


## Bachelor Thesis

<b>Title of Bachelor Thesis (english)</b>	Leveraging AI-Tools for Video Translation in Higher Education
<b>Title of Bachelor Thesis (german)</b>	Einsatz von KI-Werkzeugen für die Videoübersetzung in der Hochschulbildung
<b>Author (last name, first name):</b>	Böck, Simon
<b>Student ID number:</b>	12115066
<b>Degree program:</b>	Bachelor of Science (WU), BSc (WU) 
<b>Examiner (degree, first name, last name):</b>	Univ.-Prof., Reka Marta, Sabou, Ph.D.


I hereby declare that:

1. I have written this Bachelor thesis myself, independently and without the aid of unfair or unauthorized resources. Whenever content has been taken directly or indirectly from other sources, this has been indicated and the source referenced.
2. This Bachelor Thesis has not been previously presented as an examination paper in this or any other form in Austria or abroad.
3. This Bachelor Thesis is identical with the thesis assessed by the examiner.
4. (only applicable if the thesis was written by more than one author): this Bachelor thesis was written together with

The individual contributions of each writer as well as the co-written passages have been indicated.

20/09/2024

Date

  
Unterschrift

Bachelor Thesis

# Leveraging AI-Tools for Video Translation in Higher Education

Simon Böck

Date of Birth: 15.11.2002

Student ID: 12115066

**Subject Area:** Information Systems

**Studienkennzahl:** UJ033561

**Supervisor:** Univ.-Prof., Reka Marta Sabou, Ph.D.

**Co-Supervisor:** Michael Feurstein, MSc.

**Date of Submission:** 23. September 2024

*Department of Information Systems & Operations Management, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria*

# Contents

<b>1</b>	<b>Introduction</b>	<b>6</b>
1.1	Background and Motivation of the Study . . . . .	6
1.2	Statement of the Problem . . . . .	6
1.3	Research Objectives . . . . .	7
1.4	Structure of the Thesis . . . . .	7
<b>2</b>	<b>Related Work</b>	<b>9</b>
2.1	Audiovisual Translation . . . . .	9
2.1.1	Dubbing . . . . .	10
2.1.2	Framework: TRAVID . . . . .	11
2.2	Evaluation Methods for Machine Translation Algorithms . . .	11
2.2.1	Manual Evaluation Metrics . . . . .	12
2.2.2	Automatic Evaluation Metrics . . . . .	13
<b>3</b>	<b>Methodology</b>	<b>14</b>
3.1	Data Source . . . . .	14
3.2	Research Design . . . . .	15
3.2.1	Online Service Provider . . . . .	17
3.2.2	Custom-Built Pipeline . . . . .	18
3.2.3	Human Translation . . . . .	18
3.2.4	Machine Translation Evaluation . . . . .	19
<b>4</b>	<b>Results</b>	<b>24</b>
4.1	Video Output . . . . .	24
4.2	Metric Scores . . . . .	24
<b>5</b>	<b>Discussion</b>	<b>28</b>
5.1	Practical Assessment of the Dubbing Output . . . . .	28
5.1.1	Tonal Quality . . . . .	29
5.1.2	Hallucinations: Content Deviation and Terminology . .	30
5.2	Implications for Higher Education . . . . .	31
5.2.1	Pricing and Scalability . . . . .	32
5.2.2	State-of-the-Art Usability and Integration . . . . .	32
<b>6</b>	<b>Legal Considerations</b>	<b>33</b>
6.1	Data Privacy . . . . .	33
6.2	Intellectual Property . . . . .	34
<b>7</b>	<b>Conclusion</b>	<b>35</b>

## List of Figures

1	Steps Involved Within This Thesis . . . . .	16
---	---	----

## List of Tables

1	Metric Scores for the Different Solutions . . . . .	25
---	---	----

## **Abstract**

This thesis investigates the use of AI-based translation tools for educational video content in higher education, focussing on the feasibility and challenges associated with machine translation and dubbing. Two translation approaches were compared: a commercially available AI-based video translation system (HeyGen) and a custom-built pipeline. The research evaluates the output of both systems using BLEU, chrF, and COMET metrics, exploring the accuracy and naturalness of the translations. Key findings indicate that HeyGen outperformed the custom pipeline in all metrics, but significant issues, such as inconsistencies in output length, unnatural tone, and hallucinations in the output, raise questions about the reliability of AI-driven solutions in their current state. Although these AI tools show promise, particularly for short, prerecorded videos, they remain unsuitable for full-length lecture recordings due to unpredictability and error proneness. The study further explores legal and ethical considerations, such as data privacy and intellectual property, providing recommendations for the responsible use of AI translation technologies in educational settings.

# 1 Introduction

## 1.1 Background and Motivation of the Study

In recent years, the rise of machine learning (ML) algorithms has been particularly noticeable. With technologies and tools evolving faster than ever and the rising popularity of Artificial Intelligence (AI), many new ways of using such technology are unfolding. This change is influencing not only the daily lives of many or the way we carry out our work and commitments, but also the education system [21].

Talking about universities, another important niche, which is also affected by the rapid rise of AI, is the video translation industry. Initially, video translation might be considered similar to traditional dubbing in television, where the focus is on overlaying the target language audio, often without concern for lip sync or precise alignment with the original visual elements [26].

However, advances in AI and related technologies have revolutionised this process, enabling automated lip-syncing that closely matches the mouth movements of the speakers with the translated dialogue [33]. This ensures that the translated content appears more natural and seamless. On several online platforms such as YouTube, many videos are circulating, indicating that the use of such AI tools for video translation is as simple as ever and that there are little to no negative side effects [23]. Since this technology sounds very promising, especially in regard to its application in higher education, it is crucial to further investigate about this topic.

## 1.2 Statement of the Problem

The traditional approach to providing lectures in multiple languages often involves re-recording sessions with translated scripts, which is both time-consuming and costly. This method requires additional resources, including studio time and post-production efforts, which can quickly escalate expenses. Alternatively, leveraging AI-driven tools potentially offers a more efficient solution by utilising existing video content and automatically translating it. However, while AI is often touted as a simple and resource-efficient option, the actual process and results may involve complexities and limitations that have not been fully explored. This thesis aims to dive into these aspects, evaluating the feasibility and effectiveness of AI to support multilingual education while addressing the potential challenges that arise during its implementation.

### 1.3 Research Objectives

In this Bachelor thesis, we will analyse how feasible it is to support higher education with the resources currently available on the market. In an educational setting, multilingual teaching, in particular, is important as many societies become increasingly interconnected with each other and the adoption of languages, especially English, becomes crucial for universities [3].

In order to shed light on this topic, we will carry out a comparative analysis. For this, we will compare the video translation output of an online service provider and a custom-created pipeline. The research objectives for this thesis are defined as the following:

1. Assess the feasibility of AI-based video translation solutions in higher education.
2. Evaluate and compare the AI-based video translation output against the custom-created pipeline translation using relevant metrics.
3. Explore the impact of AI translation tools on scalability and cost-effectiveness in higher education and address potential limitations.
4. Evaluate legal and ethical considerations for AI-driven translations in higher education.

### 1.4 Structure of the Thesis

The thesis begins with a Related Work chapter, which reviews the existing literature and frameworks relevant to video translation and machine translation evaluation, providing the necessary context for the research. Following this, the Methodology chapter details the research design, including the use of a custom-built pipeline as well as the HeyGen selection process for video translation.

The Results chapter presents the findings of the evaluation of the translated video segments, using metrics such as BLEU, chrF, and COMET as quality measurement. In the Discussion chapter, the video results are analysed in depth, with topics such as scalability and pricing, while placing particular focus on the challenges related to dubbing. The chapter on legal considerations examines the implications of using AI-driven translation tools, including concerns related to data privacy and the ethical use of this technology. Finally, the Conclusion chapter summarises the key findings and reflects on the feasibility of the approaches evaluated.

The Jupyter notebooks containing the code for this thesis can be found in the following GitHub repository:

[https://github.com/simboeck/bachelor\\_thesis](https://github.com/simboeck/bachelor_thesis)



## 2 Related Work

This chapter reviews the existing literature on Audiovisual Translation (AVT) and Machine Translation (MT), focussing on the key developments and evaluation methods that have shaped these fields. We will particularly look at the context of dubbing and automated translation frameworks and further discuss the evolution of evaluation metrics used to assess machine translation quality.

### 2.1 Audiovisual Translation

Audiovisual Translation, often referred to as AVT, is a specialised field within Translation Studies that deals with the transfer of multimodal and multimedia texts from one language to another [7]. Unlike traditional text-based translation, AVT encompasses the translation of content that combines visual and auditory elements, such as films, television programmes, video games, or online media and is often used as an umbrella term to specify screen translation, multimedia translation, multimodal translation, and film translation (as cited in) [28].

In his overview of the field, Chaume [7] further emphasises that AVT is inherently more complex than traditional forms of translation due to its multimodal nature. This multimodality introduces unique challenges, such as the need to synchronise the translated dialogue with the visual elements on screen, including the lip movements of actors and scene timing, ensuring coherence between the soundtrack and the translated text. These limitations require not only linguistic expertise, but also a deep understanding of the audiovisual field.

According to Gambier [14], digital technologies made a shift to multimodality possible. With the evolution of communication shifting into integrating multiple nodes, while compared to traditional translation processes, which only take text into account, now the visual as well as the audio becomes more relevant than ever. Gambier further discusses how user behaviour shifted with content consumption (social media, videos, and streaming), indicating a strong shift to AVT becoming more important than ever for translated content. In addition, the integration of new technologies, such as artificial intelligence, is significant in shaping the future of AVT.

Granell and Chaume [15] discuss this evolution of technology and how early solutions such as MT or computer-assisted translation (CAT) are being integrated into current AVT solutions to enhance efficiency. In addition,

they outline the importance of human-machine convergence, discussing how new technologies will impact certain work fields. In addition, they highlight the increasing importance of cloud-based solutions in fostering new developments, as these platforms diminish the importance of geographical distances, encouraging a more interconnected world.

### 2.1.1 Dubbing

What separates AVT from most other translational studies is the multiple streams of input that must be considered within the translation process. One of the first studies on visual influence on speech perception was the work of McGurk and MacDonald [30], who showed that visual cues, such as lip movements, can significantly alter how speech sounds are perceived by the listener. This study introduced what is now known as the McGurk effect, where conflicting visual and auditory inputs lead to a third entirely different perceptual experience. This phenomenon underscores the complexity of AVT, where translators must account for the intricate interplay between auditory and visual components to maintain the integrity and clarity of translated content. This integration of both audio and visual elements is commonly referred to as dubbing. In order to gain an understanding of several challenges referred to dubbing, we will take a look at the framework of Wu et al. [40]:

An important aspect is the key differences between video dubbing systems and pure Speech-to-Speech solutions. Their solutions split the task of dubbing a video into three sub tasks, consisting of Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and a Text-to-Speech (TTS) process. The most important and commonly referred to as the hardest part of AVT is the length synchronisation between the input and target languages. Since the output text should match the lip movements of the original video, it is important to choose the right words, as the wrong translation of words could end in a completely desynchronised video result. This is where NMT comes into play, as explained by Wu et al. [40] seeking to find the optimal translation into the target language while also considering the length of the words. In addition, they introduce a so-called pause token [P], with which the model controls the output length. By deliberately inserting pauses when the pacing of the translated text seems to be off, they can control the rhythm as well as the pacing in order to get a natural flow as an outcome. In addition to the pause token, they also add a duration predictor, which enables them to predict the duration of each word. This greatly helps to match the output length of words with the original text. TTS is then another tool for further influencing the length of the synthesised

speech by adjusting the duration of pauses made between certain words, as well as the duration of words themselves. This is achieved by introducing the TTS model 'AdaSpeech 4'. The advantage of this model is that they do not need to scale everything down by the same factor but rather adjust the duration depending on whether it is a vowel or consonant. Due to these two approaches, they are able to control the length of output with which they are able to match the video. Preliminary results from evaluating available frameworks indicate that this approach is a robust and reliable system that outperforms other baseline models when applied to real-world data.

### **2.1.2 Framework: TRAVID**

One problem which comes with AI based solutions for audiovisual translation is that they are not traceable for the user since they function just like a black-box. Adhikary et al. [2] provide an overview to gain a deeper understanding of how these AI-based solutions operate behind the scenes. This paper discusses the complexities and methodologies involved in advanced AI systems such as TRAVID. TRAVID, an end-to-end video translation framework, exemplifies how multimodal AI systems translate and synchronise spoken language across different languages while maintaining lip synchronisation. This approach enhances the realism and effectiveness of translated communication, particularly in educational settings, by incorporating visual cues that are essential for understanding nuanced speech. When we examine such systems, we can better appreciate the intricate processes and challenges involved in creating AI-driven translations that go beyond traditional text-to-text translation [2].

## **2.2 Evaluation Methods for Machine Translation Algorithms**

As interest in Machine Translation (MT) has grown over the last few decades, different ways of evaluating it have developed. What were initially rather static systems using statistical methods have evolved into advanced approaches, particularly in modern times, using neural network-based techniques.

The first reports on the evaluation of MT underline the importance of (1) intelligibility and (2) fidelity as the main characteristics of a machine output compared to its original text [4]. This means that the translated output should convey the same meaning as the original text, with minimal alterations or distortions. In general, there are two ways to evaluate machine-translated texts: manual and automatic evaluation methods.

### 2.2.1 Manual Evaluation Metrics

When looking at manual evaluation methods, a very common method is to evaluate machine-translated text based on fluency and adequacy [38]. These judgements can vary in all aspects, depending on specific research designs. Some studies document different machine-translated outputs, which then should be ranked from best to worst, also referred to as side-by-side comparisons, while others assess the quality of the translation based on a 5- or 7-point scale for attributes such as fluency, adequacy, or usability [9].

Han et al. [16] outline that the main problems with manual evaluation methods are time consumption, expenses, and variation within the result. The latter is understood as the assessment, or rather the sharpness of the evaluation, which is in the eye of the beholder. Koehn and Monz [25] also outline that a problem resulting from the use of this scale evaluation approach is the variance in the result. What one person may consider a poor translation in terms of fluidity of the text, another person may find more acceptable.

Another widely used method is post-editing. Post-editing can be seen as the task of correcting a machine-translated text in order to create a text of publishable quality [9]. Fischer et al. [12] outline that in addition to correcting superficial errors, post-editing also involves addressing deeper issues such as terminology consistency, omissions, and typographical errors, which can be just as common in professional human translation as in machine translation. A well-known framework for evaluating translation quality, including post-editing output, is the Localisation Industry Standards Association (LISA) model, which defines a wide range of error types and severity levels, such as minor, major and critical, allowing for a structured and systematic evaluation process [12]. Evaluation can also be performed using different measurements, such as the time it takes to edit the text or by documenting the shortest number of changes made, such as insertions, deletions, or substitutions [16]. This version of post-editing is the so-called HTER (Human Translation Error Rate), which is defined as following:

$$HTER [16] = \frac{\sum (number\ of\ editing\ steps)}{number\ of\ words\ in\ the\ acceptable\ translation} \quad (1)$$

Interestingly, studies have shown that different types of machine translation errors, such as coherence or structural issues, can significantly impact the post-editing effort, highlighting the complexity and especially the critical role of post-editing in such translation workflows [17].

### 2.2.2 Automatic Evaluation Metrics

The increasing amount of data and scalability has led to the development of automatic evaluation methods [39]. When looking at different automatic evaluation metrics, they typically fall into two main categories: lexical similarity metrics and linguistic feature-based metrics [16].

Lexical similarity methods, including metrics such as BLEU<sup>1</sup> (bilingual evaluation understudy) and chrF<sup>2</sup> (CHaRacter-level F-score), compare machine-generated translations against human reference translations based on word overlap, word order, and editing distance [31] [32]. Although these methods are efficient, they have limitations, particularly in handling linguistic variability and capturing the semantic nuances of translations [16].

Linguistic feature-based metrics, on the other hand, offer a more nuanced approach to evaluating machine translations. Unlike traditional metrics such as BLEU, linguistic feature-based metrics dive deeper into the syntactic and semantic qualities of the translation, going beyond surface-level word matching [16]. These metrics evaluate how well the translated text captures the meaning, structure, and grammatical correctness of the source text. An example of such a metric is COMET<sup>3</sup> (Crosslingual Optimized Metric for Evaluation of Translation). COMET uses neural networks to model complex linguistic features, including context, word order, and semantic alignment, to provide better evaluation results for translation quality [34].

---

<sup>1</sup><https://github.com/neural-dialogue-metrics/BLEU>

<sup>2</sup><https://github.com/m-popovic/chrF>

<sup>3</sup><https://github.com/Unbabel/COMET>

## 3 Methodology

The Methodology section of this thesis outlines the research design and methods used to evaluate different approaches of video translation in higher education. It details the data sources, the translation tools utilised (including both an online service provider and a custom-built pipeline), and the human translation used as a benchmark. In addition, this section explains the evaluation metrics used to assess the quality of the translation output, providing a comprehensive framework for the analysis conducted within this study.

### 3.1 Data Source

The data for this project are provided by the Institute for Data, Process, and Knowledge Management of the Wirtschaftsuniversität Wien. The data comprise nine MP4 files with a length of roughly ten hours, serving as a video-based lecture series, meaning that they are professionally produced recordings of teaching sessions in higher education. The specific course is called 'Grundlagen der Wirtschaftsinformatik' and is equated with 'the foundations of Management Information Systems.' The language in which the lectures are held is German. The video maintains a simple style with minimal cuts, featuring the lecturer projected on the screen alongside the slides of each unit.

The costs of making these lecture recordings are associated with an extreme effort. In order to produce a 45 minute lecture recording, the actual recording phase in the studio takes approximately five hours, generating video material, which still has to go through post-production (viewing, cutting, etc.). In addition, other costs in the production and post-production have to be considered. Additionally, if we decide to make this lecture international, we would have to translate every script into the target language, meaning increased costs for translation and adapting the content to fit different cultural and educational contexts, further increasing the production expenditures. To ensure that the analysis is manageable and meaningful for this thesis, the relevant video data are limited to a 30-minute segment from the first lecture in the series. The first 10 minutes of this lecture will be used to evaluate the translation output, while the 30-minute segment will serve as a basis for discussing dubbing challenges and other relevant issues that may arise. This selection is intentional, as the first lecture is foundational and designed to introduce key concepts without assuming prior knowledge within this specific field. It provides a clear and constructive starting point that is accessible to a broader audience, making it ideal for our purposes. Focussing

on this specific segment allows us to avoid the complexities and specialised terminology introduced in subsequent lectures, which could complicate the evaluation process. Finally, this does not mean that the selection is entirely free from bias. However, for the purposes of this thesis, it suffices to provide initial analysis results.

## 3.2 Research Design

The research design of this thesis is structured into four key components to evaluate and compare different methods of video translation. The process begins with the translation of an educational video in German through three distinct approaches: (1) a service provider’s translation tool, (2) a custom-built pipeline utilising Application Programming Interfaces (APIs), and (3) a manual translation by a professional translator. Each of these translation processes is performed simultaneously, ensuring a consistent comparison. The final component (4) involves the evaluation<sup>4</sup> of the machine translation output using predefined metrics to assess their quality.

In Section 5.1, we will discuss the limitations associated with the video output of such online service providers.

---

<sup>4</sup>Note that for this study, the author has explicitly chosen to focus on comparing the translation output and not the video output. Tools provided by online service providers offer further services such as voice cloning and lip-syncing. Since these techniques cannot be easily replicated in a custom pipeline, the scope of this study is limited to comparing the output of text-to-text translation.

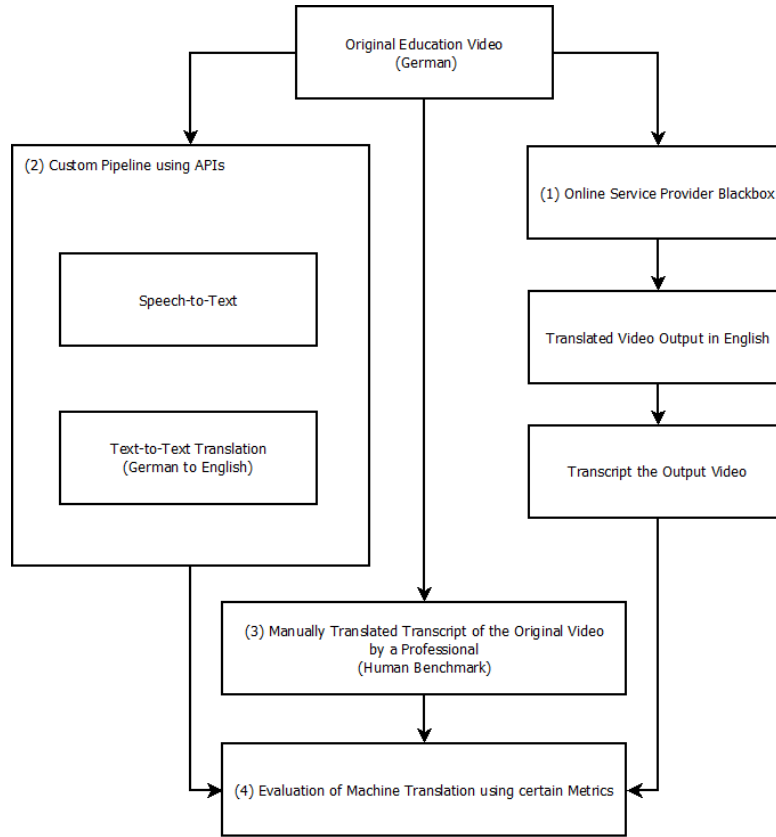


Figure 1: Steps Involved Within This Thesis

Figure 1 represents an illustration of the research design process, showing the steps involved in translating and evaluating the original educational video. The process includes an online service provider, the use of a custom pipeline, and a professional translator, followed by the evaluation of the translation output using specific metrics.



### 3.2.1 Online Service Provider

In order to evaluate the feasibility of outsourcing the process of creating the educational video in English, we need to identify a suitable provider. When searching for the term 'video translation AI'<sup>5</sup> we are presented with multiple online providers that sell similar solutions of translating videos:

- RaskAI<sup>6</sup>
- Dubly.AI<sup>7</sup>
- Wavel AI<sup>8</sup>
- HeyGen<sup>9</sup>

Of these options, we found HeyGen (HeyGen Los Angeles, California, USA) to be the most suitable for our needs. HeyGen stands out for several reasons. Not only do they advertise high-quality video translation with AI-driven accuracy, but also additional features that enhance the overall video production process. These features include voice cloning, lip-syncing, and avatar creation, which should ensure that the translated video maintains a natural and engaging presentation. Among all providers that we evaluated, HeyGen was selected for its particularly impressive demonstration of results and the features it offers.

Moreover, compared to other providers, HeyGen appears to deliver the most promising results based on the information available on their website. The website's advanced AI capabilities and user-friendly interface suggest that it can produce videos that closely match the quality and tone of the original content, making it an ideal choice for this thesis.

We now turn to what we were able to accomplish with it given the limitations of the pricing model. Despite these constraints, we managed to translate one 30-minute video with a US accent and two 10-minute videos with a UK accent. Apart from the length, the input video content was identical for all outputs. While the 30-minute US-accented video quality will be evaluated using our predefined metrics, the two UK-accented videos were produced to explore whether the output from HeyGen's black-box changes when the same input is provided multiple times.

---

<sup>5</sup><https://www.google.com/search?q=video+translation+AI>

<sup>6</sup><https://de.rask.ai>

<sup>7</sup><https://dubly.ai/>

<sup>8</sup><https://de.wavel.ai/>

<sup>9</sup><https://www.heygen.com>

### 3.2.2 Custom-Built Pipeline

In order to validate the result of the online provider, we created a pipeline consisting of two API accesses: the Speech-to-Text API by Google Cloud<sup>10</sup> and the DeepL API, provided by DeepL Translate<sup>11</sup>.

APIs play a crucial role in modern software development, allowing different software systems to communicate with each other by defining a set of rules and protocols. APIs enable the automation of complex tasks that would otherwise require significant manual effort, making them useful tools for streamlining workflows and ensuring accuracy in various tasks.

Although a deep understanding of APIs is not essential to follow the methodology of this thesis, an overview is provided here for those interested in the underlying technical details of how these services interact based on real-world implementations [36].

The goal is to extract the audio of the data, which is then transcribed and translated by the API services. After the process, we should be left with a translated text, correlating to what is said in the original video.

### 3.2.3 Human Translation

In order to let certain automatic evaluation metrics classify both the quality of the provider’s output and the output of our custom pipeline, we need a human translated text as a reference point [29].

Given that we had the 30-minute output video provided by HeyGen, our original plan was to transcribe and translate the entire 30 minutes for evaluation. However, due to the limitations inherent in this study, we decided to focus on the first 10 minutes for the automatic evaluation metrics. This shorter segment was chosen to ensure a manageable scope for the detailed analysis of the translation quality.

However, the full 30-minute video will remain part of the discussion in terms of visual evaluation and will be referenced when addressing specific challenges and issues related to dubbing in the video output.

The video section relevant for our purpose was transcribed and then handed over to a professional translator (BA). The German transcript comprised a total of 1288 words, which resulted in an English output text of

---

<sup>10</sup><https://cloud.google.com/speech-to-text/docs/speech-to-text-requests>

<sup>11</sup><https://www.deepl.com/en/pro-api/>

1153 words after the translation process. The result of this translation process now serves as a human benchmark for the automatic evaluation metrics, which can now be computed.

### 3.2.4 Machine Translation Evaluation

The last but also most important aspect of our research design is the evaluation of these translation outputs. As discussed in Section 2.2, there are many different ways of evaluating machine translations. With the available resources and time constraints we have agreed to evaluate the output of both the service provider’s translation tool, as well as the custom-built pipeline translation purely based off automatic evaluation metrics.

In order to provide a broad insight, we agreed on using three metrics, which will cover both the lexical similarity metrics and the linguistic feature-based metrics [16]:

#### 1. BLEU<sup>1</sup> (bilingual evaluation understudy):

The BLEU score is a widely used metric for evaluating the quality of machine-translated text by comparing it to one or more human reference translations [39]. The score ranges from 0 to 1, with higher scores indicating closer alignment between the machine translation and the reference translations. BLEU is calculated based on n-gram precision, with adjustments made for sentence length through a brevity penalty [31]. To illustrate, consider the following example:

- **Candidate Text:** "The cat is on the mat."
- **Reference Text 1:** "There is a cat on the mat."
- **Reference Text 2:** "A cat is sitting on the mat."

In this case, the BLEU score would be computed by comparing the n-grams in the candidate text against those in the reference texts. For instance, unigram ('the', 'cat', 'is', 'on') precision and bigram (pair of words: 'the' and 'cat', 'cat' and 'is') precision are calculated as follows:

- **Unigram Precision:**  $\frac{5}{6} \approx 0.833$
- **Bigram Precision:**  $\frac{3}{5} = 0.6$

The overall BLEU score is then given by:

$$BLEU [31] = BP \times \exp \left( \frac{1}{N} \sum_{n=1}^N \log p_n \right)^{12} \quad (2)$$

The BLEU score benefits from having multiple reference translations, meaning multiple human translations as benchmark, as this increases the likelihood of matching n-grams, leading to a potentially higher score [39]. For a more detailed explanation of BLEU and its computation, refer to the official documentation [31].

BLEU scores, in our case, were calculated using a single reference translation due to the availability of only one human-translated text. Although this provides a useful measure of translation quality, the BLEU score does not fully capture the range of possible valid translations, which could be addressed by using multiple reference translations [31]. Therefore, the results should be interpreted with this limitation in mind.

Reiter’s [35] review of BLEU argues that while BLEU is valuable for diagnostic evaluations of machine translation systems, it should not be relied upon as the sole metric for evaluating translation quality. The validity and reliability of BLEU is further argued to be questionable due to its inconsistent correlation with human evaluations in different contexts, its inability to measure real-world outcomes, and its potential technological biases, particularly as new methods such as neural networks become more prevalent [35].

However, BLEU is considered to be one of the most basic and fundamental methods in this field, which is why we integrate it into our framework [34].

---

<sup>12</sup>where  $N$  is the maximum n-gram length (standard is set to 4),  $p_n$  is the precision for each n-gram length, and BP is the brevity penalty.

## 2. chrF<sup>2</sup> (CHaRacter-level F-score):

Recent studies, including the work of Mathur et al. [29], have highlighted that BLEU, although useful, is not always reliable for fine-grained evaluation, especially when dealing with high-quality translation systems where its correlation with human judgment can be inconsistent or even negative. As a result, there has been a push towards alternative metrics that can address these shortcomings more effectively.

One such metric is chrF, which operates at the character level and has been shown to provide more stable and reliable correlations with human evaluations, particularly for languages with rich morphology [32]. Unlike BLEU, which focuses on word-level n-grams, chrF evaluates character-level n-grams, making it more sensitive to small but meaningful differences in translation quality [32]. For example, consider the same candidate and reference texts used earlier:

- **Candidate Text:** "The cat is on the mat."
- **Reference Text 1:** "There is a cat on the mat."
- **Reference Text 2:** "A cat is sitting on the mat."

Precision (**chrP**) measures how many of the character n-grams in the machine translation are also present in the reference translations, while recall (**chrR**) measures how many of the character n-grams in the reference translations are present in the machine translation [32]. The overall chrF score is computed as the F-score, which balances precision and recall:

$$chrF_{\beta}[32] = (1 + \beta^2) \cdot \frac{chrP \cdot chrR}{\beta^2 \cdot chrP + chrR}^{13} \quad (3)$$

It is language-independent and tokenisation-independent, which means that it can be applied to any language without requiring specific pre-processing steps [32]. For a more detailed explanation of chrF and its computation, refer to the official documentation [32].

---

<sup>13</sup>where  $\beta$  is a parameter that allows assigning more weight to recall than precision. When  $\beta = 1$ , precision and recall are weighted equally.

### 3. COMET<sup>3</sup> (Crosslingual Optimized Metric for Evaluation of Translation):

COMET is a neural network-based framework specifically designed to evaluate machine translation quality by training multilingual models that closely align with human judgment. Unlike traditional metrics such as BLEU and chrF, which rely heavily on n-gram matching, COMET uses cross-lingual embeddings to capture semantic meaning more effectively [34]. This makes COMET particularly powerful in evaluating translations in which word order or specific lexical choices may differ, but the overall meaning is preserved [34].

COMET models are trained using various types of human judgments, such as Direct Assessments (DA), Human-mediated Translation Edit Rate (HTER, 2.2.1), and Multidimensional Quality Metrics (MQM) [34]. The framework incorporates not only the machine-translated text and the reference translation but also the original source text. Using this triplet (source, hypothesis, and reference), COMET can better predict translation quality by evaluating how well the hypothesis aligns with both the source and the reference [34]. For example, if we consider the same candidate and reference texts used in earlier examples:

- **Candidate Text:** "The cat is on the mat."
- **Reference Text:** "A cat is sitting on the mat."
- **Source Text:** "Die Katze ist auf der Matte."<sup>14</sup>

COMET evaluates quality by generating cross-lingual embeddings for each segment (source, hypothesis, reference) and then calculates a quality score based on how closely the hypothesis aligns with the embeddings of the source and reference [34]. The COMET score is calculated using the following model:

$$COMET[34] = f(source, hypothesis, reference)^{15} \quad (4)$$

COMET has consistently shown state-of-the-art performance in machine translation evaluation tasks, outperforming traditional metrics like BLEU and chrF in both segment-level and system-level evaluations. For a more detailed explanation of COMET and its implementation, refer to the official documentation [34].

---

<sup>14</sup>German: "The cat is on the mat."

<sup>15</sup>where  $f$  represents the neural network model that outputs a quality score based on the cross-lingual embeddings of the input.

Evaluating the outcome of machine translation with the three metrics presented above is highly feasible for this thesis because of the complementary strengths and comprehensive coverage that each metric provides. BLEU, as one of the most established metrics, is highly valued for its light-weighted and fast computation [34]. However, as highlighted by Mathur et al. [29], BLEU’s reliance on n-gram overlap can sometimes lead to inconsistencies, particularly in high-quality translations where word order and lexical choices may differ without impacting the overall meaning. ChrF addresses some of these limitations by operating at the character level, allowing it to capture finer linguistic nuances, making it an ideal complement to BLEU [32].

Finally, COMET represents the latest advancement in translation evaluation, leveraging neural networks and cross-lingual embeddings to model semantic similarity more accurately. As Reiter [35] notes, metrics that incorporate deeper linguistic features, such as COMET, are crucial to achieving evaluations that correlate more closely with human judgments, especially in the context of modern neural machine translation systems [34].

Using these three metrics, this thesis ensures a robust and well-rounded evaluation of the translation output, covering a wide range of linguistic aspects from surface-level precision to deep semantic alignment. This particular combination of metrics was also used in several other studies such as that of Jon and Bojar [20] or Salesky et al. [37]. Based on the results, we can further draw conclusions on which kind of metrics is best suited for evaluation when it comes to the field of machine translation in higher education.

## 4 Results

### 4.1 Video Output

Before we move on to the section where the quality of the translation is evaluated, it is important to mention the video solutions produced by HeyGen. This is crucial because the evaluation of the translation quality is just one aspect of our assessment, with the visual elements also playing a significant role in the overall measurement.

Given the budgetary constraints of this study, we were able to produce three distinct video outputs: one 30-minute video with a US accent and two 10-minute videos with a UK accent. Although the content of the input video was identical across these outputs, the videos differed in length. This approach allowed us not only to assess the translation quality of a full-length video but also to investigate whether HeyGen’s system produces consistent results when processing the same input multiple times. The following videos were produced:

- Input video (length: 30:58) → output length: 28:08 (US)
- Input video (length: 09:56) → two outputs lengths: 09:14/09:06 (UK)

After presenting and evaluating the results of the proposed evaluation metrics, we will go more into depth about the problems and limitations that occur regarding the HeyGen video output. In Section 5.1, we will specifically discuss the current situation in the field of dubbing and highlight the issues with the corresponding video clips of the output.

### 4.2 Metric Scores

The evaluation metrics were calculated using the code available in the GitHub repository (see Section 1.4). The only preprocessing required involved aligning the lines in the text files with the human reference text file. This alignment was necessary because the evaluation scores compare the machine translation directly with the human-translated text [31]. Since machine translations occasionally split or merge sentences differently than human translation, we ensured that the text files were split by paragraphs. This approach allowed us to maintain consistency in the meaning of each paragraph being compared.



The following table presents the metric scores for the different solutions<sup>16</sup>.

	HeyGen	Custom Pipeline
<b>BLEU</b>	0.29	0.23
<b>chrF</b>	0.646	0.597
<b>COMET</b>	0.7715	0.7075

Table 1: Metric Scores for the Different Solutions

What immediately stands out is that (1) the HeyGen translation outperforms the custom pipeline in all evaluation metrics and (2) that the BLEU score in particular seems to have a relatively low score compared to the other two. To gain a deeper understanding of the reasons behind this result, we first need to discuss what constitutes a good score. This involves understanding the thresholds that define the quality of translation for each of the different evaluation metrics.

According to Reiter [35] the thresholds for the BLEU score are defined as follows:

- *High*: Correlation is  $\geq 0.85$
- *Medium*: Correlation is  $0.70 \leq x < 0.85$
- *Low*: Correlation is  $0 < x < 0.70$
- *Negative*: Correlation is  $< 0$

This would classify our result as a low correlation between the machine translation and the human translation. Although this classification sounds relatively harsh, it should be interpreted with caution. The BLEU metric is designed to be used with multiple reference translations for each sentence, increasing the likelihood of capturing a correct translation [29]. In our study, we used only a single reference text, which is acceptable but not optimal.

Furthermore, as demonstrated in the BLEU paper [31], even a professional translator can achieve a score as low as 0.2571 when compared against two other reference translations. This highlights that the scores we obtained in our evaluation of 0.29 and even 0.23 are not as poor as they might initially appear.

---

<sup>16</sup>The scores from HeyGen reflect the video computed with an US accent. Only the transcript of the first 10 minutes was used for the evaluation in this test.

To better understand the differences in the evaluation metrics, it is important to note that ChrF is a metric that acts very similarly to the BLEU metric, but instead of comparing the candidate and reference texts at the word level, it does so at the character level. The chrF documentation by Popović [32] also points out that the metric performs better in languages with rich morphology and outperforms metrics such as BLEU. Another important aspect is that chrF is completely language and tokenisation independent. In languages with rich morphology, such as German, character-level models just like chrF have been shown to outperform word-level models due to their ability to better capture the nuances of morphological variations and partial matches [24]. This characteristic of chrF could explain why it produced significantly higher scores than BLEU in this study.

A study by Mathur et al. [29] underscores that chrF consistently outperforms BLEU, particularly when evaluating translation quality across different systems. The study suggests that chrF should be preferred over BLEU due to its better alignment with human judgments, as it reduces the likelihood of errors that BLEU might introduce, especially in scenarios where translation quality is high. This shows that the significantly higher chrF scores observed in this study are an indication of its robustness and reliability in handling complex linguistic variations, further confirming its use alongside or in place of BLEU.

When we examine the COMET score, we immediately see that it outperformed the other scores by a relatively high margin. COMET uses a cross-lingual transformer model, such as XLM-RoBERTa, that encodes both the source text and the machine translation into a shared embedding space. This approach allows COMET to capture more nuanced aspects of translation quality, including semantic similarity and contextual appropriateness, which are often missed by surface-level metrics such as BLEU and chrF [34].

Salesky et al. [37] study highlights the superiority of COMET over BLEU, particularly in multilingual and complex translation scenarios. The closer alignment of COMET with human judgment and its ability to account for semantic nuances and contextual information make it a more reliable metric for evaluating translations in diverse language pairs. These findings help explain why, in our experiment, COMET not only surpassed BLEU but also outperformed chrF, suggesting that COMET’s advanced model-based approach provides a more comprehensive evaluation of the translation quality in our specific context of German to English.

Furthermore, the recent study by Fernandes et al. [11] underscores the importance of using more sophisticated metrics like COMET, particularly in neural machine translation (NMT). The study highlights that COMET, when used in quality-aware decoding, not only aligns more closely with human evaluations but also significantly improves translation quality across different language pairs. The research findings suggest that COMET should be chosen over BLEU due to its better correlation with human judgments and its ability to capture the context of the translated text more effectively.

The results of our study further reinforce the effectiveness of the HeyGen translation system, as it outperformed our custom pipeline in all evaluation metrics. However, it should be noted that our custom pipeline only focuses on the text-to-text translation output and does not go further into features provided by HeyGen, such as voice cloning or lip-syncing. Taking this into account, comparing HeyGen and our custom pipeline, one has to consider that based on technological infrastructure, the custom pipeline performs nearly as well as an industry service. The custom pipeline still delivered respectable results, which is commendable. This suggests that while HeyGen provides superior translation quality, simpler and more accessible systems can also deliver competitive results, making it a viable option in contexts where resources are limited.

In summary, the findings of our study reveal that the translation scores generated align well with current research in the field of machine translation. The chrF and COMET scores notably outperformed BLEU, which is consistent with the broader body of research that indicates that these more advanced metrics offer better alignment with human judgments, particularly in complex linguistic scenarios such as German to English translation and when performed with a single reference text as human benchmark [37] [11] [29]. The ability of chrF to capture morphological nuances and the semantic and contextual sensitivity of COMET provide plausible explanations for their higher scores in our evaluations [24]. Despite these encouraging results, it is important to acknowledge that the overall translation quality still falls short of being classified as high or near-perfect, underscoring the ongoing challenges and areas for improvement in machine translation systems. However, the correlation between our findings and established research supports the validity of our results and suggests that, while progress has been made, further advances are necessary to achieve consistently high-quality translations.

In Section 5.2, we will discuss the implications of these findings for settings in higher education, specifically addressing whether the current quality of machine translation is sufficient to meet the demands of academic contexts and if not, what improvements are needed to reach that standard.

## 5 Discussion

In the last chapter, we analysed and interpreted the evaluation results of the translations. In order to provide a complete picture, we must also analyse the visual part. In the following sections, we will look at any issues that we noticed when analysing the video output. We will also cover topics such as the costs resulting from this output and the general pricing model of HeyGen.

Moreover, we will assess the implications for higher education and consider whether outsourcing audiovisual translation to systems like HeyGen is a viable alternative to traditional methods.

### 5.1 Practical Assessment of the Dubbing Output

As we were unable to find an evaluation method for dubbing that could be realised within the constraints, we agreed to evaluate the video manually. We watched the video sections and noted fundamental errors in the output. To illustrate this, we will reference the following sequences and use the available literature to find possible explanations for these errors.

First of all, we will discuss concerns around the two video outputs generated with a UK accent. Although the input video, which was processed by the HeyGen system, was identical in length, the resulting output videos differed by a total of eight seconds. This discrepancy may appear relatively minor. However, upon closer examination, it raises significant questions about the consistency and reliability of the translation process within such a framework. In a system designed for professional use, such variability is unexpected and problematic. The fact that one input video can produce outputs of varying length suggests an underlying inconsistency in the translation algorithm or processing mechanics. This inconsistency could lead to unpredictable results, where the quality of the translation could vary with each run. This behaviour is particularly concerning for businesses and professional users who rely on these tools for accurate and stable translations. It introduces an element of uncertainty, almost like a lottery, where the outcome may or may not meet the necessary standards.

The issues observed in the HeyGen outputs are symptomatic for the broader challenges associated with black-box AI systems, where the decision-making process of the model is not transparent and often not fully understood by the users. As highlighted by Castelvechi [6], black-box systems can lead to unexpected and unexplained outcomes because knowledge is "baked into the network" rather than being accessible by humans. This unpredictability challenges the credibility of the service and raises the question of its suitability for professional applications where precision is essential.

### 5.1.1 Tonal Quality

The discrepancies extend beyond the eight-second difference. One of the videos exhibits an odd tonal quality that reduces the overall naturalness of the output. In the following video section<sup>17</sup>, the system attempts to mimic the natural human cadence of handling sentences. Typically, when a sentence is prolonged, the pitch of the voice may increase slightly as the speaker runs out of breath, a subtle detail that adds to the authenticity of the speech, one of many so-called prosodic features [1] [13]. However, the system seems to fail in recognising the end of a sentence. This results in a strange sequence lasting more than a minute, where the pitch of the voice continues to rise unnaturally, without the expected pauses or drops in tone that indicate sentence endings. This escalating pitch creates an increasingly odd and artificial sound, which not only lowers the quality of the translation but also calls into question the system's ability to accurately process and reproduce human speech patterns. This flaw could significantly affect the user experience, particularly in professional settings where the naturalness and clarity of speech are crucial. This issue is consistent with the findings of Cutler's research [8], which highlights how errors in stress and intonation, particularly the failure to apply proper intonation contours at sentence endings, can lead to artificial and disjointed speech, significantly affecting the perceived naturalness of the output.

Another notable issue related to prosodic features is the unnatural pronunciation observed in a section of the video with a US accent<sup>18</sup>. In this section, the system struggles to produce a natural-sounding speech, leading to a series of mispronunciations that lower the overall quality. This problem is particularly distinctive in the section where the lecturer's face is not visible and only the slides are displayed. Lack of visual impressions may increase the problem, as the system seems to lose context without facial expressions

---

<sup>17</sup><https://www.youtube.com/watch?v=fBCE4tCF01Q>

<sup>18</sup>[https://www.youtube.com/watch?v=\\_kGIx32Czfs](https://www.youtube.com/watch?v=_kGIx32Czfs)

or lip movements of the speaker to guide speech synthesis [30]. According to Bohacek and Farid [5], models that rely on synchronising audio and video can encounter significant challenges when the speaker’s lips are not clearly visible, resulting in mismatches that lead to unnatural or disjointed speech output. In addition, there are long unnatural pauses between sentences, as well as sentences that are unnaturally rushed without appropriate pauses. These issues contribute to an artificial sound, which can significantly worsen user experience, especially in professional or educational settings where clear and natural communication is essential [8].

### 5.1.2 Hallucinations: Content Deviation and Terminology

Another observation involves the system experiencing hallucinations, where the generated speech content deviates clearly from the original input, in a sense degenerating the input [19]. For example, in one video sequence<sup>19</sup>, the word "systems" was prolonged, while in another instance the number "2001" was randomly inserted without any basis in the source material. These types of error are a known issue in Neural Machine Translation (NMT) systems, often referred to as hallucinations. As highlighted in the research by Lee et al. [27], these errors occur when the translation is untethered from the input, resulting in an output that is fluent but incorrect or entirely fabricated. Hallucinations can significantly undermine the reliability of the system, leading to mistrust and potential misuse of the translated content. The presence of such errors in the HeyGen output indicates potential weaknesses in the system’s ability to maintain fidelity to the input content, especially under challenging conditions.

The last issue we will discuss relates to specific terminology, which, although not as significant as the other problems, still deserves attention. For instance, the term "Grundlagen der Wirtschaftsinformatik" was inaccurately translated to "fundamentals of business informatics" instead of the more contextually appropriate "management information systems." According to Dinu et al. [10], this misrepresentation of domain-specific terminology highlights a weakness in the translation process, as neural machine translation systems can struggle to accurately incorporate specialised terms, especially when they are not part of standard training data. The paper furthermore highlights that training neural machine translation systems to properly apply terminology constraints can mitigate such issues by teaching models to integrate specific terms more accurately during translation. Although this error is relatively minor compared to other issues such as prosody or hallucinations, it can still

---

<sup>19</sup><https://www.youtube.com/watch?v=V1MEusLkhz8>

lead to confusion in academic or professional contexts. Fortunately, this specific issue can be addressed using HeyGen’s proofreading feature, available in their enterprise plan, which allows users to review and correct the output, ensuring that such terminological errors are corrected before the final translation is delivered. The different HeyGen plans will be discussed in Section 5.2.1.

In summary, our practical assessment of HeyGen dubbing outputs has revealed several critical challenges that need to be addressed to ensure the reliability and effectiveness of such systems. The most important issues include inconsistencies in output length, unnatural prosody, and pronunciation, as well as the occurrence of hallucination errors in which the generated content deviates significantly from the original input. These findings highlight the limitations of current neural machine translation systems, particularly in professional and educational contexts. The variability in output length, despite identical inputs, showcases the system’s consistency, making it unsuitable for scenarios where uniformity is crucial. The unnatural tonal qualities and prosodic errors suggest that the system struggles to authentically replicate human speech patterns, especially when visual impressions are limited.

Furthermore, the hallucinations observed, in which the system produced content that was either prolonged or entirely fabricated, pose a serious risk to the accuracy and trustworthiness of the translations. These challenges underscore the need for careful consideration when implementing such tools in higher education, where accuracy and clear communication are important. In Section 5.2.2, we will discuss what needs to be considered to effectively integrate these technologies into academic settings as of now, ensuring that they meet the necessary standards of reliability and quality.

## 5.2 Implications for Higher Education

Based off the findings presented above, it is clear that a complete outsourcing of translation and dubbing of entire lecture series using the current state of AI-based solution from HeyGen is not feasible. The output is too unpredictable and filled with errors, as demonstrated by inconsistencies in video length, prosodic issues, and hallucinations in translated content. These challenges make it difficult to rely on these systems for professional and academic applications where accuracy and naturalness are critical. In the following, we will explore additional limitations, such as the costs involved, and consider settings where the current framework could still add value in higher education.

### 5.2.1 Pricing and Scalability

The pricing structure for HeyGen has recently undergone significant changes. Previously, their pricing model was based on a credit system, where one credit equals one minute of translated video material. The cost of these credits varied depending on the plan chosen, and for this project, we spent \$60 for 30 credits. However, due to an issue we encountered while processing, we were granted an additional 20 credits, which allowed us to produce more content. Without this circumstance, our budget would have only covered 30 minutes of video material. The updated pricing model now moves away from a credit-per-minute system and instead limits users based on the length of videos they can produce under each plan. More details on the updated structure can be found on their pricing page<sup>20</sup>.

Despite the new flexibility in producing video content, there were some noticeable shortcomings in our output, specifically related to incorrect usage of terminology. These issues can be addressed with proofreading, but this feature is only available with their enterprise plan, which costs \$10,000 per year. Although the enterprise plan now offers unlimited video output, the high price point makes it less accessible to smaller educational institutions or individual users. Thus, while the scalability of the system is promising, the high costs and ongoing challenges with output quality pose limitations on its practicality for widespread use in higher education.

### 5.2.2 State-of-the-Art Usability and Integration

To ensure practical application of these advanced dubbing tools in higher education, one approach could be the strategic use of preplanned short videos. These so-called knowledge pills or micro-lessons would be designed with the specific intention of leveraging machine translation and dubbing technology [33]. By keeping videos concise and structuring them to minimise the complex linguistic features that typically pose challenges for translation, educators can reduce the risks of errors such as prosodic mistakes, unnatural pronunciation, and hallucinations.

Short videos that are tailored for this purpose can be used effectively to supplement larger lecture content. These segments are less prone to variability in translation quality, which is a significant concern in longer lecture recordings. This allows for more predictable outputs and better integration into multilingual learning environments.

---

<sup>20</sup><https://www.heygen.com/pricing>



As highlighted by Pérez et al. [33], such technologies are already being successfully implemented in institutions to support blended and online learning by producing multilingual educational content at scale. Although not yet suitable for full-length lecture recordings due to challenges in maintaining consistency and accuracy, these short, tailored videos provide a promising solution to improve the internationalisation in higher education.

## 6 Legal Considerations

The implementation of AI-based translation tools in higher education implies several legal considerations that institutions must address to ensure compliance with relevant laws and regulations. As AI technologies continue to evolve, issues related to data privacy, intellectual property, and liability become increasingly complex. Educational institutions must ensure that AI systems used for translating and content production operate within legal frameworks that protect both proprietary content and user data [22]. In addition, institutions must carefully assess their liability when using AI-generated outputs, particularly in cases where inaccuracies or biases could affect the quality and accuracy of educational materials. In the following, we will explore these legal concerns in greater detail, focussing on the implications of data privacy and intellectual property in the context of AI-driven translation tools.

### 6.1 Data Privacy

One of the key concerns when using AI-based translation tools in higher education is the protection of sensitive data related to educational content and institutional resources. AI systems, especially those used for tasks such as translating and dubbing, handle exclusive data, including lecture content, intellectual property, and confidential internal communications. This raises several privacy concerns, particularly when processing institutional content. According to Kamocki and O'Regan [22], many free machine translation services operate on a model where the input data is processed not only to deliver a translation but also for secondary purposes, such as improving the translation model or for further commercial usage. This data processing represents a risk to the confidentiality of educational material, as there is little control over how the input data is stored or reused by the provider. In addition, Huang [18] also criticises AI systems that often collect and process large amounts of sensitive information, further underlining the risks of data exploitation and breaches.

To mitigate these risks, institutions must implement robust data privacy frameworks that ensure compliance with regulations such as the General Data Protection Regulation (GDPR<sup>21</sup>). Educational institutions must ensure that any AI translation tools used do not compromise the security of the content being processed. Increased transparency in how these AI systems handle, store, and protect data is essential to protect both the institution's proprietary content and the integrity of the educational process.

## 6.2 Intellectual Property

The question of intellectual property (IP) is another important topic when using AI-based translation tools in higher education, as these systems process significant amounts of proprietary educational content, including lectures, research, and academic resources. In this context, authorship and ownership of the translated content must be carefully considered. As highlighted by Gambier [14], the process of producing translated works involves a network of contributors, making it essential to protect the intellectual, moral, and financial rights of all parties involved. This notion extends to the use of AI tools, where multiple stakeholders, including educators, translators, and institutions, contribute to the creation of translated material.

In various international frameworks, including the Berne Convention and more recent directives such as the European Directive of 2019, authors, educators, and institutions are protected by intellectual property rights, which ensure that their work cannot be used, reproduced, or adapted without proper authorisation [14]. Institutions must ensure that AI tools used for translation do not violate on these rights by transferring ownership or allowing unauthorised use of translated materials. Moreover, as AI-generated content may not have a clearly defined "author," institutions must establish clear guidelines on ownership and usage rights of translated content to prevent potential conflicts over intellectual property.

---

<sup>21</sup><https://gdpr-info.eu/>

## 7 Conclusion

This study aimed to explore the feasibility of using AI-driven video translation tools, particularly within the context of higher education. The comparison between HeyGen’s translation system and a custom-built pipeline revealed that while HeyGen outperformed the custom solution across all evaluation metrics, it still exhibited significant issues such as inconsistencies in prosody, random hallucinations, and inaccurate terminology. These shortcomings indicate that current AI-based translation technologies, though promising, are not yet reliable enough to translate entire video-based lecture recordings. However, they hold potential for applications involving shorter, tailored videos, where the risks of translation errors are reduced. Furthermore, this study discussed important legal considerations, including data privacy and intellectual property, underscoring the need for robust frameworks to ensure that these tools are used responsibly.

The findings suggest that while AI-driven solutions offer significant advantages, especially in terms of scalability, further advancements are needed to improve translation quality, particularly for the use in professional and educational settings. In future work, these tools can become the key to multilingual education, provided that the identified limitations are addressed.

We conclude this thesis with two recommendations for higher education settings: (1) fully outsourcing the translation and dubbing of entire lecture recordings using current AI-based solutions, such as HeyGen, is not feasible; and (2) strategically using AI for preplanned short videos (e.g. knowledge pills, micro lessons) is a more viable approach.

## References

- [1] Jordi Adell, Antonio Bonafonte, and David Escudero. Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech. In *Procesamiento del Lenguaje Natural*, pages 277–283. Sociedad Española para el Procesamiento del Lenguaje Natural, Jaén, España, 2005.
- [2] Prottay Kumar Adhikary, Bandaru Sugandhi, Subhojit Ghimire, Santanu Pal, and Partha Pakray. Travid: An end-to-end video translation framework, 2023.
- [3] Philip G. Altbach and Jane Knight. The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education*, 11(3-4):290 – 305, 2007.
- [4] Automatic Language Processing Advisory Committee. Language and machines: Computers in translation and linguistics. Technical report, National Academy of Sciences, Washington, D.C., 1966.
- [5] Matyas Bohacek and Hany Farid. Lost in translation: Lip-sync deepfake detection from audio-video mismatch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4315–4323, 2024.
- [6] Davide Castelvecchi. Can we open the black box of ai? *Nature*, 538(7623):20 – 23, 2016.
- [7] Frederic Chaume. An overview of audiovisual translation: Four methodological turns in a mature discipline. *Journal of Audiovisual Translation*, 1(1):40–63, 2018.
- [8] Anne Cutler. Errors of stress and intonation. In Victoria A. Fromkin, editor, *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand*, pages 67–80. Academic Press, New York, 1980.
- [9] Daems, Joke. *A translation robot for each translator? : a comparative study of manual translation and post-editing of machine translations: process, quality and translator attitude*. PhD thesis, Ghent University, 2016.
- [10] Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. Training neural machine translation to apply terminology constraints, 2019.

- [11] Patrick Fernandes, António Farinhas, Ricardo Rei, José G.C. de Souza, Perez Ogayo, Graham Neubig, and André F.T. Martins. Quality-aware decoding for neural machine translation. page 1396 – 1412, 2022.
- [12] Lukas Fischer and Samuel Läubli. What’s the difference between professional human and machine translation? a blind multi-language study on domain-specific MT. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 215–224, Lisboa, Portugal, November 2020. European Association for Machine Translation.
- [13] Robert W. Frick. Communicating emotion. the role of prosodic features. *Psychological Bulletin*, 97(3):412 – 429, 1985.
- [14] Yves Gambier. Audiovisual translation and multimodality: What future? *Media and Intercultural Communication: A Multidisciplinary Journal*, 1(1):1–16, 2023.
- [15] Ximo Granell and Frederic Chaume. Audiovisual translation, translators, and technology: From automation pipe dream to human–machine convergence. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 22, Dec. 2023.
- [16] Lifeng Han, Alan Smeaton, and Gareth Jones. Translation quality assessment: A brief survey on manual and automatic methods. In Yuri Bizzoni, Elke Teich, Cristina España-Bonet, and Josef van Genabith, editors, *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online, May 2021. Association for Computational Linguistics.
- [17] Bettina Hiebl and Dagmar Gromann. Quality in human and machine translation: An interdisciplinary survey. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 375–384, 2023.
- [18] Lan Huang. Ethics of artificial intelligence in education: Student privacy and data protection. *Science Insights Education Frontiers*, 16(2):2577–2587, Jun. 2023.
- [19] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of

- hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 2023. Cited by: 793; All Open Access, Green Open Access.
- [20] Josef Jon and Ondřej Bojar. Character-level nmt and language similarity. volume 1, page 360 – 371, 2023.
  - [21] M.I. Jordan and T.M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255 – 260, 2015.
  - [22] Pawel Kamocki and Jim O’Regan. Privacy issues in online machine translation services - European perspective. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4458–4462, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).
  - [23] Stratvert Kevin. Best ai dubbing | elevenlabs. YouTube, 2023. Accessed: 2024-08-22.
  - [24] Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. Character-aware neural language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.
  - [25] Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121. Association for Computational Linguistics, 2006.
  - [26] Cees M. Koolstra, Allerd L. Peeters, and Herman Spinhof. The pros and cons of dubbing and subtitling. *European Journal of Communication*, 17(3):325 – 354, 2002.
  - [27] Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in neural machine translation, 2019.
  - [28] Jeniffer Lertola. From translation to audiovisual translation in foreign language learning. *Trans*, (22):185 – 202, 2018.
  - [29] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the*

- Association for Computational Linguistics*, pages 4984–4997, Online, July 2020. Association for Computational Linguistics.
- [30] Harry Mcgurk and John Macdonald. Hearing lips and seeing voices. *Nature*, 264(5588):746 – 748, 1976.
  - [31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
  - [32] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
  - [33] Alejandro Pérez, Gonçal Garcés Díaz-Munío, Adrià Giménez, Joan Albert Silvestre-Cerdà, Albert Sanchis, Jorge Civera, Manuel Jiménez, Carlos Turró, and Alfons Juan. Towards cross-lingual voice cloning in higher education. *Engineering Applications of Artificial Intelligence*, 105:104413, 2021.
  - [34] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics.
  - [35] Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, September 2018.
  - [36] Carlos Rodríguez, Marcos Baez, Florian Daniel, Fabio Casati, Juan Carlos Trabucco, Luigi Canali, and Gianraffaele Percannella. Rest apis: A large-scale analysis of compliance with principles and best practices. In Alessandro Bozzon, Philippe Cudre-Maroux, and Cesare Pautasso, editors, *Web Engineering*, pages 21–39, Cham, 2016. Springer International Publishing.

- [37] Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. page 62 – 78, 2023.
- [38] Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. Fluency, adequacy, or hter? exploring different human judgments with a tunable mt metric. In *Proceedings of the fourth workshop on statistical machine translation*, pages 259–268, 2009.
- [39] Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. Multimodal machine translation through visuals and speech. *Machine Translation*, 34(2):97–147, 2020.
- [40] Yihan Wu, Junliang Guo, Xu Tan, Chen Zhang, Bohan Li, Ruihua Song, Lei He, Sheng Zhao, Arul Menezes, and Jiang Bian. Videodubber: Machine translation with speech-aware length control for video dubbing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13772–13779, Jun. 2023.