

КОРПУСНІ ТЕХНОЛОГІЇ В ЗАБЕЗПЕЧЕННІ ЯКОСТІ АНТРОПОГЕННОГО ТА ПОСТРЕДАГОВАНОГО МАШИННОГО ПЕРЕКЛАДУ ПРИ ЛОКАЛІЗАЦІЇ АУДІОВІЗУАЛЬНИХ ПРОЄКТІВ

CORPUS TECHNOLOGIES FOR ENSURING THE QUALITY OF HUMAN AND POST-EDITED MACHINE TRANSLATION IN AUDIOVISUAL LOCALIZATION

Бондаренко К.Л.,

orcid.org/0000-0002-2964-5123

кандидат філологічних наук,

доцент кафедри перекладу, прикладної та загальної лінгвістики

Центральноукраїнського державного університету імені Володимира Винниченка

Бондаренко О.С.,

orcid.org/0000-0002-3148-8753

кандидат філологічних наук,

доцент кафедри перекладу, прикладної та загальної лінгвістики

Центральноукраїнського державного університету імені Володимира Винниченка

Статтю присвячено вивченню можливостей використання корпусних технологій для забезпечення якості антропогенного перекладу та постредагованого машинного перекладу при локалізації аудіовізуальних продуктів (АВП). Для досягнення поставленої мети було сформульовано та реалізовано низку дослідницьких завдань: створення паралельного корпусу англійських та українських субтитрів; автоматичне видобування ключових лексем та оцінка їхньої частотності й значущості в межах вихідного тексту; аналіз та типологізація перекладацьких відповідників; виявлення загальних стратегій передачі частотної лексики в локалізованому продукті; формулювання практичних рекомендацій для перекладачів та постредакторів, які працюють з подібними типами контенту (за жанром, стилістикою, цільовою аудиторією тощо). Емпіричним матеріалом слугували субтитри англійською та українською мовами до телевізійного серіалу "Brassic" (5 сезон, серії 1–7), загальним обсягом 40 940 слів для англійської та 28 549 слів для української версії. Застосування дистрибутивно-статистичного аналізу до двомовного паралельного корпусу, сформованого у Sketch Engine, дало змогу виявити найчастотніші лексеми та їхні типові колокації. Отримані статистичні дані дозволили встановити релевантні контексти використання ключових слів та з'ясувати їх функціональне навантаження. Метод "coupled pair analysis", реалізований через функцію "Parallel Concordance", дав змогу визначити регулярні українські відповідники до сленгових одиниць, що належать до найуживаніших у корпусі. Дослідження має на меті сформулювати базу перекладацьких рішень, що може бути інтегрована в локалізаційні глосарії, настанови зі стилю та довідкові матеріали для роботи з діалогами, насиченими сленговою лексикою. Аналіз відповідників довів, що основною перекладацькою стратегією роботи з обценною лексикою є повне увідповіднення як на семантичному, так і на соціолінгвістичному рівнях. Цензурування чи самоцензурування у процесі подальшої роботи над локалізацією такого продукту через нейтралізацію обценної лексики не відповідає запиту глядачів та не може бути рекомендованим локалізаторам. Розвідка поєднує теоретичний і прикладний підходи до забезпечення якості антропогенного перекладу та постредагованого машинного перекладу в динамічному сегменті локалізації аудіовізуальних продуктів.

Ключові слова: управління якістю перекладацького проєкту, локалізація, аудіовізуальний продукт, постредагування машинного перекладу, антропогенний переклад, корпусні технології, соціолінгвістичні особливості локалізації, сервісорієнтованість, ринок лінгвістичних послуг.

This article explores the potential of corpus-based technologies to ensure the quality of human and post-edited machine translation in audiovisual localization. To achieve this goal, a series of research tasks were outlined: creation of a bilingual parallel corpus of English and Ukrainian subtitles; automatic extraction of key lexical items and assessment of their frequency and significance within the source text; analysis and typologization of translation equivalents; identification of general strategies for rendering frequent lexical items in the localized product; and development of practical recommendations for translators and post-editors working with similar types of content (in terms of genre, style, and target audience). The empirical data consisted of English and Ukrainian subtitles for the British TV series Brassic (Season 5, Episodes 1–7), totaling 40,940 words in English and 28,549 words in Ukrainian. The use of a distributional-statistical method applied to the bilingual parallel corpus, built in Sketch Engine, made it possible to identify the most frequent lexical items and their typical collocations. Statistical data allowed the researchers to determine relevant usage contexts and the functional roles of key words. The "coupled pair analysis" method, implemented via the "Parallel Concordance" function, enabled the identification of regular Ukrainian equivalents for frequent English slang items in the corpus. The study aims to establish a database of translation solutions that can be integrated into localization glossaries, style guides, and reference materials for working with slang-rich dialogue. The analysis of equivalents has shown that the primary strategy for rendering obscene language is full equivalence on both semantic and sociolinguistic levels. Censorship or self-censorship through the neutralization of such language during the subsequent localization process does not meet audience expectations and cannot be recom-

mended to localizers. Thus, this research combines theoretical and applied approaches to improving the quality of human and machine-assisted translation in the fast-evolving field of audiovisual localization.

Key words: translation project quality management, localization, audiovisual product, post-editing of machine translation, human translation, corpus technologies, sociolinguistic aspects of localization, service-orientation, language services market.

Постановка проблеми. Стрімкий розвиток ринку локалізації аудіовізуальних продуктів, пов'язаний із цифровізацією, яка визначає як специфіку виробництва та поширення, так і безпосереднього споживання продуктів аудиторією [1], викликає гостру потребу в науковому переосмисленні діяльності лінгвістів, залучених до процесу локалізації аудіовізуальних продуктів (АВП). За даними загальноєвропейського опитування представників галузі лінгвістичних послуг ELIS, у 2023 році секторі аудіовізуальної локалізації розглядався як такий, що демонстрував найбільш активне зростання [2].

Розгляд традиційної послуги АВП у контексті локалізації, де антропогенний переклад чи постредагування машинного перекладу – лише один із етапів виробництва продукту, – достатньо стійка тенденція як серед відомих дослідників [1, 3], так і серед практиків та аналітиків ринку [2]. Під локалізацією розуміють лінгвістичну та культурну адаптацію цифрового продукту до вимог та особливостей певної локалі, а також послуги та технології, покликані забезпечити багатомовність світового інформаційного потоку [4].

Актуальність дослідження. Власне лінгвістична складова адаптації аудіовізуальних продуктів для української локалі потребує додаткових наукових розвідок. Традиційний для українського перекладознавства розгляд локалізованих українською аудіовізуальних продуктів в контексті теорії трансформацій [5] не завжди бере до уваги специфіку самих продуктів, які, на відміну від традиційних перекладацьких проєктів (текстів художньої літератури, юридичних, медичних тощо) характеризуються мультимодальністю, що є не факультативною, а ключовою характеристикою аналізованих продуктів [6; 7]. Можливим кроком у напрямку виявлення певних тенденцій і готових перекладацьких рішень, системно пропонованих при адаптації аудіовізуальних продуктів, могла б стати обробка значних масивів паралельних субтитрів за допомогою інструментів корпусної лінгвістики. За умови аналізу значного обсягу текстової інформації, отримані результати можуть бути використані в галузі, наприклад, для укладення чи оновлення настанов зі стилю (style guides).

Аналіз останніх досліджень і публікацій. Враховуючи специфіку аудіовізуального дискурсу

і особливості його адаптації для різних локалей, підтримуємо думку про недостатність використання класичних перекладознавчих теорій, які “представляють вади, непослідовності, втрати як випадковості, що загалом не порушують цілісності системи” [8, с. 234] для аналізу субтитрованого продукту, характеристиками якого якраз і є “втрати” – результат конденсації вихідного продукту [6]. Робочим визначенням локалізації аудіовізуального продукту як послуги вважатимемо адаптацію аудіовізуального контенту (художніх, документальних фільмів, реклами, популярних відео) задля їх повного розуміння аудиторією цільової локалі. Процес значно відрізняється від перекладу загальних чи спеціалізованих текстів, оскільки передбачає врахування візуальної та звукової складових для відтворення художньої та комунікативної цілісності оригіналу... Субтитрування передбачає відтворення діалогів цільовою мовою зі збереженням оригінальної звукової доріжки. Субтитрування передбачає лінгвальну та темпоральну адаптацію задля увідповіднення цільового тексту зображенню на екрані [9].

Метою нашої розвідки стало визначення потенціалу застосування інструментів корпусної лінгвістики для забезпечення стійкої якості локалізованого аудіовізуального продукту. Досягнення мети передбачало виконання низки **завдань**: створення двомовного паралельного корпусу англійських та українських субтитрів; видобування ключових слів, частотність яких свідчить про їхню важливість для вихідного продукту; виявлення та систематизація англо-українських відповідників; встановлення загальних стратегій відтворення ключових слів у аналізованому продукті; розробка рекомендацій та допоміжних матеріалів (references) для антропогенного перекладу та постредагування машинного перекладу схожих (за змістом, цільовою аудиторією тощо) аудіовізуальних продуктів. **Матеріалом дослідження** стали англійські та українські субтитри до серіалу “Brassic” (5 сезон, 1–7 серії) загальним обсягом 40940 та 28549 слів.

В дослідженні було використано якісний та кількісний (дистрибутивно-статистичний) методи корпусного аналізу, релевантність якого для перекладознавства підтверджена схожими дослідженнями в галузі локалізації [10], в тому

числі, й аудіовізуальних продуктів [11; 12; 13; 14]. Застосування дистрибутивно-статистичного методу до сформованого дослідниками у програмі Sketch Engine двомовного паралельного корпусу дозволило виявити частотні слова та їхні колокати. На основі даних кількісного аналізу було встановлено статистично значущі контексти, а відповідно, функції ключових слів. Застосування методу “coupled pair analysis” [15] до запропонованих програмою Sketch Engine зіставлених сегментів (parallel concordance) дозволило виявити регулярні відповідники до сленгових англійських лексем – найчастотніших ключових слів аналізованого контенту.

Аналіз англо-українських відповідників мав декілька цілей. Першою з них було виявлення загальних тенденцій відтворення обценізмів і перевірка гіпотези щодо цензури (чи самоцензури лінгвістів) [13]. Цілком очевидно, що якщо в подальшій роботі над проектом (наприклад, наступним сезоном), лінгвіст застосовуватиме інші стратегії (нейтралізування або евфемізації), клієнт відчуватиме себе ошуканим, оскільки персонаж в українській версії “використовуватиме” відмінний від попереднього сезону ідіолект [16]. Другою ціллю було створення переліку регулярних відповідників, які можна було б використовувати в якості довідникового матеріалу, що може бути використаний як для антропогенного перекладу субтитрів, так і для постредагування машинного перекладу, деякі рушії якого (осо-

бливо нетреновані) демонструють певну “цнотливість” відносно сленгової обценної лексики, евфемізуючи або нейтралізуючи її.

Виклад основного матеріалу. Як зазначалося вище, матеріалом для дослідження стали англійські та українські субтитри до аудіовізуального продукту “Brassic” (5 сезон, 2023 р.), зокрема, серії «Eddie», «Missing Body», «Hans», «Sweet Sixteen», «Ratcatcher», «Caravan of Courage», «Albert». На платформі SketchEngine було створено паралельний (англо-український) корпус субтитрів аналізованого матеріалу, а функціонал програми дозволяє переглянути докладну аналітику щодо обсягу, складу, типу текстів тощо.

Видобування ключових слів відбувалось шляхом автоматичного опрацювання англійської складової корпусу програмою і подальшим критичним переглядом результатів видобування з фокусом на ті частотні лексичні одиниці, які становитимуть потенційну проблему для лінгвістів. Функція видобування ключових слів (Keywords) реалізується в варіантах “Single words” та Multiword terms. Результати видобування можуть бути збережені в кількох форматах задля подальшої обробки фахівцями.

Як бачимо, серед видобутих однослівних варіантів на другому (після власних імен) місці – сленгова обценна лексика, яка й буде в фокусі нашої уваги. Її частотність і розмаїття функцій, у яких її було вжито опосередковано підтверджує її фіксація і серед багатослівних

Таблиця 1

Нейтралізація обценної лексики в необробленому машинному перекладі

English	Автоматичний	Google Translate	DeepL Translate
He's your boy, Tommy – there's nothing more likely to cause a fucked up adult than an absent father.	Він твій син, Томмі, і для дорослого немає нічого більш шкідливого, ніж відсутність батька.	Він твій хлопець, Томмі — немає нічого більш шкідливого для дорослого розумування, ніж відсутність батька.	Він твій син, Томмі, і ніщо так не шкідливе для дорослого, як відсутність батька.
Daffine? Daffine!	Дайфно? Дайфно!	Дайфно? Дайфно!	Дайфно? Дайфно!
What the fuck's this Kenny Diamond shit?	Що за жорня з Кенні Даймондом?	Що за лайно з Кенні Даймондом?	Що це за фігня з Кенні Даймондом?
Hans, that's not even the best bit.	Ганс, це ще не найкраще.	Ганс, це навіть не найкраще частини.	Ганс, це ще не найкраще.
Hiya, where's Vinnie?	Здаров, де Вінні?	Привіт, де Вінні?	Привіт, а де Вінні?
Is Mami around, please?	А Мамі вдома?	Мамі вдома поруч, будь ласка?	Мамі тут, будь ласка?
Right, very good. – Carl!	Фіно, дуже добре, Карл!	Правильно, дуже добре, Карл!	Так, дуже добре, Карл!
A dickhead with a fuck truck.	Улюбок з автойОбузом.	Придурак з вантажником для сексу.	Мудак з вантажником.
I'm buying a fucking Nine Swinger.	Я купив мінімунуку, бляха.	Я купую великий секер.	Я купую дитячий мінімунука.
German love sky diving.	Німецьколюбство, стрибати з парашутом.	Німецьколюбство, стрибки з парашутом.	Німецьколюбство, стрибки з парашутом.
Fucking hell...	Бляха-аука...	Чорт, айбири...	Трешка твій антирт...



Рис. 1. Візуалізація статистичних даних паралельного корпусу в програмі SketchEngine (функція Corpus Info)

Рис. 2. Візуалізація видобування однокомпонентних та багатокомпонентних ключових слів /словосполук у програмі SketchEngine (фрагмент)

Серед запропонованою системою ключових слів нами було обрано наступні сленгові лексеми: *fuck* (963 вживань), *cunt* (24), *dickhead* (16), *shite* (9), *fucker* (11), *shit* (101), *dude* (56), *bastard* (23),

На першому етапі було застосовано функцію “Concordance”, яка дозволяє побачити контекст ужитку лексичних одиниць, які були предметом нашої розвідки. Застосована в подальшому функція “Parallel concordance” дозволила візуалізувати англо-українські відповідники в контексті. Функція “GDEX” (good dictionary examples) дозволила виокремити найбільш характерні приклади відповідників у контексті.

Найчастотнішими відповідниками до сленгового обценізму *arsehole* є *мудило*, *мудло*, *пиздобол*, *сука*. За допомогою функції “Parallel Concordance” можна візуалізувати всі варіанти відповідників у контексті.



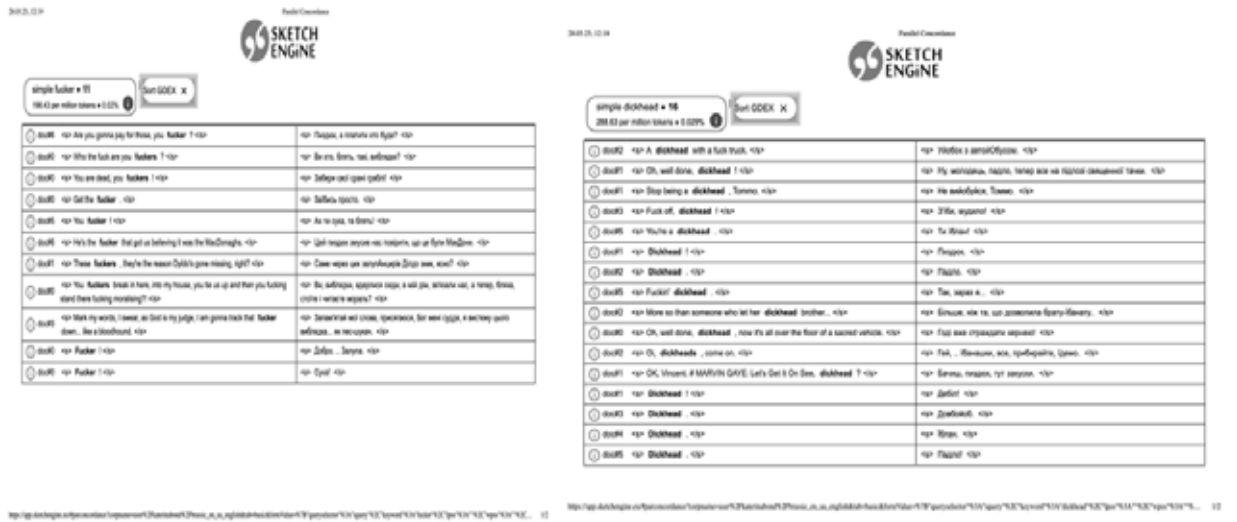


Рис. 3. Візуалізація зіставлення відповідників у вихідних та локалізованих субтитрах засобами програми SketchEngine

Відповідниками до *bastard* в адаптованій для української локалі версії серіалу є *байстрюк*

(2), виблядок (3), падло (2), виродок (2), лайно, вилупок, гівнюк, покидьок, козяра, сука.



За наявності кількох синонімічних відповідників можна встановити певні закономірності колокації кожного з них. За допомогою функції “Word

Sketch Difference” візуалізується статистична інформація (кількість уживань) і особливості взаємодії аналізованих лексем.

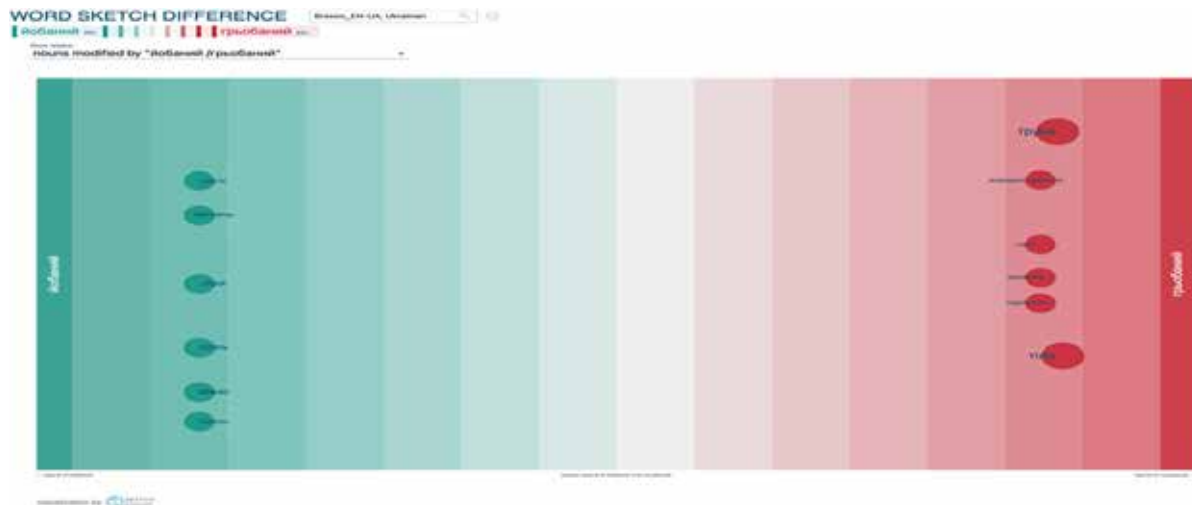
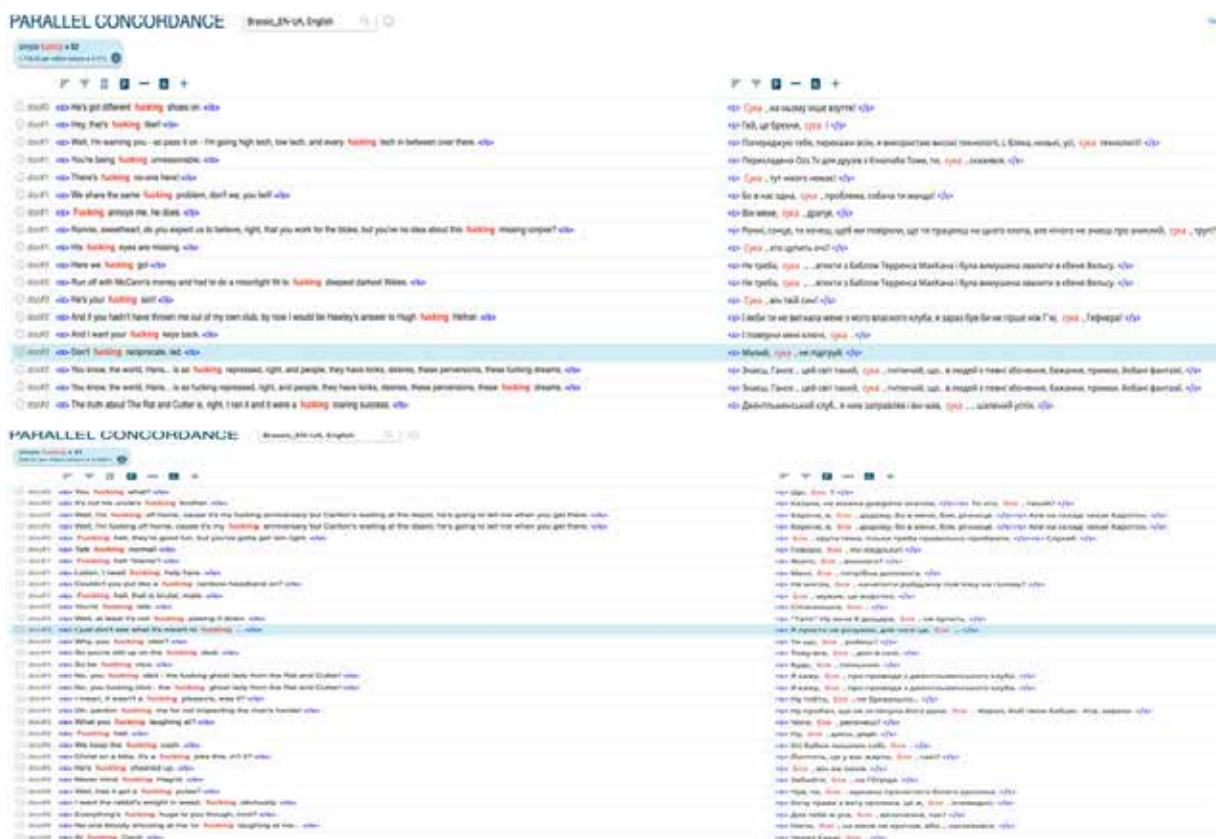


Рис. 4. Візуалізація зіставлення колокатів синонімів у локалізованих субтитрах засобами програми SketchEngine

Зауважимо, що коли *Fucking* використано з емоційною метою, відповідниками будуть переважно *бляха*, *бля*, *сука* тощо.



В незначній кількості випадків (3,08%) українські відповідники тотожні за соціостилістичним регістром, але відмінні за семантикою. Як пра-

вило, в локалізованій версії фіксуємо обценну одиницю з конкретнішим у порівнянні з вихідною лексемою значенні.

i doc#1 <s> Fucking giving it big licks. </s>

<s> Он як заманділа. </s>

i doc#2 <s> Why fucking lie to him? </s>

<s> Нащо ти йому пиздиш? </s>

У понад 50% випадків обценна лексика має відповідниками українські обценізви, притому, значно різноманітніші, аніж у вихідному тексті. Це спростовує поширену думку про неперекладність сленгу. За нашою гіпотезою, висунутою в результаті аналізу української адаптації творів К.Тарантино для української локалі [17] і підтвердженою аналізованим матеріалом “Brassic”, найбільший відсоток випущень /нейтралізації обценізмів пояснюється не цензурою (бо в інших парах відповідників впливу

цензури не простежується), а іншими чинниками, зокрема специфікою субтитрування, конденсація неважливої інформації в яких є характеристикою цієї лінгвістичної послуги.

Вважаємо подану нижче інформацію щодо виявлених в результаті корпусного зіставного аналізу такою, що може бути використана як довідковий матеріал, який забезпечуватиме стійку якість локалізації аудіовізуальних проєктів схожого змісту.

Таблиця 2

Англо-українські відповідники обценної лексики в оригіналі та локалізованій версії аудіовізуального продукту “Brassic”

Англ. обценізм	Значення	Українські обценізви
arsehole	An insult used to describe someone who is very unpleasant, annoying, or mean-spirited.	мудло, мудило, пиздобол, сука
bastard	A very unpleasant, cruel, or despicable person. Used as an insult.	виблядок, байстрюк, падло, виродок, лайно, вилупок, гівнюк, покидьок, козляра, сука
cunt	Used as a deeply offensive insult to describe a person (most often a man) as hateful, mean, or despicable.	пиздюк, сука, хмир, хуйло, суче вим'я, курка патлата, гад
dickhead	Insult for a person who is behaving stupidly, rudely, or obnoxiously.	падло, уйобок, мудило, пиздюк, їбанат, їбанашка, їблан, дебіл, довбойоб
fucking	Used before a noun to emphasize negative quality or contempt, displeasing, despicable	йобаний, сраний, гробаний, ссаний, сучий, їбучий, заїбенний, злоїбучий, блядський, плюгавий, кончений.
	Used to add strong emotion (positive or negative) to a statement.	бляха, сука, бля, блять, бляха-муха

Висновки. Застосування корпусних інструментів (зокрема, Sketch Engine) довело свою ефективність для виявлення лінгвальної специфіки проєкта, в тому числі, регулярних відповідників у перекладі сленгової та обценної лексики, що особливо важливо для підтримання стилістичної послідовності між сезонами аудіовізуальних продуктів. У понад 50% випадків англословна обценна лексика у субтитрах була передана українськими обценізмами тотожної семантики, що спростовує поширену думку про неперекладність сленгу. Українські переклади демонструють значно ширший спектр варіантів, ніж англословний оригінал. Це свідчить про творчу свободу перекладачів або постредакторів, а також – про

відсутність жорсткої цензури. Відповідники до слів *arsehole*, *bastard*, *cunt*, *dickhead*, *fucking* тощо мають чітку залежність від контексту – або зберігають функціональне навантаження (інвектива, зневага, емоційна інтенсивність), або ж нейтралізуються через особливості субтитрування. Виявлені відповідники можуть бути використані як основа для створення глосаріїв або довідкових матеріалів для фахівців з перекладу і постредагування, особливо при роботі з діалогами в АВП. Отримані результати сприяють стандартизації перекладацьких рішень, підвищенню якості локалізації та послідовності в адаптації мовленнєвих характеристик персонажів, що є критично важливим для збереження ідіолекту.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Chaume F. Localizing Media Contents: Technological Shifts, Global and Social Differences and Activism in Audiovisual Translation. *The Routledge companion to global television*. Routledge, 2019. С. 320–331.
2. European Language Industry Survey – ELIS 2023 [Електронний ресурс]. – Режим доступу: https://ec.europa.eu/info/resources-partners/translation-and-language-resources/language-industry_en

3. Jiménez-Crespo M. A. Localization in translation. Abingdon, Oxon; New York NY: Routledge, 2024. 336 с.
4. Schäler R. Translators and localization: Education and training in the context of the Global Initiative for Local Computing (GILC). *The Interpreter and Translator Trainer*. 2007. С. 119–135.
5. Хорошун О. О. Застосування лексичних трансформацій в українському перекладі мультиплікаційного фільму «Монстри на канікулах». *Studia methodologica*. 2024. Вип. 58. С. 191–197.
6. Díaz-Cintas J. Audiovisual translation. *The Bloomsbury companion to language industry studies*. 2020. С. 209–230.
7. Лук'янова Т. Г. Мультиmodalьне конструювання емоцій: інтерсеміотичний переклад (на матеріалі англійської та української мов). *Науковий вісник Херсонського державного університету. Серія: Германістика та міжкультурна комунікація*. 2020. № 2. С. 122–128.
8. Carreres Á. Cruzando límites: la retórica en la traducción de Jacques Derrida. Oxford: Peter Lang, 2005.
9. GALA Global [Електронний ресурс]. – Режим доступу: <https://www.gala-global.org/knowledge-center/about-the-industry/language-services>
10. Jiménez-Crespo M. A. Conventions in localisation: a corpus study of original vs. translated web texts. *Jostrans: The Journal of Specialized Translation*. 2009. № 12(1). С. 79–102.
11. Bywood L., Volk M., Fishel M., Georgakopoulou P. Parallel subtitle corpora and their applications in machine translation and translatology. *Perspectives*. 2013. Т. 21. № 4. С. 595–610.
12. Abu-Rayyash H., Haider A. S., Al-Adwan A. Strategies of translating swear words into Arabic: a case study of a parallel corpus of Netflix English-Arabic movie subtitles. *Humanities and Social Sciences Communications*. 2023. Т. 10. № 1. С. 1–13.
13. Zakuan T. I. M. Subtitling profanity into Malay: Balancing formal and creative taboo expressions. *Salud, Ciencia y Tecnología-Serie de Conferencias*. 2024. Т. 3. С. 1244.
14. Jiang Q., Doherty S. A corpus-based study on the English-Chinese swearwords produced by amateur (fansubbing) and professional (prosubbing) subtitling. *Perspectives*. 2024. С. 1–27.
15. Toury G. Descriptive Translation Studies – and Beyond. Amsterdam: John Benjamins, 2012. С. 116.
16. Chaume F. Is audiovisual translation putting the concept of translation up against the ropes? *The Journal of Specialised Translation*. 2018. № 30(1). С. 84–104.
17. Bondarenko K., Bondarenko O. Translating English slang into Ukrainian: A study of Pulp Fiction and Once Upon a Time in... Hollywood. *SHS Web of Conferences*. 2021. Т. 105. С. 04001.