Enhancing Learning Objects Accessibility Through Speech-To-Text Based Architecture: A Comprehensive Triangulation Study

1st Venilton FalvoJr
University of São Paulo (ICMC-USP) and DIO (dio.me)
São Carlos, São Paulo, Brazil
falvojr@usp.br

2nd Anderson da Silva Marcolino
Federal University of Paraná (UFPR)
Palotina, Paraná, Brazil
anderson.marcolino@ufpr.br

3rd Diego Renan Bruno
University of São Paulo (ICMC-USP)
São Carlos, São Paulo, Brazil
diego_renan_bruno@hotmail.com

4th Catherine Helen Martins Falvo
Pontifical Catholic University (PUC)
Belo Horizonte, Minas Gerais, Brazil
cattmartins@gmail.com

5th Fernando Santos Osório
University of São Paulo (ICMC-USP)
São Carlos, São Paulo, Brazil
fosorio@icmc.usp.br

6th Ellen Francine Barbosa
University of São Paulo (ICMC-USP)
São Carlos, São Paulo, Brazil
francine@icmc.usp.br

***Abstract*—This full research paper expands on the results and discussions from a case study conducted within a Brazilian EdTech company. It focuses on the accuracy of AI-based speech recognition models for automatic transcription of video lectures, aiming to improve the accessibility of Learning Objects (LOs), especially the audible ones. Previous work assessed the quality of automatic transcriptions in English, Portuguese, and Spanish from Speech-To-Text (STT) services provided by major market players (Amazon, Google, IBM, Microsoft, and OpenAI), using a quantitative approach based on lexical similarity algorithms. Statistically significant differences were identified between the providers and languages evaluated, indicating that STT service can affect the quality of audible OAs enriched with transcriptions or subtitles. This study introduces a new dimension of analysis by incorporating the perceptions of technology students on the accuracy of automatic transcriptions, gathered through an anonymous survey with 56 participants. The survey captured mainly quantitative data using a Likert scale on the accuracy of transcriptions. To provide a comprehensive analysis, this study employs a data triangulation approach, integrating (1) quantitative data from lexical similarity methods, (2) quantitative data from the survey, and (3) findings from a complementary review of the literature. The qualitative analysis, applying Grounded Theory principles to both the survey data and the bibliographic review, enables the exploration of emerging themes and enriches the understanding of factors influencing the quality of automatic transcriptions. This effort underscores the importance of a user-centered perspective and demonstrates the complexity of evaluating STT technologies, pointing to the necessity for future research that combines quantitative and qualitative methods. Additionally, it highlights the relevance of STT in various educational contexts and the need to align such technologies with principles of accessibility and inclusion. By aiming to create more accessible LOs, this work emphasizes the need to develop solutions that ensure inclusion and equity in educational access, reflecting a concerted effort to meet the diverse needs of learners and promote a more welcoming and inclusive learning environment.**

Brazilian funding agencies – São Paulo Research Foundation (FAPESP) under grant #2018/26636-2; CNPq and CAPES.

***Index Terms*—Speech-To-Text (STT), Software Architecture, Learning Objects (LOs), Accessibility, Data Triangulation**

I. Introduction

In the digital age, ensuring accessibility in learning environments is increasingly imperative. Technology, particularly Automatic Speech Recognition (ASR) and Speech-To-Text (STT) techniques (treated as synonyms in this work), can play a pivotal role in making educational content accessible to all students, regardless of physical or sensory individualities. Incorporating these technologies into Learning Objects (LOs) has shown promising potential to enhance student inclusion and engagement [1]–[3].

The concept of LOs is central to current pedagogical practices. Defined as any digital resource that can be reused to support the teaching-learning process, LOs are adaptable to various formats and contexts [4]. In this context, the *Speech2Learning* Architecture proposed by FalvoJr et al. (2024) [3] serves as the primary related study. This architecture enhances audible LOs through STT to generate subtitles, transcriptions, or translations as metadata, thereby making video lessons accessible in multiple languages and even enabling signage through text-based sign language avatars. Additionally, the architecture adheres to metadata standards, guided by recent findings in software engineering [5].

Moreover, the *Speech2Learning* Case Study identified statistically significant differences in the accuracy of automatic transcriptions generated by services from Amazon, Google, IBM, Microsoft, and OpenAI. These differences were quantified using lexical analysis algorithms across three languages: English, Portuguese, and Spanish [3]. Despite these important findings, the study revealed a substantial gap in understanding the real-world impact of automatic transcriptions on learners' experiences.

This research aims to investigate additional sources of evidence beyond the findings of FalvoJr et al. (2024) [3]. Toward this goal, we have designed a Data Triangulation approach [6], drawing on the following sources:

1) *Lexical Similarity Dataset*: Quantitative data on the accuracy of automatic transcription using lexical similarity methods, available in the FalvoJr et al. study [3].
2) *Survey Responses*: Anonymous survey with technology learners focusing on quantitative data on the accuracy of automatic transcriptions (the same ones explored in the lexical dataset).
3) *Documentary Research*: Analysis of relevant studies on ASR and STT for qualitative evidence on the use of these technologies in the teaching-learning process.

Thus, the objective of this study is to expand our understanding of the role of automatic transcriptions in a more accessible and empathetic teaching-learning process, discussing the convergence of quantitative and qualitative results [7].

These insights are crucial for understanding how automatic transcriptions can be improved to better meet the educational needs of learners. While the survey utilized a Likert scale to gather general perceptions of transcription accuracy, the primary focus of qualitative insights will be derived from the documentary research, which applies Grounded Theory principles to explore emergent themes and insights into the factors influencing transcription quality [8].

Therefore, this work not only emphasizes the importance of a user-centered perspective in evaluating STT technologies but also highlights the need for further research combining quantitative and qualitative methods to provide a more comprehensive analysis [6].

By triangulating data from lexical analysis, survey responses, and documentary research, we aim to offer a holistic view of the challenges and opportunities in utilizing automatic transcriptions to enhance the accessibility of learning materials. Additionally, this approach underscores the relevance of STT in various educational contexts, stressing the imperative to align such technologies with accessibility and inclusion principles [2], [3].

In short, our study underscores the critical importance of developing solutions that ensure inclusion and equity in educational access by creating more accessible LOs. This research contributes to the ongoing discourse on leveraging technology to advance educational accessibility and inclusivity.

The paper is structured as follows: Section II provides a detailed description of our methodology, including the Data Triangulation approach. In Section III, we present the results obtained from each data source. Section IV engages in a discussion, synthesizing insights from the triangulation process. Finally, Section V encapsulates the key findings and implications of this study.

II. Methodology

This section outlines the methodology devised for this research to investigate the role of ASR and STT in enhancing the accessibility of educational content.

TABLE I: Overview of the 15 Video Lessons Used for Lexical Similarity and Survey Analysis

Id	Lang	Accent	Gen.	Educational Topic	Time
1	pt-BR	BRA	M	Android Apps	0:17
2	pt-BR	BRA	M	SCRUM	0:26
3	pt-BR	BRA	F	Selenium WebDriver	0:23
4	pt-BR	BRA	F	Blockchain	0:20
5	pt-BR	BRA	M	Hybrid Kernel	0:24
6	en-US	BRA	F	Transit Visa	0:29
7	en-US	USA	Both	Job Interview	0:16
8	en-US	BRA	F	Job Opportunities	0:20
9	en-US	BRA	F	Servant Leadership	0:15
10	en-US	BRA	M	Goroutines	0:15
11	es-AR	ARG	F	Programming Logic	0:12
12	es-AR	ARG	F	Programming Languages	0:21
13	es-AR	ARG	F	Data Types in Python	0:14
14	es-AR	ARG	F	Hello World with Python	0:17
15	es-AR	ARG	F	String Slicing with Python	0:26

To ensure consistency with related work by FalvoJr et al. (2024) [3], we maintained a partnership with EdTech DIO (https://dio.me), which granted access to the same video lessons explored in the previous case study, available on its e-learning platform.

This collaboration allowed us to maintain uniformity in the scope of automatic transcriptions, analyzing identical LOs in our Survey. Consequently, both our Lexical Similarity analysis and Survey Responses assessed a common set of 15 video lessons (Table I), strategically chosen to represent an equal distribution across English, Portuguese, and Spanish. This selection enabled the collection of two distinct quantitative perspectives on automatic transcript quality: one based on lexical similarity algorithms and the other on learners' perceptions.

Moreover, our methodology integrates a third aspect of data collection through Documentary Research. This involves a literature review focused on ASR and STT, employing Grounded Theory [9] to ensure a rigorous content investigation process. Therefore, our documentary analysis aims to provide qualitative insights into the observed phenomenon [7].

By triangulating data from the Lexical Similarity Dataset, Survey Responses, and Documentary Research (as depicted in Figure 1), we strive to offer a comprehensive understanding of the challenges and opportunities associated with utilizing ASR technologies to augment the accessibility of educational materials within the educational domain.

In the following subsections, we outline each source of evidence and its collection techniques, clarifying the entire triangulation process and its contribution to a robust understanding of the impact of ASR technologies, such as STT, on educational accessibility.

A. 1st Source: Lexical Similarity Dataset

In the Case Study conducted by FalvoJr et al. (2024) [3], the lexical similarity dataset was meticulously constructed to evaluate the transcription quality of STT services from leading providers such as Amazon, Google, IBM, Microsoft, and OpenAI. This evaluation utilized three lexical similarity methods: Cosine Similarity (CS), Jaccard Index (JI), and Levenshtein

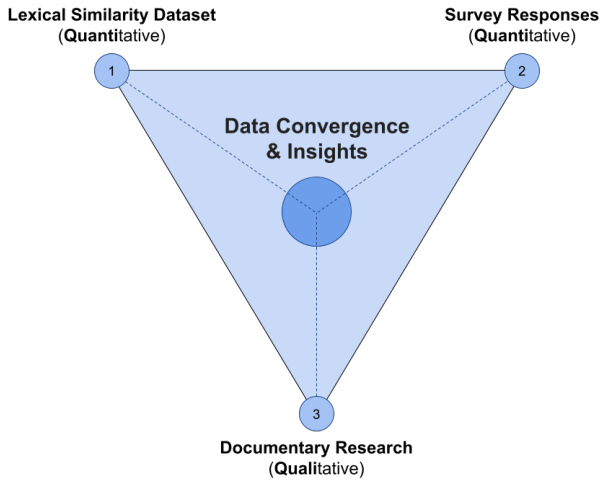


Fig. 1: Triangulation Data: Sources of Evidence.

Distance (LD), with a particular focus on the Jaccard Index due to its robust performance across data samples.

The Jaccard Index measures the similarity between two sets of data by calculating the size of the intersection divided by the size of the union of the sample sets. In this context, it compares the words in the automatic transcription with those in the human-reviewed reference, providing a straightforward quantitative measure of transcription accuracy.

The dataset comprised 15 short video lessons, carefully selected to include five lessons in each of three languages: Portuguese, English, and Spanish. This strategic selection ensured a comprehensive analysis across different linguistic and dialectical nuances, providing a broad assessment of each STT provider's capabilities in handling diverse linguistic content.

This approach not only highlighted the varying degrees of effectiveness among the different STT technologies but also underscored the importance of linguistic diversity in the evaluation of transcription services in educational settings. The choice of employing lexical similarity metrics in this study aligns with the recommendations of Sommerville (2015) [10] for rigorous empirical evaluation in software engineering research.

B. 2nd Source: Survey Responses

We conducted a Survey to evaluate learners' perceptions of the quality of automatic transcription provided by ASR/STT. This Survey collected quantitative and qualitative feedback from participants with experience in Information Technology (IT) or Linguistics and Education.

The Survey utilized a mix of Likert-scale and open-ended questions. Likert-scale questions aimed to quantitatively assess the coherence/quality of transcriptions across three languages (English, Portuguese, and Spanish). This consistency in content allowed for a direct comparison of technical and user-centric evaluations. The open-ended questions sought qualitative insights into the perceived effectiveness of ASR solutions. However, due to ethical compliance, we will not explore qualitative data in this study.

Participants evaluated transcriptions from the same set of 15 video lessons (five per language) and the same five providers (Amazon, Google, IBM, Microsoft, and OpenAI) as used in the lexical similarity study. This alignment between the two sources ensured that the Survey Responses directly complemented the technical analysis, providing a holistic view of transcription performance.

To mitigate bias, the providers were anonymized and assigned a random numerical identifier rather than being named, ensuring that the evaluations reflected genuine user experiences without influence from brand recognition. This methodological approach aligns with best practices in empirical research methodology, particularly in gathering user feedback and perceptions in software engineering studies [10], [11].

C. 3rd Source: Documentary Research

A literature review complements our empirical findings by identifying relevant studies on ASR and STT technologies. This review supports the contextualization and enrichment of quantitative data, providing a deeper understanding of the potential and limitations of current ASR-based technologies.

We employed Grounded Theory [9] as a guiding framework for our documentary analysis. Grounded Theory is a systematic methodology for analyzing qualitative data to develop theories, allowing for the emergence of themes and patterns from the data itself rather than imposing preconceived notions.

Our Documentary Research involved a comprehensive review of relevant literature, including academic articles, books, and reports, focusing on ASR and STT technologies in the context of educational accessibility. Through this analysis, we aimed to identify recurring themes, theoretical frameworks, and gaps in existing research, contributing to a nuanced understanding of the subject matter.

By integrating insights from Lexical Similarity Dataset, Survey Responses, and Documentary Research, we aimed to provide a comprehensive and multi-dimensional exploration of the role of ASR and STT in enhancing educational accessibility.

D. Our Data Triangulation

Triangulation, which began as a geometric technique for determining location, has evolved into a metaphor for research methods that integrate various approaches, theories, or data sources to comprehensively understand a phenomenon, particularly in qualitative research. This concept enhances the validity of a study by employing multiple methods to verify or support a particular event, description, or fact, thus increasing the study's credibility and reliability [6], [8].

For this study, we adopted a triangulation approach that incorporates multiple sources of evidence and their respective techniques. This mixed-methods approach uses qualitative and quantitative results to improve the analysis and understanding of the data collected. Specifically, we employ data triangulation, using multiple sources of evidence to corroborate the same fact or phenomenon [8].

In software engineering, data triangulation strengthens the rigor and validity of research. By integrating multiple methods, such as quantitative surveys, qualitative interviews, and documentary analysis, researchers can cross-verify findings and obtain a more nuanced understanding of complex phenomena. This approach is particularly valuable in examining the efficacy of software development practices, user experience, and technology adoption [12].

Figure 2 provides a detailed illustration of our comprehensive data triangulation methodology. This diagram delineates the study's object, the three distinct sources of evidence, their respective collection techniques, and the process of data convergence that generates our data corpus. This thorough and multifaceted approach ensures robust validation and offers an in-depth understanding of the ways in which ASR technologies can enhance educational accessibility.

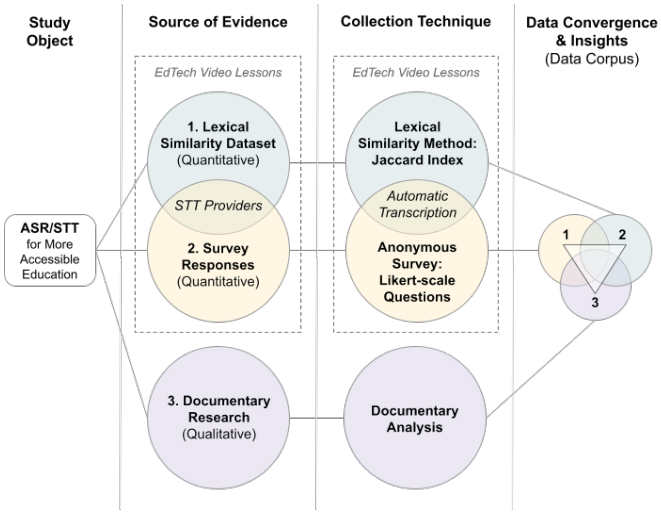
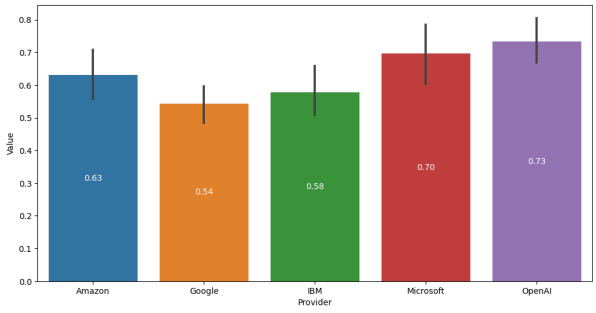
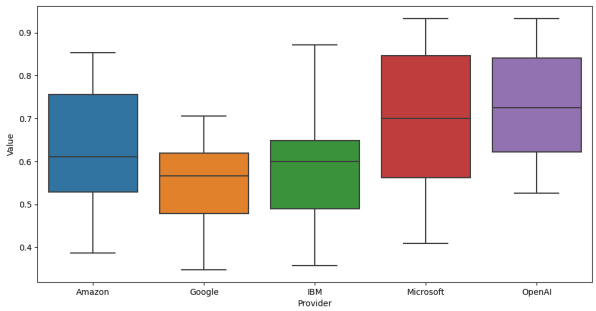


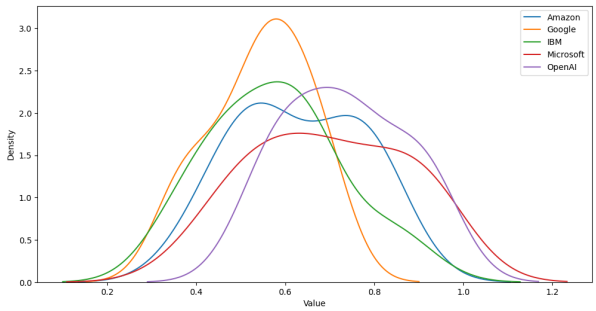
Fig. 2: Our Triangulation Data Methodology.



(a) Jaccard Index Bar Plot.



(b) Jaccard Index Box Plot.



(c) Jaccard Index KDE Plot.

Fig. 3: Jaccard Index Similarity Plots.

III. Results

A. Lexical Similarity

This Case Study investigated the lexical similarity across various language service providers using the Jaccard Index, focusing on automatic transcriptions [3]. These transcriptions serve not only as quality metrics for captioning but also as foundational text for applications in diverse areas such as live captioning, document transcription, and text-based inputs for sign language avatars. This multifaceted utility underscores the broad applicability and significance of the study's findings.

The initial findings of the study were visualized through a comprehensive set of plots (Figure 3), including a Bar Plot that displays the average Jaccard Index for each provider, a Box Plot that details the data distribution, and a KDE Plot that illustrates the probability density of the scores. These visualizations prominently featured OpenAI, which demonstrated the highest score, suggesting its superior performance in capturing lexical similarity across various transcription applications.

Statistical analyses detailed in Table II supported these observations. Normality tests, including Kolmogorov-Smirnov [13], [14] and Shapiro-Wilk [15], affirmed the data's normal distribution, enabling further analysis via One-Way ANOVA [16]. This analysis revealed significant differences among providers, with subsequent Tukey HSD [17] and Bonferroni [18] tests highlighting notable disparities, especially between OpenAI and other providers like Google and IBM. These differences suggest variability in provider efficiency at handling lexical similarity across different transcription contexts.

Moreover, the analysis extended to evaluate provider performance across multiple languages—English, Spanish, and Portuguese—as depicted in Figure 4. This linguistic breakdown revealed persistent performance disparities, which are critical for global applications that rely on accurate and reliable automatic transcriptions.

Corresponding statistical evaluations (Table III) included normality tests and One-Way ANOVA for each language. The findings highlighted significant differences in how providers handle English transcriptions, as evidenced by the English subgroup ANOVA.

TABLE II: Jaccard Index Statistical Tests.

Tests of Normality				
Test	**Kolmogorov-Smirnov**		**Shapiro-Wilk**	
	Statistic	**Significance** (p)	**Statistic**	p
Jaccard	0.057	**0.200**	0.973	**0.111**
Interpretation				
Kolmogorov-Smirnov: If $p > 0.05$, sample follow the same statistical distribution.				
Shapiro-wilk: If $p > 0.05$ is normal.				
One-Way ANOVA (One-Way Analysis of Variance)				
Test	**F**		p	
Jaccard	4.562		**0.002**	
Interpretation				
If $p < 0.05$, there are significant differences between at least two groups.				
Tukey HSD (Honest Significant Difference)				
Providers (Groups from 1 to 5)			p	
OpenAI (Group 5) to Google (Group 2)			**0.005**	
OpenAI (Group 5) to IBM (Group 3)			**0.032**	
Microsoft (Group 4) to Google (Group 2)			**0.037**	
Interpretation				
If $p <> 0$ and $p < 0.05$, there is a significant difference among the groups.				
Bonferroni				
Providers (Groups from 1 to 5)			p	
OpenAI (Group 5) to Google (Group 2)			**0.006**	
OpenAI (Group 5) to IBM (Group 3)			**0.041**	
Microsoft (Group 4) to Google (Group 2)			**0.047**	
Interpretation				
If the adjusted p ($\alpha' = \alpha/n$), where n is the number of comparisons, is < 0.05, then there is a statistical difference among the groups.				

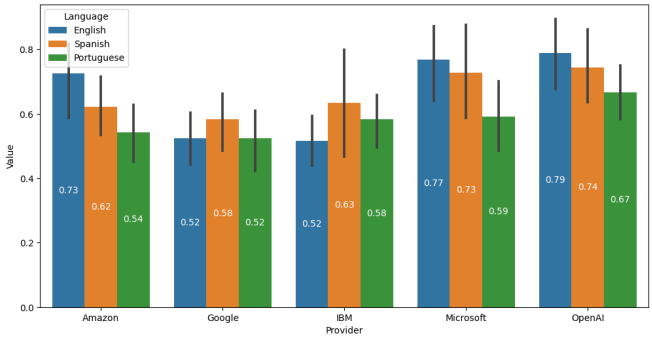


Fig. 4: Jaccard Index Similarity by Language Plot.

Specific disparities between providers were further outlined through post-hoc analyses such as Tukey HSD and Bonferroni for the English language, indicating that performance can significantly vary depending on the language of the transcriptions.

These insights not only highlight the differences in provider capabilities concerning lexical similarity but also set the stage for further exploration into how these capabilities can impact user experiences across various transcription applications, including learner satisfaction and the perception of transcription quality in different linguistic contexts, as will be discussed in the following subsection based on Survey results.

B. Survey

Building on the insights from the lexical similarity study, this section presents results from an anonymous survey exploring learner perceptions of the quality of automatic transcriptions derived from the same dataset of video lessons. The survey aimed to correlate lexical similarity data with learners' subjective quality assessments of transcription across different STT service providers.

TABLE III: Jaccard Index Statistical Tests by Language.

Tests of Normality				
Test	**Kolmogorov-Smirnov**		**Shapiro-Wilk**	
	Statistic	**Significance** (p)	**Statistic**	p
EN	0.121	**0.200**	0.961	**0.427**
PT	0.114	**0.200**	0.942	**0.169**
ES	0.951	<0.001	0.951	**0.264**
One-Way ANOVA				
Test	**F**		p	
EN	4.88		**0.007**	
PT	1.058		0.403	
ES	0,931		0,466	
Tukey HSD (Honest Significant Difference) - EN				
Providers (Groups from 1 to 5)			p	
OpenAI (Group 5) to Google (Group 2)			**0.041**	
OpenAI (Group 5) to IBM (Group 3)			**0.034**	
Bonferroni - EN				
Providers (Groups from 1 to 5)			p	
OpenAI (Group 5) to IBM (Group 3)			**0.047**	

Visualizations of the survey outcomes were provided to illustrate the average ratings and distribution of responses. The bar plot displays the average ratings for each provider, with OpenAI receiving the highest score, suggesting a strong preference among learners (Figure 5a). The box plot details the spread and central tendency of ratings for each provider, highlighting the variance in user satisfaction (Figure 5b), while the KDE plot provides a visual estimation of the distribution density of the ratings (Figure 5c).

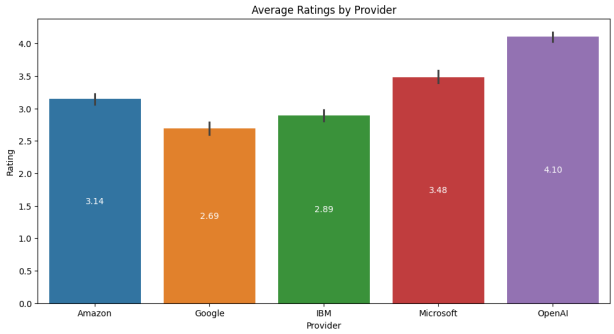
An additional Chart details reviews by language (English, Portuguese and Spanish), indicating significant variations in user satisfaction based on transcription language (Figure 6).

Statistical analyses indicated that the ratings did not follow a normal distribution, as confirmed by the Kolmogorov-Smirnov [13], [14] and Shapiro-Wilk [15] tests (Table IV).

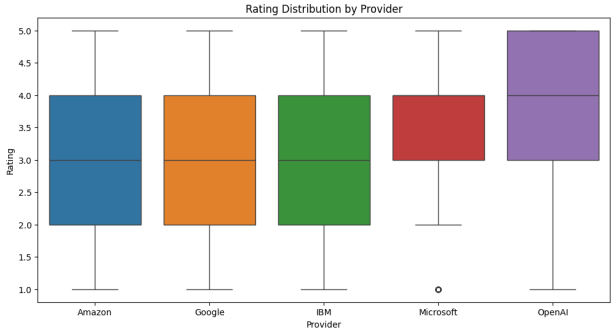
Consequently, non-parametric tests were required for further analysis. The Kruskal-Wallis [19] test, a non-parametric alternative to ANOVA, was employed to determine if there were significant differences in median scores across different groups. This test showed highly significant differences among providers ($p < 0.05$), suggesting that learners' perceptions varied markedly depending on the provider.

Given the non-normal distribution of data, the Dunn's [20] test was chosen over the Conover [21] test for post-hoc analysis because of its less stringent assumptions about data distribution and its ability to handle data with outliers effectively (Table IV). Dunn's test compared pairs of providers, indicating significant differences between almost all pairs. Notably, all comparisons involving OpenAI and other providers demonstrated significant differences, with OpenAI typically favored over others.

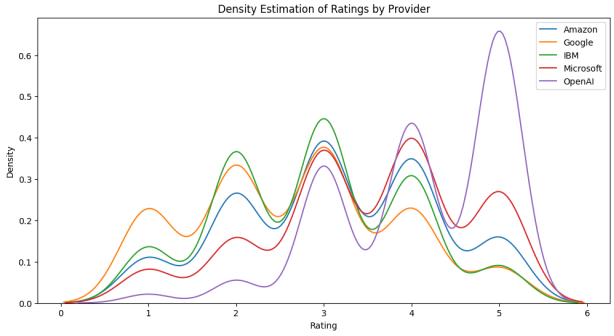
The results stratified by language reinforced these findings, with all language groups showing significant differences among providers in how the automatic transcriptions were rated. This suggests that the quality perceived by users is not



(a) Survey Answers/Ratings Bar Plot.



(b) Survey Answers/Ratings Box Plot.



(c) Survey Answers/Ratings KDE Plot.

Fig. 5: Survey Answers/Ratings Plots.

only provider-dependent but also varies with language, highlighting the challenges in providing uniformly high-quality transcriptions across different linguistic contexts (refer to Table V).

The survey analysis effectively bridges the objective measures of lexical similarity and the subjective perceptions of transcription quality. The correlation between the high lexical similarity scores and higher user satisfaction ratings, particularly for OpenAI, underscores the practical relevance of lexical accuracy in user satisfaction. These insights are crucial for providers aiming to optimize their transcription services for educational content, as they illustrate the importance of both linguistic precision and user perception in evaluating transcription quality. The findings suggest a roadmap for future improvements and the potential customization of services to meet diverse linguistic needs more effectively.

TABLE IV: Survey Statistical Tests.

Tests of Normality				
Test	**Kolmogorov-Smirnov**		**Shapiro-Wilk**	
	Statistic	**Significance** (p)	**Statistic**	p
Ratings	0.887	<0.001	0.973	<0.001
Interpretation				
Kolmogorov-Smirnov: If $p > 0.05$, sample follow the same statistical distribution.				
Shapiro-wilk: If $p > 0.05$ is normal.				
Kruskal-Wallis (Non-Parametric equivalent to ANOVA)				
Test	**Statistic**		p	
Ratings	536.167		**<0.001**	
Interpretation				
If $p < 0.05$, there are significant differences among the groups.				
Dunn (Non-Parametric Post-Hoc equivalent to Tukey HSB)				
Providers (Groups from 1 to 5)			p	
Amazon (Group 1) to Google (Group 2)			**<0.001**	
Amazon (Group 1) to IBM (Group 3)			**<0.001**	
Amazon (Group 1) to Micorsoft (Group 4)			**<0.001**	
Amazon (Group 1) to OpenAI (Group 5)			**<0.001**	
Google (Group 2) to IBM (Group 3)			**0.010**	
Google (Group 2) to Microsoft (Group 4)			**<0.001**	
Google (Group 2) to OpenAI (Group 5)			**<0.001**	
IBM (Group 3) to Microsoft (Group 4)			**<0.001**	
IBM (Group 3) to OpenAI (Group 5)			**<0.001**	
Micorsoft (Group 4) to OpenAI (Group 5)			**<0.001**	
Interpretation				
If $p < 0.05$, it indicates significant pairwise differences between groups.				

TABLE V: Survey Statistical Tests by Language.

Tests of Normality				
Test	**Kolmogorov-Smirnov**		**Shapiro-Wilk**	
	Statistic	**Significance** (p)	**Statistic**	p
EN	0.897	<0.001	0.897	<0.001
PT	0.847	<0.001	0.912	<0.001
ES	0.959	<0.001	0.893	<0.001
Kruskal-Wallis (Non-Parametric)				
Test	**Statistic**		p	
EN	244.355		**<0.001**	
PT	355.323		**<0.001**	
ES	37.615		**<0.001**	
Dunn (Non-Parametric) - EN				
Providers (Groups from 1 to 5)			p	
Amazon (Group 1) to Google (Group 2)			**<0.001**	
Amazon (Group 1) to IBM (Group 3)			**<0.001**	
Amazon (Group 1) to OpenAI (Group 5)			**0.002**	
Google (Group 2) to Microsoft (Group 4)			**<0.001**	
Google (Group 2) to OpenAI (Group 5)			**<0.001**	
IBM (Group 3) to Microsoft (Group 4)			**<0.001**	
IBM (Group 3) to OpenAI (Group 5)			**<0.001**	
Micorsoft (Group 4) to OpenAI (Group 5)			**0.005**	
Dunn (Non-Parametric) - PT				
Providers (Groups from 1 to 5)			p	
Amazon (Group 1) to Microsoft (Group 4)			**<0.001**	
Amazon (Group 1) to OpenAI (Group 5)			**<0.001**	
Google (Group 2) to IBM (Group 3)			**0.017**	
Google (Group 2) to Microsoft (Group 4)			**<0.001**	
Google (Group 2) to OpenAI (Group 5)			**<0.001**	
IBM (Group 3) to Microsoft (Group 4)			**<0.001**	
IBM (Group 3) to OpenAI (Group 5)			**<0.001**	
Micorsoft (Group 4) to OpenAI (Group 5)			**<0.001**	
Dunn (Non-Parametric) - ES				
Providers (Groups from 1 to 5)			p	
Amazon (Group 1) to OpenAI (Group 5)			**<0.001**	
Google (Group 2) to Microsoft (Group 4)			**0.002**	
Google (Group 2) to OpenAI (Group 5)			**<0.001**	
IBM (Group 3) to OpenAI (Group 5)			**<0.001**	

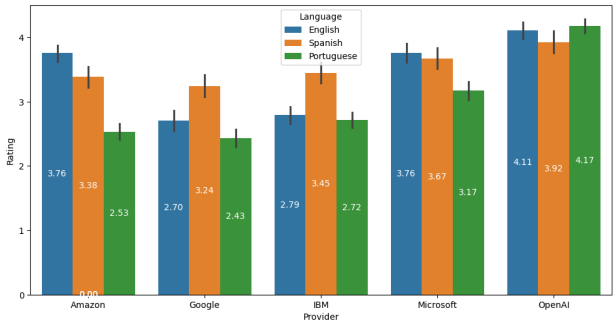


Fig. 6: Survey Answers/Ratings by Language Plot.

C. Documentary Analysis

In this subsection, we extend our exploration into ASR/STT technologies by reviewing relevant literature, aiming to contextualize our quantitative findings with qualitative insights. This documentary analysis delves into various research works that have significantly contributed to the field of ASR and STT, providing a broader understanding of how these technologies, powered by advancements in Machine Learning (ML) and Artificial Intelligence (AI), enhance the accessibility of LOs.

The study by Ferraro et al. (2023) [22] presents an extensive investigation into the transcription of spoken language using ML models. Their analysis compares open-source and paid services for STT transcription, focusing on quality and input variety. The research employs diverse datasets from interviews, lectures, and speeches, utilizing metrics such as the Word Error Rate (WER) for evaluation. Ferraro et al.'s work provides a benchmark for assessing transcription accuracy, laying a solid foundation for our study.

Bengesi et al. (2024) [23] offer a comprehensive review of recent advancements in Generative Artificial Intelligence (GenAI), highlighting its potential applications in automatic transcription processes. While not directly targeting speech-to-text conversion, their exploration of state-of-the-art models, including Generative Adversarial Networks (GANs), Generative Pre-trained Transformers (GPT), autoencoders, and diffusion models, sets the stage for understanding how GenAI can enhance transcription accuracy.

Homburg and Thieme (2019) [24] investigate the use of humanoid robots as avatars for sign language translation, aiming to improve inclusivity for the deaf community. Unlike previous research focused solely on sign language recognition, their study explores a novel approach by employing robots as intermediaries for sign language communication. By surveying 50 deaf participants, they ascertain the perceived effectiveness of humanoid robots in facilitating sign language translation, offering innovative insights into accessibility solutions.

AlShaikh et al. (2024) [25] delve into the integration of GenAI in education through the development and evaluation of an AI Educational Video Assistant. Grounded in the Cognitive Theory of Multimedia Learning (CTML), their tool, equipped with modules for Transcription, Engagement, and Reinforcement, leverages ASR technologies to enhance the learning experience. By focusing on multimodal learning experiences, their study showcases the potential of advanced AI techniques, including STT, to improve educational outcomes.

Cao et al. (2022) [26] address the limitations of traditional audio transcription tools in noisy real-world classrooms. Their research emphasizes the crucial role of effective intelligent learning systems in collaborative environments, particularly in analyzing and comprehending conversations among learners. By exploring the influence of ASR errors on conversation models, highlighting the challenges and opportunities for improving ASR accuracy in educational contexts.

Drawing upon the insights from these studies, we aim to apply Grounded Theory [9] principles to categorize and analyze the qualitative data gathered from our research. Grounded Theory provides a systematic approach for exploring emerging themes and enriching our understanding of factors influencing the quality of automatic transcriptions. By integrating results collected from our evidence sources (Lexical Similarity Dataset, Survey Responses, and Documentary Research), we seek to uncover patterns and relationships that inform the optimization of ASR and STT technologies in enhancing learning object accessibility.

To facilitate this analysis, we categorize the findings from each study within the Grounded Theory into distinct themes, allowing for a comprehensive understanding of the technological and educational implications of ASR and STT tools. These categories include:

- **Transcription Quality Assessment (TQA)**: Evaluating the performance and quality of ASR technologies in transcription tasks.
- **Integration of GenAI in Education (IGE)**: Exploring the integration of GenAI in educational settings to enhance learning experiences.
- **Inclusive Technologies for Accessibility (ITA)**: Investigating innovative approaches, such as humanoid robots and multimodal learning tools, to improve accessibility for diverse learners.
- **Challenges and Opportunities in ASR (COA)**: Identifying challenges and opportunities for improving ASR accuracy and effectiveness in educational contexts.

In Table VI, we present the categorization of each paper within the Grounded Theory, providing a detailed overview of the themes addressed in each study and their relevance to our research objectives.

TABLE VI: Categorizing Studies Using Grounded Theory

Category	Description	Papers
TQA	Evaluating ASR performance and quality in transcription tasks.	[22]
IGE	Exploring the integration of GenAI in educational settings.	[23], [25]
ITA	Investigating inclusive technologies for accessibility.	[24], [25]
COA	Identifying challenges and opportunities in ASR for educational contexts.	[26]

IV. Discussion

This research seeks to clarify the complexities of ASR and STT technologies in improving the accessibility of audible LOs by triangulating data from three distinct sources of evidence. Each source contributes uniquely to a comprehensive understanding of how automatic transcription impacts learner interaction with LOs.

The Lexical Similarity Dataset, highlighted significant variances in transcription accuracy among leading STT providers (Amazon, Google, IBM, Microsoft, OpenAI). The data elucidated not only the technical capabilities of these systems but also underscored the linguistic nuances that affect their performance across different languages [3].

The Survey extends our analysis by incorporating the subjective perceptions of learners regarding the quality of these transcriptions. Interestingly, the feedback from the Survey often aligns with the technical metrics from the Case Study *Speech2Learning*, reinforcing the importance of accuracy in user satisfaction. However, it also brings to light user-centric aspects of STT services—such as ease of understanding and practical integration into learning environments—that are not captured by lexical similarity alone.

This powerful quantitative data, based on the Jaccard Index metric and Survey Likert-scale Responses, serves as a fundamental aspect of our triangulation, providing a baseline for assessing the accuracy of transcriptions.

The Documentary Research enriches quantitative perception by introducing qualitative elements from existing literature. This analysis builds on the findings of other sources of evidence, as well as exploring broader themes, such as the impact of ASR/STT technologies on education and their potential to promote accessibility. This complementary literature review identified interesting gaps and opportunities in the current use of STT, suggesting areas for future work in both technological and pedagogical fields.

A. Triangulation Data Convergences

The integration of these three sources of evidence (Lexical Similarity Dataset, Survey Responses, and Documentary Research) provides a holistic view of the impact of speech recognition technologies in education. This data triangulation confirms the importance of transcription accuracy, as well as revealing the need for a user-centered, context-aware approach to maximizing the effectiveness of these technologies. Below is a summary of the convergence of triangulated data:

- **Lexical Similarity vs. User Perceptions**: The Case Study conducted by FalvoJr et a. (2024) [3] revealed that different STT providers exhibit statistically significant differences in transcription accuracy, as measured by lexical similarity metrics such as the Jaccard Index. Notably, OpenAI showed superior performance across languages compared to other major providers like Google and IBM. However, the survey results, which focused on quantitative assessments using Likert scale ratings, indicate that user satisfaction is influenced by more than just lexical accuracy. Although OpenAI also ranked highest in learner satisfaction, the statistical differences in survey responses suggest that user perceptions are influenced by factors such as context understanding and error tolerance, which are not fully captured by lexical similarity metrics alone.
- **Qualitative Insights from Documentary Research**: The integration of Grounded Theory in our documentary analysis allowed for a deeper understanding of the themes affecting transcription quality. Studies discussed in the documentary research, such as those by Ferraro et al. (2023) and AlShaikh et al. (2024), emphasize the importance of considering diverse data types and learning environments when evaluating STT technologies [22], [25]. These findings suggest that the educational effectiveness of STT applications extends beyond mere transcription accuracy, encompassing ease of integration, customization capabilities, and the technology's ability to adapt to diverse educational needs.
- **Educational Context and Technological Inclusivity**: The triangulated data underscore the need for STT technologies to align with the principles of accessibility and inclusivity. The discrepancies between the mechanical accuracy of transcriptions and the qualitative satisfaction of users highlight a gap in the current capabilities of STT. This gap points to the necessity for providers to innovate beyond traditional metrics of accuracy and incorporate user feedback into the development of more context-aware and inclusive transcription solutions.

B. Implications for Future Research and Development

The identified convergences reveal that, although the lexical accuracy of transcriptions is relevant, user satisfaction and educational effectiveness also depend on other factors, such as ease of use and the ability to integrate/adapt to specific contexts. This multifaceted overview highlights the importance of an integrated approach in the development and implementation of ASR and STT technologies, with implications for future research as follows:

- **Enhanced Model Training**: The findings indicate the potential for improving STT technologies by training models on more diverse linguistic datasets, which could help in understanding context and reducing errors in transcriptions that affect user satisfaction.
- **Customization for Educational Use**: Providers should consider customization options that allow educational institutions to tailor STT features to their specific needs, such as adjusting for different accents, dialects, and technical vocabulary specific to courses or subjects.
- **User-Centered Design Approaches**: Incorporating user feedback into the development process can ensure that future improvements in STT technologies align more closely with the needs and expectations of end-users, particularly in diverse educational settings.

The convergence of data from lexical analysis, user surveys, and scholarly research paints a complex picture of the current state and potential of ASR and STT technologies in education.

While notable advances in transcription accuracy have been made, significant work remains to fully realize the potential of these technologies in enhancing educational accessibility and inclusion. Future research should focus on bridging the gap between technical proficiency and user satisfaction, emphasizing the development of adaptable, user-friendly, and context-aware systems that can support a wide range of educational environments and learning needs.

V. Conclusion

This study offers a comprehensive exploration of the role of STT technologies in enhancing the accessibility of LOs through a triangulated data analysis approach. By integrating quantitative data from lexical similarity metrics, learner perceptions from surveys, and qualitative insights from documentary research, we provide a nuanced understanding of the strengths and limitations of current STT solutions.

Our findings indicate significant variability in transcription accuracy among leading STT providers, with OpenAI demonstrating superior performance across multiple languages. However, the learner surveys reveal that user satisfaction is influenced not only by transcription accuracy but also by contextual understanding and error tolerance. This underscores the need for STT providers to consider user-centric factors in their service improvements.

The documentary research further contextualizes these findings by highlighting the importance of aligning STT technologies with principles of accessibility and inclusivity. It emphasizes that the educational effectiveness of STT applications extends beyond technical accuracy to include factors such as ease of integration, customization capabilities, and adaptability to diverse educational needs.

In conclusion, while notable advancements in STT technologies have been made, significant work remains to fully realize their potential in enhancing educational accessibility and inclusion. Future research should focus on bridging the gap between technical proficiency and user satisfaction, emphasizing the development of adaptable, user-friendly, and context-aware systems. By addressing these areas, STT technologies can better support a wide range of learning environments and meet the diverse needs of learners, thereby promoting a more inclusive and equitable educational landscape.

Acknowledgment

The authors would like to thank the Brazilian funding agencies – São Paulo Research Foundation (FAPESP) under grant #2018/26636-2; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001; and CNPq. We extend our gratitude to EdTech DIO for the partnership on both the PoC and the Case Study, enabling market insights through one of the largest e-learning platforms in Latin America.

References

[1] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[2] R. E. Mayer and L. Fiorella, *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press, 3rd ed., 2021.

[3] V. FalvoJr, A. Marcolino, D. Bruno, C. M. Falvo, F. Osório, and E. Barbosa, "Lexical Analysis of Automatic Transcriptions Using Speech-to-Text Services: A Statistically Evaluated Case Study," in *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*, pp. 5317–5326, 2024.

[4] A. Parakh, M. Subramaniam, and P. Chundi, "A framework for incorporating serious games into learning object repositories through experiential learning," 01 2022.

[5] F. A. Santana, A. F. R. Cordeiro, and E. OliveiraJr, "Use of the dublin core standard to express open metadata related to software engineering experiments," *Anais do III Workshop de Práticas de Ciência Aberta para Engenharia de Software (OpenScienSE 2023)*, 2023.

[6] J. Farquhar, N. Michels, and J. Robson, "Triangulation in industrial qualitative case study research: Widening the scope," *Industrial Marketing Management*, vol. 87, pp. 160–170, 2020.

[7] E. B. Lima Junior, G. S. de Oliveira, A. C. O. dos Santos, and G. F. Schnekenberg, "Análise documental como percurso metodológico na pesquisa qualitativa," *Cadernos da Fucamp*, vol. 20, no. 44, 2021.

[8] R. K. Yin, *Qualitative research from start to finish*. Guilford publications, 2015.

[9] K. Charmaz, *A Construção da Teoria Fundamentada: Guia Prático para Análise Qualitativa*. Penso, 1 ed., 2009.

[10] I. Sommerville, *Software Engineering*. Pearson Education, 10th ed., 2015.

[11] R. S. Pressman and B. Maxim, *Engenharia de Software*. McGraw Hill Brasil, 8th ed., 2016.

[12] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical Software Engineering*, vol. 14, no. 2, pp. 131–164, 2009.

[13] A. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, pp. 83–91, 1933.

[14] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions," *Ann. Math. Stat.*, vol. 19, pp. 279–281, 1948.

[15] S. S. Shapiro and M. B. Wilk, "Analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[16] R. A. Fisher, "Statistical methods for research workers," *Genesis Publishing Pvt Ltd*, 1925.

[17] J. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949.

[18] C. E. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilità," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 3–62, 1936.

[19] W. H. Kruskal and W. Wallis, "The use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[20] O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, no. 3, pp. 241–252, 1964.

[21] W. Conover, "Practical nonparametric statistics," *Wiley New York*, vol. 3, 1999.

[22] A. Ferraro, A. Galli, V. La Gatta, and M. Postiglione, "Benchmarking open source and paid services for speech to text: an analysis of quality and input variety," *Frontiers in Big Data*, vol. 6, p. 1210559, Sep 2023.

[23] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers.," *IEEE Access*, pp. 1–1, 2024.

[24] D. Homburg, M. S. Thieme, J. Völker, and R. Stock, "Robotalk - prototyping a humanoid robot as speech-to-sign language translator," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[25] R. AlShaikh, N. Al-Malki, and M. Almasre, "The implementation of the cognitive theory of multimedia learning in the design and evaluation of an ai educational video assistant utilizing large language models," *Heliyon*, vol. 10, p. e25361, Feb 2024.

[26] J. Cao, A. Ganesh, J. Cai, R. Southwell, E. M. Perkoff, M. Regan, K. Kann, J. H. Martin, M. Palmer, and S. D'Mello, "A comparative analysis of automatic speech recognition errors in small group classroom discourse," in *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23, pp. 250–262, June 2023.