Enhancing Player Immersion: Automatic AI Localisation of Romanian Dialogue in Video Games

VALERIU-LIVIU NIȚU, University of Twente, The Netherlands

Video game dubbing using AI (artificial intelligence) involves using technology to automate and enhance the availability of voice and text regionalisation in an efficient, easy-to-use, and cost-effective manner. The approach addresses the challenges associated with traditional dubbing, including, but not limited to, hiring voice actors, scheduling recording sessions, translators, etc., by reducing the costs and overhead while also retaining the original voice performance quality. The Romanian language faces unique challenges regarding game localisation. Localisations are often overlooked because the language has relatively low usage outside of Romania (and subsequently the Republic of Moldova) combined with limited market share.

This paper investigates the application of AI-driven techniques for automatic text translation and voice dubbing from English to Romanian in video games. We present our findings on this approach's effectiveness, challenges, and player reception, focusing on two case studies: Pokémon FireRed and Fallout 4. Our research explores the potential of AI to make game localisation more accessible for languages with smaller market shares while maintaining game immersion.

Additional Key Words and Phrases: AI-Powered Localisation, Romanian Text-to-Speech (TTS), Voice Cloning, Game Immersion, AI Translation, AI Dubbing

1 INTRODUCTION

The global video game market size is currently estimated at $217 billion and is expected to grow at a 12% compound annual growth rate (CAGR) by 2030 [24]. As the market expands into multiple countries, localisation becomes increasingly essential to increase immersion and enjoyment for players in different regions.

As noted by Bernal-Merino [2], video game translation does not differ fundamentally from other types of audio-visual translation: from voice acting extensive lines for the main character to many, short lines for NPC (non-playable characters) that can be found in the game world. This cannot be done lightly since it has been shown that audio stimuli are the driving factors of immersion in video games, as portrayed by Jaewhan Byun [3]. If a game contains auditory dialogue, work on voice recording starts as soon as script writing begins. The recording continues during the game's entire development, and localisation for different languages begins once it is done. This is also a very time-consuming and costly development process, as words can have different meanings in other languages: the complex cultural environment of various languages, combined with the intricacy of slang and language-specific expressions, can not only lead to poor localisations but can also detour developers for pursuing specific languages altogether.

Recent advancements in artificial intelligence (AI) models, such as large language models (LLMs) like GPT-3, can help automate translation while preserving the original, intended meaning[10]. LLMs can capture long-range dependencies between words and phrases, allowing them to handle complex sentence structures and act as universal translators.

In addition to text translation, voice dubbing requires synthesising the original actor's voice to maintain immersion. Specific particularities, like voice depth, tone, softness, and hardness of the voice, need to be considered when trying to carry over a voice in another language altogether. With continuous developments in the AI domain, some great models have emerged, for example, Tacotron 2 [18], a neural network architecture for speech synthesis that can comprehensively mimic the original voice actors' characteristics by predicting spectrograms based on the original recording. WaveNet is another deep generative model that can capture raw waveforms, including pitch, tone, and subtle variation in the voice [13].

However, there is a lack of dedicated models or solutions for automated voice translation, cloning, and synthesis for languages like Romanian, which have limited representation in the video game market. This research paper aims to investigate and develop a framework to automate the extraction of voices and text from video games, capture voice specificities and translate from English to Romanian. Ultimately, the results will be evaluated via a public opinion survey that will rank the quality of the translated text and the translated and synthesised voice.

2 PREVIOUS WORK

The use of AI models in video game localisation, particularly in text and voice dubbing, has been explored in an academic context. Rea Tot and Ivan Dunder's [22] work on the application and significance of the Video Game Localisation Process delves into the practical aspects of localisation, emphasising the need for a thorough understanding of the target audience's language and cultural context to ensure the game's success in international markets.

Moreover, it is also noted that video games are becoming a "medium for cultural expression and communication". Shira Chess and Mia [4] mention in their paper that this perspective is crucial in understanding how to localise video games using artificial intelligence while ensuring that the resulting translation is not only efficient but also culturally sensitive by keeping the meaning of the original language without introducing new terms and concepts that can be confusing in the translated language.

From a practical standpoint, Rask.ai [17] currently offers an AI-powered video translation service that can translate video content into Romanian while trying to adapt to the original dubbing, targeting the same voice tone and style. This gives out conflicting results, as the TTS (text-to-speech) voices are not always similar to the original voices.

TScIT 41, July 5, 2024, Enschede, The Netherlands

3 PROBLEM STATEMENT

While recent advancements in AI models have shown promise in automating translation and voice synthesis, there is a lack of dedicated models and voices trained specifically for the Romanian language. Moreover, game development companies and large game publishers refrain from localising video games in smaller demographics/smaller markets due to the increased costs associated with traditional voice dubbing [12].

Existing AI-powered video translation services like Rask.ai attempt to translate content into Romanian while adapting to the original video nuances. Still, the TTS (text-to-speech) voices are not always similar to the original voices.

To address these challenges, we aim to answer the following research question: **How can AI be leveraged to create an efficient, accurate, and complete solution for Romanian video game localisation?**

- How can existing AI models for text translation and voice synthesis be used to handle the specific challenges of the Romanian language, such as its complex grammar, unique phonology, and context?
- What techniques can be employed to improve the matching of synthesised Romanian voices to the original voice actors' characteristics (e.g., voice depth, tone, subtle variations) to maintain immersion and consistency in the localised video game dialogue?
- Does automated AI dubbing in Romanian enhance the user experience in video games? (i.e. does the user feel more immersed, is the gameplay experience consistent, does it make the game easier to play)

By investigating and developing a framework or pipeline that can automate the extraction of voices and text from video games, capture voice specificities, perform automated translation from English to Romanian paired with voice synthesis in Romanian, and generate subtitles/text, we aim to provide a standardised approach for Romanian video game localisation using AI technologies together with an analysis on the effect of such localisation among video game players.

4 METHODOLOGY

In this subsection, the methodology that will be used to answer each sub-research question is described as follows:

4.1 Using Existing AI Models

As a starting point, we will be using `Helsinki-NLP/opus-mt-tc-big-en-ro` [8], which is a neural machine translation model specifically trained to translate text from English to Romanian. The model is based on the transformer-big architecture and has been trained on the OPUS Corpus, which is composed of a collection of translated texts from various internet sources, including subtitles, news, legal texts, and web content across multiple languages[21]. Given that the model has been specifically trained on Romanian data, it helps tremendously with integrating linguistically accurate text, ensuring immersion. It scored 48.6 on Tatoeba-test-v2021-08-07 and 40.4 on Flores101-devtest on the BLEU (Bilingual Evaluation Understudy)[7], which are good scores for an off-the-shelf translation solution. For reference, a percentage between 40-50 on the BLEU scale is considered a High-Quality Translation. Any result over 60 is considered to be better than human translation.

We are also using `whisper-large-v2`[14] which is an automatic speech recognition (ASR) model for converting the audio into text. Given that most games come with subtitles that differ from the actual transcription of the character's voices, it is critical to obtain the specific text that should be translated. This model is used in the process of translating spoken words into written text with high precision, and it is very suitable for most languages and their dialects. Among all the other tested models (Facebook Wav2Vec 2. 0, Kaldi, Nvidia NeMo, and SpeechBrain), whisper is one of the most stable and reliable models as it can distinguish onomatopoeia and translate the spoken text from speech components with the particularities of the given text.

To conclude our pipeline, we are using `ElevenLabs TTS` [20], which is a state-of-the-art text-to-speech (TTS) model. This model can produce natural and expressive speech from the text with high quality and it is best suited for multilingual applications. Compared to other tested models, including Google Text-to-Speech, Amazon Polly, IBM Watson and Microsoft Azure TTS, ElevenLabs can be considered the most stable. It created the most realistic speech that could imitate human intonation and emotions and was very accurate in the Romanian language.

By using the three solutions provided above, we ensure that our pipeline overcomes technical and time-constrained challenges of developing a solution firsthand while also providing a base that can be developed for the sub-research questions.

4.2 Creating a translation pipeline

Text to speech and text translation:

By combining the models mentioned above for automatic speech recognition (ASR) and machine translation, we create a translation pipeline that converts spoken language into translated text. A diagram of the entire pipeline can be seen in Figure 1. This pipeline leverages the capabilities of OpenAI's `Whisper` [14] for ASR and `Helsinki-NLP/opus-mt-tc-big-en-ro`[8] for translation. All of this is done in a Python script that processes .wav and .mp3 files and exports the translation in either a .csv file with the original name of the files and translation in two separate columns or as a .txt with the original name of the file and the Romanian translation inside.

Voice Synthesis Enhancement: For this part, our focus is on improving the synthesised voices' emotional expressiveness and naturalness. This will be done by training `ElevenLabs'`[20] solution with 10-20 (depending on the number of recorded voices one character has) voice snippets. The resulting trained model is then fine-tuned/tweaked for each generated voice line in order to ensure that it retains the emotion of the original character.

4.3 Performance & quality

We will conduct surveys with native Romanian speakers. Participants will listen to and read segments of the AI-generated dubbing and translations. They will provide feedback on various aspects, such as the accuracy of the translation, the quality of the dubbing,

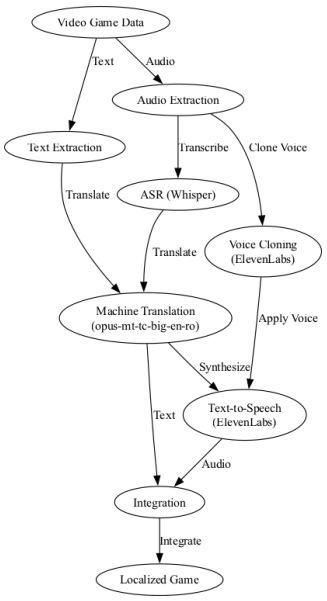

Figure 1. Localisation Pipeline Diagram

and how natural the translation sounds, in addition to the immersiveness and naturalness of the translated text. All of the survey participants are required to be proficient English speakers and to have played video games in the past.

The questions in the survey are nuanced based on the work of Mangiron and O'Hagan [11], in which they explain how game localisation is dependent on preserving the game atmosphere and cultural impact, as well as the challenges of adapting the cultural references. Moreover, the questions have also been heavily influenced by the work of Hinterleitner et al. [9], in which the authors explore the importance of naturalness/human-likeness of the synthetic voice as well as prosody and intonation. Combining the information from those papers, 32 questions have been developed that target not only how the participants perceive the localisations but also the quality and naturalness of the text and voice translation/synthetisation.

5 EXPERIMENTS

5.1 Extracting Video Game Data

For the purpose of this research, we have chosen two video games: Pokémon FireRed and Fallout 4. These games were selected based on the availability of tools for data extraction, which aid in the process of obtaining the necessary text and audio files for translation and dubbing.

5.1.1 Pokémon FireRed. Pokémon FireRed is a classic role-playing game with extensive text-based conversations. This forms a good basis for our text translation part since Pokémon does not contain any voiced characters. For the extraction of the text data, we used tools such as `Pokémon FireRed/LeafGreen Decompilation Project` [16] and `AdvanceText` [1], which allow for the extraction and modification of in-game text. Once the text has been extracted in a .txt format, it is then translated using the mentioned aforementioned `Helsinki-NLP/opus-mt-tc-big-en-ro`[8] model using a Python script detailed in Section 5.2.

5.1.2 Fallout 4. Fallout 4 is a modern open-world game with a significant amount of voice-acted dialogue. For extracting audio data, we utilised tools like `BSA Browser` [6], a tool designed to unarchive .ba2's (Bethesda Archives) and extract the `.fuz (Bethesda Softworks Voice File)` audio files, `YakitoriAudioConverter`[25] which can convert .fuz files to .wav, in turn, allowing us to extract the voice files from the game's archives. These extracted audio files will be processed using the mentioned ASR model to convert them into text, which will then be translated and synthesised back into Romanian using the TTS model. Once this process is complete, we used Bethesda's archiver packaged into Fallout 4's files, `Archive2`, to zip the data back into a format that is recognised by the game.

By using the tools mentioned above, both for Pokémon and Fallout, we ensure that the extraction process is efficient and that we have access to high-quality data for our experiments. The extracted data will be used to test the effectiveness of our AI models in translating and dubbing video game content from English to Romanian.

5.2 Translation and Voice Synthesis

5.2.1 Translation Pipeline. The translation pipeline involves several steps to ensure accurate and contextually appropriate translations from English to Romanian. The process is as follows:

Text Extraction. For both video games, the text is saved in a .txt format that is then parsed by the custom Python code to be processed and translated.

Automatic Speech Recognition (ASR). Using the `whisper-large-v2` [14] model, we transcribe the extracted audio files from Fallout 4 into text. All the data is then saved as .txt (for both versions) and as .csv for the Fallout 4 version (labelled for English version and Romanian Version) in order to be used for training for voice synthesis models. The speech recognition is able to process around 2.9 it/s on an RTX 2070.

Machine Translation. The transcribed text from Fallout 4 and the extracted text from Pokémon FireRed is then fed into the `opus-mt-tc-big-en-ro`[8] model for translation. The Python algorithm runs this on the CPU, managing to complete around 1.7 it/s on a Ryzen 7 5800x CPU.

5.2.2 Voice Synthesis. Once the text has been translated, the next step is to synthesise the translated text into speech. This involves the following steps:

Voice Cloning. Using the `ElevenLabs` TTS model, we clone the original voice actors' characteristics. This involves training the model with 10-20 voice snippets from each character to capture their unique vocal traits, such as tone, pitch, and emotional expressiveness. `ElevenLabs`[20] requires the user to input specificities about the voice of the character in order to ensure voice tone accuracy. An LLM (large language model) is used to define specificities of the voice of the character; in this case, `Phind 70b` [15] is used to generate specific characteristics of the voice; in our experiment, the voice of the Vault-Tec employee (1960 salesman in his 40's, persuasive but friendly voice).

Figure 2. Romanian Localisation of Pokémon FireRed

Text-to-Speech (TTS). The translated text is then synthesised into speech using the trained `ElevenLabs`[20] model. During this process, the voice of the character is synthesised multiple times by changing the parameters (voice similarity, voice stability and exaggeration) in order to match the expressiveness and the tonality of the English version of the dubbing. Since we are not using a complete voice transfer technique (see 7.2), this part is done and tested manually in order to ensure the quality and accuracy of the voice of the character by listening to each audio snippet, modifying the specifics of the audio generation (tone, depth etc.) then regenerating accordingly. Currently, `ElevenLabs`[20] does not offer an API for voice synthesis.

Integration. The final synthesised voice files are integrated back into the game. For Fallout 4, this involves using `Archive2` to package the synthesised voice files into the game's archive format. For Pokémon FireRed, the translated text is reinserted into the game using the `Pokémon FireRed/LeafGreen Decompilation Project` and `AdvanceText` tools based on how many lines of text have been changed.

6 RESULTS AND EVALUATION

6.1 Preconditions:

A survey has been conducted on the quality and immersiveness of the translation for both games. A total of 62 responses have been collected from the survey. After data cleaning and normalisation, 57 responses have been processed. The critical factor in this selection was the initial requirement of playing video games beforehand.

From the researcher's standpoint, good results have been achieved. The introductory part of Pokémon FireRed has been completely translated into Romanian, including conversations with Mom, Prof. Oak, and the various NPCs around the town. A screen capture of the Romanian translation is showcased in Figure 2.

Regarding Fallout 4, the initial conversation between the main character and the Vault-Tec employee has been dubbed in Romanian. This includes the part where both characters ask and answer questions. The main UI of the game, specifically the main menu UI, has also been translated into Romanian in order to showcase the capabilities of text translation. A screenshot of the UI is showcased in Figure 3.

6.2 Survey Results

6.2.1 Demographics. To understand the demographics of our survey participants, we collected data on their gender, age, and gaming habits. Below are the visual representations of the collected data.

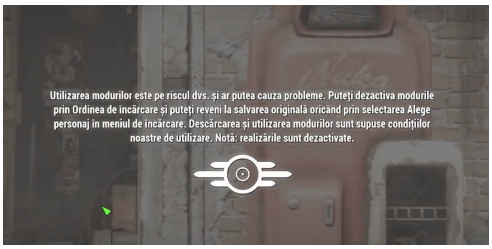

Figure 3. Romanian Localisation of the UI of Fallout 4

Gender Distribution. Out of all the participants, 75% were male, 23.2% were female and 1.8% preferred not to specify their gender.

Age Distribution. On average, the participants were 24 years old.

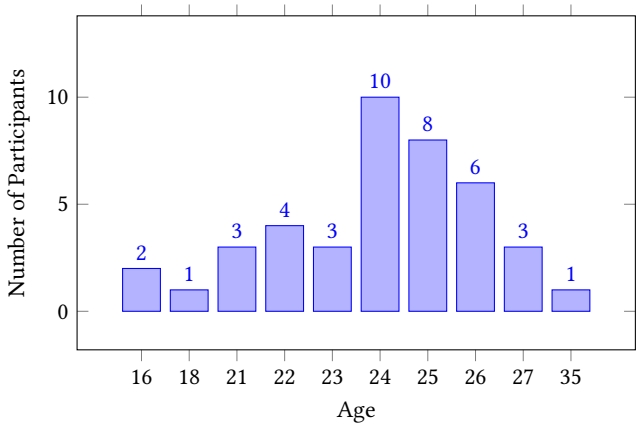

Figure 4. Age Distribution of Survey Participants

Gaming Habits. The survey also collected information on the participants' gaming habits, including the frequency of playing video games and the languages in which they enjoy playing. As seen in Figure 5, the majority of participants play video games daily or several times a week, and most prefer to play in English. A significant portion of participants also indicated they would enable Romanian text and voice dubbing if available, which showcases the interest the Romanian market has in localising videogames in their own language.

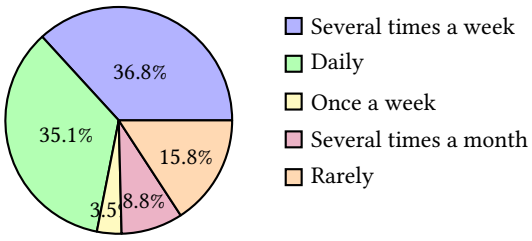

Figure 5. How often do you play videogames

- **Preferred Languages for Playing Video Games:**
  - English: 90%
  - Romanian: 10%
- **Activation of subtitles when playing videogames:**
  - Yes: 73.7%
  - No: 26.3%

Preferences for Romanian Localisation. When asked if they would enable Romanian text and voice dubbing if available, the responses were as follows:

- Yes: 22.8%
- Sometimes: 31.5%
- No: 45.7%

Other spoken languanges: Some users reported that they also speak Russian, Dutch, Hungarian, German, and Italian fluently or near natively, in addition to English and Romanian.

6.2.2 ***Pokémon FireRed***. For Pokémon FireRed, the participants were presented with a 3-minute video showcasing the initial part of the game, translated into Romanian. This includes the initial conversation with Prof. Oak and the interaction with the various NPCs and objects. The following results have been collected:

Font Consistency and Readability. The users were presented with the Romanian and the English version of the first dialogue in the game. The following results were recorded:

- 68.4% of respondents found the font style and size to be identical
- 29.8% found it very similar
- Only 1.8% reported it as somewhat different

This is a very interesting finding since the font style, size, and position are identical to the English version. This can be traced to the concept that participants perceive things differently in different languages. The specific phenomenon can be related to the linguistic relativity hypothesis, which suggests that the structure of a language can influence its speakers' cognitive processes and perception of the world, as presented by Cook V. [5].

Regarding readability and understanding:

- 68.4% of participants said the Romanian version makes the text easier to read and understand
- 31.6% disagreed, suggesting the English version was more readable

Dialogue Naturalness. The Romanian translation was generally perceived as natural:

- 84.2% found it natural or very natural
- 14% were neutral
- Only 1.8% found it unnatural

Impact on Storytelling. The impact of the Romanian translation on storytelling was mixed:

- 38.6% felt it enhanced the storytelling experience
- 57.9% felt it neither enhanced nor hindered the experience
- 3.5% felt it hindered the storytelling

Gameplay Experience. The majority of respondents (63.2%) reported that the Romanian translation did not affect their gameplay experience in a bad way, while 31.6% said it did, and 5.2% were unsure. The participants who were unsure explained that some details might be lost during translation, such as undertones, but they also believed that the Romanian translation would help young children learn the native language.

Immersion. When asked which version they found more immersive:

- 50.9% found both versions equally immersive
- 33.3% found the English version more immersive
- 15.8% found the Romanian version more immersive

Unique Perspective. 50.9% of respondents felt that the Romanian version offered a unique perspective on the game's story and characters, while 45.6% disagreed, and 3.5% were unsure.

Overall Quality. The overall quality of the Romanian translation can be considered highly rated. Figure 6 showcases the results. It is worth noting that no participants considered the translation to be "poor":

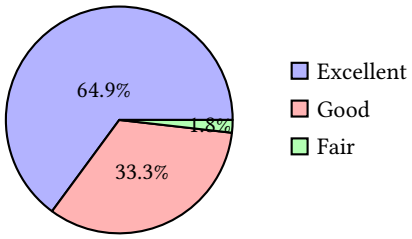

Figure 6. Rating of Romanian Localisation of Pokémon FireRed

This can be attributed to the simplicity of the text, given that the target audience of Pokémon FireRed is mainly composed of children & teenagers. Moreover, the lack of audio dubbing can also influence the opinions of the users, as we later see in the study. People seem to prefer the game voice to be in English and the text to be in Romanian.

Additional Observations. Some participants noted that the Romanian text sometimes overflowed the textbox or overlapped with existing text. However, most felt that these issues did not significantly impact the game's immersiveness. A few respondents suggested that adding diacritics and using a font that supports them would improve readability.

6.2.3 ***Fallout 4***. For Fallout 4, the participants were presented with a 3-minute video showcasing the conversation between the main character and the Vault-Tec employee at the beginning of the game, fully dubbed in Romanian. They were also presented with the same scene in the game, dubbed in English, for comparison reference. Moreover, at the end of the video, the users were presented with the entire main menu UI, fully translated into Romanian.

Character Portrayal. When comparing the Vault-Tec Sales Employee's portrayal in English and Romanian:

- 59.6% felt the English version was better or much better
- 19.3% found both versions equally good
- 21.1% found the Romanian version better or much better

Voice Similarity. Regarding the similarity of voices between versions:

- 47.4% said the voices sounded the same
- 36.8% said they did not sound the same
- 15.8% were unsure

Voice Actor Retention. 73.7% of respondents felt that retaining (synthesising) the original voice actor's voice added value to the immersion, while 19.3% were neutral, and 7% disagreed.

Dialogue Naturalness. The naturalness of the Romanian dialogue was perceived as follows:

- 63.2% found it natural or very natural
- 12.3% were neutral
- 24.5% found it unnatural or very unnatural

Translation Accuracy. 50.9% of respondents noted instances where the translation felt awkward or misplaced, while 49.1% did not.

UI Readability. The readability and intuitiveness of the UI in Romanian were rated as follows:

- 87.7% clear or very clear
- 10.5% neutral
- 1.8% unclear or very unclear

57.9% of respondents felt that the translation of UI elements enhanced their understanding of game mechanics.

Gameplay Experience. When asked if these changes affected their gameplay experience:

- 45.6% said yes
- 33.4% said no
- 21% were neutral

Immersion. Regarding which version was more immersive:

- 50.9% found the English version more immersive
- 24.6% found the Romanian version more immersive
- 24.6% found both equally immersive

Unique Perspective. 56.2% felt that the Romanian version offered a unique perspective on the game's story and characters, while 24.6% disagreed, and 19.3% were neutral.

Dubbing Preference. When asked about their preferred dubbing option:

- 31.6% preferred English dubbing with Romanian subtitles
- 29.8% preferred Romanian dubbing with optional Romanian subtitles
- 15.8% wanted both options available
- 22.8% felt no Romanian dubbing/translation was necessary for immersion

Overall Quality. The overall quality of the Romanian translation of Fallout 4 was rated as:

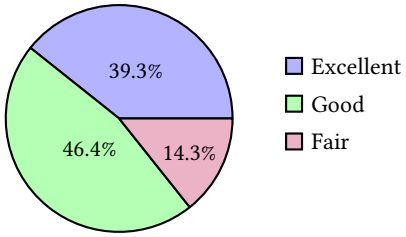

Figure 7. Rating of Romanian Localisation of Fallout 4

In conclusion, the Romanian translation and dubbing of Fallout 4 received mixed but generally positive feedback. While some aspects like UI readability and overall quality were highly rated, there were concerns about dialogue naturalness and translation accuracy, with some participants arguing that some phrases were not part of the "spoken Romanian". The AI-generated voices were recognised by many, with opinions split on whether this enhanced or detracted from the experience, as can be seen in the Impact on Storytelling part. The results suggest that while there is potential for AI dubbing in games, there is still room for improvement in naturalness and accuracy.

Taking into account the written feedback, most participants suggested that a complete dubbing would be possible with some minor tweaking and further model development.

6.3 Discussion

The findings of this research work on Romanian Pokémon FireRed and Fallout 4 games demonstrate the viability and limitations of game localisation processes, especially when employing AI-translated and dubbed methods. Below is a detailed analysis for each game:

Pokémon FireRed. The Romanian translation was received rather positively in Pokémon FireRed. It can be concluded that text-based localisation can be successfully implemented even in the games of previous generations. As reflected in the results section, the three aspects of the translation: consistency, naturalness, and quality (98.2% of the participants rated it as good or excellent) show that the translation maintained the game's immersion while making it more accessible to the Romanian-speaking users. This aligns with Mangiron and O'Hagan's emphasis on preserving the game atmosphere and cultural impact during localisation[11].

The observation that 30% of the participants perceived differences in font style and size (very similar), despite these elements being identical to the English version, aligns with the linguistic relativity hypothesis [5]. This finding shows that there could be a link between language and perception in the gaming context.

However, it is worth noting that in the written feedback given by participants, most would focus on how accessible the games would now be for people who are not proficient in English and for young people who are just learning English.

Fallout 4. The more mixed reception of Fallout 4's Romanian localisation, particularly its AI-generated dubbing, reveals both the potential and limitations of current AI voice synthesis technologies in game localisation. While 85.7% of participants rated the overall quality as excellent or good, concerns about dialogue naturalness and translation accuracy suggest that AI dubbing still has room for improvement.

Since 47.4% of participants found the synthesised voices similar to the original, and 63.2% felt that retaining the original voice actor's characteristics added value to immersion, we cannot state for sure that the voice cloning was successful. However, when asked about the synthesis of the voice (and subsequently, the retention of the character's voice), 73.7% answered yes, which aligns with the work of Hinterleitner et al. on the importance of naturalness and human likeness in synthetic voices [9].

6.4 Conclusion

Our findings suggest that while there is an interest in Romanian localisation among players (54.5% would enable Romanian text and voice dubbing at least sometimes), the approach may need to be tailored to the specific game and target audience. This supports O'Hagan's assertion that game localisation faces unique challenges due to the interactive nature of games [12].

For text-heavy games like Pokémon FireRed, a well-executed text translation could be sufficient and cost-effective. For more complex games like Fallout 4, a hybrid approach might be most effective, involving professional translation of text elements combined with AI-assisted dubbing, with ongoing refinement based on player feedback.

The study revealed interesting cultural and linguistic phenomena. Some participants identified the accent of the Vault-Tec Employee as being specific to their region, highlighting the potential of AI-generated voices to capture linguistic characteristics, even from cloned voices. This observation, paired with the fact that some users found the Romanian version offered a unique perspective on the game's story and characters (50.9% for Pokémon FireRed and 56.2% for Fallout 4), adds to the cultural value of localisation beyond a simple translation. From my point of view, this could add replayability value to the game, which was also mentioned by some of the participants in the survey.

7 LIMITATIONS

7.1 Dataset size and data formatting

It's important to note that our study focused on specific segments of two games and had a limited sample size of games and users. Future research could explore a broader range of games and genres and include a larger, more diverse participant pool. Additionally, longitudinal studies could provide insights into how players' perceptions of localised content change over extended gameplay periods.

Moreover, our participant pool was composed of users who are proficient in English, for which such localisations might not be as crucial as they would be for users who only speak Romanian. A complete analysis specifically tailored for users who do not speak English that well would probably consist of entirely different findings.

Fallout 4. Although Fallout 4 is considered one of the most modifiable games currently on the market, it posed some challenges regarding the extraction of data and its labelling. The game's extensive modifiability, while advantageous, presented challenges in processing the extracted data. In particular, the labelling of the data obtained in Fallout 4 was far from obvious, making it difficult to distinguish and classify different UI items. To circumvent this issue, the French localisation of the game was utilised as a workaround; however, this approach required additional efforts to modify the UI components, which are not stored in plain text formats. For instance, the Heads-Up Display (HUD) within the game worked on Adobe Flash, which needs additional effort for the data manipulation to be effective. In addition, the lack of easily identifiable terms like the "main menu interface" or "Pip-Boy interface" meant that more manual work had to be taken to correctly identify, translate and then re-implement the data in the game context.

Moreover, while the audio files contained the expressiveness of the character, some extra information on the background of the file (like the tonality, where the character is using sarcasm, joy, or frustration) was lost. This limitation affected the rate of voice synthesis and also made it difficult to switch from one voice style to another in the subsequent audio files. Every sound file was correlated to one line of text within the game, which resulted in losing the context and making the task of attaining a consistent voice style even more difficult. Adding to the issues, the disorganised arrangement of voice lines, which were neither sorted by chronological appearance nor usage in the game files, complicated the synchronisation of synthesised voices with the original performances.

Pokémon FireRed. Despite the age of Pokémon FireRed, exceeding two decades, the problems of text data extraction and processing were rather significant. The first challenge was due to the unavailability and archaic nature of most community tools that are either classified as abandonware, no longer available online or simply non-operational. This was the case, especially in the data preparation phase, where some of the transformations had to be done manually because of formatting peculiarities of texts in the game. Notably, each line of dialogue must conclude with a newline character (\n") to denote a new line, a convention that the translation model struggled to accommodate without manual adjustment. Consequently, a significant portion of the game's conversations required manual reintegration to prevent text overflow or unintended replacement of existing text within the game's elements.

Moreover, some textual components that are considered part of the gaming context, like menu messages, specific interfaces, and stylized texts, are inserted as images in the game files. This embedding raised challenges to the process of automated data extraction, which in turn placed extra pressure on the automation of the translation pipeline. The fact that these elements had to be tweaked by hand added more challenges to the localisation process, pointing to the inefficiencies of current approaches in dealing with differences in data formatting within the older games" code.

An additional complication arose from the linguistic characteristics of the Romanian language, which incorporates special characters (ș, ț, â, ă, î) absent from the original font set of Pokémon

FireRed. Due to the poor type support of the game font, these characters have been replaced with similar-looking characters (s, t, a, a, i). Although semantically and syntactically similar, this adaptation may interfere with the game's' realism for those users who are used to the standard Romanian orthography. This modification highlights the broader implications of character encoding limitations in game localisation efforts, particularly for languages with specialised alphabets or diacritics not universally supported across software platforms.

7.2 Synthesis using Tacotron 2

One of the main internal goals of this study was to deploy a functional implementation of Tacotron 2 [19]. Despite efforts, this goal remained unfeasible due to persistent challenges encountered during the integration and deployment of the Tacotron 2 model alongside Apex on the available computing platforms (AWS, University of Twente's Jupyter Lab as well as my personal computer) paired with the short timeframe of the project. These difficulties were also linked to limitations in computational resources and discrepancies between library dependencies, needing complex adjustments such as version upgrades or downgrades. Furthermore, Tacotron 2's inability to undergo effective training added to these issues.

If the integration of the Tacotron 2 model was successful, then it would have enabled the realisation of a complete pipeline of voice processing. This pipeline would have assured the "twinning" of all the voice specific characteristics, including tones, intonations, and the ability to convey feelings such as irony, joy or sorrow. Some of these advancements would have enriched the scope of analysis and data collection and would have provided a more comprehensive understanding of the subject matter.

8 FUTURE WORK

The information presented in this paper was gathered over the course of 6 weeks and can be regarded as a foundation for AI-based Romanian localisation in video games. However, there are several aspects for future work that could significantly enhance the capabilities and outcomes of the approach presented in this paper:

Tacotron 2 Implementation. : Based on the difficulties observed in the current study, future research could focus on realising the implementation/ deployment of Tacotron 2 for voice synthesis. This would enable the transfer of voice in a more natural manner with details such as intonations and other essential expressions that are necessary for immersion in video games. The specific technical problems and compatibility issues identified in this study could be solved in order to produce a more extensive and completely autonomous pipeline.

Fine-tuning Translation Models. : The suggestion for further research would be to apply the proposed method in translating selected topic-specific Romanian texts into English to achieve better accuracy and perceived realism of the translations. This could involve developing a corpus using Romanian TV shows that are subtitled, for example, the T.V. show "Umbre" [23] which contains a lot of Romanian jargon and phenomenal English subtitles that capture the essence of the Romanian language. Including this data into the model, the latter can reflect all the specifics of the Romanian language as used in commonly spoken language, which in turn will make the translations more natural and appealing.

Expanding Game Selection. : Based on the preferences of the author, the research was carried out on Pokémon FireRed and Fallout 4; further studies can be made on a different selection of games from various genres and trends. This would help in knowing the effectiveness of the AI localisation strategy depending on the type of gaming content and story, and can further research which games are more suitable for this method and which are not.

Improving Data Extraction and Processing. : Due to the issues experienced while performing data extraction and formatting, especially for games such as Pokémon FireRed, future research could be directed towards creating better and more dynamic tools for game data extraction. This could include developing tools that are freely available and capable of dealing with any format of a game or its data.

REFERENCES

[1] *Advance Text – Hack Rom Tools*. en-US. URL: https://www.hackromtools.info/advance-text/.

[2] Miguel Á Bernal-Merino. "Challenges in the translation of video games". en. In: ().

[3] JaeHwan Byun and Christian Loh. "Audial engagement: Effects of game sound on learner engagement in digital game-based learning environments". In: *Computers in Human Behavior* (May 2015). DOI: 10.1016/j.chb.2014.12.052.

[4] Shira Chess and Mia Consalvo. ""The future of media studies is game studies"". In: *Critical Studies in Media Communication* 39.3 (May 2022). Publisher: Routledge _eprint: https://doi.org/10.1080/15295036.2022.2075025, pp. 159–164. ISSN: 1529-5036. DOI: 10.1080/15295036.2022.2075025. URL: https://doi.org/10.1080/15295036.2022.2075025.

[5] Vivian Cook. "The language in language and thinking". In: *Vigo International Journal of Applied Linguistics* 18 (Jan. 2021). Publisher: University of Vigo, pp. 35–58. DOI: 10.35869/vial.v0i18.3364.

[6] Alexander Ellingsen. *AlexxEG/BSA_Browser*. original-date: 2014-10-31T12:58:09Z. June 2024. URL: https://github.com/AlexxEG/BSA_Browser.

[7] *Evaluating models | AutoML Translation Documentation*. en. URL: https://cloud.google.com/translate/automl/docs/evaluate.

[8] *Helsinki-NLP/opus-mt-tc-big-en-ro · Hugging Face*. URL: https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-ro.

[9] Florian Hinterleitner, Christoph Norrenbrock, and Sebastian Möller. "Is intelligibility still the main problem? a review of perceptual quality dimensions of synthetic speech". In: *The Eighth ISCA Tutorial and Research Workshop on Speech Synthesis, Barcelona, Spain, August 31-September 2, 2013*. ISCA, 2013, pp. 147–151. URL: http://www.isca-speech.org/archive/ssw8/ssw8_147.html.

[10] *LLMs are universal translators: on building my own translation tools for a foreign language conference | Library Innovation Lab*. en. Nov. 2023. URL: https://lil.law.harvard.edu/blog/2023/11/29/llms-are-universal-translators/.

[11] Carme Mangiron and Minako O'Hagan. "Game Localisation: Unleashing Imagination with 'Restricted' Translation". In: *JOURNAL OF SPECIALISED TRANSLATION* 6 (July 2006).

[12] Minako O'Hagan and Carme Mangiron. *Game Localization: Translating for the Global Digital Entertainment Industry*. Aug. 2013. DOI: 10.1075/btl.106.

[13] Aaron van den Oord et al. *WaveNet: A Generative Model for Raw Audio*. arXiv:1609.03499 [cs]. Sept. 2016. DOI: 10.48550/arXiv.1609.03499. URL: http://arxiv.org/abs/1609.03499.

[14] *openai/whisper-large-v2 · Hugging Face*. Sept. 2023. URL: https://huggingface.co/openai/whisper-large-v2.

[15] *Phind*. URL: https://www.phind.com/blog/introducing-phind-70b.

[16] *pret/pokefirered*. original-date: 2017-12-31T04:07:51Z. June 2024. URL: https://github.com/pret/pokefirered.

[17] *Romanian Video Translator - Translate Videos To and From Romanian with AI*. en. URL: https://www.rask.ai/tools/video-translator/romanian-video-translator.

[18] Jonathan Shen et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. arXiv:1712.05884 [cs]. Feb. 2018. DOI: 10.48550/arXiv.1712.05884. URL: http://arxiv.org/abs/1712.05884.

[19] *Tacotron 2*. en. URL: https://pytorch.org/hub/nvidia_deeplearningexamples_tacotron2/.

[20] *Text to Speech & AI Voice Generator | ElevenLabs*. URL: https://elevenlabs.io/.
[21] Jörg Tiedemann and Santhosh Thottingal. "OPUS-MT – Building open translation services for the World". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Ed. by André Martins et al. Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61.
[22] Rea Tot and Ivan Dunđer. "Primjena i značaj procesa lokalizacije video igaraApplication and significance of the video game localization process". In: *Elektronički zbornik radova Veleučilišta u Šibeniku* 15 (July 2021), pp. 93–105. DOI: 10.51650/ezrvs.15.1-2.6.
[23] *Umbre*. en. Page Version ID: 1200941476. Jan. 2024. URL: https://en.wikipedia.org/w/index.php?title=Umbre&oldid=1200941476.
[24] *Video Game Market Size, Share And Growth Report, 2030*. en. URL: https://www.grandviewresearch.com/industry-analysis/video-game-market.
[25] *Yakitori Audio Converter - Convert fuz-xwm-wav-various audio files*. en. Oct. 2020. URL: https://www.nexusmods.com/skyrim/mods/73100.

A AI

During the preparation of this work, the author used Grammarly, Phind 70b, Claude 3.5 Sonnet, and Github Copilot to correct the grammar of the text, format the text in LaTEX, find abstracts and information about specific academic papers, and fix bugs in the pipeline code. After using this tool/service, the author(s) reviewed and edited the content as needed and took (s) full responsibility for the content of the work.