Quality and Comprehensibility of Interlingual Subtitles Produced by Humans or with Machines

Lara Shoana Schlüter[1] **Ekaterina Lapshinova-Koltunski**[1] **Sylvia Jaki**[2]
[1]University of Hildesheim [2]KU Leuven
[1]schueter, lapshinovakoltun@uni-hildesheim.de [2]sylvia.jaki@kuleuven.be

Abstract

The present paper focuses on the analysis of automatic subtitles produced with three different systems (HappyScribe, CapCut and Amberscript). We compare the outputs among each other paying attention to the categories of quality derived from audiovisual translation quality research. Besides that, we consider the comprehensibility of the produced subtitles. Additionally, we use automatic evaluation scores from BLEU, BLEURT and BERTScore to assess the overall quality. Our results show that automatically generated subtitles remain below human standards in quality and comprehensibility.

1 Introduction

The rapid development of language technology, including automatic speech recognition (ASR) and machine translation (MT), is profoundly reshaping subtitling workflows and professional practices (Schmalz, 2019; Tardel, 2023). While ASR has long been used in intralingual subtitling (Künzli, 2024b), demand is increasingly shifting towards fully automatically generated subtitles produced without direct human intervention (Agnetta, 2025). This trend is closely linked to the ongoing digitalisation of media and the exponential growth of audiovisual content: Between 2014 and 2019 alone, the number of hours of "Netflix original content" increased by 2,400% (Papi et al., 2023), exemplifying the scale of production that challenges traditional subtitling capacities.

To accommodate the increase, the audiovisual translation (AVT) industry is integrating technologies driven by Artificial Intelligence (AI), see e.g. (Bolaños García-Escribano, 2025). Recent advances in large language models (LLMs) show that these models are also well-suited for handling subtitling tasks (Tang et al., 2026). AI thus occupies a structural position within contemporary AVT ecosystems, shaping expectations and practices among both professionals and audiences (Jaki et al., 2024). However, subtitling remains a cognitively and multimodally complex task (Jüngst, 2020).

While subtitling is inherently a multimodal practice, this study adopts a text-centred analytical approach to systematically evaluate linguistic quality and comprehensibility as key components of multimodal reception. Although multimodality is not directly operationalised, several analysed features such as omissions, segmentation, and reading speed (CPS/CPL), are functionally linked to multimodal processing and cognitive load. We therefore frame this study as an investigation of linguistic performance under multimodal constraints rather than as a comprehensive multimodal analysis. High-level competences are therefore necessary at every stage of the process. In automated subtitling, these stages are distributed across technological components, including ASR, automatic spotting and segmentation, and MT (Karakanta et al., 2022).

In this paper, we analyse automatically generated interlingual subtitles from English into German, comparing subtitles produced with three different systems. Besides this, we compare machine-generated subtitles with those produced by humans. Our aim is to systematically evaluate the linguistic performance of AI-based subtitle gener-

ation systems under multimodal constraints, focusing on quality, comprehensibility and recurring error patterns. We specifically pay attention to the quality measured with the framework established in audiovisual translation, the FAR model. We also aim to find out if the automatically generated subtitles achieve a high level of comprehensibility exploring possible comprehensibility issues that occur in automatically generated subtitles.

The remainder of the paper is organised as follows. Section 2 describes the main challenges of subtitle production and gives an overview of related work. Section 3 presents our research design. In Section 4, we outline our results, before drawing conclusions in Section 5.

2 Theoretical Background and Related Work

2.1 Challenges in subtitling

From a technical perspective, subtitling requires precise synchronisation between image, source text audio and subtitle display, a process referred to as *spotting* (Díaz Cintas, 2015; Jüngst, 2020). Spotting ensures that subtitles appear and disappear in alignment with the audiovisual signal and verbal content (ibid.). This process helps determining the duration of the subtitles, i.e., how long the subtitles are displayed on screen for the audience (Kurch et al., 2015, 155).

Display time is generally constrained by minimum and maximum thresholds. A subtitle should remain on screen for at least one second (Kurch et al., 2015) and no longer than six seconds (Diaz Cintas and Remael, 2007). Readability guidelines specify quantitative parameters: Characters per second (CPS) should not exceed 15 (Mälzer and Wünsche, 2018), and characters per line (CPL) range between 37 and 42 (Jüngst, 2020; Papi et al., 2023). Limits in CPL vary depending on the target audience and institutional guidelines (Künzli, 2017; Mälzer and Wünsche, 2019).

Across academic research, guidelines and public broadcasting standards, there is consensus that subtitles should not exceed two lines (Bolaños García-Escribano, 2025; Jüngst, 2020). To ensure comprehensibility when transferring longer utterances into subtitle format, segmentation plays a central role (Jüngst, 2020). During segmentation, sentences are divided into logical semantic units, enabling efficient processing and (information) reception by the target audience (Kurch et al., 2015). Segmentation thus supports readability under strict temporal and spatial constraints. These constraints show that subtitling is not merely a linguistic transfer, but a regulated multimodal practice governed by cognitive, temporal and spatial parameters.

Given the limitations imposed by characters per second (CPS), characters per line (CPL), and display time, spoken dialogue cannot be transcribed verbatim (Kurch, 2018; Mälzer and Wünsche, 2018). Subtitlers therefore apply reduction and abbreviation strategies to ensure clarity and readability (Kurch, 2018). Common techniques include condensation, summarisation and paraphrasing (Papi et al., 2023; Wünsche, 2022). Further strategies involve omission or deletion, which affect non-essential elements of the source text – such as repetitions or interjections (Jüngst, 2020) – rather than content-bearing units (Kurch, 2018). The reduction is possible because audiovisual texts distribute information across both auditory and visual channels (Nardi, 2016). Subtitlers can therefore omit or shorten elements already conveyed visually without compromising comprehension (Korycińska-Wegner, 2024).

Formal design conventions further structure subtitle presentation. For example, speaker changes should be indicated by hyphens to ensure clear attribution (Brambilla et al., 2016).

2.2 Quality measures in subtitling

Subtitle quality has been addressed in numerous studies with regard to the subtitling process, the final product and reception. For instance, drawing on a survey with professional subtitlers, Künzli (2017) systematises four overarching quality categories: meaning, style, linguistic correctness and technical/visual aspects. High-quality subtitles are semantically coherent and accurately convey the source text in a comprehensible manner. Technical quality further depends on correct segmentation and synchronisation between image, sound and text (Künzli, 2017). Readability and timing are also integral to quality assessment.

Quality requirements, however, may vary depending on the genre of the source text, the service provider and the target audience (Papi et al., 2023). Despite such variability, the criteria mentioned have become widely established and serve as guiding principles for ensuring readability and overall subtitle quality (Ludera et al., 2024).

Owing to their intersemiotic and multimodal embedding, subtitles differ fundamentally from written text translations (Künzli, 2024b; Künzli, 2025). Consequently, evaluation models developed for text translation can only be applied to subtitling to a certain extent (Bolaños García-Escribano, 2025). Instead, AVT research has established assessment models of their own, which define explicit criteria and examine how these are implemented in subtitle output (Künzli, 2024a). They address intralingual, interlingual, live and pre-produced subtitles (Ludera et al., 2024).

For live subtitling, Romero-Fresco and Pöchhacker (2017) propose two models: NER (Number of words, Edition errors, Recognition errors) and NTR (Number of words, Translation errors, Recognition errors). While the NER model is designed for intralingual live subtitling, NTR evaluates interlingual live subtitling. Both incorporate the total number of words per subtitle and errors caused by automatic speech recognition. They differ, however, in their focus on Edition versus Translation errors. In intralingual contexts, Edition errors refer to inaccuracies introduced during the reformulation of spoken discourse (e.g., omissions or additions during respeaking). In interlingual contexts, Translation errors are assessed, ranging from minor inaccuracies to substantial meaning loss. Errors are weighted according to severity (0.25, 0.5, or 1 point) and calculated using a standardised formula (Romero-Fresco and Pöchhacker, 2017; Davitti et al., 2024).

For pre-produced interlingual subtitles for hearing audiences, Künzli (2020) proposes the CIA model (Correspondence, Intelligibility, Authenticity), which conceptualises quality as a multidimensional construct defined by specific parameters. High-quality subtitles should enable a seamless viewing experience during which audiences remain immersed in the film without distraction. This is described as a "flow" experience(Hof, 2025).

A comparable notion underpins the FAR model (Functional Equivalence, Acceptability, Readability) developed by Pedersen (2017) for pre-produced interlingual subtitles. Pedersen (2017) refers to the viewer's immersion as a "contract of illusion", sustained when subtitles meet predefined criteria within the dimensions of functional equivalence, acceptability and readability. Each dimension comprises operationalisable parameters, grounded either in established guidelines or target-language norms, which serve as analytical benchmarks for systematic quality evaluation. In this paper, we adopt the FAR model for our analysis, and we provide more details on the model in Section 3.2.

2.3 Related work on quality in machine-translated subtitles

Existing works on subtitle quality address various aspects. Mondello et al. (2025) relate quality with productivity and analyse subtitling workflows with technology. However, the authors do not go into detail about the quality criteria used and show just the final score. The quality issues are also linked to the research of post-editing in subtitling. Koponen et al. (2020) show that subtitle post-editing generally requires fewer keystrokes than human translation from scratch, but also that there are considerable differences when it comes to language pairs. At the same time, human subtitle translation outperforms subtitle post-editing in terms of perceived quality as shown for Finnish and German by Schierl (2023).

Most case studies assessing the quality of automatically produced subtitles compare automatically produced subtitles with those created by humans. Earlier works include (Martínez Martínez and Vela, 2016) who provide insights into the nature of human and machine translation in subtitling. Although Hagström and Pedersen (2022) do not directly compare human- and machine-generated subtitles, they generally postulate that current subtitles have decreased in quality due to the increased use of MT. Other authors, in contrast, point to the good quality of MT subtitles (Bellés-Calvera and Caro Quintana, 2021).

Some studies show that although human and machine-generated subtitles are hard to distinguish, there are still differences in terms of linguistic properties and also in terms of quality (Lapshinova-Koltunski et al., 2025; Calvo-Ferrer, 2023). However, no comprehensive analysis of quality following an established model as described in Section 2.2 is provided. One of the few studies using the FAR model is Al Sawi and Allam (2024), comparing human and machine-generated Arabic subtitles. Likewise, there are not many works that compare the usage of various models or systems: Fernandes and Lopes (2024) analyse the

differences in performance of open-source LLMs versus traditional NMTs for translating subtitles from English into Brazilian Portuguese. Mondello et al. (2025) also compare several systems (Amazon Translate, Chat-GPT and Google Translate).

Further research includes focus on quality of subtitles and post-editing effort. Karakanta et al. (2022) conduct a study with 22 subtitles and deliberately select source videos containing challenges such as background noise, slang or multiple speakers. Main sources of error in automatically generated interlingual subtitles lie in the speech recognition, MT, automatic spotting and segmentation. Fast or unclear speech, dialects and the general quality of the medium lead to errors. Automatic spotting is rated as non-synchronous, and in some instances the spoken word is not captured at all in the subtitles. The authors identify incorrect segmentation (under- or over-segmentation) in automatically generated subtitles. Some segmentations do not adhere to target language norms and units of meaning (Karakanta et al., 2022).

Other error sources in ASR include missing punctuation marks or the omission of upper and lower case letters as pointed e.g. by Rei et al. (2020). ASR systems also easily fail to correctly distinguish between similar sounding words such as *environmentally* and *mentally* as found by Davitti et al. (2024) in Amberscript's speech recognition output. Research thus shows that automatically generated interlingual subtitles need to be post-edited to achieve the quality standards of human-generated subtitles in terms of technical, content-related and linguistic realisation (Künzli, 2017; Davitti et al., 2024).

In our study, we undertake the analysis of differences between human and machine-generated subtitles using the FAR model and look into the performances of several systems. To our knowledge, there are not so many studies integrating the FAR model to compare automatic subtitles. For instance, Hof (2025) examines the quality of interlingual subtitles generated by YouTube and Whisper in German-English. The two systems differ in their architecture: While YouTube is based on a pipeline system, in which subtitles are created using the individual steps mentioned above (ASR, segmentation and MT), Whisper is based on an end-to-end architecture, ensuring a direct transformation of audio to subtitle (Hof, 2025). The author reports errors in the segmentation and separation of sentences; contrary to subtitling guidelines, some subtitles exceed two lines; punctuation marks are missing, and utterances are omitted occasionally. Hof (2025) also highlights unidiomatic translations, which are considered errors in the dimension of acceptability. All of the errors mentioned can have a negative impact on the comprehensibility of the product, which is essential for high-quality subtitles.

2.4 Subtitling and comprehensibility

Comprehensibility is defined as characteristics that ensure or promote the construction of meaning (Christmann and Groeben, 2018). Comprehensibility according to Künzli's CIA model lies within the quality dimension of intelligibility (Künzli, 2020), which states that subtitles should be short and legible, both in their formal presentation as well as in their wording. The FAR model also establishes a link to comprehensibility in the dimension of readability, which refers to formal and technical aspects (Hof, 2025).

Readability research has established the Reading Ease Formula (Christmann and Groeben, 2018), which allows for readability to be determined on the basis of characteristics such as word and sentence length (Wünsche, 2022; Christmann and Groeben, 2018). Rudolf Flesch's Reading Ease Formula captures the average sentence length in words and the number of syllables per 100 words and provides a value between 0 and 100 – a low value indicating texts that are difficult to understand (Dziuk Lameira, 2023). According to readability research, comprehensibility is achieved through the use of short, simple and common words and sentences. The formula does not include individual requirements of recipients and focuses solely on text characteristics (Christmann and Groeben, 2018).

The Hamburg Comprehensibility Concept is one of two distinct research methodologies that have emerged in the field of text comprehensibility research. Similar to the Reading Ease Formula, it evaluates the texts characteristic rather than the situational environment (Wünsche, 2022). Groeben's comprehensibility model is the second model and was developed with a greater degree of participation from recipients. In spite of contrasting methodologies, both models possess comparable and mutually reinforcing "dimensions" that promote comprehensibility. The Hamburg Compre-

hensibility Concept features the following comprehensibility dimensions: simplicity, structure and order, brevity and conciseness, and stimulating additions (Ballod, 2020; Christmann and Groeben, 2018). Simplicity involves the use of short sentences, familiar and concrete words as well as active verbs.

3 Research Design

3.1 Data

For the empirical analysis, we use a publicly available dataset from the International Conference of Spoken Language Translation (Salesky et al., 2024). We select one video for our analysis (Video 1 of *Aquarius*, Season 2, Episode 1). The segment analysed covers the first eight minutes of the episode (00:00–08:00), which includes varying speech rates, multiple speaker changes, overlapping audio cues, and background music, providing a representative sample of typical subtitling challenges. The analysed materials are referenced using the following abbreviations:

- **AT** – AusgangsTEXT, English source video
- **UT-h-en** – human-produced English subtitles
- **UT-h-de** – human-produced German subtitles
- **UT-KI_NAME** - German subtitles generated by the respective AI system (HappyScribe, CapCut, Amberscript).

The abbreviations used in this study follow conventions commonly applied in subtitling research (Wünsche, 2022). The human-produced subtitles are initially reviewed in synchrony with the source video using subtitling software. The English subtitle files contain typical formatting features of professional subtitles such as markers for italic text. They also include descriptions of non-verbal information typical of subtitles for the deaf and the hard of hearing, for example descriptions of sounds or actions (e.g., music or background noises). The German subtitles also include displays such as introductory captions. For the purposes of the present analysis, formatting markers, sound descriptions, and other display elements were removed during preprocessing. This decision reflects the focus of the study on subtitles intended for a hearing audience and ensures comparability between human-produced and AI-generated subtitle output (Künzli, 2024b).

To generate AI-produced subtitles, three web-based tools are selected: HappyScribe, CapCut and Amberscript. All three systems provide automatic subtitle generation based on ASR and MT. The selected systems represent widely used, publicly available AI subtitling tools. Their inclusion ensures comparability and reproducibility of the analysis. The systems differ in the extent to which subtitle-specific constraints are implemented and made accessible.

For UT-KI_HappyScribe, the trimmed source video is uploaded to the HappyScribe platform[1] with English as the source and German as the target language. The system generates time-coded subtitles based on ASR and subsequently translates them into German. According to the provider, the system uses proprietary AI to segment and translate subtitles in a way that supports readability. The eight-minute video is processed in approximately four minutes, producing 93 subtitle segments. Users can further influence output quality by integrating glossaries or style guides during processing. In addition, the subtitle editor flags excessive CPS through visual warnings, allowing users to adjust subtitle length and timing accordingly. However, the underlying implementation of subtitle-specific rules (e.g. segmentation norms or CPS limits) remains only partially transparent.

For UT-KI_CapCut, the same source video is uploaded to CapCut's online subtitle translation tool[2]. The system first generates English subtitles within seconds using ASR. These are then translated into German via the built-in translation function. The resulting subtitles are exported as an srt file for further analysis.

For UT-KI_Amberscript, the video is uploaded to the Amberscript platform[3]. After selecting the source language and speaker configuration, the system automatically generates subtitles through ASR and allows for subsequent translation into German. The resulting subtitles are exported as an srt file.

Detailed information on the internal architectures of these systems is not publicly available. Consequently, it was not possible to determine whether the tools rely on direct speech-to-translation models or cascaded systems. For analytical purposes, the generated subtitles are there-

[1] Accessed in May, 2025.
[2] Accessed in September, 2025.
[3] Accessed in October, 2025.

fore treated as outputs of a pipeline process consisting of ASR, subtitle segmentation and subsequent machine translation (Hof, 2025).

3.2 Evaluation of subtitles

As mentioned above, we evaluate automatically generated subtitles using established quality and comprehensibility models. We follow a mixed-methods design, combining quantitative and qualitative approaches (Kelle, 2022).

FAR quality analysis As already mentioned above, we use the FAR model (Pedersen, 2017) for analysis. The model has the following dimensions: `Functional equivalence`, `Acceptability` and `Readability` (see Figure 1 for an overview). Functional equivalence includes semantic and stylistic errors. Acceptability includes grammar, spelling and idiomaticity errors. Readability includes segmentation and spotting, punctuation and graphics as well as reading speed and line length.

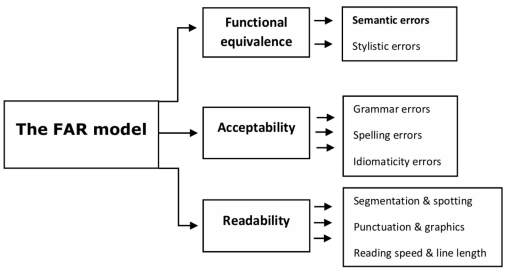

Figure 1: FAR model with its dimensions.

The subtitles are compared segment by segment across three layers: the English source subtitles (UT-h-en), the German human-produced subtitles (UT-h-de) and the AI-generated German subtitles (UT-KI_HappyScribe, UT-KI_CapCut, UT-KI_Amberscript). Errors are annotated and scored according to the FAR framework following the scoring approach of Ludera et al. (2024) and Hof (2025), with weighted scores reflecting the severity of functional, linguistic or readability-related issues. The FAR analysis is conducted by a single trained evaluator with an academic background in audiovisual translation. To reduce subjectivity, the data are annotated in multiple rounds with temporal distance between evaluations, ensuring consistency in decision-making.

Because segmentation differs between human-produced and AI-generated subtitles, some subtitle units are manually aligned to allow for meaningful comparison. In addition, `Readability` is examined quantitatively by calculating characters per line (CPL) and characters per second (CPS). These values are automatically computed in a spreadsheet software based on subtitle length and display time. The analysis follows the commonly cited guidelines of approximately 37–42 characters per line and 12–15 characters per second (Ludera et al., 2024; Papi et al., 2023; Jüngst, 2020) .

Assigning errors to a single FAR category is not always straightforward. Certain issues, such as literal translations, may simultaneously affect functional equivalence and acceptability, and punctuation errors can influence both acceptability and readability. Similarly, the weighting of error severity remains partly interpretative despite the model's scoring guidelines. These limitations are acknowledged as methodological considerations and are discussed further in the interpretation of the results.

For clarity and comparability, the results of the FAR analysis are presented in a standardised format. Each example is structured according to the following sequence: the English source subtitle with its segment number, the corresponding German human-produced subtitle, and the AI-generated German subtitle. The segment number refers to the numbering used in the first column of the analysis tables. The examples are therefore displayed as follows: UT-h-en (Seg. of the subtitle): source subtitle >UT-h-de: human-produced subtitle >UT-KI_Name: AI-generated subtitle.

To facilitate comparison across the AI systems, the results are additionally presented following the comparative structure proposed by Nikoloudaki (2024), where the outputs of the individual AI tools are displayed consecutively.

Automatic metrics Additionally to the manual evaluation, we obtain automatic quality indicators for the generated subtitles. For this, the evaluation platform MATEO Evaluate (Vanroy et al., 2023) is used. The subtitle files (UT-h-en, UT-h-de, and UT-KI_NAME) are first preprocessed by removing time codes and aligning the number of segments across files, so that we have 81 segments in total per system output. In the analysis, the AI-generated subtitles represent the system outputs, the German human-produced subtitles are the reference, and the English subtitles the source. The tool provides several automatic metrics, including BLEU and BERTScore. In this study, particular

attention is given to BERTScore, as it accounts for semantic similarity and paraphrasing and is therefore better suited to subtitle translation than surface-based metrics such as BLEU (Zhang et al., 2020). However, we also report the other two scores in the results.

Comprehensibility analysis To assess subtitle comprehensibility, the Flesch Reading Ease index is calculated for both the human-produced German subtitles and the AI-generated subtitles. For this purpose, the subtitle files are preprocessed so that only the subtitle text remains, while time codes, formatting and segmentation markers are removed. Due to the lack of sufficient punctuation in the CapCut outputs, we additionally insert the minimal punctuation here to enable the calculation of the readability score.

In addition, elements of the Hamburg Comprehensibility Concept are applied to complement the FAR-based quality analysis. FAR assesses the severity of errors and functional appropriateness. The Hamburg Model focuses on cognitive-linguistic comprehensibility. Although this model overlaps partly with the FAR framework, it provides an additional perspective on textual properties related to understanding (Schmitz, 2016). Because the present study does not include user reception experiments, comprehensibility is examined primarily from a linguistic perspective, i.e. language features that are believed to contribute to comprehensibility. Combining the two provides complementary insights into quality and reception.

The model is applied selectively. In the analysed excerpt of *Aquarius*, frequent speaker changes, flashbacks, and topic shifts occur, which makes the dimension `structure and order` difficult to assess at the level of individual subtitles. Instead, this dimension is considered across coherent meaning units. Furthermore, subtitles cannot be fully separated from the source material, which limits the extent to which aspects such as simplicity or additional explanatory elements can be modified (Wünsche, 2022). For these reasons, the analysis focuses on the dimensions simplicity and structure or order, which are widely regarded as key factors for textual comprehensibility (Christmann and Groeben, 2018).

Each subtitle is evaluated through a simple rating scheme (+/–) indicating whether the criterion is fulfilled. `Simplicity` is assessed by examining the use of common vocabulary, syntactic complexity, and whether sentences are excessively long or structurally difficult to process within the subtitle format. `Structure and order` are evaluated by analysing whether subtitles followed a logical progression, maintained coherence across segments, and preserve references to preceding information. Particular attention is given to whether omissions or translation choices disrupt semantic continuity or obscure contextual relations between subtitle segments.

While the analysis draws on principles from text-comprehension research, it does not attempt to evaluate full linguistic cohesion independently of the audiovisual context. Instead, the focus lies on whether subtitles maintain meaningful links between segments and allow viewers to follow the narrative flow – corresponding to the "red thread" described in the Hamburg Comprehensibility Concept (Christmann and Groeben, 2018).

To avoid redundancy with the FAR analysis, the comprehensibility assessment is conducted with a stronger focus on the subtitle text itself rather than direct comparison with the source subtitles. In a second review stage, the AI-generated subtitles are therefore evaluated independently of the source material in order to identify additional comprehensibility issues that may not be visible in a direct source-target comparison. This procedure allows for further insights into the readability and internal coherence of the automatically generated subtitles.

4 Results

4.1 FAR quality analysis

The results of the FAR-based analysis are presented in Table 1. The higher the score, the more errors were observed for the corresponding dimension.

In terms of `Functional equivalence`, all AI-generated subtitles show relatively high scores. The elevated scores indicate frequent semantic deviations from the source content. For example, *she's gone* in example (1) from Seg. 4 is rendered by UT-KI_HappyScribe as *Sie ist tot*, implying death rather than disappearance. Ambiguous expressions are also problematic as illustrated by example (2): the sexual meaning of *plow* is not recognised by any system, resulting in literal or unrelated translations in (2-b) and (2-c).

(1) a. *Sam, my daughter, Emma, she's gone.*
b. *Sie ist tot.* ("She died.")

Dimension	UT-KI_HappyScribe	UT-KI_CapCut	UT-KI_Amberscript
Functional equivalence	0.574	1.071	0.871
Acceptability	0.096	0.333	0.186
Readability	0.191	0.321	0.471

Table 1: Scores according to the FAR-based analysis.

(2) a. *I plow them all eventually.*
b. *Ich werde sie alle irgendwann umpflügen.* ("I'll plough them all up one day.")
c. *Ich durchforste sie schließlich alle.* ("I'll end up going through them all.")

Proper names present additional difficulties. The character name *Cut* (Seg. 70) is translated as *Schnitt* ("a cut") or *schneiden* ("to cut"), and *Ron Kellaher* (Seg. 63) appears as *Juan Kealoha* in the UT-KI_Amberscript output. In contrast, some names (e.g. *John Wayne*) are reproduced correctly. Overall, the results show that ASR errors, lexical ambiguity and incorrect handling of proper names are the main sources of semantic deviations in the AI-generated subtitles.

UT-KI_CapCut obtains the highest F-Score, largely due to ASR errors that propagate into translation as illustrated in example (3) from Seg. 3. Similar recognition errors affect several segments, leading to omissions or incoherent outputs.

(3) a. *That'd be illegal. I'm a cop.*
b. *Und ein bisschen neuer Look* ("And a slightly new look")

The AI-generated subtitles also exhibit unidiomatic renderings and word-for-word translations, particularly in UT-KI_CapCut. For example, UT-h-en (Seg. 71) "Everybody's doing the funny." is translated as "alle machen das lustige", which is unnatural compared to more idiomatic renderings. Literal translations of idiomatic expressions are common. In Seg. 67, "orchestrated" is rendered as "orchestriert" by UT-KI_CapCut and UT-KI_Amberscript, whereas UT-KI_HappyScribe correctly conveys the meaning with "inszeniert." Similarly, "Now that's what I'm talking about." becomes "das ist was ich spreche" in UT-KI_CapCut.

Judging semantic similarity from the BERTScore (see Figure 2), we see that UT-KI_HappyScribe achieves the highest score (>80), whereas UT-KI_CapCut scores around 70. Short, unambiguous sentences (e.g. *Look at this* or *I'm Charlie Manson*) receive scores of 100. This either indicates high quality or the fact that the sentences are contained in the training data. At the same time, UT-KI_CapCut contains multiple segments with a score of 0, reflecting omissions or segmentation failures. Stylistic inconsistencies also occur. For instance, in segments 75–78, *Sie* (formal *You*) and *du* (informal *you*) alternate in all AI system outputs, although the dialogue consistently requires the informal register. Such inconsistencies further affect `Functional equivalence`.

In terms of `Acceptability`, the scores also reflect a number of issues. UT-KI_CapCut output shows the highest error score, partly due to capitalisation, spelling and segmentation errors. It also contains fragmented sentences that lack syntactic coherence, often linked to missing punctuation.

The `Readability` dimension includes missing or incorrect punctuation, segmentation errors and missing speaker-change markers (Pedersen, 2017). As already mentioned above, punctuation is largely absent inUT-KI_CapCut: Commas, full stops and question marks rarely occur. The lack of punctuation in UT-KI_CapCut constitutes a major quality and comprehension issue, as subtitles without punctuation are often perceived as incomplete sentences and make it difficult for viewers to identify sentence boundaries. This also affects the "contract of illusion" described by Pedersen (2017), since the audience cannot reliably distinguish between statements and questions. Similar limitations have been observed in ASR systems, which often fail to generate punctuation or capitalisation reliably (Rei et al., 2020).

The higher `Readability` score of UT-KI_Amberscript is mainly caused by segmentation errors. Although two-line subtitles are standard, meaning units are frequently split across segments. For example, in Seg. 1 the phrase *Zuvor bei Aquarius* ("Previously on Aquarius") is followed by the beginning of the next sentence in the same subtitle, disrupting the semantic unit. According to subtitling guidelines, segmentation should follow

syntactic and semantic boundaries to facilitate information processing (Kurch et al., 2015; Nardi, 2016).

A positive feature of UT-KI_Amberscript is the use of dashes to mark speaker changes. However, this strategy is not always applied correctly. In some cases, a single utterance is divided into two speaker turns, falsely creating the impression of a dialogue and potentially confusing the alignment between audio, image and subtitle.

Overall, the results indicate that literal translations, ASR errors and segmentation problems negatively affect the idiomaticity and readability of AI-generated subtitles.

4.2 Automatic scores

An overview of the automatic scores are presented in Figure 2. Overall, the HappyScribe output achieves the highest scores in all three metrics, suggesting the strongest semantic similarity and translation quality relative to the reference subtitles. It reaches the best results for BERTScore (ca. 81), BLEU (ca. 24), and BLEURT (ca. 61) indicating that its outputs most closely match the reference in terms of both lexical overlap and contextual meaning. Amberscript ranks second across the metrics, with slightly lower values (BERTScore ca. 79, BLEU ca. 17, BLEURT ca. 59). This suggests relatively good semantic alignment with the reference subtitles, although with somewhat weaker lexical correspondence compared to HappyScribe. CapCut shows the weakest performance in all metrics (BERTScore ca. 70, BLEU ca. 8, BLEURT ca. 43). The substantially lower BLEU and BLEURT scores indicate weaker lexical overlap and lower contextual similarity with the reference subtitles. This result aligns with the qualitative analysis, which identifies frequent ASR errors, omissions and segmentation issues in UT-KI_CapCut. In summary, the automatic evaluation metrics consistently show HappyScribe as the strongest system, Amberscript as intermediate, and CapCut as the weakest, reflecting differences in the overall translation quality.

4.3 Comprehensibility

The comprehensibility of subtitles is quantitatively assessed using the Flesch Reading Ease score. The results are shown in Table 2. All scores fall within a range equivalent to a standard comprehensibility level (e.g., instructions or content accessible to ages 9+). While the indices suggest high comprehensibility, the Flesch metric primarily evaluates sentence and word length, neglecting semantic coherence (Ballod, 2020). Notably, UT-KI_CapCut requires intuitive punctuation insertion for calculation, potentially skewing its score. Consequently, these metrics provide only a surface-level analysis, necessitating a qualitative review using the Hamburg Comprehensibility Concept. Our qualitative analysis reveals that AI-generated subtitles, particularly UT-KI_CapCut, frequently employ complex compounds and nested structures. This contrasts with the `Simplicity` dimension of the Hamburg Comprehensibility Concept, which prioritises common, concrete lexis (Schmitz, 2016).

AI systems often opt for formal register compound terms, e.g. *Innenrevision*, *Dienstaufsichtsbehörde* as in example (4), which increase cognitive load (Deilen, 2022). In contrast, the human reference (UT-h-de) resolves these into more accessible phrases like in example (4-a).

(4) a. UT-h-de: *interne Ermittlung* (“internal investigation”)
b. UT-KI_HappyScribe: *Innenrevision* (“internal audit”)
c. UT-KI_CapCut/Amberscript: *Dienstaufsichtsbehörde* (“disciplinary authority”).

Research on accessible communication (such as `Easy German`) suggests that hyphenating long compounds unburdens working memory (Deilen, 2022). This was observed in UT-KI_CapCut and UT-KI_Amberscript with the term *Motorrad-Entführung* (“motorcycle kidnapping”).

Also, AI systems struggle with syntactic brevity. For instance, UT-KI_HappyScribe produces segmented, paratactic structures (e.g., *Ein Mann, ein Motorrad, eine Entführung*) which, while following the source text, increase visual processing time. To ensure comprehensibility, subtitles should avoid nested syntax in favour of precise, information-dense formulations (Ballod, 2020).

5 Conclusion and Discussion

This study focuses on the analysis of automatic subtitles produced with three systems. We compare their performance evaluating their outputs in terms of quality and comprehension.

Our results show that automatically generated subtitles are not yet comprehensive, high-quality and fully comprehensible. For instance, content-

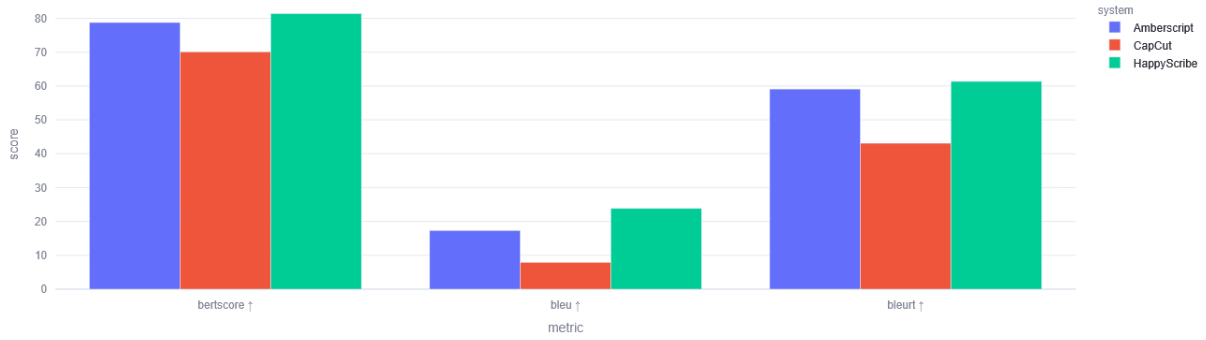

Figure 2: Automatic scores for UT-KI_HappyScribe, Capcut and Amber.

	UT-h-de	UT-KI_HappyScribe	UT-KI_CapCut	UT-KI_Amberscript
Flesch Reading Ease	72	73	73	72

Table 2: Comprehensibility measured with Flesch Reading Ease.

and meaning-bearing words are often omitted or misrecognised by ASR, leading to a high number of errors in the Functional Equivalence dimension. Omissions are permissible due to the multimodal nature of subtitles, which convey information through multiple channels, including visual images, spoken dialogue, subtitles and background sounds. Human subtitlers are able to determine which elements must be verbalised to avoid cognitive overload. In contrast, the automatically generated subtitles frequently omit words or entire sentences, preventing the audience from processing information across all sensory channels. Additionally, substitutions caused by ASR errors introduce semantically distorted content, which compromises functional equivalence. Literal, word-for-word translations – most evident in UT-KI_CapCut – reduce readability and naturalness. High reading speed, long words and sentence complexity further impair comprehension.

The analysis of segmentation and line breaks reveals that repeated or misaligned segments, particularly in UT-KI_Amberscript, disrupt the logical flow and synchronous alignment with the source text, a core feature of subtitles. Named entities and idiomatic expressions are frequently mistranslated or omitted across all the systems under analysis, echoing known issues in MT (Ottmann and Canfora, 2020; Loock et al., 2022). The analysis with the automatic evaluation metrics confirms that automatically generated subtitles can achieve human-level quality, but only in short segments.

Overall, AI-generated subtitles remain below human standards in quality and comprehensibility. They cannot fully replicate the skills of human subtitlers, who combine auditory sensitivity, editorial judgment and aesthetic sense to produce coherent, synchronised subtitles.

Limitations We recognise that focusing on publicly accessible AI tools and other more dedicated systems may yield higher quality. Also, insufficient transparency about the system architectures prevents conclusive attribution about the observed differences across the systems, e.g. between UT-KI_HappyScribe and UT-KI_CapCut. The absence of interrater validation is acknowledged as a further limitation as the annotation and scoring with the FAR model is inherently subjective. Therefore, future research could benefit from multiple raters or alternative metrics, such as the NTR model. Besides that, post-editing or audience evaluation would be an asset too. While the inclusion of additional language pairs would provide a broader perspective, this study focuses on a single language pair to ensure depth and detailed qualitative analysis. Future work will expand the scope to include additional language pairs, post-editing scenarios and user reception studies.

Outlook As automatically generated subtitles cannot yet deliver broadcast-ready quality without human intervention, research should continue to explore AI potential, particularly regarding iterative improvements in ASR, segmentation and translation. Ethical implications of AI integration in audiovisual translation remain an important area for future study, especially regarding professional practice and human subtitler roles.

References

Agnetta, Marco. 2025. On the assessment and acceptance of ai technologies in the avt industry: The associations' perspective. *Lebende Sprachen*, 70(1):149–176.

Al Sawi, Islam and Rania Allam. 2024. Exploring challenges in audiovisual translation: A comparative analysis of human- and ai-generated arabic subtitles in birdman. *PLOS ONE*, 19(10):1–20, 10.

Ballod, Matthias. 2020. *Klar-text in Organisationen: Ein Ratgeber zur Optimierung administrativer Informationen*. Springer Fachmedien Wiesbaden, Wiesbaden.

Bellés-Calvera, Lucía and Rocío Caro Quintana. 2021. Audiovisual translation through NMT and subtitling in the netflix series 'cable girls'. In Mitkov, Ruslan, Vilelmini Sosoni, Julie Christine Giguère, Elena Murgolo, and Elizabeth Deysel, editors, *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 142–148, Held Online, July. INCOMA Ltd.

Bolaños García-Escribano, Alejandro. 2025. *Practices, Education and Technology in Audiovisual Translation*. Routledge.

Brambilla, Marina, Valentina Crestani, and Fabio Mollica. 2016. *Untertitelung: interlinguale, intralinguale und intersemiotische Aspekte*. Peter Lang Verlag, Berlin, Deutschland.

Calvo-Ferrer, José Ramón. 2023. Can you tell the difference? A study of human vs machine-translated subtitles. *Perspectives*, 32(6):1115–1132.

Christmann, Ursula and Norbert Groeben. 2018. Verständlichkeit: die psychologische perspektive. In Maaß, Christiane and Isabel Rink, editors, *Handbuch Barrierefreie Kommunikation*, volume 3 of *Kommunikation – Partizipation – Inklusion*, pages 123–146. Frank & Timme, Berlin.

Davitti, Elena, Annalisa Sandrelli, Tomasz Korybski, Yuan Zou, Constantin Orasan, and Sabine Braun. 2024. Using asr tools to produce automatic subtitles for tv broadcasting: A cross-linguistic comparative analysis. *Journal of Audiovisual Translation*, 7(2):1–35, Dec.

Deilen, Silvana. 2022. *Optische Gliederung von Komposita in Leichter Sprache*. Frank and Timme, Berlin.

Diaz Cintas, Jorge and Aline Remael. 2007. *Audiovisual Translation: Subtitling*. Kinderhook: St. Jerome Publishing, Manchester.

Díaz Cintas, Jorge. 2015. Technological strides in subtitling. In Chan, Sin-wai, editor, *The Routledge Encyclopedia of Translation Technology*, pages 632–643. Routledge, London and New York.

Dziuk Lameira, Katharina. 2023. Zur komplexität von texten. von der lesbarkeitsformel zur textlinguistischen komplexität. In Schrott, Angela, Johanna Wolf, and Christine Pflüger, editors, *Textkomplexität und Textverstehen: Studien zur Verständlichkeit von Texten*, pages 69–98. De Gruyter, Berlin, Boston.

Fernandes, Rafael and Marcos Lopes. 2024. Open-source LLMs vs. NMT systems: Translating spatial language in EN-PT-br subtitles. In Martindale, Marianna, Janice Campbell, Konstantin Savenkov, and Shivali Goel, editors, *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 152–153, Chicago, USA, September. Association for Machine Translation in the Americas.

Hagström, Hanna and Jan Pedersen. 2022. Subtitles in the 2020s: The influence of machine translation. *Journal of Audiovisual Translation*, 5(1):207–225, Dec.

Hof, Lea. 2025. Maschinelle untertitelung mittels pipeline- und end-to-end-ansatz. ein vergleich von youtube und whisper. In Agnetta, Marco, Astrid Schmidhofer, and Alena Petrova, editors, *Bild – Ton – Sprachtransfer: Neue Perspektiven auf Audiovisuelle Translation und Media Accessibility*, TransKultur, pages 39–68. Frank & Timme, Berlin.

Jaki, Sylvia, Maren Bolz, and Sophie Röther. 2024. Ki-technologien in der audiovisuellen translation. *trans-kom*, 17:320–342.

Jüngst, Heike Elisabeth. 2020. *Audiovisuelles Übersetzen. Ein Lehr- und Arbeitsbuch*. Narr Francke Attempto, Tübingen, 2. überarbeitete und erweiterte auflage edition.

Karakanta, Alina, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. Post-editing in automatic subtitling: A subtitlers' perspective. In Moniz, Helena and Lieve et al. Macken, editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 261–270, Ghent, Belgium. European Association for Machine Translation. hdl:1854/LU-8756355.

Kelle, Udo. 2022. Mixed methods. In Baur, Nina and Jörg Blasius, editors, *Handbuch Methoden der empirischen Sozialforschung*, pages 163–177. Springer VS, Wiesbaden.

Koponen, Maarit, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. MT for subtitling: User evaluation of post-editing productivity. In Martins, André, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124, Lisboa, Portugal, November. European Association for Machine Translation.

Korycińska-Wegner, Małgorzata. 2024. Linguistische zugänge. In Künzli, Alexander and Klaus Kaindl, editors, *Handbuch Audiovisuelle Translation. Arbeitsmittel für Wissenschaft, Studium, Praxis*, pages 61–74. Frank & Timme, Berlin.

Künzli, Alexander. 2017. *Die Untertitelung – von der Produktion zur Rezeption*, volume 90 of *Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens, TRANSÜD*. Frank & Timme, Berlin.

Künzli, Alexander. 2020. From inconspicuousness to flow – the CIA model of subtitle quality. *Perspectives*, 29(3):326–338.

Künzli, Alexander. 2024a. Normen und qualität in der AVT. In Künzli, Alexander and Klaus Kaindl, editors, *Handbuch Audiovisuelle Translation. Arbeitsmittel für Wissenschaft, Studium, Praxis*, pages 345–360. Frank & Timme, Berlin.

Künzli, Alexander. 2024b. Untertitelung. In Künzli, Alexander and Klaus Kaindl, editors, *Handbuch Audiovisuelle Translation. Arbeitsmittel für Wissenschaft, Studium, Praxis*, pages 107–122. Frank & Timme, Berlin.

Künzli, Alexander. 2025. Untertiteln im streamingzeitalter – überlegungen von untertitelexpertinnen im deutschsprachigen raum zum kompetenzprofil. *Fachsprache. Journal of Professional and Scientific Communication*, 47(3–4):129–144.

Kurch, Alexander, Nathalie Mälzer, and Katja Münch. 2015. Qualitätsstudie zu live-untertitelungen – am beispiel des "tv-duells". *trans-kom*, 8(1):144–163.

Kurch, Alexander. 2018. Produktionsprozesse der hörgeschädigten-untertitelungen und audiodeskription: Potenziale teilautomatisierter prozessbeschleunigung mittels (sprach-)technologien. In Maaß, Christiane and Isabel Rink, editors, *Handbuch Barrierefreie Kommunikation*, volume 3 of *Kommunikation – Partizipation – Inklusion*, pages 437–454. Frank & Timme, Berlin.

Lapshinova-Koltunski, Ekaterina, Sylvia Jaki, Maren Bolz, and Merle Sauter. 2025. Human- or machine-translated subtitles: Who can tell them apart? In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 496–505, Geneva, Switzerland, June. European Association for Machine Translation.

Loock, Rudy, Benjamin Holt, and Sophie Léchauguette. 2022. The use of online translators by students not enrolled in a professional translation program: beyond copying and pasting for a professional use. In Macken, Lieve and et al., editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 23–29, Ghent, Belgium, June. European Association for Machine Translation. hdl:1854/LU-8756355.

Ludera, Ewa, Agnieszka Szarkowska, and David Orrego-Carmona. 2024. Expertise in interlingual subtitling: applying the FAR model to study the quality of subtitles created by professional and trainee subtitlers. *The International Journal of Translation and Interpreting Research*, 16(1):55–75.

Mälzer, Nathalie and Maria Wünsche. 2018. Untertitelung für Hörgeschädigte (SDH). In Maaß, Christiane and Isabel Rink, editors, *Handbuch Barrierefreie Kommunikation*, volume 3 of *Kommunikation – Partizipation – Inklusion*, pages 327–344. Frank & Timme, Berlin.

Mälzer, Nathalie and Maria Wünsche. 2019. Handlungsempfehlungen für die TV-untertitelung für gehörlose und schwerhörige kinder zwischen acht und zwölf jahren. Technical report, Universität Hildesheim, Institut für Übersetzungswissenschaft & Fachkommunikation, Hildesheim.

Martínez Martínez, José Manuel and Mihaela Vela. 2016. SubCo: A learner translation corpus of human and machine subtitles. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2246–2254, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Mondello, Ashley, Romina Cini, Sahil Rasane, Alina Karakanta, and Laura Casanellas. 2025. Using AI tools in multimedia localization workflows: a productivity evaluation. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Samuel Läubli, Martin Volk, Miquel Esplà-Gomis, Vincent Vandeghinste, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 2*, pages 1–7, Geneva, Switzerland, June. European Association for Machine Translation.

Nardi, Antonella. 2016. Sprachlich-textuelle faktoren im untertitelungsprozess: Ein modell zur übersetzerausbildung deutsch-italienisch. *trans-kom*, 9(1):34–57.

Nikoloudaki, Dionysia. 2024. Rendering patriarchy through gendered translator gaze in Romeo and Juliet. *inTRAlinea*. Special Issue: Translating Threat.

Ottmann, Angelika and Carmen Canfora. 2020. Risiken und haftungsfragen bei neuronaler maschineller übersetzung. In Bund der Dolmetscher und Übersetzer (BDÜ), editor, *Maschinelle Übersetzung für Übersetzungsprofis*, pages 171–184. BDÜ Fachverlag, Bonn.

Papi, Sara, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023. Direct speech translation for automatic subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

Pedersen, Jan. 2017. The FAR model: assessing quality in interlingual subtitling. *The Journal of Specialised Translation*, 28:210–229.

Rei, Ricardo, Nuno Miguel Guerreiro, and Fernando Batista. 2020. Automatic truecasing of video subtitles using BERT: A multilingual adaptable approach. In Lesot, Marie-Jeanne et al., editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, volume 1237 of *Communications in Computer and Information Science*, pages 708–721, Cham. Springer.

Romero-Fresco, Pablo and Franz Pöchhacker. 2017. Quality assessment in interlingual live subtitling: The NTR model. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 16:149–167.

Salesky, Elizabeth, Marcello Federico, and Marine Carpuat, editors. 2024. *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, Bangkok, Thailand (in-person and online), August. Association for Computational Linguistics.

Schierl, Frederike. 2023. Reception of machine-translated and human-translated subtitles – a case study. In Yamada, Masaru and Felix do Carmo, editors, *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 42–53, Macau SAR, China, September. Asia-Pacific Association for Machine Translation.

Schmalz, Antonia. 2019. Maschinelle übersetzung. In Wittpahl, Volker, editor, *Künstliche Intelligenz*, pages 194–211. Springer Vieweg, Berlin, Heidelberg.

Schmitz, Anke. 2016. *Verständlichkeit von Sachtexten. Wirkung der globalen Textkohäsion auf das Textverständnis von Schülern*. Psychologische Genese der Sprach- und Schreibkompetenz. Springer VS, Wiesbaden.

Tang, Yunlong, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. 2026. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376.

Tardel, Anke. 2023. A proposed workflow model for researching production processes in subtitling. *trans-kom*, 16(1):140–173.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland, June. European Association for Machine Translation.

Wünsche, Maria. 2022. *Untertitel im Kinderfernsehen: Perspektiven aus Translationswissenschaft und Verständlichkeitsforschung*, volume 46 of *TransKultur*. Frank & Timme, Berlin.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.