, отримавши кілька текстів в одному запиті, схильна «усереднювати» стиль між ними, що могло б спотворити стилістичну специфіку окремих записів (особливо телепатичних із їхнім нестандартним регістром). Для забезпечення чистоти подальшої валідації

кожен запис перекладався окремим, незалежним запитом із ідентичною інструкцією. Це повільніше, але методологічно коректніше.

Технічна реалізація передбачала стійкість до збоїв і обмежень API. Оскільки добовий ліміт у п'ятсот запитів означав, що повна генерація розтягнеться на кілька діб, критичним було поступове збереження: кожен отриманий переклад одразу дописувався до проміжного файлу gemini_results.jsonl, ще до завершення всього процесу. Це гарантувало, що жоден збій – вичерпання денного ліміту, переривання з'єднання, перезапуск комп'ютера – не призведе до втрати вже виконаної роботи. При повторному запуску система зчитувала вже наявні в файлі ідентифікатори й пропускала їх, продовжуючи з місця зупинки. Завдяки цьому механізму процес фактично виконувався автономно: щойно денний ліміт вичерпувався й запити починали повертати помилку перевищення квоти, процес «засинав» у циклі очікування з експоненційними затримками, а наступної доби, коли ліміт оновлювався, автоматично відновлював роботу. Помилки оброблялися диференційовано: тимчасові (перевищення ліміту, серверні збої) спричиняли повтор із наростанням паузи, тоді як критичні (некоректний запит, проблема автентифікації) зупиняли процес із чітким повідомленням у журналі.

Окремої згадки заслуговує дрібний, але показовий технічний інцидент, пов'язаний із зберіганням ключа доступу до API. Ключ зберігався в окремому конфігураційному файлі .env, що не входить до системи контролю версій. При першому запуску система не могла знайти ключ, попри його наявність у файлі. Причиною виявилася позначка порядку байтів (BOM) на початку файлу, додана текстовим редактором при збереженні в кодуванні UTF-8: бібліотека читання конфігурації сприймала назву першого ключа як такий, що містить невидимий символ U+FEFF перед «GEMINI_API_KEY», унаслідок чого ключ не розпізнавався. Файл було перезбережено без BOM, після чого ключ зчитувався коректно. Цей епізод ілюструє типову складність роботи з кодуваннями текстових файлів у крос-платформному середовищі.

Обсяг генерації визначався стратифікованою вибіркою, описаною в підрозділі 4.5: 4 992 записи, розподілені на сім функціональних класах із

навмисним перепредставленням рідкісних класів (телепатичне мовлення, книжкові тексти, прив'язки клавіш) для забезпечення статистичної придатності їх окремого аналізу. Уся вибірка була успішно перекладена – 4 992 із 4 992 записів (стовідсоткове покриття) – за приблизно десять діб роботи в межах безкоштовного ліміту. Згенеровані переклади дописувалися до фінального файлу корпусу corpus.jsonl у поле uk_gemini, тоді як решта записів корпусу (поза вибіркою) зберегли це поле порожнім із відповідною позначкою в метаданих. Отриманий матеріал становить основу для валідаційного аналізу, що його викладено в наступному підрозділі. Приклад такого запису з трьома колонками наведено в Додатку Б

4.5. Загальна характеристика укладеного паралельного англо-українського корпусу текстів відеогри Baldur's Gate

Загальний обсяг матеріалу, витягнутого з англомовного файлу локалізації, становить 232 876 записів. Кожен запис відповідає одному елементу content у вихідному XML, що зазвичай становить одну смислову одиницю тексту – репліку діалогу, назву предмета, абзац опису, статтю інтерфейсу. З українського файлу витягнуто 218 232 записи, що на 14 644 менше, ніж з англомовного. Ця різниця в 6,3% відповідає текстам, які ще не отримали офіційного українського перекладу. Окрема перевірка показала, що в українському XML немає записів з порожнім вмістом для тих самих contentuid, що присутні в англомовному – тобто всі 14 644 розбіжності зумовлені саме фізичною відсутністю запису, а не порожнім перекладом. Цей підкорпус англомовних записів без українського відповідника зберігається окремо у файлі en_only.jsonl і не входить до основного паралельного корпусу.

Після вирівнювання за спільним ключем contentuid отримано 218 232 паралельних пар EN-UK – це базовий обсяг до подальшої обробки. На етапі фільтрації виконувалося очищення хвостових технічних артефактів (тегів br, зайвих пробілів) та видалення порожніх записів: виявлено й вилучено два записи з порожнім текстовим умістом, унаслідок чого залишилося 218 230 пар. Проте

основною операцією, що визначила фінальний обсяг, стала дедуплікація точних повторів – пар (en, uk_human), де обидві колонки повністю ідентичні. Знайдено й видалено 31 919 таких дублікатів (14,6% від базового обсягу), причому перший зразок кожного дубліката збережено, а інформація про частоту повторень кожного унікального запису зафіксована в метаданих у полі duplicate_count. Унаслідок цього кінцевий обсяг корпусу становить 218 230 − 31 919 = 186 311 унікальних паралельних пар. Структуру та приклади записів корпусу наведено в Додатках А та Б.

Аналіз топ-20 найчастіших дублікатів підтвердив природний характер цього явища: переважна більшість – короткі функціональні рядки, які гра використовує в багатьох контекстах. Лідером є рядок *«Leave.»* з 1 679 входженнями – стандартна репліка завершення розмови, що повторюється в діалогах з кожним другорядним персонажем. Серед інших найчастіших – *«Broken Barricade»* (250 входжень), *«Attack.»* (236), *«Who are you?»* (156), *«Tombstone»* (155). Жоден з топ-20 не перевищує 100 символів довжиною, що підтверджує функціональну природу дублікатів – це не повтори книжкових текстів чи унікальних реплік, а саме UI-елементи й шаблонні фрази системи. Серед наративних рядків з обгорткою «*…*» теж знайшлися повтори: чотири системних рядки, що показуються над кожним трупом або після завершення заклинання (наприклад, *«*The corpse regards you lifelessly.*»* – 134 входження), є нормальними технічними дублікатами без втрати інформаційної цінності.

Новий корпус – це повний набір унікальних EN-UK-записів, готовий для класифікації, генерації третьої колонки та валідації. Розподіл записів за функціональним типом тексту, отриманий через евристичну класифікацію, наведено в таблиці 4.1.

Таблиця 4.1. Розподіл записів корпусу за функціональним типом тексту

Тип тексту	Записів	% корпусу
dialogue (репліки діалогів)	150 837	81,0%
ui_short (короткі елементи інтерфейсу)	22 429	12,0%
narrative (наративний текст)	7 920	4,3%
mechanic_description (описи механік)	4 179	2,2%

book_or_document (книги, листи, документи)	668	0,4%
telepathic (телепатичне мовлення)	186	0,1%
ui_keybind (підказки прив'язок клавіш)	92	0,05%
Разом	186 311	100%

Виокремлення семи функціональних класів дозволяє трактувати корпус не як гомогенний масив текстів, а як структурований набір страт, кожна з яких має власні стилістичні особливості й, відповідно, потребуватиме окремого аналізу при валідації якості машинного перекладу. Особливо цінні рідкісні класи: 186 телепатичних реплік представляють унікальну стилістику з навмисною ламаною граматикою; 668 книжкових записів – основний матеріал для оцінки якості перекладу довгого художнього тексту; 92 keybind-записи – спеціалізований формат з технічними плейсхолдерами, що тестують здатність моделі зберігати незмінні шаблонні елементи.

Розподіл записів за довжиною англомовного тексту показує переважно фрагментарний характер ігрового матеріалу. До класу «short» (менше 100 символів) потрапляє 154 070 записів – 82,7% корпусу. Це здебільшого UI-елементи, короткі репліки, назви предметів. Клас «medium» (100–500 символів) налічує 31 380 записів – 16,8%, що відповідає типовим діалогам середньої довжини, описам механік. До «long» (500–2 000 символів) належить лише 853 записи – 0,5%; до «extreme» (понад 2 000 символів) – 8 записів, тобто 0,004%. Медіана довжини запису становить 50 символів, середнє арифметичне – 61,3 символи, дев'яносто п'ятий перцентиль (значення, нижче якого лежать 95% записів) – 140 символів. Це означає, що типовий запис ігрового корпусу значно коротший за типове речення стандартних паралельних корпусів (на кшталт OPUS чи Europarl), де переважають повноскладні речення з 15–30 слів. Корпус Baldur's Gate 3 – це передусім корпус коротких, контекстно-залежних висловлювань.

Розподіл за довжиною тексту має методологічне значення для подальшої валідації. Метрики автоматичної оцінки якості перекладу, що були розроблені й первинно валідовані на матеріалі довгих новинних і парламентських текстів, можуть поводитися інакше на коротких ігрових записах. Зокрема, BLEU оперує

n-грамами і чутлива до довжини; ChrF++ на рівні символів менш залежна від цього параметра. У підрозділі 4.6 ми побачимо, як саме цей розподіл впливає на агрегатні оцінки якості колонки згенерованих текстів. Усі записи зберегли свої тексти повністю – фільтрації за довжиною не застосовувалося; натомість кожен запис отримав мітку length_class у метаданих, що дозволяє в подальшому проводити аналіз за окремими стратами.

Окрему групу характеристик корпусу становлять метадані щодо внутрішньої розмітки записів. Чотири типи тегів супроводжуються відповідними булевими прапорами в metadata: has_lstag (наявність LSTag-посилань), has_italic (наявність i-тегів), has_bold (наявність b-тегів), has_br (наявність br-розривів рядка в середині тексту). За результатами обробки, тег i наявний у 22 246 записах (11,9% корпусу) – це найпоширеніший тег, що використовується для виокремлення курсивом цитат, іншомовних слів, навмисних емфаз у репліках. Тег br зустрічається у 2 219 записах (1,2%) – переважно у книжкових і довгих документах для розбивки на абзаци. Тег LSTag присутній у 4 215 записах (2,3%); саме цей тег несе функціональне навантаження – посилання на ігрові механіки, статуси, ресурси. Найрідкіснішим є тег b, що зустрічається лише у 742 записах (0,4%), переважно для виділення критично важливих слів у репліках.

Особливим явищем корпусу є асиметрія тегів між англомовним оригіналом і українським перекладом. Якщо припустити, що локалізатори при перекладі повністю зберігають структуру розмітки оригіналу (тобто кожен тег у EN має відповідник у UK з тими ж атрибутами), то таких розбіжностей не повинно бути. Однак аналіз показав, що 333 записи (0,2% корпусу) містять щонайменше одну тегову розбіжність. Розподіл цих записів за функціональними класами нерівномірний: найбільше їх у класі dialogue – 250 записів, що становить 0,2% від усіх діалогових пар; помітна частка у mechanic_description – 46 записів (1,1% класу) і book_or_document – 22 записи (3,3% класу, найвища відносна частота); решта припадає на ui_short (8) і narrative (7). У класах telepathic та ui_keybind асиметрії не виявлено зовсім. Підвищена відносна частота асиметрії саме в книжкових і механічних текстах закономірна: це найдовші й найбагатші на розмітку записи, де перекладач найчастіше адаптує абзацний поділ (теги br)

або скорочує опис, прибираючи теги-посилання LSTag, тоді як в українському варіанті курсивне виокремлення і подеколи замінюється маркером «зірочки» або вилучається як надлишкове.

Кожен запис з асиметрією тегів отримує в metadata булевий прапор tag_asymmetry зі значенням true, а також структуровану деталізацію tag_asymmetry_details, де для кожного типу тегу зберігаються пари «en_count, uk_count». Ця деталізація відкриває окреме поле досліджень – функціональні трансформації розмітки при професійному людському перекладі ігрового домену. У контексті нашої роботи цей факт має два методологічні наслідки: по-перше, він уточнює положення підрозділу 1.2 про дотримання локалізаторами принципу збереження структури розмітки (structure-preserving translation) – насправді цей принцип діє приблизно у 99,8% випадків, з невеликою, але стилістично значущою долею функціональних адаптацій; по-друге, при обчисленні метрик автоматичної оцінки якості (підрозділ 4.6) ми зможемо окремо проаналізувати, як себе поводять згенеровані тексти на записах з симетричною і асиметричною розміткою.

Окремої уваги заслуговують технічні артефакти у вихідних даних. Аналіз тегів виявив принаймні два типи одруківок у файлах локалізації, створених самою компанією Larian Studios. По-перше, тег br з крапкою замість слешу зустрічається в одному записі – очевидно, друкарська помилка при формуванні файлу. По-друге, атрибут «Tootip» (замість «Tooltip») зустрічається у 9 записах, а «Tooltips» – в 1 записі. Ці артефакти зберігаються в корпусі «як є», без виправлення – вони відображають реальний стан вихідних даних і можуть бути предметом окремих досліджень якості внутрішніх локалізаційних процесів великих ігрових студій. Виявлення подібних артефактів є додатковим побічним результатом методології, що демонструє цінність систематичного аналізу для розкриття раніше непомічених проблем у відкритих даних.

Окрема метаінформація стосується якості людського перекладу. Аналіз 186 телепатичних записів виявив один випадок з порушенням маркера – у записіу з ідентифікатором hc487b8fe український варіант не містить відкривального символу телепатичного маркера, що порушує стилістичну цілісність

телепатичного коду. Цей запис зафіксовано в metadata з полем uk_human_quality зі значенням marker_mismatch. Окрім нього, ще два записи мають технічний хвіст br після завершального маркера – це не дефект перекладу, а технічний артефакт, проте формально вирівнювання маркерів між EN і UK у них теж порушено. Усі інші 183 телепатичні записи мають коректний маркер і отримали значення clean. Загалом частка дефектних записів становить близько 1,6% від класу telepathic, що свідчить про високу якість локалізації, виконаної студією Шлякбитраф.

Окремо варто розглянути ще один аспект якості людського перекладу – стилістичну нормалізацію. У підрозділі 1.1 ми згадували навмисну ламану граматику телепатичного мовлення в англомовному оригіналі – відсутність заголовних літер на початку речень, тенденцію до фрагментарного, телеграфного синтаксису. Аналіз вибірки телепатичних записів показав, що українські перекладачі систематично нормалізують цей стиль, повертаючи заголовні літери й типову українську пунктуаційну структуру. Так, англомовний оригінал *«go - rest at your camp. my servant is not yet ready»* у людському перекладі набирає вигляду *«Іди, відпочинь у таборі. Моєму слузі ще потрібен час.»* – з великих літер, нормальним розділовим знаком, типовим українським синтаксисом. Це не дефект перекладу, а свідоме стилістичне рішення локалізаційної команди: пріоритет читабельності українського тексту над буквальним відтворенням стилістичного коду оригіналу. У контексті нашої валідації це означає, що людський еталон уже містить певне «стилістичне коригування», і Gemini, потенційно навчена на нормативних текстах, може давати ще ближчий до людського перекладу результат саме на телепатичних записах – за рахунок паралельної нормалізації.

На рівні плейсхолдерів і шаблонних елементів корпус містить кілька структурно значущих груп. Класичні рантайм-плейсхолдери у формі квадратнодужкового числового маркера зустрічаються у 2 550 записах (1,1% корпусу). Це шаблони для підстановки числових значень, імен, кількостей, що формуються динамічно під час виконання гри. Технічні плейсхолдери прив'язок клавіш мають три префікси: [IE_*] (Input Event) у 92 записах, [GLO_*] (Global) у

63 записах, [GEN_*] (Generic) у 12 записах. Усі ці записи зібрані в окремому класі ui_keybind. Для цих типів плейсхолдерів критично важливе збереження точної структури при перекладі – будь-яке порушення формату призведе до некоректного відображення в грі. Це робить ui_keybind важливою тестовою стратою для оцінки здатності машинного перекладача зберігати незмінні шаблонні елементи.

Підкорпус англомовних записів без українського перекладу (en_only.jsonl) налічує 14 644 записи. Аналіз їх класового розподілу і стилістичного характеру виходить за межі основної мети цього дослідження, але сам факт існування й обсяг цього підкорпусу є важливим методологічним результатом. Він кількісно характеризує неповноту української локалізації Baldur's Gate 3 станом на момент виходу Patch 7. Зазначений підкорпус потенційно цінний для майбутніх досліджень – як матеріал для оцінки покриття автоматизованих перекладацьких систем, як тестова множина для систем-передбачень, що могли б закрити цю прогалину автоматизовано, або як база для подальшої роботи україномовних локалізаторів.

Підсумовуючи опис, можна стверджувати, що отриманий корпус є структурованим, збагаченим метаданими паралельним ресурсом обсягом 186 311 пар, що покриває сім функціональних класів ігрового тексту, чотири основні типи внутрішньої розмітки, кілька систем плейсхолдерів і явища тегової асиметрії, які раніше не описувалися систематично для англо-українських ігрових корпусів. Структура корпусу дозволяє проводити багатоаспектний аналіз – як зведений по всьому масиву, так і за окремими функціональними чи довжинними стратегіями, з урахуванням якості людського перекладу й технічних особливостей вихідних даних. Репрезентативні приклади записів корпусу різних функціональних класів – з усіма трьома колонками й метаданими – наведено в **Додатку Б**. Це створює базу для коректної валідації якості згенерованої колонки, опис якої міститься у наступному підрозділі.

4.6. Валідація: оцінка якості згенерованого перекладу

Перед обчисленням метрик (BLEU, METEOR, TER, ChrF++, BERTScore) з обох колонок (згенерованої та еталонної) видалялися теги розмітки – <LSTag>, <i>, <b>,
. Це принципове методологічне рішення: оскільки в підрозділі 4.5 встановлено, що теги загалом зберігаються однаково в обох колонках, їх присутність у тексті при обчисленні метрик лише штучно завищувала б збіг за рахунок ідентичних службових елементів, а не за рахунок якості власне перекладу. Натомість рантайм-плейсхолдери ([1], [2]) та стилістичні маркери (*наратив*, ((*телепатія*))) зберігалися, оскільки вони є значущою частиною тексту. Для метрики BERTScore використано багатомовну модель xlm-roberta-base, що підтримує українську мову, – застосування англоцентричної моделі дало б некоректні результати для україномовного тексту.

Зведені результати валідації по всій вибірці наведено у таблиці 4.2.

Таблиця 4.2. Зведені дані метрик якості згенерованого перекладу (n = 4 990)

Метрика	Значення	Шкала
BLEU	23,16	0–100, вище краще
METEOR	49,57	0–100, вище краще
TER	73,16	0+, нижче краще
ChrF++	46,75	0–100, вище краще
BERTScore F1	91,93	0–100, вище краще
BERTScore P / R	91,77 / 92,09	0–100, вище краще

Уже на рівні зведених показників проступає центральний результат усього дослідження – разюча розбіжність між поверхневими лексичними метриками та семантичною метрикою. BLEU, що оцінює буквальний збіг послідовностей слів (n-грам), показує лише 23,16 зі 100 – значення, яке за традиційною шкалою інтерпретації машинного перекладу відповідало б «зрозумілому, але неякісному» перекладу. Натомість BERTScore, що оцінює семантичну близькість на рівні контекстних векторних представлень, показує 91,93 – значення, яке свідчить про майже повну смислову еквівалентність. Така розбіжність не є суперечністю чи помилкою вимірювання – вона є ключовим змістовним висновком: згенерований переклад передає той самий зміст, що й людський, але робить це іншими словами, з іншим порядком слів, з іншим добором синонімів і синтаксичних конструкцій.

Цей результат має пряме методологічне значення в контексті всієї роботи. На етапі попередніх досліджень, що передували цій роботі, було виявлено обмеження автоматичних метрик при оцінюванні якості машинного перекладу ігрових текстів – проблема, що й мотивувала створення повноцінного паралельного корпусу. Тепер, на масштабі 4 990 записів, ця проблема отримала систематичне емпіричне підтвердження: BLEU і BERTScore дають фактично протилежну оцінку того самого матеріалу. Метрики, побудовані на буквальному збігу слів (BLEU, ChrF++, частково TER), систематично недооцінюють якість машинного перекладу художнього й діалогічного тексту, де природна варіативність формулювань є нормою, а не дефектом. Метрики, побудовані на семантичній близькості (BERTScore), відображають реальну якість значно адекватніше. Це не означає, що BLEU непридатна взагалі – вона залишається корисною для оцінки технічно регламентованих текстів, де варіативність небажана, – але для ігрового домену з його стилістичним розмаїттям опора виключно на лексичні метрики методологічно хибна.

Аналіз метрик у розрізі функціональних класів виявляє чітку й змістовну закономірність. Результати наведено в таблиці 4.3.

Таблиця 4.3. Дані метрик якості згенерованого перекладу за функціональними класами (за спаданням BLEU)

Тип тексту	**n**	**BLEU**	**METEOR**	**TER**	**ChrF++**	**BERTScore**
ui_keybind (підказки прив'язок клавіш)	92	49,81	66,36	48,96	64,53	94,11
telepathic	98	39,72	73,76	67,11	49,45	93,46
narrative (наративний текст)	800	23,81	56,05	73,59	46,28	92,75
dialogue (репліки діалогів)	2 500	23,26	50,42	73,25	45,89	92,15
ui_short (короткі елементи інтерфейсу)	700	22,76	36,96	60,79	47,29	90,70
book_or_document (книги, листи, документи)	200	21,61	46,64	72,66	48,53	91,41

mechanic_description (описи механік)	600	19,50	46,58	77,89	44,23	90,92

Найвищі показники демонструє клас ui_keybind (підказки прив'язок клавіш): BLEU 49,81, що вдвічі перевищує середній по корпусу. Пояснення просте: ці записи здебільшого складаються з технічних плейсхолдерів і коротких стандартних формулювань ([IE_ToggleInventory] Open Inventory), де простір для варіативності перекладу мінімальний, а модель фактично копіює структуру. Це очікуваний результат, що підтверджує методологічну валідність метрик на регламентованому матеріалі.

Найнижчий BLEU має клас mechanic_description (19,50) – описи ігрових механік. Це теж пояснювано, але причина протилежна. Механічні описи насичені усталеною ігровою термінологією, для якої українська локалізація має конкретні відповідники, що часто не збігаються з вибором моделі. Так, англійський термін Disadvantage офіційний переклад передає як «завада», тоді як Gemini обирає «Недолік» або «Перешкода»; Proficient – «спеціалізація» проти «володіє»; Hidden – «скрадаєтеся» проти «Прихований». Жоден із цих варіантів не є помилкою – вони стилістично й семантично прийнятні, – але вони не збігаються з еталоном, і BLEU за це штрафує. Окрему роль відіграє систематична капіталізація: модель послідовно пише ігрові терміни з великої літери («Опір», «Перевірки», «Характеристику»), наслідуючи англійську традицію, тоді як українська локалізація використовує малу. До того ж саме в цьому класі зосереджені фактичні семантичні помилки моделі – згадана в підрозділі 4.4 інверсія числових співвідношень (*«halved»* передається як *«збільшується вдвічі»*), що становить реальний дефект перекладу, а не стилістичну розбіжність.

Варто уточнити, що до вибірки відібрано 100 телепатичних записів, проте в обчисленні метрик задіяно 98: два записи з дефектами телепатичних маркерів (позначені в метаданих як marker_mismatch, див. підрозділ 4.5) виключено, щоб показники відбивали якість згенерованого перекладу, а не похибки еталона. Особливої уваги заслуговує несподівано високий результат класу telepathic (BLEU 39,72 – другий за величиною, METEOR 73,76 – найвищий по корпусу).

На етапі проєктування промпта існувало побоювання, що модель нормалізує навмисно «ламану» граматику телепатичного мовлення й тим самим віддалиться від оригінального стилю. Натомість результат виявився протилежним і пояснюється тонким спостереженням, зафіксованим ще в підрозділі 4.5: українські локалізатори самі систематично нормалізують стиль телепатичних реплік, повертаючи заголовні літери й типову пунктуацію. Отже, і людський перекладач, і машинна модель застосовують до цього регістру схожу стратегію нормалізації – обидва «вирівнюють» ламаний оригінал, – унаслідок чого їхні результати зближуються. Це показовий випадок, коли висока метрика зумовлена не стільки якістю моделі, скільки збігом перекладацьких стратегій людини й машини.

Ілюстративним є приклад телепатичної репліки flesh-walker, tongue-talker. you hail from the Dark – yet far you've come, де обидва перекладачі вдалися до креативного словотвору для передачі складених прізвиськ: людський варіант – *«плоть ходить, язик мовить. походиш ти з Пітьми, та блукаєш далеко»*, згенерований – *«плотеходцю, мовомовцю. ти родом із Темряви – але ж далеко ти зайшов»*. Обидва варіанти стилістично винахідливі, обидва передають той самий образ, але роблять це різними словотвірними засобами – що знову ілюструє розрив між семантичною еквівалентністю й лексичним збігом.

Найнижчі показники серед діалогічно-наративних класів має ui_short за метрикою METEOR (36,96) – попри прийнятний BLEU. Причина в тому, що короткі інтерфейсні елементи (назви предметів, кнопки) часто є однослівними, і METEOR, яка враховує точність зіставлення основ слів і синонімію, на таких коротких рядках стає нестабільною: один незбіг різко знижує показник. Це методологічне нагадування, що на дуже коротких текстах (медіана довжини запису в корпусі – лише 50 символів) усі метрики поводяться менш надійно, ніж на повноскладних реченнях, для яких їх первісно розробляли.

Окремий аналіз записів із асиметрією тегів підтверджує гіпотезу, висунуту в підрозділі 4.5 (повні показники наведено в **Додатку В**). Для 4 975 записів із симетричною розміткою BLEU становить 23,20, тоді як для 15 записів з асиметрією – лише 13,63; TER зростає з 73,09 до 91,70, BERTScore знижується з

91,93 до 89,15. Тобто записи, де людський перекладач свідомо переструктурував розмітку або сам текст, є найважчими для машинного відтворення. Показовий приклад – репліка Dust. On. My. Tongue!, яку офіційний переклад повністю переосмислив як *«Повний рот... піску!»* (передавши прагматику ситуації, а не буквальний зміст), тоді як модель переклала дослівно: *«Пил. На. Моєму. Язиці!»*. Тут людина-локалізатор застосувала глибоку прагматичну адаптацію, на яку машинна модель середнього класу не здатна, – і саме такі випадки формують найбільший розрив між машинним і людським перекладом. Подібно, у механічному описі з умовою On a successful save, the target still takes half damage офіційний переклад радикально скоротив текст до *«Завдає 3к4 отрутної шкоди за хід»* (через обмеження простору ігрового інтерфейсу), тоді як модель переклала повний опис дослівно – приклад, де розбіжність зумовлена не якістю, а різними завданнями людського й машинного перекладу.

Узагальнюючи кількісний і якісний аналіз, можна сформулювати кілька висновків щодо придатності великих мовних моделей середнього класу для перекладу ігрових текстів українською мовою. По-перше, на рівні передавання змісту модель Gemini 3.1 Flash-Lite демонструє високу якість (BERTScore близько 92) – згенерований переклад майже завжди семантично еквівалентний людському. По-друге, на рівні точного відтворення формулювань модель значно відрізняється від професійного перекладу (BLEU близько 23), що зумовлено природною лексичною й синтаксичною варіативністю, а не помилками. По-третє, виявлено три класи реальних обмежень моделі: систематичні семантичні помилки в числових співвідношеннях механічних описів; нездатність до глибокої прагматичної адаптації, доступної людському локалізаторові; розбіжності в усталеній ігровій термінології та капіталізації. По-четверте, ефективність моделі суттєво залежить від типу тексту: вона найкраща на регламентованих коротких рядках і, парадоксально, на телепатичному мовленні (через збіг стратегій нормалізації), і найслабша на технічно насичених механічних описах. Ці висновки окреслюють реалістичну межу застосовності доступних великих мовних моделей у спеціалізованому ігровому перекладі: вони придатні для чорнового перекладу й передавання змісту, але потребують

людського постредагування там, де критичні точність термінології, числових даних і прагматична адаптація.

4.7. Експертна оцінка якості перекладу

Кількісні метрики, розглянуті в підрозділі 4.6, дають суперечливу картину: BLEU близько 23 балів свідчить про низьку поверхневу подібність, тоді як BERTScore близько 92 балів вказує на високу семантичну еквівалентність згенерованого й людського перекладів. Така розбіжність є симптомом обмеженості самих метрик, проте, спираючись лише на них, неможливо встановити, який показник ближчий до реальної якості: усі вони є машинними оцінками, і одна з них не може слугувати незалежним арбітром для іншої. Щоб розірвати це коло, у цьому підрозділі застосовано зовнішній орієнтир – сліпе експертне оцінювання, у якому якість перекладу визначає людина-оцінювач, а не алгоритм.

Методологія. З вибірки у 4 990 пар відібрано 70 записів, розподілених стратифіковано за сімома функціональними класами пропорційно до їхньої частки в корпусі, з мінімальним представленням рідкісних класів для забезпечення їх присутності в оцінці. Для кожного запису оцінювачеві надано оригінал і два переклади – людський (uk_human) і машинний (uk_gemini) – у вигляді знеособлених «Варіанта А» і «Варіанта В». Порядок варіантів визначався випадково для кожного запису й не розкривався під час оцінювання, що унеможливлювало упередженість на користь того чи іншого джерела. Кожен варіант оцінювався за двома незалежними шкалами від 1 до 10: адекватність (повнота й точність передавання змісту оригіналу) і флюентність (природність і граматична правильність української мови незалежно від оригіналу). Оцінювання проводив один оцінювач (рівень володіння англійською мовою – C1; обізнаний з дискурсом відеоігор); це обмеження структури окремо обговорюється нижче. Повний протокол оцінювання – зведені таблиці результатів і представницькі записи з оцінками, приклади – наведено в **Додатку Ґ**.

Структура опитувальника була такою: для кожного з 70 записів оцінювач бачив англійський оригінал і два знеособлені варіанти українського перекладу, позначені нейтрально (наприклад, «Варіант А» і «Варіант В»), причому відповідність варіанта людському чи згенерованому перекладу та порядок їх подання рандомізувалися задля сліпого режиму. Кожен варіант оцінювався окремо за двома шкалами від 1 до 10 – адекватність (наскільки точно передано зміст оригіналу) і флюентність (наскільки природним є українське формулювання); додатково оцінювач міг залишити короткий коментар.

Результати. Усупереч очікуванню, що офіційна професійна локалізація отримає вищі оцінки, у сліпому режимі вищі бали систематично діставав машинний переклад. Зведені середні наведено в таблиці 4.4.

Таблиця 4.4. Середні експертні бали (шкала 1–10, n = 70)

Вісь	**Людський переклад**	**Gemini**	**Різниця (Л−G)**
Адекватність	6,76	8,60	−1,84
Флюентність	8,27	8,79	−0,51
Загальна (Суб'єктивна)	7,51	8,69	−1,18

Перевага машинного перекладу за адекватністю проявилася у 54 записах із 70 і є статистично значущою: критерій знакових рангів Вілкоксона дає $p < 0{,}001$ для адекватності та $p \approx 0{,}009$ для флюентності, парний t-критерій підтверджує цей результат. Тобто спостережена різниця не є випадковою. Розбивку за функціональними класами наведено в таблиці 4.5.

Таблиця 4.5. Середні бали адекватності та BERTScore за класами

Клас	**n**	**Адекв. (Л)**	**Адекв. (G)**	**BERTScore**
dialogue (репліки діалогів)	24	7,3	9,0	92,2
narrative (наративний текст)	10	7,4	8,5	93,3
ui_short (короткі елементи інтерфейсу)	8	6,1	9,8	88,6
mechanic_description	8	6,2	8,8	90,4
book_or_document (книги, листи, документи)	7	6,1	8,6	92,7
telepathic (телепатичне мовлення)	7	6,7	8,9	94,2

ui_keybind (підказки прив'язок клавіш)	6	5,7	5,3	96,4

Інтерпретація: розбіжність перекладацьких стратегій. Перевагу машинного перекладу за адекватністю не слід тлумачити як свідчення вищої якості Gemini порівняно з людським перекладом – таке узагальнення суперечило б і обсягу вибірки, і самій природі виявленого ефекту. Аналіз записів, де людський переклад отримав низькі бали за адекватністю, показує закономірність: це переважно випадки творчої адаптації. Так, у книжковому записі офіційний переклад відходить від оригіналу заради збереження каламбуру й колориту назв; у репліках персонажів людина-локалізатор переосмислює фразу заради природного звучання, жертвуючи буквальною відповідністю джерелу. Офіційна локалізація Baldur's Gate 3 є транскреацією – свідомим відступом від оригіналу заради функціональної та емоційної дії на гравця. Оскільки адекватність у цьому оцінюванні визначалася як відповідність вихідному англійському тексту, буквальний машинний переклад за таким критерієм отримує перевагу майже за визначенням, тоді як творча адаптація «штрафується» за відхилення, яке насправді є її метою. Отже, виявлений ефект говорить не про те, що машина перекладає краще за людину, а про те, що результат оцінювання вирішально залежить від того, як операціоналізовано поняття «адекватність», і що буквальна вірність джерелу й функціональна якість локалізації – різні, подеколи протилежні величини.

Головний результат: невідповідність метрик експертній оцінці. Ключова цінність експертного оцінювання полягає не в порівнянні людини й машини, а в тому, що воно слугує зовнішнім орієнтиром для перевірки самих автоматичних метрик – і ця перевірка виявляє їхню системну ваду. Експертні оцінки адекватності машинного перекладу практично не корелюють із BERTScore (коефіцієнт Пірсона $r \approx -0{,}24$), тобто метрика, яку в підрозділі 4.6 визнано найнадійнішою, насправді не відстежує те, що оцінювач сприймає як якість. Ще показовішим є зв'язок розриву оцінок із BERTScore: що вищий BERTScore запису, то більший розрив на користь машинного перекладу (рангова кореляція Спірмена $\rho \approx 0{,}40$, $p < 0{,}001$). Усі наведені

коефіцієнти кореляції (Пірсона та Спірмена) обчислено автором роботи за допомогою стандартних статистичних функцій на основі отриманих експертних оцінок. Усі коефіцієнти кореляції та критерії статистичної значущості (Пірсона, Спірмена, Вілкоксона) обчислено програмно (скриптом), а не вручну. Іншими словами, висока метрика семантичної подібності трапляється саме там, де людський і машинний переклади якісно найбільше розходяться, – метрика не розрізняє два різних за стратегією, але обидва прийнятних переклади.

Найвиразніше обмеженість метрик ілюструє клас ui_keybind. Це єдиний клас, де машинний переклад поступився людському за адекватністю (5,3 проти 5,7), – і водночас саме він має найвищий середній BERTScore з усіх класів (96,4). Причина розходження в тому, що в кількох записах цього класу модель пошкодила службову розмітку рядка, замінивши функціональний атрибут тегу й зробивши рядок технічно некоректним. Для гри такий рядок є непрацездатним, проте BERTScore, який вимірює семантичну схожість тексту після відсікання тегів, цієї поломки «не бачить» і присвоює запису високий бал. Це наочний випадок, коли метрика винагороджує функціонально зламаний переклад, – пряме емпіричне підтвердження центральної тези роботи: автоматичні метрики не вловлюють низку аспектів якості, критичних для ігрової локалізації, і потребують доповнення експертною оцінкою.

Обмеження. Отримані результати слід інтерпретувати з урахуванням обмежень дослідження. По-перше, оцінювання проводив один оцінювач, тож оцінки відображають індивідуальне судження й не дають змоги обчислити узгодженість між анотаторами; залучення кількох незалежних оцінювачів із підрахунком коефіцієнта каппа є очевидним напрямом подальшої роботи. По-друге, обсяг вибірки (70 записів) достатній для виявлення статистично значущих тенденцій, але замалий для надійних висновків на рівні окремих рідкісних класів. По-третє, дослідження ґрунтується на локалізації однієї гри з виразно транскреативним стилем перекладу, тому виявлена розбіжність стратегій може бути менш вираженою для текстів з вимогою буквальної точності. З огляду на ці обмеження висновки підрозділу формулюються не як твердження про перевагу

машинного перекладу, а як діагностика валідності автоматичних метрик, для якої наявних даних достатньо.

4.8. Поліфункціональність перспектив використання паралельного корпусу текстів відеоігор

Хоча первинне призначення корпусу – валідація якості згенерованої колонки, отриманої через модель Gemini 3.1 Flash-Lite (підрозділ 4.6), архітектура й структура корпусу свідомо розроблені так, щоб ресурс міг служити різноаспектним інструментом для подальших досліджень у галузях машинного перекладу, корпусної лінгвістики й перекладознавства відеоігор. Перспективи використання можна згрупувати в три основні напрями: технологічні, лінгвістично-дескриптивні, освітні.

Перший напрям – технологічне застосування – стосується безпосереднього використання корпусу для покращення систем машинного перекладу в ігровому домені. Найочевидніший варіант – донавчання (fine-tuning) відкритих LLM на парах (en, uk_human). Сучасні відкриті моделі на кшталт Mistral 7B, LLaMA 3, Gemma 2 та їхні похідні демонструють значне покращення якості після донавчання на доменно-специфічних даних, особливо коли загальні моделі недостатньо орієнтовані в специфічній термінології чи стилістиці. Для англо-українського ігрового напряму такий ресурс є особливо цінним, оскільки сучасні комерційні моделі (GPT-4, Claude, Gemini) тренуються переважно на новинних, технічних, наукових та літературних текстах, де ігровий контент представлений нерівномірно. Зокрема, специфіка перекладу імен персонажів, локацій, ігрових механік, стилістичних кодів (телепатичне мовлення, наративні марковані фрагменти) у нашому корпусі представлена систематично і з метаданими, що робить можливим створення доменно-адаптованих моделей.

Другий технологічний сценарій – тренування моделей автоматичного постредагування (Automatic Post-Editing, APE). На відміну від класичних МП-моделей, APE-системи приймають на вхід пару (вихідний текст, чорновий машинний переклад) і навчаються виправляти чорновий переклад до якості людського еталона. Наявність трьох колонок у нашому корпусі – оригіналу,

людського перекладу й згенерованого – створює природний навчальний набір для такого типу систем. Зокрема, пари (uk_gemini, uk_human) можна використати безпосередньо як вхід і цільовий вихід для моделі постредагування, тоді як паралельний англомовний оригінал служить додатковим контекстом. Це особливо актуально для практичної локалізаційної індустрії, де редагування чорнового машинного перекладу є реальним робочим процесом, що поєднує продуктивність автоматизації зі стилістичною точністю людської роботи.

Третій технологічний напрям – використання корпусу для бенчмаркінгу й порівняння різних систем машинного перекладу. У нашій валідації (підрозділ 4.6) ми оцінюємо одну модель – Gemini Flash-Lite – як представника LLM середнього класу. Корпус, однак, придатний для аналогічної оцінки будь-якої іншої системи: достатньо згенерувати ще одну згенеровану колонку через альтернативну модель (DeepL, GPT-4, Claude, спеціалізовані NMT-системи) і порівняти метрики за тими самими стратами. Це дозволить будувати порівняльні таблиці, що показують відносні переваги різних систем у конкретних типах текстів (UI, діалоги, наратив, телепатичні репліки). Така інфраструктура корпусу прокладає шлях до систематичних бенчмарків для українського ігрового домену, яких на момент проведення нашого дослідження не існувало.

Другий напрям перспективного використання – лінгвістично-дескриптивний. Корпус, збагачений метаданими про тип тексту, наявність розмітки, асиметрії тегів, дефекти маркерів, є цінним матеріалом для досліджень самого феномену відеоігрового перекладу – без прямого зв'язку з машинним перекладом. Зокрема, на матеріалі корпусу можна досліджувати такі питання: яким чином українські локалізатори обирають між форенізацією й доместикацією при перекладі ігрових власних назв; як змінюється стилістична палітра української мови залежно від функціонального типу тексту (діалог, наратив, інтерфейс); якими прагматичними засобами передаються нестандартні мовні коди (телепатичне мовлення, ламана граматика, архаїзми) при перекладі. Кожне з цих питань може стати темою окремого перекладознавчого або соціолінгвістичного дослідження, і наявність структурованого корпусу істотно знижує бар'єр входу для таких робіт.

Окремим лінгвістично-дескриптивним напрямом є дослідження стилістичних трансформацій, виконуваних професійними локалізаторами при перекладі. У підрозділі 4.5 ми згадували виявлений факт систематичної нормалізації стилістики телепатичного мовлення в українських варіантах – повернення заголовних літер, типової пунктуаційної структури, повних граматичних форм. Це лише один з багатьох можливих об'єктів дослідження. Аналіз асиметрій тегів (333 записи з tag_asymmetry=true) відкриває можливості для опису функціональних трансформацій розмітки. Дослідження дефектів маркерів і опечаток у вихідних даних (як в українському, так і в англомовному файлах) формує самостійний предмет – оцінку якості локалізаційних процесів великих ігрових студій.

Особливо цінним для майбутніх досліджень є підкорпус англомовних записів без українського відповідника (en_only.jsonl, 14 644 записи). Він є кількісним показником поточного стану локалізаційного покриття BG3 – 6,3% контенту не перекладено офіційно станом на момент дослідження. Цей підкорпус може використовуватись як тестова множина для систем автоматичного передбачення українських перекладів, які б могли закрити прогалину; як матеріал для майбутніх ітерацій офіційної локалізації командою Шлякбитраф; як предмет соціолінгвістичного аналізу – який саме контент залишається без перекладу (можливо, новіший контент пізніших патчів, або, навпаки, технічні чи маргінальні рядки). Систематичний аналіз цього підкорпусу виходить за межі нашої роботи, але самий факт його ідентифікації й обсягу формує підставу для майбутніх досліджень.

Третій напрям – освітній. Корпус може використовуватись як навчальний матеріал у курсах перекладознавства, прикладної лінгвістики, корпусних методів. Поєднання в одному ресурсі трьох колонок – оригіналу, професійного людського перекладу, згенерованого машинного перекладу – дозволяє будувати практичні завдання різного рівня складності. Студенти можуть аналізувати, як саме професійний перекладач трансформує оригінал; порівнювати людський і машинний переклад за обраними критеріями; самостійно оцінювати якість обох варіантів за обраними метриками; досліджувати конкретні стилістичні питання

(форенізація/доместикація на матеріалі імен, прагматика в діалогах, передача гумору і культурних алюзій). Метадані щодо типу тексту і наявності розмітки спрощують відбір ілюстративних прикладів для конкретних навчальних задач.

Окрім трьох основних напрямів, корпус може мати застосування у суміжних практичних галузях. Локалізаційні студії можуть використовувати його для оцінки відповідності претендентів на ролі перекладачів – даючи кандидатам тестові завдання, відібрані з конкретних страт корпусу, і порівнюючи їхні результати з еталоном Шлякбитрафа. Розробники інструментів автоматичного контролю якості перекладу можуть валідувати свої системи на нашому корпусі. Дослідники прагматики, культурної адаптації, термінотворення в українській мові можуть знаходити в корпусі ілюстративний матеріал для своїх теоретичних побудов.

Окремо слід згадати обмеження, які звужують перспективи використання. По-перше, корпус створено в межах конкретної гри – Baldur's Gate 3, RPG у жанрі темного фентезі. Жанрова специфіка позначається на стилістичному профілі корпусу – переважання діалогів і фентезійної термінології. Перенесення моделей, тренованих на цьому корпусі, на інші ігрові жанри (шутери, симулятори, головоломки) може супроводжуватися деградацією якості – це слід враховувати при практичному використанні. По-друге, корпус створено для академічних цілей у межах кваліфікаційної бакалаврської роботи і не передбачає публічного поширення через захист авторських прав Larian Studios. Це обмежує його доступність для широкої наукової спільноти, хоча методологія побудови, описана в розділі 3, переноситься на будь-яку гру, локалізатори якої погодяться на створення подібного відкритого ресурсу. По-третє, згенерована колонка отримана через одну конкретну модель (Gemini Flash-Lite) і відображає її поведінку станом на момент дослідження. Швидкий розвиток LLM-технологій означає, що через рік-два моделі того ж класу можуть демонструвати істотно інакшу якість – тому корпус як ресурс для оцінки сучасних МП-систем потребуватиме оновлення.

Підсумовуючи, можна стверджувати, що створений корпус становить ресурс, цінність якого виходить далеко за межі завдань цього кваліфікаційного

дослідження. Технологічні, лінгвістично-дескриптивні й освітні напрями використання, окреслені вище, формують програму потенційних досліджень, кожне з яких може дати самостійний науковий чи практичний результат. Обмеження, природно, теж існують – як специфічні для матеріалу, так і обумовлені правовим режимом доступу. Однак методологічна основа, розроблена в межах цієї роботи, є відтворюваною й здатною масштабуватися на інші ігри, мовні пари, моделі машинного перекладу. Це робить нашу роботу не лише локальним дослідженням однієї конкретної гри, а й внеском у розбудову методологічної інфраструктури для досліджень ігрової локалізації в українському мовознавстві.

Висновки до розділу 4

У цьому розділі викладено практичну реалізацію методології, розробленої в розділі 3, та результати валідації створеного корпусу. Програмну систему побудовано як модульний пайплайн із дев'яти послідовних етапів – від інвентаризації розмітки й витягування текстів з ігрових файлів до генерації перекладу й обчислення метрик якості. Така архітектура забезпечила відтворюваність кожного етапу й стійкість до збоїв, критичну при тривалій генерації через API з обмеженим доступом.

Сформований корпус містить 186 311 унікальних паралельних пар «оригінал – людський переклад», структурованих за сімома функціональними класами й збагачених метаданими щодо довжини, розмітки, тегової асиметрії та якості перекладу. Для 4 992 записів стратифікованої вибірки додано третю колонку – згенерований переклад, згенерований моделлю Gemini 3.1 Flash-Lite. Окремо виокремлено підкорпус із 14 644 англомовних записів, що не мають офіційного українського відповідника, – кількісний показник неповноти локалізації, придатний для майбутніх досліджень.

Валідація згенерованої колонки за п'ятьма метриками виявила центральний результат дослідження: разючу розбіжність між лексичними метриками (BLEU 23,16) та семантичною метрикою (BERTScore 91,93). Ця розбіжність емпірично, на масштабі майже п'яти тисяч записів, підтверджує тезу

про неадекватність метрик буквального збігу при оцінюванні художнього й діалогічного ігрового перекладу – тезу, що мотивувала створення корпусу. Згенерований переклад виявився семантично еквівалентним людському, але лексично й синтаксично відмінним, що є наслідком природної варіативності, а не дефектом.

Аналіз за класами окреслив реалістичну межу застосовності великих мовних моделей середнього класу в ігровому перекладі: модель найефективніша на регламентованих коротких рядках і телепатичному мовленні (через збіг стратегій нормалізації з людським перекладачем) і найслабша на технічно насичених механічних описах, де виявлено систематичні семантичні помилки (інверсія числових співвідношень) та розбіжності в усталеній термінології. Записи з асиметрією розмітки, де людський локалізатор вдавався до глибокої прагматичної адаптації, виявилися найважчими для машинного відтворення. Загальний висновок практичної частини: доступні великі мовні моделі придатні для чорнового перекладу й передавання змісту, але потребують людського постредагування там, де критичні точність термінології, числових даних і прагматична адаптація.

Нарешті, сліпе експертне оцінювання 70 записів (підрозділ 4.7) дало корпусу зовнішній, не–машинний орієнтир і підтвердило центральну тезу роботи з боку, недосяжного для самих метрик: експертні оцінки якості машинного перекладу практично не корелюють із BERTScore ($r \approx -0{,}24$), а в класі ui_keybind метрика присвоїла найвищі бали записам із функціонально пошкодженою розміткою. Тобто автоматичні метрики подекуди винагороджують зламаний переклад – найпряміше емпіричне свідчення їхньої обмеженості. Виявлена в оцінюванні перевага машинного перекладу за формальною адекватністю-до-джерела інтерпретується не як вищість моделі над людиною, а як наслідок транскреативної природи офіційної локалізації, що свідомо відходить від буквального тексту; з огляду на

однооцінювачну структуру і обсяг вибірки цей результат сформульовано саме як діагностику валідності метрик.

ВИСНОВКИ

Метою цього дослідження було створення паралельного корпусу для машинного перекладу на основі текстів відеоігор та оцінюванню на його основі якості згенерованого перекладу, отриманого через велику мовну модель Gemini 3.1 Flash-Lite.

У межах першого завдання проаналізовано теоретичні засади перекладу відеоігор, сучасний стан машинного перекладу та метрик оцінки його якості. Встановлено, що локалізація відеоігор є окремим складним різновидом перекладацької діяльності, у якому поєднуються ознаки художнього, технічного й аудіовізуального перекладу, а провідною стратегією виступає скопос-орієнтована адаптація, спрямована на збереження ігрового досвіду, а не буквальної відповідності. Простежено еволюцію машинного перекладу від систем на правилах через статистичні системи до нейронних і великих мовних моделей. Розглянуто п'ять метрик автоматичної оцінки якості (BLEU, METEOR, TER, ChrF++, BERTScore) та окреслено їхні теоретичні обмеження, які згодом підтвердилися емпірично.

У межах другого завдання розроблено оригінальну методологію побудови паралельного корпусу з ігрових локалізаційних файлів і реалізовано програмну систему її втілення. Методологія охоплює витягування текстів із власних бінарних форматів гри, вирівнювання паралельних пар за спільними ідентифікаторами, евристичну класифікацію за функціональними типами, багаторівневу фільтрацію й дедуплікацію, а також збагачення записів метаданими. Програмна система реалізована як модульний пайплайн мовою Python.

У межах третього завдання сформовано корпус, що систематизує текстові дані за трьома колонкамии: англомовний оригінал, офіційний український переклад студії Шлякбитраф і згенерований переклад моделі Gemini 3.1 Flash-Lite. Кінцевий обсяг корпусу становить 186 311 унікальних паралельних пар, з яких 4 992 записи стратифікованої вибірки мають усі три колонки. Корпус

структуровано за сімома функціональними класами й забезпечено детальними метаданими, що робить його придатним для багатоаспектного аналізу. Це перший, наскільки відомо авторові, систематично задокументований паралельний англо-український корпус ігрового домену такого обсягу й структурованості.

У межах четвертого завдання оцінено якість згенерованих перекладів у порівнянні з людським еталоном за допомогою автоматичних метрик і якісного аналізу. Головним результатом стала разюча розбіжність між лексичними метриками (BLEU 23,16) і семантичною оцінкою (BERTScore 91,93), що емпірично підтвердила центральну тезу дослідження: метрики буквального збігу неадекватно оцінюють якість художнього й діалогічного ігрового перекладу, систематично занижуючи її через природну варіативність формулювань. Якісний аналіз за класами виявив, що велика мовна модель середнього класу передає зміст із високою точністю, але має реальні обмеження в точності ігрової термінології, числових співвідношень і прагматичній адаптації, доступній професійному локалізаторові.

Наукова новизна роботи полягає в трьох аспектах. По-перше, розроблено й задокументовано відтворювану методологію побудови паралельного корпусу з локалізаційних файлів відеоігор, що переноситься на інші ігри, мовні пари й перекладацькі системи. По-друге, створено сам корпус як інфраструктурний ресурс, цінність якого виходить за межі цього дослідження. По-третє, отримано емпіричне підтвердження неадекватності лексичних метрик в ігровому домені на репрезентативному масштабі – результат, що має значення для методології оцінювання машинного перекладу загалом.

Практичне значення роботи визначається можливістю використання корпусу для донавчання відкритих моделей машинного перекладу, тренування систем автоматичного постредагування та бенчмаркінгу перекладацьких систем у спеціалізованому ігровому домені, а також як навчального й дослідницького матеріалу в перекладознавстві та прикладній лінгвістиці. Перспективи подальших досліджень охоплюють розширення корпусу на інші ігри й мовні пари, порівняльне оцінювання кількох моделей машинного перекладу на

спільній основі, а також залучення експертної оцінки якості людьми-перекладачами для зіставлення з автоматичними метриками.

Підсумовуючи, проведене дослідження не лише досягло поставленої мети, а й створило інфраструктурну основу – паралельний корпус і відтворювану методологію його побудови – для подальшого розвитку досліджень машинного перекладу ігрових текстів в українському мовознавстві.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Корунець, І.В. (2003). Теорія і практика перекладу (аспектний переклад): підручник. Вінниця: Нова Книга. 448 с.
2. Шлякбитраф (SBT Localization) (2024). Українська локалізація Baldur's Gate 3 у партнерстві з Larian Studios. Available at: https://sbt.localization.com.ua/ (Accessed: 20 April 2026).
3. Arnold, D., Balkan, L., Meijer, S., Humphreys, R.L., & Sadler, L. (1994). Machine Translation: An Introductory Guide. London: NCC Blackwell.
4. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473. Available at: https://arxiv.org/abs/1409.0473 (Accessed: 22 April 2026).
5. Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72.
6. Berger, A.L., Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Gillett, J.R., Lafferty, J.D., Mercer, R.L., Printz, H., & Ureš, L. (1994). The Candide System for Machine Translation. In: Proceedings of the ARPA Workshop on Human Language Technology, pp. 157–162.
7. Bernal-Merino, M.Á. (2014). Translation and Localisation in Video Games: Making Entertainment Software Global. New York: Routledge.
8. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., & Mercer, R.L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2), pp. 263–311.
9. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078. Available at: https://arxiv.org/abs/1406.1078 (Accessed: 22 April 2026).

10. Costales, A.F. (2012). Exploring Translation Strategies in Video Game Localisation. MonTI: Monografías de Traducción e Interpretación, 4, pp. 385–408.
11. Drugan, J. (2013). Quality in Professional Translation: Assessment and Improvement. London: Bloomsbury.
12. Halliday, M.A.K., & Hasan, R. (1976). Cohesion in English. London: Longman.
13. Hansen, D., Bhattacharyya, A., & Roturier, J. (2022). A Snapshot into the Possibility of Video Game Machine Translation. In: Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (AMTA), Vol. 2, pp. 257–269.
14. Hatim, B., & Munday, J. (2004). Translation: An Advanced Resource Book. London: Routledge.
15. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), pp. 1735–1780.
16. Hutchins, J. (2005). The History of Machine Translation in a Nutshell. Available at: http://www.hutchinsweb.me.uk/Nutshell-2005.pdf (Accessed: 15 April 2026).
17. Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proceedings of MT Summit X, pp. 79–86.
18. Koehn, P. (2009). Statistical Machine Translation. Cambridge: Cambridge University Press.
19. Mangiron, C., & O'Hagan, M. (2006). Game Localisation: Unleashing Imagination with 'Restricted' Translation. The Journal of Specialised Translation, 6, pp. 10–21.
20. McEnery, T., & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press.
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems 26.
22. Munday, J. (2016). Introducing Translation Studies: Theories and Applications. 4th edition. London: Routledge.
23. Nirenburg, S. (Ed.) (1993). Progress in Machine Translation. Amsterdam: IOS Press.

24. Nugroho, A.B. (2017). Video Game Localization: A Case Study of the Translation of Bully. Journal of English Language and Culture, 7(2), pp. 33–44.
25. O'Hagan, M., & Mangiron, C. (2013). Game Localization: Translating for the Global Digital Entertainment Industry. Amsterdam/Philadelphia: John Benjamins Publishing Company.
26. Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318.
27. Popović, M. (2015). chrF: Character n-gram F-score for Automatic MT Evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 392–395.
28. Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In: Proceedings of the Third Conference on Machine Translation, pp. 186–191.
29. Pym, A. (2010). Exploring Translation Theories. Abingdon and New York: Routledge.
30. Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.
31. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pp. 223–231.
32. Sutskever, I., Vinyals, O., & Le, Q.V. (2014). Sequence to Sequence Learning with Neural Networks. In: Advances in Neural Information Processing Systems 27, pp. 3104–3112.
33. Šiaučiūnė, V., & Liubinienė, V. (2011). Video Game Localization: The Analysis of In-Game Texts. Kalbų studijos / Studies About Languages, 19, pp. 46–55.
34. Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, pp. 2214–2218.

35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. In: Advances in Neural Information Processing Systems 30, pp. 5998–6008.
36. Venuti, L. (1995). The Translator's Invisibility: A History of Translation. London: Routledge.
37. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. arXiv preprint arXiv:1904.09675. Available at: https://arxiv.org/abs/1904.09675 (Accessed: 25 April 2026).
38. Google AI Studio (2026). Gemini API Documentation. Available at: https://ai.google.dev/gemini-api/docs (Accessed: 18 May 2026).
39. Larian Studios (2023). Baldur's Gate 3. Microsoft Windows, Mac, PlayStation 5, Xbox Series X/S. Available at: https://baldursgate3.game/ (Accessed: 10 April 2026).
40. Norbyte (2024). LSLib: A Library for Working with Larian Studios Games. GitHub Repository. Available at: https://github.com/Norbyte/lslib (Accessed: 12 April 2026).
41. Shinyhobo (2024). BG3 Modder's Multitool. GitHub Repository. Available at: https://github.com/ShinyHobo/BG3-Modders-Multitool (Accessed: 12 April 2026).

ДОДАТКИ

Додаток А. Технічний опис програмного пайплайну

Програмний пайплайн побудови корпусу реалізовано мовою Python (версія 3.11) як послідовність дев'яти автономних модулів. Кожен модуль приймає на вхід результат попереднього етапу у форматі JSONL і видає новий JSONL-файл, що забезпечує відтворюваність і можливість перезапуску з будь-якого етапу. Нижче наведено стислий опис архітектури, модулів і форматів даних; повний вихідний код із детальними сигнатурами функцій доступний у відкритому репозиторії (див. Додаток Г).

А.1. Послідовність обробки

Потік даних має такий вигляд: вихідні XML-файли локалізації → витягування (extract) → вирівнювання пар (align) → класифікація за типами (classify) → фільтрація й дедуплікація (filter) → генерація перекладу (gemini) → збірка фінального корпусу (build) → обчислення статистики (stats) і валідація (validate). Окремий нульовий етап (scout_tags) виконує одноразову розвідку структури XML перед побудовою основного конвеєра.

А.2. Опис модулів

scout_tags.py – одноразова розвідка структури XML-файлу. Інвентаризує всі теги розмітки всередині елементів <content>, підраховує їх входження, збирає атрибути й приклади контексту. Використовує ітеративний парсинг (lxml.etree.iterparse) із негайним звільненням оброблених елементів, що дозволяє обробити файл обсягом близько 150 МБ без завантаження всього дерева в пам'ять. Час виконання – близько 25–40 секунд.

extract.py – парсинг XML-файлів локалізації Larian і збереження у JSONL. Для кожного елемента <content> витягує ідентифікатор (contentuid), версію й текст. Текст зберігається «як є» – з тегами, переносами рядків і пробілами, щоб

не втратити семантично значущу розмітку. На виході – 232 876 англомовних записів і 218 232 українських.

align.py – вирівнювання паралельних пар за спільним ідентифікатором contentuid. Записи, наявні лише в англомовному файлі, виокремлюються в окремий підкорпус (14 644 записи). Результат – 218 232 вирівняні пари.

classify.py – призначення одного з семи функціональних типів кожному запису на основі евристик за англомовним текстом. Класифікація саме за оригіналом забезпечує незалежність від якості людського перекладу. Сім класів: dialogue, ui_short, narrative, mechanic_description, book_or_document, telepathic, ui_keybind.

filter.py – очищення й збагачення корпусу перед перекладом. Виконує чотири послідовні операції: відсікання хвостових тегів, видалення порожніх записів, дедуплікацію за парою (оригінал, людський переклад) і обрахунок метаданих (клас довжини, наявність тегів, асиметрія розмітки). Дедуплікація саме за парою, а не лише за оригіналом, свідомо зберігає випадки, коли той самий англійський рядок має різні українські переклади в різних контекстах. Результат – 186 311 унікальних пар.

gemini.py – автоматичний переклад стратифікованої вибірки через Google Gemini API. Підтримує відновлення роботи між сесіями (resume), диференційовану обробку помилок та інкрементальне збереження після кожного запиту. Деталі моделі й параметрів описано в підрозділі 4.4.

build.py – збірка фінального корпусу: об'єднання людських перекладів із згенерованими для записів вибірки, формування підсумкового файлу corpus.jsonl на 186 311 пар.

stats.py – обчислення описової статистики корпусу: розподіли за класами, довжиною, наявністю тегів, асиметрією розмітки. Результат зберігається у форматі Markdown.

validate.py – обчислення п'яти метрик якості (BLEU, METEOR, TER, ChrF++, BERTScore) для пари «згенерований переклад – людський еталон». Деталі методології обчислення описано в підрозділі 4.6.

А.3. Технічний стек

Основні бібліотеки: lxml (ітеративний парсинг XML), google-genai (офіційний SDK Google Gemini), tenacity (експоненційні затримки при повторних запитах), nltk (токенізація й обчислення METEOR), sacrebleu (BLEU, TER, ChrF++), bert-score (BERTScore на основі багатомовної моделі xlm-roberta-base), torch (залежність bert-score). Слід зауважити, що стара бібліотека google-generativeai застаріла з 2026 року; у роботі використано новий несумісний з нею SDK google-genai. Окрім того, у версії sacrebleu 2.x функцію обчислення METEOR вилучено, тому METEOR обчислюється засобами NLTK.

А.4. Формат запису корпусу

Кожен запис корпусу – це JSON-об'єкт із трьома текстовими полями (en – оригінал, uk_human – офіційний український переклад, uk_gemini – згенерований переклад) і блоком метаданих (тип тексту, клас довжини, ознаки наявності розмітки й тегової асиметрії, належність до перекладеної вибірки). Записи поза вибіркою мають порожнє поле згенерованого перекладу з відповідною позначкою в метаданих.

Додаток Б. Приклади записів корпусу

У цьому додатку наведено представницькі приклади записів корпусу, згруповані за функціональними класами. Кожен приклад містить англійський оригінал (EN), офіційний український переклад (UK-люд.) і згенерований переклад моделі Gemini (UK-Gemini), а також показник семантичної близькості BERTScore F1. Приклади дібрано так, щоб проілюструвати як випадки високої відповідності, так і характерні розбіжності, проаналізовані в підрозділі 4.6.

Б.1. Діалоги (dialogue)

Перші два приклади демонструють повний збіг перекладів (BERTScore 100). Третій ілюструє показовий артефакт: нерозпізнаний технічний ідентифікатор озвучення (CombatVocals...), який людський переклад залишив незмінним, а модель помилково сприйняла як змістовний рядок і «переклала» вигуком.

EN: `The drow was a cruel witch, but a clever one.`
UK-люд.: `Дроу була жорстокою відьмою, але розумною.`

UK-Gemini: Дроу була жорстокою відьмою, але розумною.
BERTScore F1: 100.0

EN: Nature's might is always ready.
UK-люд.: Сила природи завжди напоготові.
UK-Gemini: Сила природи завжди напоготові.
BERTScore F1: 100.0

EN: CombatVocals_Generic_Dwarf_Female_01
UK-люд.: CombatVocals_Generic_Dwarf_Female_01
UK-Gemini: За Азгард!
BERTScore F1: 78.26

Б.2. Телепатичне мовлення (telepathic)

Цей клас демонструє конвергенцію перекладацьких стратегій: і людина, і модель удаються до креативного словотвору для передання складених прізвиськ. Зверніть увагу на різні словотвірні рішення для flesh-walker, tongue-talker.

EN: ((*flesh-walker, tongue-talker. you hail from the Dark - yet far you've come*))
UK-люд.: ((*плоть ходить, язик мовить. походиш ти з Пітьми, та блукаєш далеко*))
UK-Gemini: ((*плотеходцю, мовомовцю. ти родом із Темряви - але ж далеко ти зайшов*))
BERTScore F1: 87.26

EN: ((*cut the blight away. watch it wither*))
UK-люд.: ((*відріжте гниль. хай чахне*))
UK-Gemini: ((*виріж гниль. дивись як вона в'яне*))
BERTScore F1: 87.95

EN: ((*no peace*))
UK-люд.: ((*миру немає*))
UK-Gemini: ((*жодного спокою*))
BERTScore F1: 87.78

Б.3. Наратив (narrative)

Наративні ремарки в зірочках; обидва приклади показують повний збіг, зокрема й у передаванні звуконаслідування (KRST → КРСТ).

EN: *The flames sputter away. The sword is complete.*
UK-люд.: *Полум'я згасає. Меч готовий.*
UK-Gemini: *Полум'я згасає. Меч готовий.*
BERTScore F1: 100.0

EN: *KRST!*
UK-люд.: *КРСТ!*

UK-Gemini: *КРСТ!*
BERTScore F1: 100.0

Б.4. Описи механік (mechanic_description)

Найскладніший клас. Розбіжності зумовлені різним вибором усталеної термінології (Proficient → «спеціалізація» / «володіє»; Checks → «перевірок» / «Перевірок») та систематичною капіталізацією ігрових термінів у згенерованому перекладі.

EN: Not <LSTag Tooltip="Proficiency">Proficient</LSTag> with [1]
UK-люд.: Немає <LSTag Tooltip="Proficiency">спеціалізації</LSTag> ([1])
UK-Gemini: Не <LSTag Tooltip="Proficiency">володіє</LSTag> [1]
BERTScore F1: 80.76

EN: to Wisdom <LSTag Tooltip="AbilityCheck">Checks</LSTag>
UK-люд.: до <LSTag Tooltip="AbilityCheck">перевірок</LSTag> мудрості
UK-Gemini: до Перевірок <LSTag Tooltip="AbilityCheck">Мудрості</LSTag>
BERTScore F1: 83.03

Б.5. Записи з асиметрією розмітки (tag_asymmetry)

Записи, де людський перекладач радикально переструктурував текст. У першому прикладі офіційний переклад скоротив повний опис механіки до стислого ефекту (через обмеження простору інтерфейсу), тоді як модель переклала дослівно. Це найважчі для машинного відтворення випадки.

EN: Deals 1d4 Poison damage per turn. Creatures must succeed a Constitution Saving Throw or take an additional 2d4 Poison damage. On a successful save, the target still takes half damage.
UK-люд.: Завдає 3к4 отрутної шкоди за хід.
UK-Gemini: Завдає 1d4 шкоди отрутою щоходу. Істоти повинні успішно пройти Рятувальний кидок на Витривалість, інакше отримають додаткові 2d4 шкоди отрутою. У разі успішного рятувального кидка ціль все одно отримує половину шкоди.
BERTScore F1: 84.83

EN: Gain Resistance to all damage.
UK-люд.: Дає опір усій шкоді. Можна використовувати лише 1 еліксир за раз.
UK-Gemini: Отримайте Опір до всієї шкоди.
BERTScore F1: 84.56

Б.6. Короткі інтерфейсні рядки (ui_short)

Короткі інтерфейсні рядки демонструють кілька типових патернів: тривіальні позначки збігаються повністю, а назви предметів і кольорів виявляють розбіжність між буквальним перекладом моделі та транскреативними чи варіативними рішеннями офіційної локалізації.

EN: `Red 3`
UK-люд.: `Червоний 3`
UK-Gemini: `Червоний 3`
BERTScore F1: 100

EN: `Moon Tone 8`
UK-люд.: `Бузковий відтінок 8`
UK-Gemini: `Місячний тон 8`
BERTScore F1: 86.28

EN: `Newborn Mind Flayer`
UK-люд.: `Новонароджений мізкожер`
UK-Gemini: `Новонароджений ілітид`
BERTScore F1: 87.42

Б.7. Книги й документи (book_or_document)

Книжкові й документальні тексти належать до стилістично найскладніших записів корпусу. Наведені приклади ілюструють відмінності у відтворенні друкарських умовностей (тип лапок) і власних назв між офіційною локалізацією та згенерованим перекладом.

EN: `[An elegiac fugue in the key of D minor, entitled 'Ode to the Oppressed.']`
UK-люд.: `[Елегійна фуга у тональності ре-мінор під назвою "Ода пригнобленим".]`
UK-Gemini: `[Елегійна фуга в тональності ре мінор під назвою «Ода пригнобленим».]`
BERTScore F1: 97.36

EN: `[The Guild's dossier on a group of unidentified travellers arriving in Baldur's Gate from the Risen Road.]`
UK-люд.: `[Досьє "Гільдії" на невідомий гурт мандрівників, що прямує Високим трактом до Брами Балдура.]`
UK-Gemini: `[Досьє Гільдії щодо групи невпізнаних мандрівників, які прибули до Балдурової Брами з Піднятого Шляху.]`
BERTScore F1: 91.1

Додаток В. Метрики залежно від симетрії розмітки

Основні таблиці метрик валідації – зведені показники (Таблиця 4.2) і метрики за функціональними класами (Таблиця 4.3) – наведено в підрозділі 4.6 основного тексту. У цьому додатку подано додатковий розріз, який не дублюється в тілі роботи: порівняння метрик для записів із симетричною та

асиметричною теговою розміткою. Методологію обчислення описано в підрозділі 4.6: теги розмітки відсікалися перед обчисленням, плейсхолдери й стилістичні маркери зберігалися, BERTScore обчислювався на основі багатомовної моделі xlm-roberta-base. Обчислення виконано на 4 990 парах стратифікованої вибірки (після виключення двох записів із дефектами маркерів).

Таблиця В.1. Метрики залежно від симетрії тегової розмітки (n = 4 990)

Група	n	BLEU	METEOR	TER	ChrF++	BERTScore
Симетрична розмітка	4 975	23,20	49,61	73,09	46,77	91,93
Асиметрична розмітка	15	13,63	36,85	91,70	40,45	89,15

Додаток Г. Доступ до програмного коду

Повний вихідний код програмного пайплайну (дев'ять модулів мовою Python, інструкція-промпт для моделі Gemini, допоміжні скрипти формування вибірки) разом із технічною документацією доступний у відкритому репозиторії за адресою: https://github.com/ZavarOvek/crispy_invention. Код супроводжується файлом технічного опису з детальними сигнатурами функцій кожного модуля, описом форматів проміжних даних і послідовності запуску, а також невеликими ілюстративними фрагментами даних, достатніми для розуміння структури, але недостатніми для відтворення повного корпусу.

Дані корпусу – оригінальні англомовні тексти та офіційний український переклад – не публікуються у відкритому доступі, оскільки є інтелектуальною власністю компанії Larian Studios. Вони використовуються виключно в межах академічного добросовісного використання (fair use) для цієї кваліфікаційної роботи. Ілюстративні фрагменти корпусу, наведені в Додатку Б, обмежено обсягом, що відповідає принципам цитування в наукових цілях. Повний масив корпусу надається членам екзаменаційної комісії окремо за посиланням: https://drive.google.com/drive/folders/1PHGYLNj4pqAisXkIUog3smQXU3fxoZIZ?usp=sharing (поза відкритим доступом).

Програмний код є оригінальним доробком автора роботи й розміщений у відкритому репозиторії, посилання на який наведено вище; він доступний науковому керівникові, рецензентові та членам екзаменаційної комісії.

Додаток Ґ. Протокол експертного оцінювання якості перекладу

У цьому додатку наведено зведені результати сліпого експертного оцінювання, описаного в підрозділі 4.7, та вибрані показові записи з протоколу. Повний протокол (усі 70 записів із оцінками за двома шкалами, мапінгом знеособлених варіантів на джерело перекладу та коментарями оцінювача) зберігається у файлі 4_7_human_eval.xlsx у репозиторії програмного коду проєкту (див. Додаток Г).

Таблиця Ґ.1. Зведені показники оцінювання (n = 70)

Показник	Людський	Gemini
Середня адекватність	6,76	8,60
Середня флюентність	8,27	8,79
Загальна оцінка	7,51	8,69
Записів з вищою адекватністю	16	54

Таблиця Ґ.2. Статистична значущість і кореляція з BERTScore

Показник	Значення
Адекватність: критерій Вілкоксона, p	< 0,001
Флюентність: критерій Вілкоксона, p	≈ 0,009
Адекватність Gemini × BERTScore (Пірсон r)	−0,24
Розрив оцінок (Л−G) × BERTScore (Спірмен ρ)	0,40 (p < 0,001)

Нижче наведено десять показових записів із протоколу, що ілюструють основні висновки підрозділу 4.7: транскреативний характер офіційної локалізації (записи 39, 70, 26), нечутливість метрики BERTScore до функціональних дефектів розмітки (запис 20), помилки моделі в іменах власних і термінології (записи 11, 57) поряд із випадками її термінологічної переваги (запис 66), нормалізацію телепатичного регістру (запис 30) та рівноцінні переклади (записи 22, 56). Позначення: А – адекватність, Ф – флюентність (шкала 1–10); Л – людський переклад, G – Gemini.

Таблиця Ґ.3. Показові записи з протоколу оцінювання

№ / клас	Оригінал та переклади	Оцінки (Л / G)	BERT	Коментар
39 book_or_document	EN: [Below is a transcript of an interview with the writer and director of 'A Pleasurable Deal', Mr Kingsley Harp] Interviewer: What was the inspiration… Л: [Це запис розмови з автором та режисером п'єси «Солодка справа» паном Кінгслі Гарпом] Репортер: Що надихнуло вас на створення такої, якщо дозволите,… G: [Нижче наведено стенограму інтерв'ю з автором і режисером «Приємної угоди»,	А: 4/8 Ф: 8/9	90.8	книжк.: А назва/нюанси точні, літературно. В 'Солодка справа', 'Баляси Балдура', 'порноп'єса', змінена кінцівка – суттєві помилки.

	паном Кінгслі Гарпом] Інтерв'юер: Що стало натхненням для цієї, якщо доз…			
70 dialogue	EN: We often gaze through a veil of roses, but love accepts both the petal and the thorn. Gale - what is his greatest flaw? Л: Кохання мов троянда, що рясніє не лише цвітом, а й шипами. І їх ми часто помічаємо не відразу. Яка найбільша вада Ґейла? G: Ми часто дивимося крізь завісу з троянд, але кохання приймає і пелюстку, і шип. Ґейл – у чому його найбільша вада?	A: 4/10 Ф: 8/10	89.2	A: точно передано метафору 'veil of roses', структуру, пряме звертання. B зовсім інша метафора, переказ а не переклад.
26 dialogue	EN: I wouldn't expect any volunteers either. Л: Не думаю, що хтось захоче дати скуштувати. G: Я б теж не очікував жодних добровольців.	A: 2/10 Ф: 7/9	85.8	A: точно, збережено 'either'. B: зовсім інший зміст ('дати скуштувати'), не переклад.
20 ui_keybind	EN: [IE_PanelSelect] Open Game Menu <LSTag Type="Image" Info="TutorialCharacterSheet"/> Open Character Sheet Select a book to read it. Л: [IE_PanelSelect] – відкрити ігрове меню <LSTag Type="Image" Info="TutorialCharacterSheet"/> – відкрити аркуш персонажа Виберіть книжку, яку хочете … G: [IE_PanelSelect] Відкрити ігрове меню <LSTag Tooltip="TutorialCharacterSheet"/> Відкрити аркуш персонажа Виберіть книгу, щоб прочитати її.	A: 6/2 Ф: 7/6	96.2	ui_keybind: A додав зайві тире, але тег правильний. B КРИТИЧНО зіпсував тег (Type/Info -> Tooltip), рядок некоректний.
11 book_or_document	EN: [A record written in Raphael's own hand, detailing the adventures of various travellers in Faerûn, including yours.] The desperate will do anything … Л: [Нотатки Рафаеля, в яких він описує пригоди різних мандрівників Фейруну, включно з вами.] Керується відчаєм і зробить усе, щоб вийти зі скрутного ст… G: [Запис, зроблений власноруч Рафаїлом, що детально описує пригоди різних мандрівників у Фаеруні, включно з вашими.] Відчайдухи підуть на все, аби виб…	A: 6/7 Ф: 6/9	90.4	A: втрачено підмет, вільне тлумачення, неприродні конструкції. B: точніше, але 'Рафаїл'/'Фаерун' – помилки в іменах; дуже плавно.
66 mechanic_description	EN: Add your <LSTag Tooltip="ProficiencyBonus">Proficiency Bonus</LSTag> to any <LSTag Tooltip="Intimidation">Intimidation</LSTag> roll. Л: Додає <LSTag Tooltip="ProficiencyBonus">бонус спеціалізації</LSTag> до всіх кидків для навички «<LSTag Tooltip="Intimidation">Залякування</LSTag>» . G: Додайте свій <LSTag Tooltip="ProficiencyBonus">Бонус майстерності</LSTag> до будь-якого кидка на <LSTag Tooltip="Intimidation">Залякування</LSTag>.	A: 6/10 Ф: 9/10	88.5	ui: A 'Бонус майстерності' стандартний D&D5e термін, точно. B 'бонус спеціалізації' застарілий/неправильний, 'для навички' зайве.
57 narrative	EN: *Minthara is supposed to be in a cell - you have some explaining to do.* Л: *Мінтара має бути в камері. Доведеться вам це пояснити.* G: *Мінтара мала бути в камері – вам доведеться пояснити, що сталося.*	A: 9/7 Ф: 9/8	94.9	A: 'має бути' (теп.час) правильно. B: 'мала бути' (мин.) змінює сенс, 'пояснити що сталося' зайва конкретика.
30	EN: ((*we continue our communion*)) Л: ((*продовжимо наше спілкування*))	A: 7/10 Ф: 9/10	95.4	telepathic: A 'продовжимо'

telepathic	G: ((*ми продовжуємо наше спілкування*))			(майб./спонук.) замість теперішнього. В 'ми продовжуємо' точний теп. час.
22 dialogue	EN: Where is Dolor now? Л: Де Долор зараз? G: Де зараз Долор?	А: 10/10 Ф: 10/10	97.0	Обидва правильні; порядок слів не впливає.
56 ui_short	EN: Newborn Mind Flayer Л: Новонароджений мізкожер G: Новонароджений ілітид	А: 10/10 Ф: 10/10	87.4	Обидва правильні: 'ілітид' книжне / 'мізкожер' розмовне; у BG3 прийнятні обидва.