Artificial intelligence-generated Arabic subtitles: insights from Veed.io's automatic speech recognition system of Jordanian Arabic

Legendas em árabe geradas por inteligência artificial: *insights* do sistema de reconhecimento automático de fala do árabe jordaniano da Veed.io

Wala' Mohammad Akasheh [*1], Ahmad S. Haider [†1,2], Bassam Al-Saideen [‡3] and Yousef Sahari [§4]

[1] *Applied Science Private University, Faculty of Arts and Humanities, Department of English Language and Translation, Amman, Jordan.*
[2] *Middle East University, MEU Research Unit, Amman, Jordan.*
[3] *Isra University, Faculty of Arts, Department of Translation, Amman, Jordan.*
[4] *University of Bisha, College of Arts and Letters, Department of English Language and Literature, Almahalah district, Bisha, Saudi Arabia.*

Abstract

This paper examines the errors that the automatic speech recognition (ASR) system of Veed.io produces when transcribing utterances spoken in Jordanian Arabic into subtitles. It attempts to propose a new classification for the subtitles that are built based on artificial intelligence technology. Through a combination of qualitative and quantitative analyses, the study examines the types of errors and their impact on comprehension. The errors observed in the generated subtitles based on linguistic and phonetic analysis are categorised into three main types: deletions, substitutions, and insertions. Furthermore, the quantitative analysis measures the word error rate (WER) and shows that the WER percentage is 38.857% revealing that deletions are the most common type of error, followed by substitutions and insertions. The study recommends conducting further research on ASR systems for Arabic language dialects and advises subtitlers to be aware of the limitations of these systems when using them, ensuring that they edit and supervise them appropriately.

Keywords: Subtitles. Auto-generated subtitles. Automatic Speech Recognition. Linguistics. Jordanian Arabic.

Resumo

Este artigo examina os erros que o sistema de reconhecimento automático de fala (ASR) do Veed.io produz ao transcrever declarações faladas em árabe jordaniano para legendas. Tenta propor uma nova classificação para as legendas construídas com base em tecnologia de inteligência artificial. Através de uma combinação de análises qualitativas e quantitativas, o estudo examina os tipos de erros e seu impacto na compreensão. Os erros observados nas legendas geradas com base na análise linguística e fonética são categorizados em três tipos principais: exclusões, substituições e inserções. Além disso, a análise quantitativa mede a taxa de erro de palavras (WER) e mostra que o percentual de WER é de 38,857%, revelando que as exclusões são o tipo de erro mais comum, seguidas pelas substituições e inserções. O estudo recomenda a realização de mais pesquisas sobre sistemas ASR para dialetos da língua árabe e aconselha os legendadores a estarem cientes das limitações desses sistemas ao usá-los, garantindo que os editem e supervisionem adequadamente.

Palavras-chave: Legendas. Legendas geradas automaticamente. Reconhecimento Automático de Fala. Linguística. Árabe jordaniano.

Textolinne
Linguagem e Tecnologia

DOI: 10.1590/1983-3652.2024.46952

Session:
Articles

Corresponding author:
Ahmad Haider

Section Editor:
Daniervelin Pereira
Layout editor:
João Mesquita

Received on:
July 22, 2023
Accepted on:
August 28, 2023
Published on:
January 6, 2024

1 Introduction

In recent years, social media and online content creators have impacted sharing knowledge by providing people with new ways to connect and communicate with each other regardless of geographical

*Email: okasha.walaa@hotmail.com
†Email: a_haidar@asu.edu.jo
‡Email: bsaidien@yahoo.com
§Email: ysahari@ub.edu.sa

boundaries or language barriers. This led to coming up with faster solutions that guarantee accessibility for different groups of people using new modes of translation, facilitating the distribution of Audiovisual content such as films, series, e-learning content, and online streaming platforms.

Audiovisual Translation (AVT) is a modern type of translation that deals with multimedia products and how to translate the meanings from a source language and culture to a target language and culture in an appropriate way, in which these languages are systems of communication that their meanings are conveyed through three main methods: spoken, written or sign. Chaume (2013, p. 105) claims that "Audiovisual translation is a mode of translation characterised by the transfer of audiovisual texts either interlingually or intralingually".

Due to globalisation, people are surrounded by different AV materials in public and private areas. As a result, it became an essential part of education, health, finance, business, and communication. Therefore, AVT is one of the most important modern tools and translation modes that makes communication effective and fast between nations (Al-Abbas; Haider, 2021; Haider; Alrousan, 2022; Haider; Saideen; Hussein, 2023; Jarrah; Haider; Al-Salman, 2023).

Due to technological innovation, audiovisual translation (AVT) has become an academic discipline under the umbrella of translation studies (Remael, 2010). One of the AVT modes is subtitling, which can be characterised as a translation technique that involves displaying a written text, typically on the bottom of the screen, which tries to recount the original dialogue of the speakers along with the other elements that appear in the image (letters, inserts, graffiti, etc.), as well as the information that is contained on the soundtrack (songs, voices off) (Díaz-Cintas; Remael, 2007).

Generally, subtitles are the graphemic reflections of a natural linguistic spoken production or a generated voice synthesis, sometimes mixed with some other extralinguistic details, in the audiovisual and media works (films, series, etc.) in which these graphemes are understandable by the receiver. These graphemes, usually numbers, letters, or words, could be typed by a human transcriber or generated by an automated speech recogniser. Recent technology has proven to be effective in facilitating communication.

Under the category of subtitling, Automatic Speech Recognition (ASR) subtitles are the intralingual subtitles generated automatically by the software that converts spoken audio into written text. ASR technology uses complex algorithms and machine learning to recognise and transcribe speech in real-time or post-processing. However, generating accurate and high-quality subtitles can be challenging, particularly for dialects that differ from the standard form of the language.

Auto-generated subtitles are provided by platforms that can be intralingual or interlingual. Usually, intralingual subtitles are generated by the ASR system, while interlingual are produced by the MT technology. ASR system is a speech recognition software that analyses human speech and changes it into text. This technology connects the most important modern fields in Audiovisual translation studies and linguistics: Subtitiling and Natural Language Processing.

ASR has become an increasingly essential tool to improve accessibility and communication between various languages and cultures. However, when it comes to the case of Jordanian Arabic, the quality of these platforms producing ASR-based subtitles can vary, and transcription errors can lead to confusion and miscommunications, in which relying on ASR-based subtitles for understanding other cultures, such as the Arab culture, can lead to misinterpretation.

This study addresses the following **two research questions:**

- What linguistic errors are most frequently noticed in ASR subtitles in JA?
- What is the Word Error Rate (WER) in ASR-based subtitles for JA in the selected AV material?

2 Literature Review

Searching for related theoretical or empirical works leads us to find that the main approaches and fields followed in handling studies of auto-generated subtitles belong to computational linguistics and phonology, Natural Language Processing, and scarce research in the field of AVT studies. This is very normal since the auto-generated subtitles topic is considered an application that requires AI techniques to deal with. However, in the AVT studies, the research deals with linguistic and translation aspects, focusing on errors and the quality of subtitles. Previous studies have focused on the quality of these

technologies for standard Arabic regardless of ASR-based subtitles for JA and MT-based subtitles written in MSA, which are part of the AV content (Almahasees, 2017; Bendou, 2021). Therefore, it is essential to conduct research that specifically investigates the accuracy of ASR-based subtitles for the JA and examines the linguistic aspects of errors in auto-generated JA subtitles.

This section is two-fold. The first part reviews the theoretical background relevant to audiovisual translation and NLP technology. In addition, it discusses subtitling as an accessibility tool and provides an overview of ASR AI-powered subtitles, mentioning the linguistic phenomena that affect them. The second part discusses some empirical studies related to the topic under study.

2.1 Theoretical Background

2.1.1 AV and Subtitling

In recent years, and due to globalisation, research on AVT and subtitling has become very widely used and important. Díaz-Cintas and Remael (2007) argue that subtitling is a translation process that involves rendering the discursive elements that appear in the image and the information that is contained on the soundtrack (songs, voices off) to a written text, which usually appears in the lower part of the screen. As per the classification of Díaz-Cintas and Remael (2007), subtitles can be categorised into three types. The first type is intralingual, which involves rendering both verbal and non-verbal signs in the same language. The second type is interlingual, which translates verbal and non-verbal signs from one language to another. The third type is bilingual, which displays verbal and non-verbal signs in two or more languages. These types are illustrated in Figure 1.

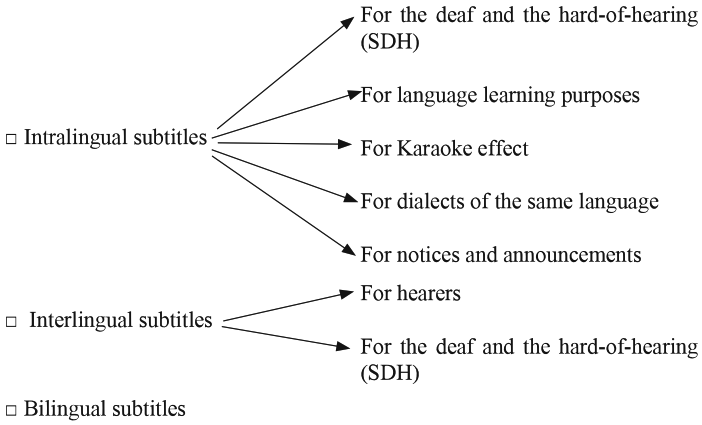

Figure 1. Classifications of subtitles based on linguistic dimension.
Source: (Díaz-Cintas; Remael, 2007, p. 14).

When considering the available time for preparation, subtitles can be categorised into various types. Díaz-Cintas and Remael (2007) classified them into pre-prepared subtitles and live or real-time subtitles, as Figure 2 shows.

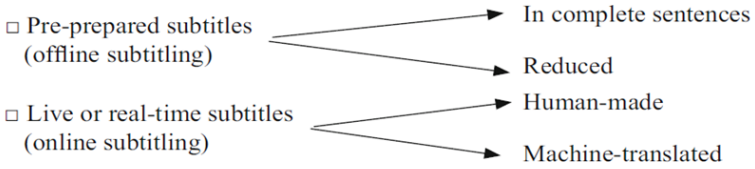

Figure 2. Classifications of subtitles based on time available for preparation.
Source: (Díaz-Cintas; Remael, 2007, p. 19).

Recent developments in voice and speech recognition have made possible the appearance of re-speaking as a practice to subtitle programmes that are broadcast (semi/real) live, such as the news

or sports (Remael, 2010). Auto-generated subtitles usually depend on two main artificial intelligence technologies: Automatic Speech Recognition and Machine Translation. Automatic speech recognition (ASR) is a machine-based method that independently decodes and transcribes oral speech (Suvorov; Levis, 2012).

Dharmale and Patil (2019) mention that "Automatic Speech Recognition permits the machine to take out oral contained from a speech signal and produce a text message by using feature extraction and classification techniques." As part of Spoken Language Translation, Machine Translation (MT) can be defined as the subfield of computational linguistics concerned with using software in translation across human languages (Almahasees, 2017).

2.1.2 AI-powered Subtitles

Through a detailed review of the relevant literature and a comprehensive analysis of various forms of AI-powered subtitles, it has become apparent that it is essential for academic research in AVT to distinguish between these different types. Therefore, to align with our research objectives and due to the insufficient literature available, a classification system has been developed for AI-powered subtitles to identify the subtitle types used in the study accurately.

Artificial Intelligence (AI) powered subtitles use machine learning algorithms and natural language processing (NLP) techniques to generate audio or video content subtitles. This approach involves using AI to analyse the audio or video content, transcribe the spoken words, sign language, or paralinguistic elements and generate accurate and synchronised subtitles.

AI-powered subtitles can be categorised based on various factors. In this study, the categorisation is based on two main factors, namely, the used technology and the target language. *The used technology* refers to some technologies used in recent years. Some of these are sound recognition (SR), automatic speech recognition (ASR), sign language recognition (SLR), machine translation (MT), and others. These technologies are all focused on language processing and understanding. *The target language,* where the process of generating subtitles may be influenced by the target audience. For instance, social media platforms in the Arab world tend to use MSA for subtitles, as it covers a wide geographical area and avoids the time-consuming task of generating subtitles for specific dialects. However, for automatic speech recognition (ASR) subtitles, Facebook generates comprehensible subtitles for the dialects that may contain orthographical errors, mainly related to glottal stop variants in Arabic, and mixes them with Alef or some phonetic-based errors like the one on the Egyptian dialect where they tend to pronounce any voiceless dental fricative consonantal sound /θ/ as voiceless alveolar fricative /s/, and therefore, the model generates the Arabic character س, so سعلب instead of ثعلب.

Drawing on the linguistic dimension classifications presented by (Díaz-Cintas; Remael, 2007), AI-generated subtitles can be categorised as Intralingual AI-powered subtitles and Interlingual AI-powered subtitles.

Intralingual AI-powered subtitles are intralingual subtitles that use different technologies to produce a written form within the same language. These could be categorised as SR-based subtitles, ASR-based subtitles, Semi-ASR subtitles, and SLR-based subtitles. *SR-based subtitles* refer to subtitles generated using a combination of automatic speech recognition and other sound recognition techniques. ASR is used to transcribe spoken language into written text, while SR is used to filter and extract relevant information from the audio signal, such as noise, music, and non-speech elements. Combining these techniques can improve the accuracy and quality of the generated subtitles, providing a complete viewing experience for people such as the Deaf and Hard of Hearing (DHH). YouTube excelled in this type of AI-powered subtitles, adding a good description for the paralinguistic elements in their auto-generated subtitles. *ASR-based subtitles*, where the spoken words in written form of an audio or video file are analysed and transcribed. ASR subtitles have become a ubiquitous feature on social media platforms such as Facebook, which generates these subtitles in multiple languages and different dialects. In addition, several platforms offer the service of auto-generated subtitles for videos for content creators and companies (paid and unpaid), such as Amara and Zubtitle. Without including any description for paralinguistic elements, ASR-based subtitles are for hearing people, which helps them to understand the spoken language better without semiotic or paralinguistic elements, i.e., the

transcription of the spoken words only. It helps people understand language and dialects and follow along with the verbal elements.

Semi-ASR subtitles are generated using a combination of automatic speech recognition (ASR) technology and manual human intervention. The initial transcription is generated by an ASR system, and then a human editor reviews and corrects the transcription to ensure accuracy and readability. Vimeo and Veed.io offer the service of ASR subtitles and enable users to edit these subtitles. *SLR-based subtitles* are generated based on SLR technology and translate sign language into a written form. SignAll and SLAIT provide their clients with a system that translates between sign language and written/spoken text in a way that the system captures signed language using a system of four cameras. It detects body movements and expressions to translate American sign language to English written form and displays it on the screen. Figure 3 illustrates the classifications of AI-powered subtitles.

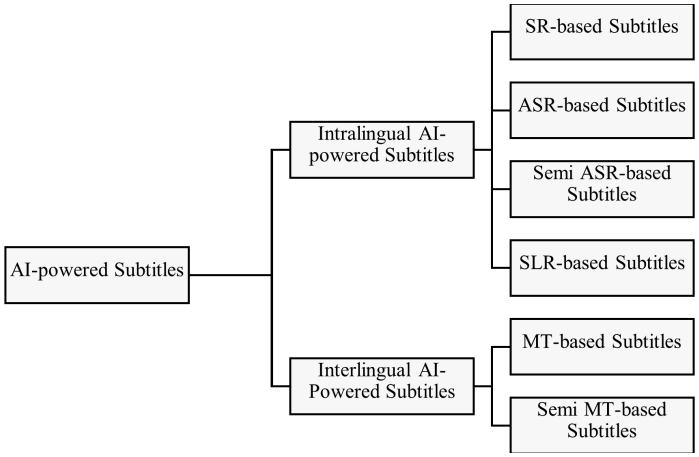

Figure 3. Classifications of Artificial Intelligence-Powered Subtitles.
Source: Own Elaboration.

Interlingual AI-powered subtitles are interlingual subtitles that use technologies to analyse a language's spoken content and produce a written form in another language. These include MT-based subtitles and Semi-MT-based subtitles. *MT-based subtitles* are generated using automated translation software rather than being translated by human translators. This technology uses algorithms to automatically translate spoken, written, or sight content from one language to another. On their videos, Facebook added the feature of choosing the closed caption in languages other than the video's original language. *Semi-MT-based subtitles* are the MT subtitles generated by the software, and a human operator edits them to level up the accuracy, i.e., the post editor. Using YouTube Studio for content creators enables them to edit captions created automatically, generating a new caption track that includes their revisions. The current study is focused on ASR-based subtitles and MT-based subtitles.

2.1.3 Arabic Language and Jordanian Arabic Dialect

Modern Standard Arabic is the language of written Arabic media, *e.g.*, newspapers, books, journals, street signs, and advertisements. It is also the language of the majority of news broadcasts on radio and television (Ryding, 2005). Sabir and Alsaeed (2014, p. 185) stated that "Arabic has 28 consonants (including two semi-vowels) and six vowels (three short vowels and three long vowels)". As a common phenomenon, diglossia in the Arab world is a normal situation where people speak and shift between the standard language and their regional dialect. Geographically speaking, Arabic is one of the most widely spoken languages in the world, and its dialects are spoken in a continuous stretch from western Iran through Mauritania and Morocco and from Oman to northern Nigeria, despite the enormous deserts and sparsely populated or uninhabited regions in between (Behnstedt; Woidich, 2013).

Arabic speakers speak a variety of mutually intelligible dialects that differ in phonology, morphology, syntactic structure, lexical content, geography, and social structure. However, they are rarely written; therefore, there are no stable writing conventions (Maamouri *et al.*, 2006). In his conclusion, Doughan

(2017) stated that "Jordanians meta-pragmatically differentiate between two registers of Arabic in Jordan: Urdunī (Jordanian) and Madanī (Urban)" (p.103).

In the case of the city of Amman, Versteegh (2006, p. 325) pointed out that "The new dialect primarily relies on elements from the madanī (Urban) dialect as well as elements from the Jordanian Bedouin dialect". In conclusion, in Jordan, most media productions are produced in the dialect that is spoken in Amman, which is the urban dialect. Therefore, it is essential to focus on choosing data spoken in this specific dialect in addition to MSA.

2.2 Empirical Studies

The field of automatic speech recognition (ASR) has seen an interest in recent years in the field of natural language processing. Yet, it is worth noting that ASR non-computational studies were lacking.

A systematic review by Alharbi *et al.* (2021) summarised the most important topics of ASR published in the last six years. published in the last six years. The study reported the most applied datasets in recent ASR research, categorising the reviewed articles based on three characteristics: domain problems, natural language pre-processing, and device efficiency. The review identified several challenges facing ASR, including speech capture issues, hardware-related problems with microphones, and speech pre-processing challenges such as dialect diversity and pronunciation problems. Finally, the study suggested future research directions for improving ASR systems.

Reviewing some specific linguistic phenomena in ASR systems has attracted much attention from research teams. In their experiment, Mustafa *et al.* (2022) measured code-switching (CS) in automatic speech recognition (ASR) systems. CS is when speech has two or more languages within an utterance. The research has identified 274 papers and selected 42 experimental papers for review covering many well-resourced and under-resourced languages and techniques to recognise CS in ASR systems, such as mapping, combining, and merging the phone sets of the languages experimented with and examined the performance of those techniques. The study found a significant variation in the performance of CS experimental papers in terms of word error rate (WER), indicating the inconsistency in the existing ASR systems' ability to handle unexpected pronunciation changes when languages are mixed.

Likewise, Sawakare, Deshmukh, and Shrishrimal (2015) discussed the techniques used in various stages of speech recognition, classifying speech recognition systems based on utterance types, vocabulary sizes, and speaker modes used. They noted that feature extraction is essential in separating relevant from irrelevant information and distinguishing one speech from another. Moreover, they concluded that feature extraction played a significant role in improving speech recognition system accuracy.

However, while many studies focused on the NLP field, Guskaroska (2019) examined the usefulness of mobile-assisted ASR dictation systems for enhancing vowel pronunciation among Macedonian EFL learners. The study utilised a mixed-methods approach, which included pre-test and post-test recordings to measure accuracy gains, a comparison of ASR written output to humans, and an analysis of learners' attitudes towards ASR through Facebook posts. The findings revealed that the experimental group (Non-Native English Speakers) had improved accuracy while the control group (Native English Speakers) did not. In addition, learners generally had positive attitudes toward ASR. The study suggested incorporating mobile-assisted ASR in EFL classrooms with careful teacher guidance and structured practice using individual words.

Some researchers in the auto-generated subtitling field are concerned about paralinguistic components being visible in the subtitles. An experimental study guided by Schlippe *et al.* (2020) evaluated a new method called WaveFont, which diversifies fonts in video captions based on voice characteristics such as loudness, speed, and pauses. The goal was to test this new method specifically for Arabic viewers and to compare it to traditional captions. The results showed that WaveFont is comprehensible and accepted by most people, including deaf and hard of hearing and normal-hearing viewers. The study suggested that this technology can revolutionise how captions and subtitles are presented, with potential applications in various fields such as video-on-demand, TV, social media, live broadcasts, and public places.

Liao *et al.* (2023) introduced a new method which is the ASR post-processing for readability

(APR) task. The goal of this task is to enhance ASR output by correcting grammatical mistakes, disfluency and making it more readable for humans. They used the Grammatical Error Correction datasets as their corpus by using TTS and ASR systems. Also, they adapt and develop evaluation metrics from related tasks. Their method proved to be effective since the human evaluation and case study further revealed the ability of the proposed model to improve the readability of ASR transcripts.

In their article, Pucci (2023) addressed some of the opportunities and challenges offered by automatic speech recognition (ASR) systems. They discussed both the advantageous aspects and challenges presented by ASR systems. The researcher pointed out that ASR technology is a valuable tool for presenting information in a multimodal manner, supporting inclusivity and communication improvement. However, they mentioned that further refinement of this technology is required before incorporating it into universally designed environments.

When it comes to Machine translation, several publications have appeared in recent years documenting the level of accuracy of MT software, making the studies in the non-computational fields, specifically AVT studies, richer and wider.

Al Mahasees (2021) conducted a comparative evaluation of the performance of three machine translation (MT) systems for Arabic: Sakhr, Google Translate, and Microsoft Translator. The study analysed the output of the three systems on both holistic analysis and error analysis (EA) scales to provide constructive feedback about their capacity. In addition, the study ranked the three systems' performance based on their adequacy, fluency scales, and error categories, including orthography, lexis, grammar, and semantic errors. Google Translate achieved the best overall performance, followed by Microsoft Translator and Sakhr.

Adopting a user-centric approach, Xie (2022) compared machine-translated subtitles on Bilibili (MTS-B) with those on YouTube (MTS-Y) and investigated the relationship between the quality, users' comprehension, and attitude towards machine-translated subtitles. The study found that quality had little impact on users' comprehension of the videos. However, accuracy had a significant effect on users' attitudes toward the quality of the translation. Participants had a better attitude towards MTS-Y as it performed better in accuracy. The study also found that MTS-Y performed better in grammar and spelling, while MTS-B showed cleaner and simpler subtitles. Overall, most participants could understand the contents of the video through the machine-translated subtitles.

Moreover, Almahasees and Jaccomard (2020) investigated the use of Facebook Translation Service (FTS) as a source of information during the COVID-19 lockdown in Jordan. The study found that Facebook and FTS became significant sources of information during the crisis, with 62.2% of participants considering Facebook as their primary source of information regarding COVID-19. Additionally, 87.1% of participants activated FTS, with 87.3% using FTS to translate English Facebook posts into Arabic. However, the majority found that FTS committed minor errors in terms of adequacy and fluency. The study suggested that health officials should create Facebook profiles with a blue tick for medical information during crises. In addition, medical specialists and translation scholars should evaluate FTS's ability to render COVID-19 medical posts fluently and adequately in Arabic.

As it is shown, for several years, different scholars have investigated ASR and MT from a computational linguistics perspective. However, to our knowledge, an evaluation of the ASR systems from a linguistic perspective is not common. This work is the first of its type, making the evaluation and analysis of AI-powered subtitles complex.

3 Methodology

This section presents an overview of the methodology used in the study. It includes details about the sample and data collection process. The study employs a combination of quantitative and qualitative approaches. The procedures followed in the study are also summarised at the end of this section.

3.1 Selected Data

3.1.1 ASR-based subtitles

To ensure the investigation's success, the selection of videos is based on various factors that mainly surround the utterances, which will be analysed and transcribed by the model. These challenging

factors could be related to the speakers' voices, such as voice quality, gender, age, breath, clicks, pauses, stress, overlapping, mumbles, and prosody. Other factors are related to the surrounding environment being part of the acoustic signal, such as background noise and music. Therefore, the chosen video is selected purposefully. The video is a Radio/TV show that is recorded with high-quality microphones and published during the year of the study. In the video, two Jordanian broadcasters use the dialect of Amman to talk about the etiquette of meals in the month of Ramadan, where they code-switch/mix, laugh, and speak fast and slow in their show. The video to be tested is shown below in Table 1.

Table 1. Selected video for the investigation of the study.

Video Type	Link	Justifications
Radio/TV shows	https://www.youtube.com/watch?v=MeDi-77lQ24	Gender, Code Switching, Mumbling, Overlapping chatter.

Source: Own elaboration.

The website that will be tested for the ASR investigation is www.veed.io. On LinkedIn[1], Veed.io define themselves as "An AI-powered online video editing platform that makes creating videos easy and accessible to everyone". This website provides content creators and businesses with tools that can help them edit their videos and add automatic subtitles in many languages and dialects. In the year of this study, before generating the subtitles, the user should select which language is spoken in the video. Among the choices, the Arabic language is available, and they can choose one of the following dialects: Jordanian, Palestinian, Lebanese, Iraqi, Saudi, Bahraini, Qatari, Kuwaiti, Omani, Egyptian, Tunisian, Algerian, and Moroccan.

The analysis will mainly focus on the linguistic aspects by categorising the errors into two main types. Errors that did not significantly affect the comprehension of the text were given a value of 0.5, while errors that affected the comprehension of the subtitles were given a value of 1. Table 2 shows the categorisation of these errors based on their types.

Table 2. Categorisation of errors based on their type.

		0.5 Errors		
Type/Category	1	2	3	4
Deletions	Affixes	Interjections	Overlapping	Vowel length
Substitutions	Affixes	Interjections	Overlapping	Pronouns
Insertions	Affixes			
		1.0 Errors		
Type/Category	1	2	3	4
Deletions	Nouns	Verbs	Function words	Foreign words
Substitutions	Nouns	Verbs	Function words	Foreign words
Insertions	Nouns	Verbs	Function words	

Source: Own elaboration.

3.2 Data Analysis Approaches

The study contains quantitative and qualitative parts. This combined approach can result in a more thorough and nuanced exploration of the analysis.

[1] Veed.io Linkedin page: https://www.linkedin.com/company/veedhq.

3.2.1 Qualitative method

Here, the analysis investigates the errors using a manual evaluation based on linguistic and lexical factors by comparing the generated transcription to the audible utterances. The study discusses the errors categorised into two values (0.5) and (1.0). Errors of Affixes, vowel length, and overlapping are given a 0.5 value, while errors of nouns, verbs, foreign words, and function words that are not affixed are given a 1.0 value.

3.2.2 Quantitative method

Word Error Rate (WER) is calculated. **WER** is a metric that measures the difference between the transcript generated by an ASR system and the actual transcript. Here, we calculate the percentage of words that were incorrectly transcribed by the ASR system. A lower WER indicates a higher-quality transcript.

3.3 Study Procedures

The procedures that are followed for the investigation of the *ASR-based subtitles* are as follows.

- **In the qualitative part:**
  1. The video is transcribed manually using Arabic Abjads in correct non-diacritised Arabic orthography (Without *ḥarakāt*).
  2. The transcription is divided into segments in an Excel sheet.
  3. The transcription is on one sheet, where this sheet represents the data and details for the website.
  4. The video is uploaded to the testing website.
  5. The website generates auto-generated subtitles.
  6. The generated transcription is extracted to txt. file, which ensures that the only data there is textual.
  7. The transcription is segmented into a column in one sheet.
  8. The errors are detected and analysed in the analysis section.
- **In the quantitative part:**
  1. The video is transcribed manually using Arabic Abjads in correct non-diacritised orthography and is considered to be the reference transcript.
  2. The transcript is divided into segments on an Excel sheet.
  3. The transcript is on one sheet, where the sheet represents the data and details for the testing website.
  4. The video is uploaded to the testing website.
  5. The website generates subtitles.
  6. The ASR-generated transcript (subtitles) is extracted to txt. file. This ensures that the only data there is textual data.
  7. The reference and ASR-generated transcripts are aligned. This determines which words in the ASR-generated transcript correspond to which words in the reference transcript.
  8. WER is calculated: The calculations of WER is computed using the following formula using Excel (Shah *et al.*, 2022):

$$\text{WER} = (S + D + I)/N$$

S = Number of substitutions (words in the ASR-generated transcript that differ from the corresponding words in the reference transcript).
D = Number of deletions (i.e., words in the reference transcript that are missing from the ASR-generated transcript)
I = Number of insertions (i.e., words in the ASR-generated transcript that are not present in the reference transcript)
N = The total number of words in the reference transcript.

  9. WER is interpreted using a percentage. A lower value indicates better accuracy. For example, a WER of 5% means that 5 out of every 100 words in the ASR-generated transcript are incorrect.

4 Findings and Discussion

4.1 Qualitative Analysis

4.1.1 Deletions

This section analyses the deletion errors. First, it discusses errors with 0.5, including affixes, interjections, overlapping, and vowel length. Second, it discusses errors with a 1.0 value, including nouns, verbs, function words, and foreign words. Table 3 below shows some examples of 0.5 deletion errors.

It is observed that the present progressive particle, which exists in the JA /b/ with its other form /bi/, is deleted, as **Example 1** shows. This particle can also be a future particle, depending on the situation. It is worth mentioning that this specific particle is weighty in JA, and sometimes when it is attached to the plural prefix /n/, it might become a nasal sound /m/ to match the nasal quality of the neighbouring sound /n/. **Example 2** shows when the coordinating conjunction /w/ or /wa/ was omitted from the subtitles. The release of this conjunction in standard Arabic is more articulate than the release of it in JA, in which the production of speech sounds is achieved with greater precision and clarity. The problem with this conjunction is that it is a semi-vowel, in which its acoustic characteristics can be influenced by the surrounding sounds and that the acoustic properties of /w/ can vary depending on the speaker, speaking style, and other factors, making it difficult for the model to consistently recognise this sound. Moreover, the definite determiner that is attached to indefinite nouns to define them was not recognised in its two forms, as **Example 3** shows. In Arabic, the lunar definite article occurs when the definite article is followed by a non-geminated consonant, and the solar definite article occurs when it is followed by a geminated consonant. Both of these should be written as "ال" in Arabic, yet they were omitted, which, according to Arabic grammar, may let the reader wait for a noun that is being described or modified by the first noun as part of "a construct phrase".

Additionally, the object pronoun that refers to the third singular masculine person /hi/, is omitted when it is attached to a preposition /fi/, as shown in **Example 4**. Deleting it could make subtitles hard to comprehend by the readers since they would assume that the word which will come after the preposition is its object, and without an object, the phrase is not completed.

Moreover, **Example 5** shows when /fa/, which can be a connective particle or coordinating conjunction, is deleted. In its function, it can help the audience in indicating relationships between various sentence elements and contributing to a clearer and more comprehensible message.

Particular attention is paid to prepositions during the investigation, which led to the assumption that three types were deleted: /b/ and its other form /bi/; /la/ and its other form /li/ and /ʕa/, in which the latter is a unique dialectal preposition that derived from the standard /ʕala/ as it shown in **Examples 6, 7 and 8**. Many reasons could be related to the fact that the dialectal versions of prepositions may have different pronunciation or acoustic properties compared to standard ones, and ASR systems may not have enough exposure to the specific dialectal pronunciations of prepositions to recognise them accurately.

The feminine markers /a/ and /a:t/ are deleted in two cases, as shown in **Examples 9 and 10**. In JA, this suffix is vital because these affixes indicate the gender of a noun, and without it, the meaning is lost. In one of the examples, "وشربا" means nothing in Arabic without "ت" and makes no sense.

The grammatical person of the subject or object is indicated through a variety of person affixes that are added to verbs, nouns, and prepositions in Arabic. Prefixes or suffixes are the most common forms of these affixes. **Examples 11,12,13,14,15, and 16** show some of these cases in which the ASR system did not recognise them when they were attached to the imperfective verb. It is important to highlight the fact that JA tends to drop vowel sounds in many affixes, i.e., vowel reduction, which leads to changes in pronunciation that may confuse the systems if they were not trained very well.

Furthermore, the accusative case marker in the form of (nunation) in Arabic is not recognised, as shown in **Example 17**. This is crucial because this specific marker added to the end of the final letter has a particular purpose: to refer to an unknown entity. Basically, "مثلاً" means "as an example", which usually indicates a pause or a stop before any coming sentence, but without the marker, it will be "For example", which lets the reader expect something to come in the utterance. It is worth mentioning that the Modern system of spelling for the Arabic language and its dialects does not

Table 3. Examples of 0.5 deletion errors in the ASR-based subtitles.

No.	Error Type		IPA (JA)	Utterance	ASR subtitle
1	Affix	Present Progressive Particle	/b/, /bi/	بتوصل، بتكون	توصل، تكون
2	Affix	Coordinating Conjunction	/w/, /wa/	و	X
3	Affix	Definite article	/ʔal/ , /ʔa/	ال ال ال ، الست ، التحية	، ست، تحية X
4	Third singular masculine pronoun (Object)		hi	فيه	في
5	Coordinating Conjunction/ Connective particle		/fa/	فانت، فبيقولك	انت، بقولك
6	Preposition		/b/, /bi/	بأريحية، بالموضوع	اريحيه، الموضوع
7	Preposition		/la/, /li/	للأشخاص، لقضية	الأشخاص، قضيه
8	Preposition		/ʕa/	عأساس	أساس
9	Suffix of singular feminine gender		/a/	الزايدة	زايد
10	Suffix of plural feminine gender		/aːt/	وشربات	وشربا
11	Imperfective second person, feminine, singular prefix		/t/	تحاولي	حاولي
12	Imperfective second person, masculine, singular prefix		/ti/	تكسر	كسر
13	Imperfective third person, masculine, singular prefix.		/j/	بيشوفك	بشوفك
14	Imperfective third person, masculine, plural prefix.		/j/	بيساعدوها	بساعدها
15	Imperfective second person, masculine, singular prefix		/t/	بتحبها	بحبها
16	Imperfective first person, masculine, plural prefix		/na/	بنطلب	بطلب
17	Accusative case marker (Nunation)		/an/	مثلا	مثل
18	Interjection		/ʔa| ʔa /	آه' آه	X
19	Interjection		/ʔe/	أي	X
20	Interjection		/ʔam/	أمم	X
21	Overlapping		/ ˈʔak.tar ˈmen ˈsit.tˤaˤʕ.ʃar ˈsaː.ʕit ˈsˤjaːm/	(هدول أصعب خمس دقايق) أكتر من ستطاشر ساعة صيام	(هدول أصعب خمس دقايق)
22	Overlapping		/' tˤaj·jib ˈ jal·laː/ /bas ˈʔak.tar min ˈheːk/	طيب يلا (أنا برجع الصحون) بس اكتر من هيك	(أنا برجع الصحون)
23	Vowel length		/mar.ˈraːt/	مرات	مره
24	Vowel length		/ ˈʔaː.lat/	قالت	قلت

Source: Own elaboration.

require the writing of diacritics, so "مثلاً" can be understandable if it is shown as "مثلا".

Moving to interjection deletion errors, these words are frequently used instantly, making them challenging for automatic subtitle-generation systems to catch, which might cause errors in the auto-generated subtitles. Surely, detecting these items may not be critical for the readers, but it would be highly important to detect for the DHH. In many cases, interjections were used in conjunction with other words in a sentence, which makes it hard to capture.

Example 18 shows a Jordanian unique interjection /ʔaː/, which is used to express agreement or affirmation, is deleted. It is important to recognise such an interjection since it is widely used in JA. Here, it is deleted when the broadcaster wanted to affirm a statement that the other broadcaster mentioned to comment.

- Utterance: أنا معاكي ، _...

  Auto-generated subtitles: أنا معاكي...

Examples 19 and 20 show examples where other types of interjections were deleted, yet it is worth mentioning that they are not crucial to be subtitled. **Examples 21 and 22** show examples of overlapping deletion errors. Overlapping deletion errors are worth mentioning since, in many times, they would result in deleting parts of the dialogue because the system may not be able to differentiate between people speaking, while in normal settings, people would turn-take to avoid interruptions. Also, when a speaker begins to talk before another speaker has finished speaking can cause overlapping, mainly when these two speakers use similar speech patterns. Yet, some overlapping types cannot be avoided in ordinary speech, and they are not important to be recognised in the transcription, such as backchannel where speakers use short interjections "Ah, mmm..." to show interest in the topic.

Vowel length errors occur because the system may fail to accurately recognise and transcribe the length of a vowel sound in a word. In JA, the results would let the words spelt inaccurately, which may hold a different meaning. **Example 23** shows the deletion of the long vowel /aː/, which is effective since deleting such a part of a morpheme lets the word become singular while it is actually plural.

Example 24 as well shows a deletion of /aː/. Deleting it resulted in changing the verb from the perfective verb of the third feminine person to the perfective verb of the first singular person.

Table 4. Examples of deletion errors with a 1.0 value.

No.	Utterance	Auto-generated subtitle	Deleted words	Type
25	وبعدين أنت كضيف رايح تفطر عند ناس	وبعدين انت ضعيف	تفطر	Verb
26	أصحابك	رايح عند اصحابك	ناس	Noun
27			هلأ أول ، العزايم	Noun
28	هلأ أول اتيكيت من اتيكيت العزايم برمضان	في رمضان	اتيكيت ، اتيكيت	Foreign word
29			من	Function word
30	لأنه أنا كتير باذلة مجهود	بعد المجهود	كتير	Noun
31	ما فش Clash	X	Clash	Foreign word
32	لأنه أنا كتير باذلة مجهود	بعد المجهود	لأنه أنا	Function word
33	Which is not nice	X	Which is not nice	Foreign words
34	يعني الspaghetti أكلة آي آي نوعا ما أكلها صعب	يعني اكله نوعا ما اكلها صعب	الspaghetti	Foreign word
35	Sorry	X	Sorry	Foreign word

Source: Own elaboration.

On the other hand, looking at errors that have a 1.0 value, cases were related to nouns, verbs, function words, and foreign words. Table 4 shows some examples.

In most cases of foreign words, the subtitles were not generated. These foreign words do not

follow Arabic language or Jordanian dialect patterns, although these words are words used by many Jordanians, such as "Etiquette", "Sorry", and "spaghetti", as shown in **Examples 28, 31, 33, 34 and 35,** so it is essential to train the systems on these words.

Nouns, verbs, and function words are also deleted on many occasions, as shown in **Examples 25, 26, 27, 29, 30 , and 32.** The verb "تفطر" is not detected in the subtitles, and a noun like "العزائم" was also not detected. Also, a function word like "من" was omitted.

Recognising the words could be challenging. These deletions result from many reasons, such as unclear speech or lack of data, which let the system struggle to recognise words spoken and can lead to words being deleted.

4.1.2 Substitutions

This section analyses the substitution errors based on their value. First, it discusses errors with 0.5, which are affixes, interjections, overlapping, and pronouns. Secondly, it discusses errors with a 1.0 value: nouns, verbs, function words, and foreign words. When looking at errors with a 0.5 value, most cases were related to affixes. Table 5 shows some examples.

Table 5. Examples of functional morphemes substitution errors with 0.5 value.

No.	Type	Utterance type	Subtitle type	Utterance	Subtitle
1	Affix	Conjunction	Conjunction	أو شكرا	وشكرا
2		Imperfective third person, masculine, plural prefix.	Imperfective second person, masculine, plural prefix.	يعرفوا	تعرفوا
3		Third singular masculine pronoun	Second masculine/feminine singular pronoun	عليه	عليك
4		Preposition	Preposition	بِ	في
5	Overlapping	Preposition + Noun	Preposition + Noun	لسبعة	بسرعة
6	Interjection	Interjection	Interjection	له	لا

Source: Own elaboration.

Example 1 shows an example when an affix was deleted. In JA, people pronounce the coordinating conjunction "و" in two ways; it could be /ʔuw/ or /wa/, where vowel length may differ depending on the speaker. The first one is widespread and connecting it to other words would be problematic and confusing for ASR systems if it was not trained well since another connective conjunction has a similar pronunciation, yet not the same function, which is "أو" /ʔaw/. Here, the conjunction was "أو", but the transcription was "و". Looking at prefixes, we can see that a change was done in some cases where in **Example 2**, the third person prefix "يعرفوا" was changed to a second person prefix "تعرفوا". Such a change is not acceptable from a grammatical perspective since the word that followed this imperfective verb was "حالهم" "Their situation", which refers to a third person. **Example 3** shows when suffixes are attached to prepositions are replaced with other suffixes. In our example, the third masculine singular pronoun was replaced with the second masculine/feminine pronoun. It is not assumed whether it is feminine or masculine since both share the same consonant but not the same vowel, where masculine is /ka/ and feminine is /ki/. This is because the modern Arabic writing system does not require diacritics, which are vowels in 3 cases. It is worth mentioning that these pronouns are attached to a preposition in the genitive case. **Example 4** shows an example of replacing /bi/ with /fi:/. We can relate this to the fact that many Jordanians do not use these two prepositions as per their usage in the standard language, which makes it confusing when do Jordanians use /bi/ and when do they use /fi/ in their dialect. Some changes are related to lenition, where people alter consonants to make them more sonorous. Here, the ASR system transcribed the stop /bi/ into a fricative /fi/, in which the ASR system predicted it – as the analysis assumes – a kind of spirantisation. Below are some samples from the uploaded video.

- *Utterance:* لطيف وخفيف لهُ علاقة برمضان
  Subtitle: لطيف وخفيف له علاقه في رمضان

- *Utterance:* لانه خصوصا بالشهر الفضيل
  Subtitle: لانه خصوصا في الشهر الفضيل

Example 5 shows when the preposition ل changed to another preposition ب, while the noun "سبعة" was changed to another word "سرعة". Due to overlapping, the ASR system here generated a new word, which is not part of the utterance. In **Example 6**, the interjection "لهْ"/lahˈ/, which means "Oh", was changed to "لا"/laːˈ/, which means "No". Yet, in that specific example, the change did not affect the message that was delivered since both the subtitle and utterance meant, "You reach there and say, oh no, guys." The Arabic utterance and subtitle are shown below.

- Utterance: بتوصل أنت له يا جماعة
  Subtitle: توصل انت لا يا جماعه

On the other hand, looking at errors that hold a 1.0 value, cases were related to nouns, verbs, function words, and foreign words. Table 6 shows some examples.

Table 6. Examples of substitution errors hold a value of 1.0.

No.	Utterance	Auto-generated subtitle	Original	Substituted
7	بعدين عفكرة عسيرة الشوربة make sure أنه الشوربة ما تكون كتير سخنة	الشوربه الشوربه ما تكون كثير سخنه	Make sure	مكشور
8	أنا للأسف sorry اللي ما بيقدر يقول عندي ارتباط آخر	اللي ما بيقدر يقول انا للاسف عندي ارتباط اخر	sorry	سري
9	أنت متأكدة unless ما تتفنني كتير كتير إنه أنا مثلا عازمة نادية	_ كثير كثير_ انت متاكده انه انا مثل	ما تتفنني unless	ماتت فنان..على الناس..
10	اعمليها مع تومة بس مثلا مش لنفترض اي نوع غريب ما عمره حدا أكله pasta	اعمليها بس مثلا نوع غريب ما حدا اكل	معصومة.	مع تومة.
11	ولا عزومة عشا ممكن حدا بيجي ماكل قبل	ولا عزومه عشاء ممكن حدا.معك الابل	قبل	الابل
12	تقعدوا تتحدثوا قبل ما تبلشوا تاكلوا كذا	تتحدث قبل ما تبلش	تقعدوا	تعد
13	اللي بدها تضيف فيهم يكونوا محطوطين عالسفرة	محطوطين على طاوله	السفرة	الصفرة

Source: Own elaboration.

Many cases show the inability of this ASR system to recognise foreign words, which corresponds with the results of Mustafa *et al.* (2022). When it comes to substitution cases, this ASR model attempts to transcribe the foreign words as words that exist in JA. **Examples 7, 8, and 9** show when "make sure"/meɪk ʃʊr/ was transcribed as "مكشور"/mak.ˈʃuːr/, "sorry" /ˈsɒri/ was transcribed as "سري"/ˈsir.riː/, and "unless" /ʌn.ˈlɛs/ was transcribed as "على الناس" /ʕa.lan.ˈnaːs/ and in JA, it would be pronounced as /ʕan.ˈnaːs/.

Moreover, when it comes to Arabic words, in some cases, such as **Examples 10 and 11**, the ASR system did not recognise the syllable breaks and word boundaries correctly, which led to generating wrong words that are similar in most of the consonants and vowels, yet not the same breaks. For example, " ما تتفنني" / maː.ˈtit.fan.na.ni / was transcribed as "ماتت فنان" /ˈmaː.tat. fan.ˈnaːn/, and "مع تومة" /maʔ. ˈtɔː.me/ was transcribed as معصومة /maʕ.ˈsˤɔː.me/.

Also, in Amman, many people -especially women- drop the consonant /q/ (ق) and replace it with a glottal stop /ʔ/ in certain word positions. Sometimes, they make pronunciation easier by replacing that voiceless uvular plosive or the glottal stop with a long or short vowel /a/. Therefore, as shown in **Example 12**, the ASR system may wrongly predict another word that exists in Arabic. For example: "قبل" /ˈʔa.bil/ where /ʔ/ is dialectal and was transcribed as الابل /ʔal.ˈʔibl/. Also, "تقعدوا", /tu.ˈʕu.duː/ was changed to "تعد" /ta.ˈʕud/.

In **Example 13**, the ASR system transcribed a consonant into a pharyngealised form, where the /s/ sound was represented as /sˤ/. This is crucial in Arabic because one simple change can lead to a change in the whole meaning if it is not transcribed well. In our case, the word "سفرة" /ˈsuf.ra/ was transcribed as "صفرة" /ˈsˤuf.ra/.

4.1.3 Insertions

This section analyses the insertion errors based on their value. First, it discusses errors with 0.5 that are affixed. Secondly, it discusses errors with a 1.0 value: nouns, verbs, and function words. When looking at errors with a 0.5 value, all cases were related to affixes that were not added to the subtitles. Table 7 shows all the examples.

Table 7. Examples of insertion errors with 0.5 value resulted in adding affixes in the subtitles.

No.	Utterance	subtitles
1	منطق	منطقه
2	فكري	تفكري
3	تحية	التحية
4	بعدين	وبعدين

Source: Own elaboration.

Due to many factors, ASR systems may insert affixes to words that were not present in the original utterance. **Examples 1, 2, 3, and 4** show samples**.** When it comes to JA, we can relate this to the complex affixation system with many affixes, which, in many cases, are similar in sound, making it hard for the system to identify them accurately. The detected insertions are few, yet they affect the meanings and the delivered message.

On the other hand, looking at errors with a 1.0 value where cases were related to nouns, verbs, and function words. These were few when compared to deletions and substitutions. **Example 5, 6, 7, and 8** in Table 8 below shows some cases.

Table 8. Examples of insertion errors with a 1.0 value resulted in adding words in the subtitles.

No.	Utterance	Subtitle	Insertion
5	مرتبة	مكتبه مرتبه	مكتبة
6	بس هاتي أقولك نكتة على سيرة رمضان (آه)	اقول لك نكته قبل ما انزل على سيره العضل	قبل ما أنزل
7	بعرف نادية بتحب مثلا أكلة معينة بعملها (صح)	طيب عرفنا ديه بتحب مثلا اكله معينه بعملها	طيب
8	يعني جلينا حلو حطينا هالصحون روحنا	هذا قصدي جالينا حلو حاطينها الصحون روحنا	هذا

Source: Own elaboration.

Regardless of the fact that these errors are the result of some technical issues, many of these insertions could be inserted as full words because the speech signals were not recognised accurately, as well as the limitations of words that these systems can detect. Therefore, the insertions exist, and it is represented in Table 8.

To sum up, affixes can significantly impact the reader's understanding of the meaning of a word or phrase. When it comes to JA, deleting, replacing, or adding them to the subtitles can change the meanings or delete crucial information about it, and sometimes may lead to creating new words. Also, while interjections were not crucial, overlapping caused many errors in subtitling that often omitted parts of the subtitles or created new words that were mixed up. Moreover, deleting, replacing, or inserting words such as nouns, verbs, foreign words, and some function words is critical and can lead to misinterpretation. Therefore, it is crucial to let the ASR systems have a large set of JA language data that would help recognise patterns and make predictions or decisions based on that set.

4.1.4 The Quantitative Analysis

This section revolves around the computation of the Word Error Rate (WER) through manual and Excel calculations. Since the focus of the study is on the JA using the unique Arabic writing system, which is Abjad, and to ensure an accurate calculation of the total number of words in the reference transcript of each cell, all punctuation marks are removed. Excel formulas are inserted, and the SUM is computed. After calculating words in both the reference transcript and auto-generated subtitles,

it is revealed that the reference transcript contains more words representing the actual utterance than the auto-generated transcript. This suggests that the ASR model failed to recognise all the words accurately. Table 9 displays the total word count for the original transcript (reference) and the auto-generated subtitles.

Table 9. Total number of words in both original transcript and Auto-generated subtitles.

Category	Volume
Total words in Reference	1943
Total words in Auto-generated Subtitles	1435

Source: Own elaboration.

4.1.5 Word Error Rate

This section discusses the word error rate. The errors were either full errors that can lead to confusion or total misinterpretation, or partial errors that would cause a little misunderstanding. Each full error is given a value of 1.0, while partial errors are given 0.5. The manual calculations involved determine the number of insertions, deletions, and substitutions. Figure 4 shows the percentage of each type of error out of the total errors.

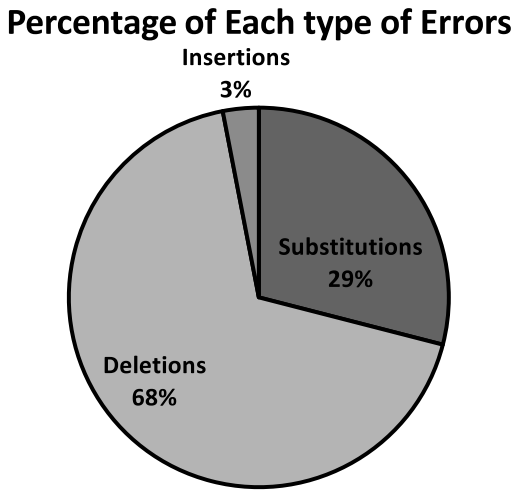

Figure 4. The percentage of each type of error out of the total errors.
Source: Own Elaboration.

The majority of errors were identified as deletions (68%), succeeded by substitutions (29%), and then insertions (3%). Upon performing calculations for substitutions, deletions, and insertions and adding up the resulting numbers, the sum is divided by the total number of words in the reference. Then, multiply the resulting number by 100% to show that the WER percentage is 38.857%. Table 10 reveals the numbers that were used in the mathematical equation.

Table 10. The total sum of the error values categorised by their type with WER percentage.

Category	Number
Deletions (D)	513
Substitutions (S)	219
Insertions (I)	23
Total words in Reference (N)	1943
Word Error Rate in Percentage (WER)	38.857%

Source: Own elaboration.

Regarding deletion errors, the primary cause appears to be the misidentification of JA affixes,

accounting for 55% of errors. This is followed by overlapping, where multiple speakers talk simultaneously, accounting for 26% of errors. Misrecognising interjections accounted for 17% of errors, while deletion of long vowels and shortening length only accounted for 2%. Figure 5 shows these percentages in a pie chart.

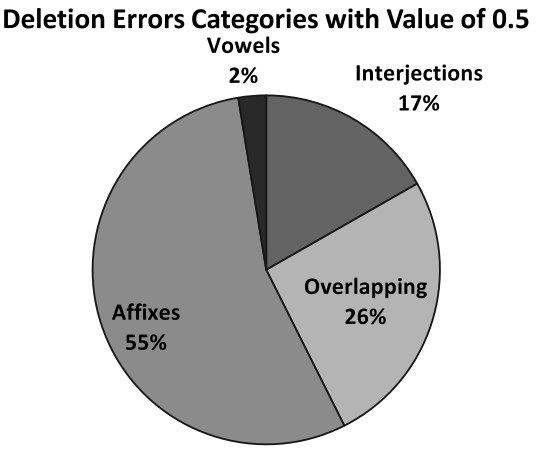

Figure 5. The percentage of each type of deletion error with 0.5 value.
Source: Own Elaboration.

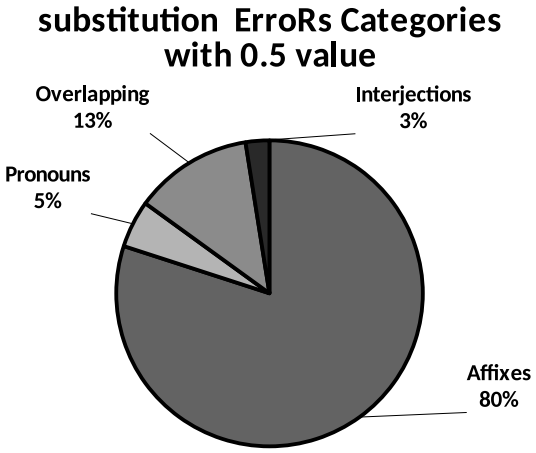

Figure 6. Percentage of each type of substitution error with 0.5 value.
Source: Own Elaboration.

When examining substitution errors, it appears that the primary cause is the same cause of deletion errors, which is the misidentification of JA affixes, accounting for 80% of errors. Overlapping errors account for 12%, errors with pronouns account for 5% of errors, while misrecognised interjections account for 3%. The percentages are depicted in a pie chart in Figure 6.

With a total of 8 errors that hold the value of 0.5, the insertion errors were all caused because of insertion for affixes not found in the original transcript, as Figure 7 shows.

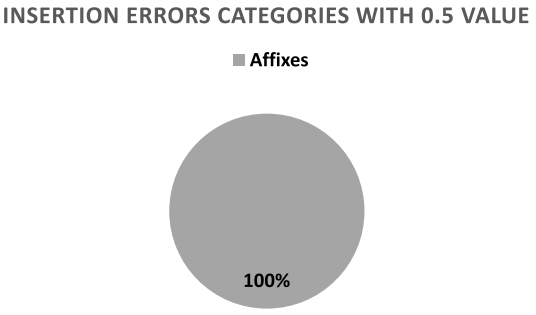

Figure 7. The percentage of each type of substitution error with 0.5 value.
Source: Own Elaboration.

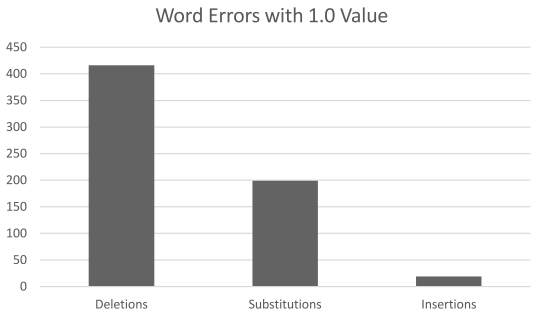

Figure 8. The Total number of errors for each type with a 1.0 value.
Source: Own Elaboration.

On the other hand, the analysis of word errors that hold a 1.0 value revealed that 416 of these errors were deletions, followed by substitutions with 199 errors, while insertions were only 19 errors. Figure 8 shows a clustered column chart that compares the values across these categories.

5 Conclusion

This research paper examines the linguistic accuracy of veed.io when automatically generating intralingual subtitles for a video in Jordanian Arabic, where two female broadcasters talk about the etiquette of meals in Ramadan. In specific, it explores the obstacles that machines may face when dealing with various linguistic and phonetic phenomena in Jordanian Arabic.

In the qualitative part, errors are categorised into three main types: deletions, substitutions, and insertions. Deletions occur when the system does not fully or partially transcribe the utterance, substitution happens when the system replaces one utterance with another, and insertion occurs when

the system inserts an item that does not exist in the original utterance. Furthermore, the analysis subcategorises errors into two main types. Errors that did not significantly affect the comprehension of the text are assigned a value of 0.5, and errors that affected the comprehension of the subtitles are assigned a value of 1.0.

Notably, the analysis revealed the deletion of 0.5 errors involves affixes, interjections, overlapping, and vowel length, while those of 1.0 include nouns, verbs, function words, and foreign words. Similarly, substitution errors of 0.5 are affixes, interjections, overlapping, and pronouns, while those of 1.0 errors are nouns, verbs, function words, and foreign words. Lastly, insertion errors of 0.5 are affixes, while those of 1.0 are nouns, verbs, and function words. These findings have implications for the readability of the comprehension of subtitles.

In the quantitative analysis, the study employed Excel formulas to calculate the word error rate (WER), which amounted to 38.857%, showing that the majority of errors are identified as deletions (68%), succeeded by substitutions (29%), and finally insertions (3%).

Furthermore, the research methodology proved to be effective in suggesting a specific classification for AI-Powered subtitles and analysing the errors. The proposed taxonomy categorises AI-powered subtitles into intralingual and interlingual types based on two factors: the used technology and the target language. The study suggests that Intralingual subtitles can be subcategorised into four main types: SR-based; ASR-based; Semi-ASR-based; SLR-based, and interlingual subtitles can be MT-based or Semi MT-based. Moreover, the methodology was developed with consideration for the unique orthographical rules and writing system of the Arabic language.

To acknowledge the limitations of our study, measuring ASR accuracy is necessary and requires tools that are specifically designed for the Arabic language. These tools should be able to differentiate between two types of errors, which can be 0.5 error or 1.0 error, along with no errors that hold a value of 0. Due to the lack of such a tool, the study used Excel formulas and manually differentiated between these errors. Also, more research is needed in the field of Jordanian dialect.

Based on the findings, the study suggests more research in the field of ASR systems when dealing with different Arabic dialects and the other types of technology that deal with language. Also, subtitlers and content creators should be aware of these challenges when using these online tools.

This research serves as a steppingstone towards development in language technology. By continuing to investigate and expand our knowledge, we can contribute to advancements and improvements in auto-generated subtitles of Jordanian Arabic and make meaningful contributions to AVT and NLP.

Acknowledgment

The authors are thankful to the Deanship of Graduate Studies and Scientific Research at University of Bisha for supporting this work through the Fast-Track Research Support Program.

References

AL-ABBAS, Linda S.; HAIDER, Ahmad S. Using Modern Standard Arabic in subtitling Egyptian comedy movies for the deaf/ hard of hearing. Ed. by Maria Del Mar And Sanchez Ramos. *Cogent Arts & Humanities*, v. 8, n. 1, p. 1993597, Jan. 2021. ISSN 2331-1983. DOI: 10.1080/23311983.2021.1993597. Available from: https://www.tandfonline.com/doi/full/10.1080/23311983.2021.1993597. Visited on: 23 Nov. 2023.

AL MAHASEES, Zakaryia. *Analysing English-Arabic machine translation: Google Translate, Microsoft Translator and Sakhr*. London ; New York: Routledge, 2021. (Routledge studies in translation technology). ISBN 9780367759117.

ALHARBI, Sadeen; ALRAZGAN, Muna; ALRASHED, Alanoud; ALNOMASI, Turkiayh; ALMOJEL, Raghad; ALHARBI, Rimah; ALHARBI, Saja; ALTURKI, Sahar; ALSHEHRI, Fatimah; ALMOJIL, Maha. Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, v. 9, p. 131858–131876, 2021. ISSN 2169-3536. DOI: 10.1109/ACCESS.2021.3112535. Available from: https://ieeexplore.ieee.org/document/9536732/. Visited on: 23 Nov. 2023.

ALMAHASEES, Zakaryia; JACCOMARD, Helene. Facebook Translation Service (FTS) Usage among Jordanians during COVID-19 Lockdown. *Advances in Science, Technology and Engineering Systems Journal*,

v. 5, n. 6, p. 514–519, Nov. 2020. ISSN 24156698, 24156698. DOI: 10.25046/aj050661. Available from: https://astesj.com/v05/i06/p61/. Visited on: 23 Nov. 2023.

ALMAHASEES, Zakaryia Mustafa. Machine Translation Quality of Khalil Gibran's the Prophet. *SSRN Electronic Journal*, v. 1, n. 4, p. 151–159, 2017. ISSN 1556-5068. DOI: 10.2139/ssrn.3068518. Available from: https://www.ssrn.com/abstract=3068518. Visited on: 23 Nov. 2023.

BEHNSTEDT, P.; WOIDICH, M. Dialectology. *In:* OWENS, J. (ed.). *The Oxford Handbook of Arabic Linguistics*. Oxford, England: Oxford University Press, 2013. p. 300–325.

BENDOU, Imane. *Automatic Arabic Translation of English Educational Content Online using Neural Machine Translation: the Case of Khan Academy*. Oct. 2021. thesis – Carnegie Mellon University. DOI: 10.1184/R1/16725304.v1. Available from: https://kilthub.cmu.edu/articles/thesis/Automatic%5C_Arabic%5C_Translation%5C_of%5C_English%5C_Educational%5C_Content%5C_Online%5C_using%5C_Neural%5C_Machine%5C_Translation%5C_the%5C_Case%5C_of%5C_Khan%5C_Academy/16725304/1. Visited on: 23 Nov. 2023.

CHAUME, Frederic. The turn of audiovisual translation: New audiences and new technologies. *Translation Spaces*, v. 2, p. 105–123, Nov. 2013. ISSN 2211-3711, 2211-372X. DOI: 10.1075/ts.2.06cha. Available from: http://www.jbe-platform.com/content/journals/10.1075/ts.2.06cha. Visited on: 23 Nov. 2023.

DHARMALE, Gulbakshee J.; PATIL, Dipti D. Evaluation of Phonetic System for Speech Recognition on Smartphone. *International Journal of Innovative Technology and Exploring Engineering*, v. 8, n. 10, p. 3354–3359, Aug. 2019. ISSN 22783075. DOI: 10.35940/ijitee.J1215.0881019. Available from: https://www.ijitee.org/portfolio-item/J12150881019/. Visited on: 23 Nov. 2023.

DÍAZ-CINTAS, J.; REMAEL, A. *Audiovisual Translation: Subtitling*. London: Routledge, 2007. Available from: https://www.amazon.com.br/Audiovisual-Translation-Subtitling-Jorge-D%5C%C3%5C%ADaz-Cintas/dp/1900650959. Visited on: 23 Nov. 2023.

DOUGHAN, Yazan. Imaginaries of Space and Language: A historical view of the scalar enregisterment of Jordanian Arabic. *International Journal of Arabic Linguistics*, v. 3, n. 2, p. 77–109, 2017. ISSN 2421-9835. Available from: https://revues.imist.ma/index.php/IJAL/article/view/11572. Visited on: 23 Nov. 2023.

GUSKAROSKA, A. *ASR as a tool for providing feedback for vowel pronunciation practice*. 2019. Master of Arts – Iowa State University, Ames, Iowa.

HAIDER, Ahmad S.; ALROUSAN, Faurah. Dubbing television advertisements across cultures and languages: A case study of English and Arabic. *Language Value*, v. 15, n. 2, p. 54–80, Dec. 2022. ISSN 1989-7103. DOI: 10.6035/languagev.6922. Available from: https://www.e-revistes.uji.es/index.php/languagevalue/article/view/6922. Visited on: 23 Nov. 2023.

HAIDER, Ahmad S.; SAIDEEN, Bassam; HUSSEIN, Riyad F. Subtitling Taboo Expressions from a Conservative to a More Liberal Culture: The Case of the Arab TV Series Jinn. *Middle East Journal of Culture and Communication*, v. 16, n. 4, p. 363–385, Mar. 2023. ISSN 1873-9857, 1873-9865. DOI: 10.1163/18739865-tat00006. Available from: https://brill.com/view/journals/mjcc/16/4/article-p363%5C_1.xml. Visited on: 23 Nov. 2023.

JARRAH, Shatha; HAIDER, Ahmad S.; AL-SALMAN, Saleh. Strategies of Localizing Video Games into Arabic: A Case Study of PUBG and Free Fire. *Open Cultural Studies*, v. 7, n. 1, p. 20220179, July 2023. ISSN 2451-3474. DOI: 10.1515/culture-2022-0179. Available from: https://www.degruyter.com/document/doi/10.1515/culture-2022-0179/html. Visited on: 23 Nov. 2023.

LIAO, Junwei; ESKIMEZ, Sefik; LU, Liyang; SHI, Yu; GONG, Ming; SHOU, Linjun; QU, Hong; ZENG, Michael. Improving Readability for Automatic Speech Recognition Transcription. *ACM Transactions on Asian and Low-Resource Language Information Processing*, v. 22, n. 5, p. 1–23, May 2023. ISSN 2375-4699, 2375-4702. DOI: 10.1145/3557894. Available from: https://dl.acm.org/doi/10.1145/3557894. Visited on: 23 Nov. 2023.

MAAMOURI, Mohamed; BIES, Ann; BUCKWALTER, Tim; DIAB, Mona; HABASH, Nizar; RAMBOW, Owen; TABESSI, Dalila. Developing and Using a Pilot Dialectal Arabic Treebank. *In:* CALZOLARI, Nicoletta; CHOUKRI, Khalid; GANGEMI, Aldo; MAEGAARD, Bente; MARIANI, Joseph;

ODIJK, Jan; TAPIAS, Daniel (eds.). *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA), May 2006. Available from: http://www.lrec-conf.org/proceedings/lrec2006/pdf/543%5C_pdf.pdf. Visited on: 23 Nov. 2023.

MUSTAFA, Mumtaz Begum; YUSOOF, Mansoor Ali; KHALAF, Hasan Kahtan; RAHMAN MAHMOUD ABUSHARIAH, Ahmad Abdel; KIAH, Miss Laiha Mat; TING, Hua Nong; MUTHAIYAH, Saravanan. Code-Switching in Automatic Speech Recognition: The Issues and Future Directions. *Applied Sciences*, v. 12, n. 19, p. 9541, Sept. 2022. ISSN 2076-3417. DOI: 10.3390/app12199541. Available from: https://www.mdpi.com/2076-3417/12/19/9541. Visited on: 23 Nov. 2023.

PUCCI, M. Towards Universally Designed Communication: Opportunities and Challenges in the Use of Automatic Speech Recognition Systems to Support Access, Understanding and Use of Information in Communicative Settings. *In:* BAMGBOJE-AYODELE, A.; PRGOMET, M.; KUZIEMSKY, C.; ELKIN, P.; NOHR, C. (eds.). *Studies in Health Technology and Informatics*. Amsterdam: IOS Press, 2023. p. 18–25.

REMAEL, A. Audiovisual translation. *In:* YVES GAMBIER, L. V. D. (ed.). *Handbook of translation studies*. [*S. l.*]: John Benjamins Publishing Company, 2010. p. 12–17.

RYDING, Karin C. *A reference grammar of modern standard Arabic*. New York: Cambridge University Press, 2005. ISBN 9780521771511.

SABIR, Iram; ALSAEED, Nora. A Brief Description of Consonants in Modern Standard Arabic. *Linguistics and Literature Studies*, v. 2, n. 7, p. 185–189, Nov. 2014. ISSN 2331-642X, 2331-6438. DOI: 10.13189/lls.2014.020702. Available from: http://www.hrpub.org/journals/article%5C_info.php?aid=1920. Visited on: 23 Nov. 2023.

SAWAKARE, Praphulla A.; DESHMUKH, Ratndeep R.; SHRISHRIMAL, Pukhraj P. Speech Recognition Techniques: A Review. *International Journal of Scientific & Engineering Research*, v. 6, n. 8, p. 1693–1698, 2015. Available from: https://www.ijser.org/researchpaper/Speech-Recognition-Techniques-A-Review.pdf.

SCHLIPPE, Tim; ALESSAI, Shaimaa; EL-TAWEEL, Ghanimeh; WÖLFEL, Matthias; ZAGHOUANI, Wajdi. Visualizing Voice Characteristics with Type Design in Closed Captions for Arabic. *In:* 2020 International Conference on Cyberworlds (CW). [*S. l.: s. n.*], Sept. 2020. p. 196–203. ISSN: 2642-3596. DOI: 10.1109/CW49994.2020.00039. Available from: https://ieeexplore.ieee.org/document/9240549/citations%5C#citations. Visited on: 23 Nov. 2023.

SUVOROV, R.; LEVIS, J. M. Automatic Speech Recognition. *In:* CHAPELLE, C. (ed.). *Encyclopedia of Applied Linguistics*. Iowa: Blackwell, 2012. p. 8.

VERSTEEGH, K. *Encyclopedia of Arabic Language and Linguistics – Brill*. Brill: Leiden, 2006.

XIE, B. A comparative study of machine translated subtitles based on the user-centered approach: a case study between Bilibili and YouTube. *Research Square*, v. 1, 2022. Available from: https://www.researchsquare.com/article/rs-2179598/v1. Visited on: 23 Nov. 2023.

Author contributions

Wala' Mohammad Akasheh: Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Validation, Writing – original draft, Writing – review and editing, Visualization; **Ahmad S. Haider**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review and editing, Visualization; **Bassam Al-Saideen**: Investigation, Resources, Writing – review and editing; **Yousef Sahari**: Conceptualization, Resources, Writing – review and editing.