Capstone Project: Literature, Data, and Technology Review

Project Title: Lecture Companion: AI-Powered Translation and Summarization tool for Burmese Learners

Group Members: Lwin Naing Kyaw, Nang Hom Paung, Sai Aike Sam, Hom Nan Thawe Htun, MangShang SauYing, Ingyin Khin

1. Introduction

As global educational resources become increasingly digitized through platforms like Coursera, edX, and MIT OpenCourseWare, access to high-quality lectures is no longer limited by geography but by language. For millions of learners in non-English-speaking regions, language remains the primary barrier to meaningful participation in global learning ecosystems. This challenge is particularly pronounced in Myanmar, where English serves as the dominant medium for advanced scientific and technical education, yet overall English proficiency remains low.

This research is important because it directly addresses this "language gap" for university students and lifelong learners in Myanmar who possess intermediate English proficiency. By developing an AI-powered system to provide real-time Burmese translations and simplified summaries of English STEM lectures, this project aims to democratize access to knowledge and empower Burmese learners to engage with international educational content without linguistic exclusion. This directly supports UN Sustainable Development Goal 4 (Quality Education) by promoting equitable and inclusive learning, and SDG 10 (Reduced Inequalities) by lowering the linguistic barriers that can limit educational opportunities.

A review of existing literature is necessary to situate this work within the current state of AI-assisted learning systems. Foundational advances in Automatic Speech Recognition (ASR), Machine Translation (MT), Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG) provide the essential technological underpinnings for our proposed system. Reviewing these works clarifies both the opportunities and gaps that our project seeks to bridge, particularly in extending cutting-edge language technologies to underrepresented, low-resource languages like Burmese.

1.1. Education and Language Landscape in Myanmar

Myanmar’s educational system faces long-standing structural and linguistic challenges that hinder equitable access to quality higher education. According to the World Bank (2023), Myanmar’s average years of schooling stands at just 4.9 years, one of the lowest in the ASEAN region. While gross primary enrollment is near universal (close to 100%), the secondary enrollment rate drops sharply, with less than 60% of adolescents continuing to upper-secondary

levels. This gap reflects persistent inequality in both access and learning quality (World Bank, 2023).

In addition, English proficiency remains among the lowest globally. The EF English Proficiency Index (2024) ranks Myanmar #92 out of 113 countries, classifying it under the “Very Low Proficiency” category. The average English score places Myanmar behind regional neighbors such as Vietnam (#58) and Thailand (#82), indicating significant barriers for students engaging with English-medium higher education or global online resources (EF EPI, 2024).

These linguistic barriers directly impact access to science and technology education. Most advanced resources including textbooks, MOOCs, and research materials are published in English. Studies by UNESCO Bangkok (2022) further highlight that Myanmar’s digital learning capacity remains limited, with less than 40% of university students reporting confidence in using English-language online learning platforms.

Against this backdrop, an AI-assisted bilingual lecture companion represents a transformative opportunity. By combining multilingual speech recognition, machine translation, and text simplification, the system provides Burmese-translated, learner-friendly summaries of English STEM lectures. This directly enhances comprehension, supports autonomous learning, and enables students to engage with global curricula in their native language, effectively bridging Myanmar’s digital and linguistic divide.

2. Organization

This review is organized thematically, focusing on the **four core AI technologies** that underpin the project pipeline:

1. Automatic Speech Recognition, for converting spoken lectures into text
2. Low-Resource Machine Translation, for translating English into Burmese
3. Text Summarization and Simplification, for making complex content accessible
4. Retrieval Augmented Generation for interactive learning experience.

3. Summary and Synthesis

Theme 1: Automatic Speech Recognition (ASR)

An essential first step in processing lecture videos is converting spoken audio into accurate textual transcripts. Recent advances in ASR have moved decisively toward large-scale, multilingual models trained on massive and diverse datasets.

- **Paper:** Radford, A., et al. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. OpenAI.
    - **Key Findings:** The "Whisper" model demonstrates remarkable accuracy and robustness in multilingual speech recognition and translation. It achieves

state-of-the-art performance without needing to be fine-tuned for specific accents, domains, or languages.
    - **Methodology:** Whisper is a Transformer-based encoder-decoder model trained on an unprecedented 680,000 hours of multilingual and multitask supervised data collected from the web. This "weak supervision" approach allows it to generalize exceptionally well.
    - **Contribution:** This work demonstrated that a single, large-scale ASR model could effectively handle heterogeneous, real-world audio. For our project, Whisper serves as an ideal foundation for the transcription module, providing a reliable fallback when official captions are unavailable and eliminating the need for multiple, domain-specific ASR systems.

Theme 2: Machine Translation for Low-Resource Languages

Once a transcript is obtained, the next challenge is translating it into a low-resource language like Burmese, for which large parallel corpora is scarce.

- **Paper:** Costa-jussà, M. R., et al. (2022). *No Language Left Behind: Scaling Human-Centered Machine Translation*. Meta AI.
    - **Key Findings:** The "No Language Left Behind" (NLLB) project successfully developed a single AI model capable of high-quality translation across 200 different languages, including many low-resource Southeast Asian languages such as Burmese.
    - **Methodology:** The NLLB team used large-scale data mining and filtering techniques to construct multilingual parallel corpora, combined with a mixture-of-experts Transformer architecture to manage linguistic diversity efficiently. Human-centered evaluation through the FLORES-200 benchmark provided a rigorous, multilingual assessment framework.
    - **Contribution:** This work demonstrated that reliable translation quality for low-resource languages is achievable with the right combination of data and modeling scale. It validates the premise of our system, that AI-based translation can support Burmese learners, and establishes an empirical benchmark for translation quality within our project's educational context.

Theme 3: Text Summarization and Simplification with LLMs

Translation alone does not ensure comprehension; educational material must also be simplified and made accessible to learners with varying proficiency levels. Modern large language models (LLMs), such as Google's Gemini series, have proven adept at these higher-order language tasks.

- **Paper:** Team, Gemini, et al. "Gemini: a family of highly capable multimodal models." arXiv preprint arXiv:2312.11805 (2023).
    - **Key Findings:** LLMs exhibit strong zero-shot and few-shot capabilities for complex, instruction-based tasks such as summarization, simplification, and

rephrasing. They can adapt content to specific target audiences such as English B1-level learners, using natural-language prompts, without explicit task-specific fine-tuning.
    - **Methodology:** These models are trained on extensive, diverse text corpora, enabling them to interpret context, maintain coherence, and generate audience-appropriate explanations. Techniques such as Reinforcement Learning from Human Feedback (RLHF) further refine their instruction-following ability and output quality.
    - **Contribution:** LLMs form the backbone of our project's simplification component. By performing cognitive transformation rather than literal translation, they distill complex technical ideas into accessible, learner-friendly formats. This capacity for adaptive simplification is central to the unique value of our project.

Theme 4: Retrieval-Augmented Learning Systems

- **Paper:** Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.
    - **Key Findings:** The paper introduced the RAG framework, which integrates a dense retriever (based on DPR: Dense Passage Retrieval) with a sequence-to-sequence generator (BART). This architecture allows the model to fetch relevant passages from an external knowledge corpus during inference, improving factual correctness and interpretability. The authors reported significant performance gains across multiple benchmarks, including Open-Domain QA, Fact Verification, and Entity Linking. Compared to purely parametric models, RAG reduced hallucination rates and demonstrated more consistent grounding in external information.
    - **Methodology:** RAG consists of two key components:
        1. **Retriever:** Uses dense embeddings (trained with DPR) to fetch the top-k most relevant documents or passages from a large external corpus.
        2. **Generator:** A pretrained generative model (BART) that conditions its output not only on the user query but also on the retrieved passages. The system is trained end-to-end, allowing the retriever and generator to optimize jointly for factual accuracy. Evaluation was conducted using datasets such as *Natural Questions (NQ)*, *TriviaQA*, and *FEVER*, with metrics including Exact Match (EM) and F1-score.
    - **Contribution:** The RAG model established a new paradigm for combining retrieval and generation in knowledge-intensive tasks. By separating knowledge storage (retriever) from reasoning and composition (generator), it addressed a key limitation of static language models, their inability to access external or updated knowledge at inference time. This work laid the foundation for subsequent developments in retrieval-enhanced educational, conversational, and tutoring systems.

Synthesis: The reviewed literature illustrates the rapid convergence of language technologies such as Automatic Speech Recognition (ASR), Machine Translation (MT), Text Simplification,

and Retrieval-Augmented Generation (RAG) toward integrated, multimodal systems capable of transforming how educational content is accessed and understood.

Across the studies, a common trajectory emerges: models have evolved from task-specific architectures toward unified, instruction-following frameworks that can adapt to varied linguistic and cognitive demands. The Whisper model (Radford et al., 2022) represents a major step in multilingual ASR, demonstrating that large-scale weak supervision can yield universal transcription capabilities across accents and domains. The No Language Left Behind (NLLB) project (Costa-jussà et al., 2022) extends this universality to translation, enabling low-resource languages such as Burmese to benefit from state-of-the-art neural machine translation.

Simultaneously, modern Large Language Models (LLMs) like Gemini (Google, 2023) introduce flexible instruction-following and simplification capabilities, moving beyond literal translation to adaptive comprehension support. Finally, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) establishes a powerful paradigm that grounds generation in external, retrieved knowledge, enhancing factual accuracy and contextual grounding.

Together, these studies form a technological foundation for our project's hybrid pipeline. The combination of ASR, MT, LLM simplification, and RAG aligns with current trends emphasizing transparency, interpretability, and learner-centered adaptation in AI-driven education. By integrating these approaches, our system bridges linguistic accessibility with cognitive scaffolding, creating a bilingual lecture companion tailored to Burmese learners.

4. Conclusion

The synthesis of these four foundational research directions confirms that the technical prerequisites for an end-to-end bilingual lecture companion are already achievable with existing state-of-the-art models.

- **ASR systems** like Whisper provide robust, multilingual transcription that eliminates the need for domain-specific speech models.
- **Multilingual MT frameworks** such as NLLB ensure high-quality translations even for low-resource languages, addressing the scarcity of Burmese educational materials.
- **LLMs** offer adaptable simplification and summarization, enabling automatic generation of B1-level learning content suitable for diverse learners.
- **RAG architectures** supply the structural backbone for building interactive, query-based learning systems grounded in verified lecture content.

Our research synthesizes these strands into a unified framework that transforms static lecture materials into interactive, bilingual, and cognitively accessible educational tools. Beyond technical innovation, the project contributes to the literature on **Responsible AI for education**, providing a replicable model for low-resource language accessibility and equitable knowledge dissemination.

Contribution to Knowledge: This project will contribute to the existing body of knowledge by providing a blueprint for an end-to-end, multi-stage language accessibility tool for education. It moves beyond theoretical models to create an integrated system that includes not only translation and summarization but also an interactive, grounded Q&A component (RAG). This practical implementation serves as a case study on how to effectively combine ASR, MT, and LLMs to enhance learning equity.

5. References

Education First. (2024). EF English Proficiency Index 2024: A ranking of 113 countries and regions by English skills. https://www.ef.com/epi/

UNESCO Bangkok. (2022). Digital learning readiness and challenges in Southeast Asia: Country case studies. https://bangkok.unesco.org/

World Bank. (2023). Myanmar Education Statistics and Indicators (EdStats Database). The World Bank Group. https://databank.worldbank.org/source/education-statistics

Radford, A., et al. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv preprint arXiv:2212.04356.

Costa-jussà, M. R., et al. (2022). *No Language Left Behind: Scaling Human-Centered Machine Translation*. arXiv preprint arXiv:2207.04672.

Team, Gemini, et al. "Gemini: a family of highly capable multimodal models." arXiv preprint arXiv:2312.11805 (2023).

Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.

Data Research

1. Introduction

Our research aims to answer the question: **"How can we leverage frontier technologies to make advanced English-language STEM lectures on platforms like Coursera, edX, and YouTube, accessible to Burmese learners with intermediate English proficiency?"** A thorough exploration of data is essential for this project. It is required to source authentic, high-quality lecture content that reflects the real-world challenges users face. This data forms the input for our entire pipeline, from transcription to summarization, and also serves as the basis for creating a small, parallel corpus to evaluate the quality of our AI-generated outputs.

2. Organization

Our data research findings are organized by the source and type of data collected and processed, moving chronologically through our project's development phases.

3. Data Description

- **Data Source:** Our primary data sources are open-access educational lectures from platforms heavily used by our target audience. We have focused on:
    - **YouTube:** Specifically, Andrew Ng's renowned Machine Learning course lectures and similar computer science related tutorial contents.
    - **MOOC Platforms (Coursera / edX):** Publicly available transcripts and accompanying PDFs from comparable courses were collected to supplement academic content diversity.
    - **Podcasts:** Audio-only content to test the robustness of our ASR pipeline.
- **Data Format:** Raw data were obtained as **video (.mp4)** or **audio (.mp3)** files. Text was derived through multiple routes:
    - **Transcripts:** Extracted using *PyPDF2* for PDF materials, *YouTube-Transcript-API* for auto-captions, and *faster-whisper* for ASR generation when captions were unavailable.
    - **Evaluation Corpus:** A small manually aligned **parallel dataset** of English–Burmese segments translated and simplified by a human annotator to benchmark model outputs.
- **Data Size:** Each lecture is typically 1-2 hours long, resulting in text transcripts of 10,000-20,000 words each. The total corpus for initial development consists of several such lectures.
- **Rationale for Source Selection:** We chose these specific data sources because they are highly representative of the content our target users wish to access. Andrew Ng's lectures, for example, are famous for their quality but are also dense with technical jargon, making them an ideal test case for our system's simplification and translation capabilities. Using a mix of video, audio, and PDF transcripts ensures our system is flexible.

4. Data Analysis and Insights

As part of our data research, we performed Exploratory Data Analysis (EDA) on the processed transcripts. In this section, we present the initial data preprocessing and analysis we have done as part of the project.

- **Key Insights & Patterns:**
    - **Caption Unavailability is Common:** A critical insight was that many high-quality educational videos on YouTube lack official, accurate English captions. This discovery validated our decision to implement a robust ASR model (faster-whisper) as an essential fallback, rather than relying solely on caption APIs.
    - **Segment Length is Suitable for APIs:** After processing, most text segments fall within a 5-20 second duration. This is an ideal chunk size for sending to the Gemini API, as it is large enough to contain semantic context but small enough to ensure low latency for translation and summarization.

- **Speaking Rate is Standard:** Our proxy analysis of Words Per Minute (WPM) showed a mean speaking rate of ~148 WPM, which is a normal conversational pace. This confirms that standard ASR models are well-suited for this data and do not need to be tuned for unusually fast or slow speech.
- **Linguistic Simplification Patterns:** Average lexical reduction from original English to summarized Burmese is roughly **65–70 %**, consistent with known readability-reduction ratios in educational simplification literature (Saggion, 2017). The CEFR histogram further demonstrates that the simplification prompts effectively controlled reading complexity.

- **Visualizations:**
  - **Segment Duration Histogram:** The Distribution of Segment Durations histogram (Figure 4, left) shows that most segments fall between 8 and 18 seconds, with a median near 13 s. This confirms that the chosen segmentation window achieves an effective trade-off between contextual completeness and responsiveness, short enough for near-real-time translation and long enough to preserve sentence-level meaning.
  - **Segment Neighborhood Cohesion**: Figure 4(right) measures Top-1 cosine similarity (excluding self) between each segment embedding and its nearest neighbor in multilingual semantic space. The majority of similarity values cluster around 0.7–0.8, indicating high topical continuity between adjacent segments. This cohesion is crucial for RAG granularity: it ensures that retrieved neighbors are semantically relevant without excessive redundancy, allowing the Gemini 2.5 Flash model to generate answers grounded in coherent local context rather than overlapping text blocks.
  - **Sentence-Length Correlation:** The first scatter plot (*Sentence Length Alignment, Figure 1-left*) shows a strong positive correlation between English and Burmese sentence lengths, indicating that Gemini's translations preserve syntactic proportionality rather than collapsing or over-expanding content. This structural alignment supports consistency across languages.
  - **Document-Level Consistency:** The middle plot in Figure 1 compares total text lengths before and after translation, while the right plot contrasts summarized English vs. summarized Burmese. The Burmese texts are slightly shorter—an expected pattern given Burmese's morphological compactness and the simplification step—but the relationship remains approximately linear, suggesting consistent content compression rather than arbitrary truncation.
  - **Distributional Analyses:** Figure 2 presents six histograms summarizing corpus properties: Original English Text Lengths center around 9 000–11 000 words, confirming uniform lecture durations.Translated Burmese Lengths show broader variance, reflecting differences in sentence segmentation and lexical density.Summarized English and Burmese Texts cluster between 400–800 words, typical of concise B1-level summaries.CEFR Level Distributions (Figure 3) reveal that most simplified outputs fall within B1–B2, with occasional C1 segments. This aligns with our design goal to target intermediate learners while preserving some higher-order content.

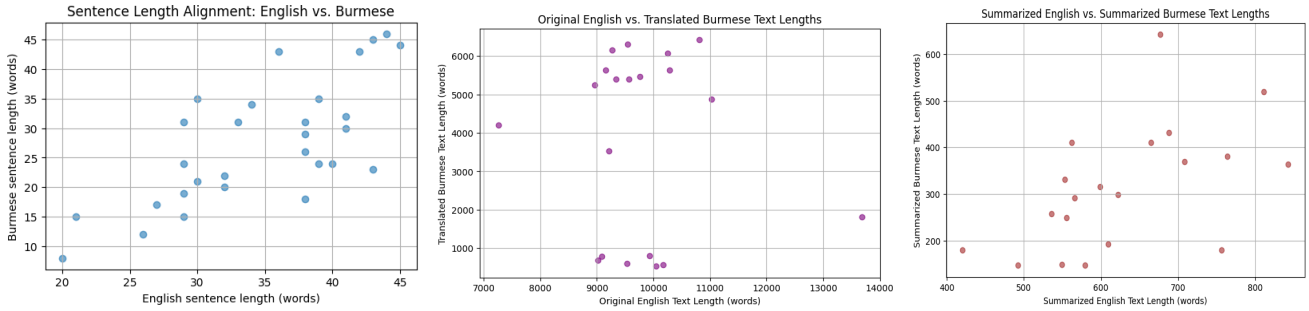

Figure 1: Scatter plots comparisons for text lengths between English vs Burmese

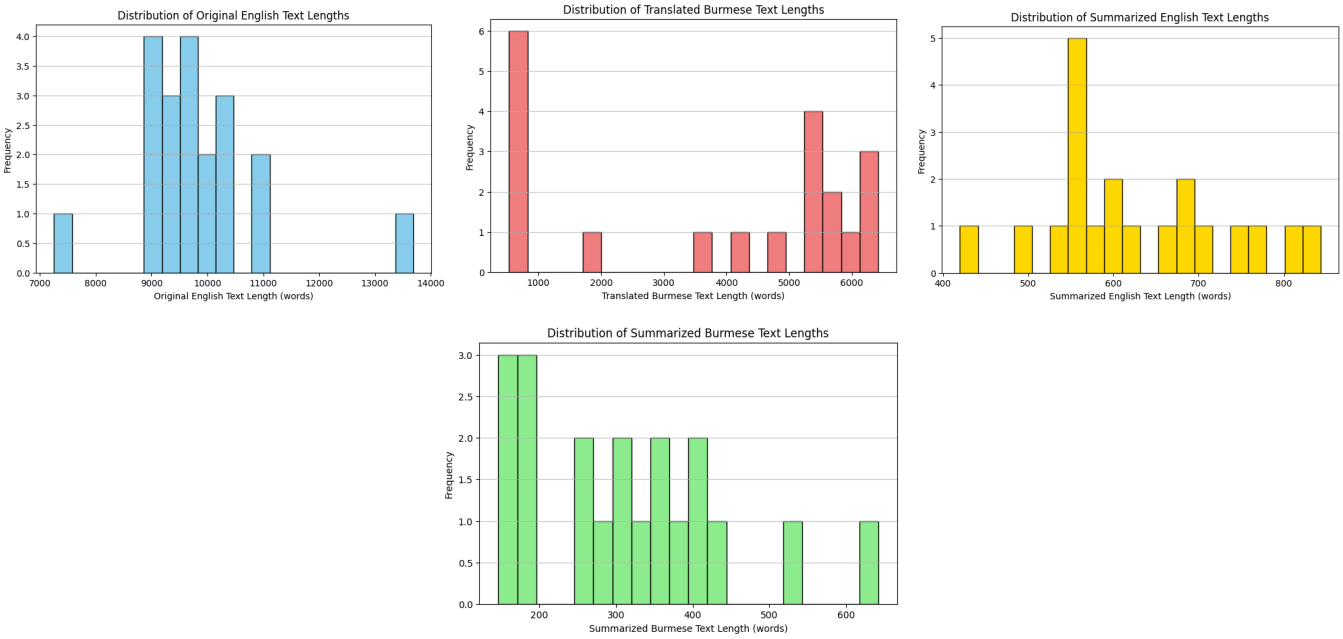

Figure 2: Distributions for documents' text lengths

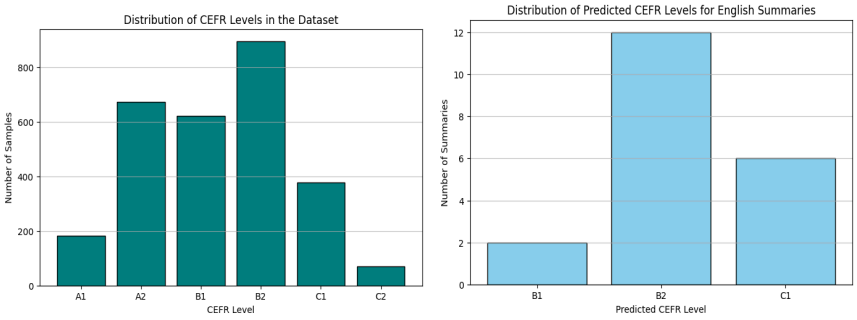

Figure 3: CEFR level analysis visualizations

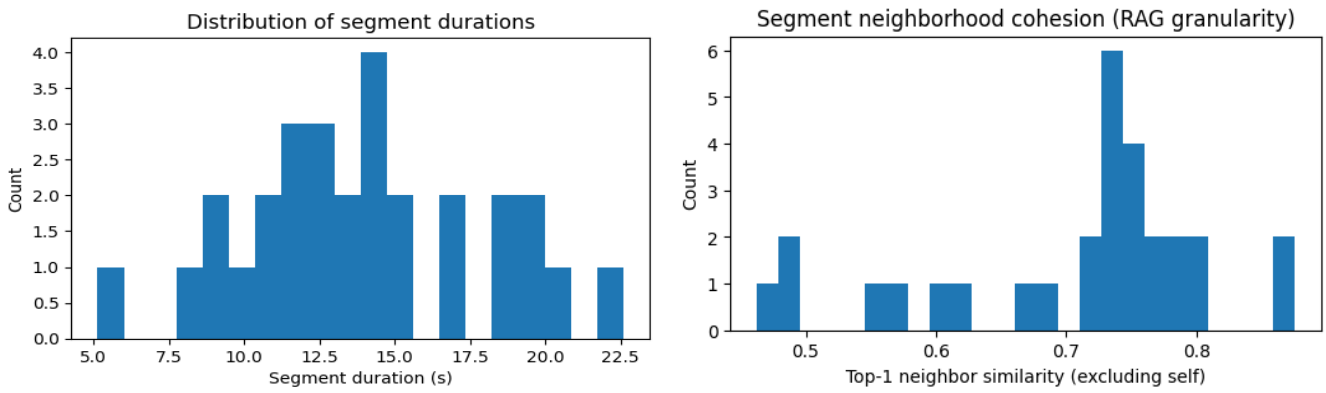

Figure 4: Segment durations

5. Conclusion

- **Key Findings:** The primary finding of our data research is that while ample source material exists, a robust and flexible ingestion pipeline is non-negotiable due to the unreliability of official captions. The data itself, once transcribed and segmented, is well-structured for sequence-to-sequence processing using modern ASR and LLM APIs.
- **Importance:** This data research was critical in shaping the technical architecture of our project. It proved the necessity of our ASR fallback system and informed our text segmentation strategy. It provides a realistic baseline corpus for developing, testing, and, most importantly, evaluating the performance of our translation and simplification models.

6. References

Ng, Andrew. "Machine Learning Course." *YouTube*, uploaded by Stanford University, 2011,playlist link.

Scarton, C. (2020). Horacio Saggion, Automatic Text Simplification. Synthesis lectures on human language technologies, April 2017. 137 pages, ISBN:1627058680 9781627058681. Natural Language Engineering, 26(4), 489–492. doi:10.1017/S1351324919000603

Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages

Technology Review

1. Introduction

This technology review provides an overview of the key tools and models selected for the "Lecture Companion" project. The importance of this review lies in justifying our choice of

technologies to solve the distinct challenges at each stage of our pipeline: transcription, translation, simplification, and retrieval. Our goal is to select a stack that is not only powerful and accurate but also efficient and well-suited for a low-resource language context like Burmese.

2. Technology Overview

Our project is built upon a stack of modern, open-source, and API-driven AI technologies.

1. **Gemini 2.5 Flash**
    - **Purpose:** Central reasoning engine for translation, simplification (B1), and answer synthesis in RAG.
    - **Key Features:** Gemini 2.x introduces 2.5 Flash as a fast, instruction-following, multimodal model; the technical report details the 2.x family (2.5 Pro & 2.5 Flash), extended context, stronger tool-use, and long-video processing capabilities (family spec), positioning Flash as the latency-/cost-optimized sibling for production.
    - **Common Use:** Used for chatbots, content generation, summarization services, and as the reasoning engine in RAG answer generations.
2. **Faster-Whisper (ASR Model)**
    - **Purpose:** High-accuracy Automatic Speech Recognition.
    - **Key Features:** An optimized implementation of OpenAI's Whisper model, offering significant speed improvements and lower memory usage. It is highly robust to accents and background noise and supports multilingual transcription. CTranslate2-based Whisper implementation with up to 4× faster inference at similar accuracy, plus CPU/GPU quantization options. Whisper’s large-scale weak supervision (680k hours) underpins cross-accent robustness; we inherit that quality while gaining speed via Faster-Whisper.
    - **Common Use:** Used in applications that require transcribing audio from files or real-time streams, such as meeting transcription tools and video subtitling services.
3. **SentenceTransformers (all-MiniLM-L6-v2)**
    - **Purpose:** To convert text sentences/paragraphs into dense vector embeddings.
    - **Key Features:** Provides a simple framework for using and training state-of-the-art sentence embedding models. We specifically use a multilingual model that can map text from different languages (like Burmese) into a shared semantic space.
    - **Common Use:** A core component in semantic search, text clustering, and Retrieval-Augmented Generation (RAG) systems.
4. **FAISS (Vector Search Library)**
    - **Purpose:** To store and efficiently search through millions of vector embeddings.
    - **Key Features:** Developed by Meta AI, FAISS (Facebook AI Similarity Search) is an open-source library optimized for fast nearest-neighbor search. It allows us to find the most semantically similar text segments to a user's query in milliseconds.

    - **Common Use:** The backbone of large-scale similarity search applications, including image retrieval, recommendation engines, and RAG databases.

3. Relevance to Your Project

Each technology directly addresses a specific project need:

- **Gemini 2.5 Flash:** Is critical for the translation and, most importantly, the *simplification* tasks. Its ability to follow nuanced instructions allows us to generate summaries tailored to a B1-level learner, a task that specialized translation models cannot perform. It also acts as the final answer-generation module in our RAG system.
- **Faster-Whisper:** Is essential for making our project robust. It solves the real-world problem of missing or inaccurate captions in our source data (YouTube videos), ensuring we always have a text transcript to work with.
- **SentenceTransformers & FAISS:** This combination forms the retrieval engine of our Q&A feature. It allows a student to ask a question in Burmese and receive an answer grounded in the lecture content by searching for conceptual meaning, not just keywords.

4. Comparison and Evaluation

- **Generalist LLM (Gemini) vs. Specialist MT Model (NLLB):**
    - **Strengths/Weaknesses:** While a specialized model like Meta's NLLB might offer slightly higher raw translation quality for Burmese, Gemini's strength is its versatility. It can handle translation, simplification, and generative Q&A within a single, unified framework, which greatly simplifies the development and deployment pipeline. For our project, this versatility outweighs a marginal potential gain in pure translation accuracy.
    - **Suitability:** Gemini is the more suitable choice for an integrated application like ours.
- **Caption API vs. Local ASR (Faster-Whisper):**
    - **Strengths/Weaknesses:** Using an official caption API is faster and less computationally expensive. However, our data research showed it is unreliable. Faster-Whisper is more resource-intensive but provides a highly accurate and universally available solution.
    - **Suitability:** Our chosen hybrid approach, try the API first, then falling back to Faster-Whisper is the most robust and practical solution.

5. Use Cases and Examples

The technology pattern we are using, known as **Retrieval-Augmented Generation (RAG)**, is a state-of-the-art approach for building knowledge-based AI systems.

- **Case Study:** Modern AI-powered search engines and enterprise chatbots use this exact architecture. A user asks a question, the system retrieves relevant documents from a vector database (like FAISS), and an LLM (like Gemini) synthesizes an answer based on

the retrieved context. Our project applies this powerful pattern to the specific domain of educational lectures.

6. Gaps and Research Opportunities

- **Limitations:** The primary limitation of using an LLM like Gemini for simplification is the risk of "hallucination" or providing an inaccurate simplification of a highly complex or niche technical term.
- **Opportunities:** A key research opportunity, as identified in our proposal, is to develop a **Learner Difficulty Prediction Model**. The output of this model could be used to dynamically adjust the simplification prompt sent to Gemini. For example, if a segment is classified as "highly technical (C1 level)," the prompt could instruct Gemini to be extra careful in defining terms and providing analogies, mitigating the risk of inaccurate simplification.

7. Conclusion

- **Key Takeaways:** The chosen technology stack—**Gemini, Faster-Whisper, SentenceTransformers, and FAISS**—provides a modern, powerful, and flexible foundation for the "Lecture Companion."
- **Importance of Chosen Technology:** This specific combination of tools is what allows our project to be an end-to-end solution. It handles messy, real-world data (with the ASR fallback) and delivers advanced, user-centric features (simplification and RAG Q&A) that go far beyond a simple translation script. The use of a multilingual RAG system is what makes the interactive learning component possible.
- **Benefit to Project:** This stack enables us to rapidly prototype and deploy a highly effective learning tool. It is both powerful enough to deliver high-quality results and flexible enough to allow for future improvements, such as the integration of a difficulty prediction model.

8. References

Comanici, G., et al. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning and multimodal understanding. arXiv preprint arXiv:2507.06261. https://doi.org/10.48550/arXiv.2507.06261

Jian, Z. (2023). *faster-whisper: Faster Whisper transcription with CTranslate2*. GitHub Repository. https://github.com/guillaumekln/faster-whisper

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv preprint arXiv:1908.10084.

Johnson, J., Douze, M., & Jégou, H. (2019). *Billion-scale similarity search with GPUs*. IEEE Transactions on Big Data.