universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Linguistic challenges in Automatic Speech Recognition (ASR): A comparison of non-native interpreted and non-interpreted speech

verfasst von | submitted by

Sandra Anita Dahl BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 070 331 342
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Translation Deutsch Englisch
Betreut von | Supervisor:	Univ.-Prof. Dragos Ioan Ciobanu PhD

Table of contents

Abstract (EN): The present research investigates the performance of Automatic Speech Recognition (ASR) systems within non-native English speech, produced in three different

modes: unmediated spontaneous monologue, planned read-aloud speech, and consecutive interpreting (CI). In the frame of a small-scale empirical study, transcripts generated by the Word Dictate Function were systematically compared to denaturalized manual transcripts produced by the researcher, maintaining all original oral characteristics the way they have perceived by the researcher. The ASR-transcription errors of all three speech tasks were identified and categorized, using both raw error counts and Normalized Error Rates per 1,000 words (NER), allowing for comparison across tasks of varying lengths. The research addresses two primary questions: What are the key differences in ASR performance when transcribing non-native English speech produced in the modes of non-interpreted speech versus interpreted speech from German to English (RQ1)? Specifically, how might Fluency Features, Syntactic Complexity, Speech errors, Phonological errors, Self-correction and Punctuation errors influence the transcription (RQ2)?

The findings reveal a performance hierarchy, where, as it has been hypothesized, the Read-aloud task (E2) exhibits the lowest number of deviations between ASR and manual transcript (NER 65.22), likely because of its predictable nature that resembles written language the most. In contrast, both unmediated spontaneous (NER 164.47) and interpreted speech (NER 181.10) resulted in significantly higher error rates. The CI task, partially due to the specific cognitive management strategies of the trainee interpreter, exhibits the highest overall NER. In contrast, the non-interpreted monologue shows structural instability, where pauses appear to be algorithmically misinterpreted as terminal boundaries, resulting in severe syntactic fragmentation and therefore more punctuation mistakes. Ultimately, this led to the conclusion that ASR actively reshapes the output to match written-style language in some cases, while literally transcribing redundant repetitions in other cases where the acoustic signal is clear.

Abstract (DE): Das Ziel dieser Masterarbeit ist die Untersuchung der Leistung automatischer Spracherkennungssysteme (ASR) bei der Transkription von nicht-muttersprachlichem Englisch in drei verschiedenen Modi: einem nicht-gedolmetschten, spontanen Monolog (E1), einer vorgeschriebenen Rede, die laut vorgelesen wird (E2) und einer Konsekutivdolmetschung dieser Rede vom Deutschen ins Englische (E3). Um lexikalische Konsistenz zu gewährleisten, wird in allen drei Phasen dieselbe Quellrede, eine Ansprache von Alain Berset (2024) genutzt. Die Studie basiert außerdem auf einem erklärenden sequenziellen Design. Sie vergleicht die ASR-Transkripte, erstellt von der Word Dictate Funktion, mit den dazugehörigen Verbatim-Transkriptionen. Diese verschriftlichen das

Gesprochene ohne jegliche Anpassungen, genau so wie es von der Forscherin wahrgenommen wurde. Dabei werden die zwei zentrale Fragen behandelt: Was sind die wesentlichen Unterschiede in der ASR-Leistung zwischen einer nicht-muttersprachlichen, Deutsch-Englisch Dolmetschung und nicht-gedolmetschter Sprache, die von Sprecher:innen ohne jeglichen akademischen Hintergrund im Sprachbereich stammt? (Forschungsfrage 1) Wie beeinflussen spezifische linguistische Faktoren wie Redefluss, die Komplexität von Satzstrukturen, phonologische Fehler, Selbstkorrektur und Interpunktion den Transkriptionsprozess? (Forschungsfrage 2)

Die Auswertung der Fehlerkategorien erfolgt mittels der Normalized Error Rate (NER), die Fehler pro Tausend Wörter darstellt. Die Ergebnisse der Studie zeigen schließlich eine deutliche Hierarchie: Das Vorlesen (E2) erzielte die höchste Übereinstimmung zwischen den ASR-Versionen und den dazugehörigen manuellen Transkripten (NER 65,22), da die Ausgangsrede vollständig vorgeschrieben war und somit eng mit schriftlichen Sprachkonventionen übereinstimmt, auf die ASR-Systeme normalerweise optimiert sind. Im Gegensatz dazu führten sowohl der nicht-gedolmetschte Monolog (NER 164,47), als auch die Konsekutivdolmetschung (NER 181,10), zu deutlich höheren Fehlerquoten. Die Dolmetschung (E3) führte, primär aufgrund adaptiver Interferenz, zu der höchsten Fehlerquote, was vermutlich aufgrund der kognitiven Strategien der Dolmetscherin zu Transkriptionsfehlern seitens ASR geführt hat. Im Gegensatz dazu ist der spontane Monolog (E1) durch strukturelle Instabilität gekennzeichnet, wobei das System funktionale Pausen als Satzgrenzen fehlinterpretierte, was zu schweren syntaktischen Fragmentierungen führte. Letztlich deuten die Ergebnisse darauf hin, dass das generische ASR-System aktiv und voreingenommen agiert, indem es die Transkription so verändert das sie bestmöglich schriftlichen Sprachkonventionen ähnelt, anstatt das Gesprochene wortwörtlich wiederzugeben.

1 Introduction

The rapid expansion and practical deployment of ASR technology have been driven by an increasing demand to streamline human–machine interaction (Alharbi et al. 2021: 131858). Because manual transcription is a highly time-consuming and costly process, ASR has emerged as an essential solution for the written documentation of large volumes of spoken language (Bazillon et al. 2008: 1). These developments are largely rooted in broader advances in Artificial Intelligence (AI) and commercial motivations for efficiency and scalability (Baumgarten & Bourgadel 2024: 508). Moreover, as Fantinuoli (2017) claims, some ASR tools now claim to surpass basic transcription by interpreting contextual meaning and aligning more closely with human judgments. Improvements in transcription accuracy may encourage a transition from fully manual transcription processes to hybrid approaches that integrate human expertise with AI-based systems (Fantinuoli 2017: 25). According to Karakasidis et al. (2024), there has been a shift from modular systems to complex End-to-End (E2E) architectures, which integrate acoustic and language modeling within a single network. While this might have led to improvements in overall performance, it also increased system reliance on large quantities of training data (Karakasidis et al. 2024: 1). Consequently, ASR performance is strictly tied to the linguistic properties of its training data, making it highly sensitive to deviations. Spontaneous, non-native interpreted speech often contains such deviations in the form of disfluencies, syntactic shifts, and phonological irregularities (Zhao et al. 2023: 1).

Generally, the use of ASR in interpreting contexts has primarily focused on the automatic transcription of source-language input to provide real-time terminological or numerical support, however, that is not what it is limited to (Fantinuoli 2017; Pisani & Fantinuoli 2021; Prandi 2020). For instance, with ASR transcribing the audio, there are additional ways to reduce the interpreter's cognitive strain of managing simultaneous listening, analyzing, translating, speaking, and memory retention terms (Tammasrisawat & Rangponsumrit 2023: 25-26) by automatically detecting specialized terms and querying databases (Fantinuoli 2017: 26). For such tools to be effective, their integration into the interpreting workflow must be carefully designed. Computer-Assisted-Interpreting (CAI) tools that incorporate ASR should offer a simple, distraction-free interface (Fantinuoli 2017: 31), as visual distractions may negatively affect interpreting performance (Tammasrisawat & Rangponsumrit 2023: 48; Prandi 2020: 2). While these applications establish the professional

utility of ASR, they do not address their performance when transcribing an interpreter's output as a linguistic product.

Interpreting, regardless of mode, is characterized by high cognitive demands, requiring the processing of incoming speech, extraction of meaning, and finally the production of target-language output under time pressure, often with limited opportunities for revision (Pöchhacker 2016: 11). Therefore, interpreted speech often exhibits hesitations, reformulations, structural adaptations, and phonological variation (Gile 2009; Bóna & Bakti 2020: 484-489). Within the framework of Gile's Effort Model (Gile 2009), such features are viewed as functional responses to cognitive load rather than indicators of deficient performance. However, for ASR systems, they represent a deviation from the fluent and more predictable speech patterns they are typically trained for (Fantinuoli 2017: 29). Additionally, it is noteworthy that CI differs substantially from simultaneous interpreting (SI) in terms of temporal structure, production conditions, and discourse organization (Setton & Dawrant 2016: 414; Pöchhacker 2016: 18-19). While the flexibility of note-taking may provide relief for consecutive interpreters (Setton 1999: 39), target-language production takes place under conditions of constrained processing capacity (Pöchhacker 2016: 10).

Existing research suggests that ASR systems are less robust when confronted with spontaneous and non-native speech, even under controlled recording conditions (Fantinuoli 2017: 28). According to Bazillon et al. (2008) and Fantinuoli (2017), prepared or read-aloud speech generally follows stable syntactic and phonetic patterns. In contrast, spontaneous speech is marked by hesitations, false starts, self-corrections, and irregular prosody (Bazillon et al. 2008: 1). Furthermore, Kuhn et al. (2023) assess that in professional and public service settings, ASR-generated transcripts are increasingly used as the basis for documentation, accessibility services, and multilingual communication. This includes the provision of live or near-live subtitles, which are particularly relevant for d/Deaf and hard-of-hearing (DHH) individuals. In such settings, transcription errors may have consequences that extend beyond technical inaccuracy, affecting accessibility, comprehension, and trust in mediated communication (Kuhn et al. 2023: 1-3). Evaluating ASR performance across different speech types is therefore, not only a technical issue but also one with direct implications for accessibility and inclusion (Kuhn et al. 2023: 20) underscoring the broader relevance of the comparative analysis applied in this thesis.

2 Spoken language, interpreting, and ASR: a theoretical framework

This chapter aims to create a foundation of the conceptual and technological aspects to understand ASR systems, as well as various forms of spoken discourse relevant for RQ1. Starting with a general discussion about CI and basic ASR technology, the chapter will follow with displaying their correlation in the utilization of ASR within interpreting contexts. Next, ASR transcription is contrasted to manual transcription, followed by an elaboration of the distinct features of interpreted and non-interpreted speech, as well as potential ASR challenges across various speech types. Lastly, the chapter will finish with an overview about the established quality assessment criteria (RQ2) to evaluate ASR performance in contrast to the corresponding Gold Standard.

2.1 Defining and theorizing interpreting

Before relating the role of ASR to interpreting settings, it appears useful to situate interpreting itself within its broader theoretical context. However, the present research does not attempt to define this complex phenomenon; rather, it aims to provide synthesis of insights derived from existing research. Interpreting Studies has emerged as a distinct domain, moving beyond its historical treatment as a variation of translation, to incorporating its own theoretical models and empirical traditions (Pöchhacker 2016: 133–134). Interpreting can be seen as a multifaceted (Pöchhacker 2016: 2), multimodal and bilingual communicative event that often remains largely unstructured (Pöchhacker 2016: 84). Interpreting is performed under specific temporal, cognitive, and interactional conditions (Pöchhacker 2016: 10). These situational factors are integrally linked to the productive processes of interpreting and directly affect the linguistic form of the target-language output (Pöchhacker 2016: 36).

Pöchhacker (2016) identifies the distinction between the two core professional interpreting modes based on time in reference to the original message: SI involves near-simultaneous reception and production, whereas CI is not being performed during, but after the delivery of the source-language input, defined by its sequential organization (Pöchhacker 2016: 18–19; Zhao et al. 2023: 5). Setton & Dawrant (2016) further explain that interpreters take notes while simultaneously listening to the source speech, often for several minutes, followed by a target-language rendition once the speaker pauses. Typical professional passage lengths in CI can range from approximately 45 seconds to several minutes. Segments exceeding five minutes are considered exceptional and tend to occur primarily in situations where speakers continue talking without noticing that their speech is being interpreted

consecutively (Setton & Dawrant 2016: 414). These observations indicate that CI is embedded in a specific rhythm of segmented delivery, where segment length may directly shape memory load, note-taking strategies, and discourse reconstruction.

The specific focus of the present study is CI, however, to further theorize the nature of the mode, it is useful to contrast the cognitive demands to those found in SI. The research of Zhao et al. (2023) suggests that the production of the target language in CI is actually more similar to spontaneous speech than SI is. This is primarily because in CI, the speaker is not forced to process new incoming messages while simultaneously speaking and translating. In SI, the time pressure and the overlap of tasks often lead to errors that arise because interpreters are rushing to keep up with the pace. In contrast, the sequential structure of CI allows for a clearer view of the distinct stages of language production and how cognitive load specifically influences the speaker's verbalization (Zhao et al. 2023: 5).

Another important factor to consider is demonstrated in Gile's (2009) Effort Model. It proposes the idea that interpreting requires the deployment of processing capacity distributed across three core efforts: listening and analysis, short-term memory, and production. Consequently, the non-automatic operations of CI draw from a limited available supply of mental energy and frequently approach the limits of available processing resources. When that capacity is exceeded, issues arise as the efficient handling of note-taking and listening becomes more challenging, and interpreters are mostly required to work on or above threshold interpreting capacity (Gile 2009: 159–161; 190). This may disrupt speech planning and execution mechanisms in terms of fluency and temporal regularity (Gile 2009: 224), resulting in the features that might challenge ASR stability: hesitations, reformulations, and structural adaptations (Pöchhacker 2016: 133–135; Zhao et al. 2023: 4). Within the framework of this study's error analysis, these linguistic features must be understood not as random errors, but as functional markers of cognitive load (Gile 2009). The specific linguistic features of interpreting will be explored further in Chapter 2.4. Additionally, according to Gile (2009), precision in interpreting is relevant to the specific communicative situation rather than a fixed quantity. As a result, interpreters may modify content if they serve the communicative goal (Gile 2009: 6; 59-62).

Beyond the characteristics outlined above, CI can be further narrowed down in terms of its communicative and discourse-related nature. Pöchhacker (2016: 133–135) conceptualizes interpreting fundamentally as an instance of oral language use, understood as natural language used for immediate interpersonal interaction rather than a product of written text generation. In this sense, interpreting is based on orality, language use realized through

speech or sign in face-to-face exchange and supported by paraverbal and nonverbal signaling systems. At the same time, in many formal organizational settings, what appears to be natural verbal communication is often actually the vocalization of pre-existing documents, meaning the underlying logic and density are rooted in the conventions of written language. However, the interpreter's strategies of selection, condensation, or reformulation as developed in training or professional expertise, influence the output as well. Professional practice suggests that interpreters tend to simplify overly complex, document-heavy source materials, while simultaneously organizing and providing cohesion to more disjointed or spontaneous verbal remarks, depending on factors such as the degree of planning of the source discourse and the communicative setting. Empirical studies confirm the difference in the amount and pattern of the said elements depending on the language direction, interpreting modes, and skill levels of the interpreter, while disfluency remains a noteworthy characteristic earmarking interpreting output itself. Accordingly, CI would represent an intermediate position on the orality-literacy continuum, revealing characteristics of both interactional spoken communication and the influence of prior discourse organization (Pöchhacker 2016: 133–135; 165–166).

Lastly, it is important to consider potential differences between professional versus trainee interpreters. In early stages, trainees are typically starting with an introduction to general CI skills such as cue-word identification, segmentation, and the initial phases of note-supported reconstruction, as well as learning to manage meaning units and produce coherent output (Setton & Dawrant 2016: 569-570). At the same time, CI is positioned as a mode requiring ongoing refinement beyond initial training. Setton and Dawrant (2016) suggest practicing interpreters to "add, upgrade or refresh" (Setton & Dawrant 2016: 570) CI skills as part of ongoing professional development. Furthermore, Shen & Liang (2021) identify self-repair as a critical strategy in CI, noting that trainee interpreters typically exhibit a higher frequency of restarts and replacements compared to professionals (Shen & Liang 2021: 768). Collectively, these descriptions highlight the foundational concepts of CI: the sequential nature, the reliance on note-supported memory, and the impact of segment pacing on linguistic output. Understanding these conditions is essential for understanding the automated processing of consecutively interpreted speech and provide the necessary contrast to non-interpreted speech (RQ1).

2.2 ASR: basics and its role in interpreting

Having established a basis on specifics of interpreting itself, the following chapter provides the basic groundwork necessary to understand ASR and its role in interpreting contexts. ASR

can be defined as the process of transforming spoken language into a written transcription by analyzing the specific features of a speech waveform (Alharbi et al. 2021: 131858). It allows users to utilize spoken language as a mode of input, which is, for many users, considered more convenient and faster than typing (Malik et al. 2021: 9412; Zapata & Søeborg Kirkedal 2015: 207).

Karakasidis et al. (2024) show that the functionality of modern ASR systems is largely reliant on E2E architectures. These systems utilize a single, large-scale neural network to execute the entire speech-to-text pipeline, in contrast to previous methods that relied on several specialized components for acoustic and language modeling. They also typically require substantially greater amounts of training data and computational resources to operate effectively. To improve the performance of these networks, researchers have employed Curriculum Learning (CL), a methodology that gradually introduces training examples of increasing difficulty to the model during training. This approach was inspired by the observation that humans intuitively acquire basic phonetic sounds before progressing toward using complex structures such as words and sentences (Karakasidis et al. 2024: 1). Difficulty in this context is often measured using metadata such as the speech's Signal-to-Noise Ratio (SNR) (Karakasidis et al. 2024: 3-4).

In Interpreting, ASR was initially integrated into Computer-Assisted Interpreting (CAI) tools to automate the querying of glossaries and the rendition of problem triggers like numbers and proper names (Fantinuoli 2017: 26). For instance**,** Pisani & Fantinuoli (2021) show that ASR tools can help interpreters manage cognitive strain during SI by comparing two groups: one with ASR support and one without. The participants were asked to interpret a numerically dense speech and reported that this high density of numbers increased their cognitive load. The participant group using ASR perceived their task as less burdensome (3.5/5) compared to those without support (4.8/5) (Pisani & Fantinuoli 2021: 193), suggesting that ASR tools can ease some of the cognitive strain associated with numerical interpretation, while minimizing phonetic perception errors. Additionally, visual aids have been shown to reduce errors, particularly when subtitles are displayed on the screen during speech (Pisani & Fantinuoli 2021: 185). Similarly, Tammasrisawat and Rangponsumrit (2023) conducted experiments with trainee interpreters using InterpretBank, an ASR-CAI tool designed for support during terminology-heavy speeches. The tool did not only transcribe the speech, but also automatically translated it. All participants reported that the support of the tool led to fewer mistranslations of specialized terms and a reduction in the use of English words in their renditions. This improvement was particularly evident in technical terms, where participants

provided correct translations when using InterpretBank, but struggled without it (Tammasrisawat & Rangponsumrit 2023: 39). However, this also highlights a cognitive trade-off: when the tool presents information overload or lacks translation suggestions, it conversely leads to increased hesitation, omissions, and errors in the interpreter's rendition (Tammasrisawat & Rangponsumrit 2023: 48). A notable limitation of their research is the controlled laboratory conditions, as they utilize pre-recorded videos, which do not appear to represent possible dynamic conditions of real-life interpreting, such as a “large audience” (Zhao et al 2023: 3), as well as the unpredictability of the dialogue (Pöchhacker 2016: 84).

The recent empirical turn in Translation and Interpreting (T&I) Studies has caused a global shift towards remote and hybrid workflows, accelerated by the COVID-19 pandemic (Wang & Zheng 2021: 1-4), suggesting that modern academic discourse must address the new complexities introduced by ASR and machine translation (MT), making technological integration a core research theme. Specifically, considering the growing reliance on AI technologies in interpretation settings, studies such as Wang & Zheng (2021) and Wang & Fantinuoli (2024) establish that there is a research gap in ASR analysis which must urgently be addressed. The recent shift toward integrating CAI tools into interpreter training can be observed at the University of Bologna, which has started reviewing and re-evaluating their understanding and role of technology in their curricula and began integrating CAI into interpreter training (Prandi 2020: 6). However, there are limitations of the inclusion of these tools as well. For instance, the linearity of ASR transcripts may hinder interpreters from understanding hierarchical units of text, a task traditionally managed through the flexible spatial layout of note-taking (Setton 1999: 39). Ultimately, Prandi (2020) identifies a long-term risk of over-relying on ASR, at the potential expense of interpreters’ primary skills. To reduce the risk of diminishing these skills, it is essential that practitioners develop a balanced skill set that utilizes ASR as a targeted aid rather than a substitute for cognitive engagement (Prandi 2020: 2; 7). While the present research does not seek to argue for the universal necessity of ASR in interpreting, establishing its relevance aims to justify the necessity for an investigation in the first place by providing empirical evidence on ASR performance under conditions that have so far received limited attention.

2.3 Transcription practices and standards: ASR versus manual transcription

To systematically evaluate system performance across distinct speech modes, it appears necessary to first define what it is being compared to. Manual transcription refers to a person typing audiovisual content without technological assistance (Tardel 2020: 81). Human

transcription is acknowledged as a more time-consuming and expensive task, while ASR is often considered to be an effective choice to reduce documentation time and costs, however, its output varies strongly depending on the input (Bazillon et al. 2008: 1). For instance, Tardel (2020) found that providing professional subtitlers and translation students with raw ASR transcripts does not reduce the overall transcription time, as the system-generated version still needs extensive human revision. However, it did reduce the technical effort (Tardel 2020: 99). In contrast, Bazillon et al. (2008) compared ASR-assisted and unassisted transcription in a ten-hour task. They found that manually transcribing spontaneous speech took about twice as long as assisted transcription of planned speech. Although their study indicates that ASR may reduce transcription time, it is important to keep in mind that ASR performance is often inconsistent. It varies greatly depending on the input, struggling particularly with spontaneous and non-native speech, challenges that do not affect human annotators to the same extent (Bazillon et al. 2008: 4). This inconsistency is particularly pronounced when processing spontaneous non-native speech, categories that are central to the comparative analysis of speech modes in this thesis. Notably, a study of the year 2008 cannot be used to generalize findings for the system quality of nowadays.

According to Tardel (2020), ASR systems are unlikely to decrease transcription effort unless they can match the quality of manual transcription (Tardel 2020: 99–100). Russell et al. (2024) second this by claiming that current ASR technology cannot replace trained human transcribers, but high-quality ASR output may nonetheless serve as a useful starting point for subsequent manual refinement (Russell et al. 2024: 13). Pentland et al. (2023) further advise researchers to manually transcribe a representative subset of their data to compare it to ASR output, acknowledging that even in controlled laboratory settings, ASR systems frequently fall short of the accuracy achieved by human annotators (Pentland et al. 2023: 671-672). These findings suggest that ASR systems fail to capture the micro-phenomena that a human transcriber could, reinforcing the need for a comparison between ASR and manual transcript, as will be the case for the present research. Furthermore, Pentland et al. (2023: 671) indicate that ASR tools tend to underestimate the total number of words in a text compared to manual transcription and frequently transcribe the same word differently, especially in unclear speech. This observation questions ASR's capacity to manage repetition and fluency, two aspects directly investigated in this research. In the broader context of qualitative research, transcription procedures play a crucial role (Russell et al. 2024: 1), however, they are inadequately documented (Davidson 2009: 47-48; Point & Baruch 2023: 7). According to Point & Baruch (2023), some publicly available research provides very little information

about the transcription process, while others do not provide any information at all (Point & Baruch 2023: 6). For this reason, the current study provides insights into the transcription process (see chapter 4.2).

Finally, the work of Bucholtz (2000) suggests that manual transcription involves interpretative decisions on the part of the transcriber, such as including the decision of whether to produce naturalized or denaturalized transcripts. For the purpose of readability, naturalized transcription transforms spoken language into more conventional written forms. As a result, the original text's linguistic shape is less accurately represented. Denaturalized transcription, on the other hand, maintains the original oral characteristics but may seem fragmented or unfamiliar to readers (Bucholtz 2000: 1461). However, is important to consider that this study from the year 2000 does not accurately represent the newest transcription practices of nowadays, as in contexts such as verbatim reporting or machine learning, transcribers may follow predefined conventions as closely as possible, reducing subjective influence on the output.

Moreover, Gile (2009: 5) proposes a distinction between “Message (which can also be referred to as ‘Primary Information’)“ and “Secondary Information”. The inherent presence of the latter in any transcription could be considered a product of linguistic or cultural norms, or individual habits such as pauses or hesitation sounds. Human transcribers might be able to recognize these two forms of information, however, the capacity of ASR to make this distinction remains uncertain. Importantly, ASR models exhibit a naturalization bias by utilizing text normalization, prioritizing written-style coherence over the faithful representation of spoken sound (Kuhn et al 2023:19). On the other hand, manual transcripts are always influenced by the linguistic, social, and political decisions made by the annotator (Bucholtz 2000: 1463). These findings are an essential part for the present research utilizing denaturalized manual transcripts to compare ASR performance with.

2.4 Features of non-interpreted and interpreted speech

Having illustrated that ASR prioritizes written-style coherence over literal phonetic transcription, it can be assumed that ASR performance depends on the linguistic and acoustic properties of the input speech. To evaluate the performance of ASR across different production conditions, it is essential to comprehend the characteristics that distinguish non-interpreted from interpreted speech, particularly the variability introduced by planning conditions and cognitive load.

However, it appears necessary to establish transparency regarding the categorical difference between planned and unplanned speech first. The degree of speech planning can be considered a significant variable that potentially impacts ASR-generated output, particularly regarding the frequency and distribution of disfluencies in non-native speech. Inspired by the work of Dufour et al. (2014), this study classifies spontaneous speech as unplanned and prepared speech as planned. Planned speech typically appears to exhibit higher grammatical correctness, whereas unplanned speech often contains pauses, repetitions, repairs, and false starts, challenging ASR recognition (Dufour et al. 2014: 1).

2.4.1 Toward a profile of non-interpreted speech: variability in planning and verbalization

Generally, speaking is considered "one of man's most complex skills" (Levelt 1998: 1). In spontaneous non-interpreted speech, speakers generate and articulate their own conceptual content (Zhao et al. 2023: 2). Importantly, cognitive load is only required for the actual language production (Bóna & Bakti 2020: 487). The basic functions of speaking can be broken down to "referring, requesting, and explaining" (Levelt 1998: 1).
Levelt (1998: 9) further elaborates:

> Talking as an intentional activity involves conceiving of an intention, selecting the relevant information to be expressed for the realization of this purpose, ordering this information for expression, keeping track of what was said before, and so on. These activities require the speaker's constant attention.

Consequently, being a state that requires ongoing attention, non-interpreted speech itself must not be treated as a homogeneous category in the frame of the present research**. T**he continuous processes of planning, monitoring, and attentional control introduce variability in fluency, structure, and lexical selection (Levelt 1998: 9) and may directly interfere with ASR recognition processes. Generally, McMillan & Corley (2010) propose to understand the connection between high cognitive demand and the resulting physical properties of speech through a cascading architecture of language production. They challenge the traditional view of seeing speech errors as simple substitution, and instead propose that the articulation of phonemes is directly influenced by activation levels within the speech plan (McMillan & Corley 2010: 243). Within this framework, the authors argue that linguistic information flows continuously from the internal planning stages to the motor system of a speaker, without prior selection (McMillan & Corley 2010: 246). Finally, when the intended and unintended forms of such activations are present at the same time, they compete for control and may produce a

signal that includes parts of both units simultaneously (McMillan & Corley 2010: 259). This overlap may manifest in non-standard articulation of speech signals.

Continuing with different speech modes, Bóna & Bakti (2020) describe spontaneous speech as a minimally constrained production setting where speakers generate content freely about familiar topics without preliminary planning. Speakers independently select ideas, lexical items, and syntactic structures and determine both the scope and structure of their utterances. Temporal analyses indicate that spontaneous speech exhibits high speech rates and fewer filled pauses, suggesting faster planning processes compared to other tasks. At the same time, filler words occur frequently, indicating that hesitation phenomena are not eliminated under lower cognitive load but instead manifest in task-specific ways (Bóna & Bakti 2020: 501). These observations are central to addressing the underlying causes of the quantitative error patterns analyzed in the parameters of RQ2.

In contrast, prepared spoken language, such as read-aloud text, can be assumed to contain more phonetic and grammatical regularity compared, which are favorable conditions for ASR transcription accuracy (Fantinuoli 2017: 28–29; Zhao et al. 2023: 1). In contrast, spontaneous speech displays more disfluencies and a decreased use of proper nouns (Bazillon et al. 2008: 1). These variations in planning and monitoring indicate that the planned speech task of this study (E2) can be seen as a condition of higher predictability, which is relevant for the comparison of errors in RQ2.

Furthermore, Fantinuoli (2017) states that in formal institutional contexts, speakers may produce hybrid forms of oral discourse that combines elements of prepared and spontaneous style. Even outside interpreting contexts, such settings do not guarantee acoustically stable speech, as variability arises from individual speaking styles, accents, and degrees of spontaneity (Fantinuoli 2017: 28). Zheng et al. (2023) further observe that speaker-related variability can substantially affect ASR performance. The authors investigated dysarthric speech, a speech impairment brought on by muscle weakness that frequently results in unclear articulation, a slower speaking pace, and uneven prosody (Zheng et al. 2023: 1; 7-10). While their study is specific to this speech condition, the importance of their results lies in the overall principle that any disruption from smooth, standard speech, such as time-variability or slurred articulation, can impact ASR transcription. This highlights the need to incorporate non-native and spontaneous speech in ASR tests and may later provide potential explanations for differences in performance in the current study.

Finally**,** Sun (2023) reports that the inherent flexibility of spontaneous speech, where learners must prioritize semantic meaning over phonetic precision, often results in lower

output accuracy for ASR systems compared to planned read-aloud tasks (Sun 2023: 10). Her study examines the influence of ASR technology on L2 (second language) pronunciation and speaking skills of intermediate-level Chinese English as a Foreign Language (EFL) learners. Through the combination of empirical data and the personal insights of the students, her study provides a complex analysis of how automated tools process various modes of verbalization**.** Specifically, her research employs various testing formats, including scripted readings and unscripted dialogues, to observe linguistic output across a range of preparation levels (Sun 2023: 1; 4). While learners found ASR beneficial for identifying mispronunciations, they also expressed frustration with the technology's inconsistent recognition of natural, unscripted talk**.** The system-generated versions of these spontaneous conversations consistently contained errors that were absent in planned readings, highlighting the machine's limited capacity to handle the fluid markers of interaction produced by EFL learners (Sun 2023: 10)**.** Even though the aim of the present study is not to argue for the necessity of ASR utilization in general, these findings are directly applicable: Sun (2023) shows that the absence of pre-planning increases the complexity of the conversational task, potentially leading to more speaker-produced errors, and finally, challenging the machine's recognition stability in ways that structured read-aloud tasks would not, making them highly relevant for RQ1 and RQ2. Collectively, the insights of this chapter demonstrate that non-mediated speech ranges from carefully prepared, read-aloud presentations to fully spontaneous, unscripted talk. As established, linguistic variability already exists in non-mediated speech, however, to fully address the research objectives, it is necessary to examine how they are exacerbated by the unique characteristics in interpreted speech.

2.4.2 The multidimensional profile of interpreted speech

As established in the previous chapter, a fundamental difference between interpreted and non-interpreted speech lies in message generation. While spontaneous non-interpreted speech involves speakers generating and articulating their own content, interpreters reformulate a pre-existing source message. Specifically in CI, they must first comprehend the source message, retain portions of information in working memory and notes, and finally reproduce the message in the target language (Zhao et al. 2023: 2-3). Consequently, even though "Interpreting, as a distinct form of language production, largely follows the same cognitive processes as regular language production" (Zhao et al. 2023: 2), it introduces additional demands that significantly differentiate it from non-mediated speech. For instance, if basic

talking already requires “constant attention” (Levelt 1998: 9), this makes bilingual mediation a state of near saturation for cognitive processes, as established by Gile (2009) in his Effort Model. It explains that interpreting involves the simultaneous coordination of non-automatic comprehension, memory, and production processes that operate close to processing capacity limits (Gile 2009: 159; 190; 224). When these processing demands outweigh available resources, performance decreases (Gile 2009: 159). Once speech production has started, consecutive interpreters have limited options for output revision, and self-repair operations could further interfere with delivery (Gile 2009: 226). Zhao et al (2023) provide similar insights, claiming that interpreters often begin planning their target rendition while still processing the incoming source message, partially rely on source-language structures during formulation to facilitate production when source and target languages share similar syntactic patterns. The Exhaustion of cognitive resources challenges predictive mechanisms and increases the likelihood of cross-linguistic interference. This results in phonological errors, syntactic errors when the transferred structure is incompatible with the target language, as well as phonological errors (Zhao et al. 2023: 2-4). Specifically, Zhao et al. (2023) analyzed three factors in their study, namely language proficiency, anxiety and working memory. Out of these, the highest rate of total errors was traced back to language proficiency, while working memory mostly impacted conceptual and phonological errors and anxiety appeared to be the reason for errors on the lexical, syntactic and phonological side (Zhao et al. 2023: 9). The high number of conceptual errors found in their study further increased in case of the target language being the interpreter’s second language (Zhao et al. 2023: 8). Because interpreting often involves public speaking under evaluative pressure, many interpreters, especially those with little experience, experience anxiety, interfering with attentional control and speech planning, though there is a lack of empirical data on its specific impacts (Zhao et al. 2023: 3-4). Wang & Zheng (2021) further explain that interpreters subjected to time pressure and high-stakes environments remain highly sensitive to emotional triggers. Sustained stress and anxiety can disrupt prosody, lexical choice, and fluency, directly influencing the output. Consequently, they recommend including the ability of emotional regulation, being a skill that can be developed over time, in professional training (Wang & Zheng 2021: 221–222).

The need to comprehend, process, and reformulate meaning across languages in interpreting is acknowledged by Fantinuoli (2017) as well. He adds that interpreted speech generally tends to be slower and more fragmented, resulting in a temporal lag that is typically accompanied by pauses, hesitations, acoustic and linguistic unpredictability, as well as

semantic inconsistencies and inaccuracies. These variations diverge from the fluent, predictable speech patterns that most ASR systems are trained to process. In contrast, faster delivery may lead to articulation problems, which are harder for ASR to process (Fantinuoli 2017: 29). Bosker et al. (2017) provide similar insights regarding temporal variation by explaining that elevated cognitive load significantly alters the temporal encoding of information, often leading to acoustic distortions such as the "shrinking of time" (Bosker et al. 2017: 174) effect. This could specifically result in a higher perceived speech rate, which might interfere with the processing logic of ASR systems through variations in segment duration, rhythm, and prosodic patterning.

Another alternative conceptualization of cognitive load in interpreting is suggested by Bóna & Bakti (2020), who organize various speech production tasks along a continuum of increasing cognitive load, CI being in an intermediate position. They demonstrate that while CI allows for a degree of macro-planning before speaking, interpreters must deliver their speech from memory without relying on a written script, thereby consistently exhibiting higher frequencies of filled pauses and overall disfluency rates compared to those engaged in extemporaneous speech (Bóna & Bakti 2020: 487). Such differences have been attributed to the additional processing demands of bilingual mediation and message reformulation, implying that similarities in planning conditions do not remove the increased cognitive load associated with interpreting (Bóna & Bakti 2020: 501-502). These patterns provide empirical evidence for a direct correlation between the degree of cognitive effort and higher probability for speaker-related errors**.**

Beyond these characteristics, Xu & Li (2024) further support the structural distinctiveness of interpreted speech. Although their study focuses on SI, their mediation-related insights are not mode-specific can be applied to interpreted speech generally. Their multidimensional analysis shows that interpreted speech exhibits a combination of features associated with both spontaneous speech and mediated texts (Xu & Li 2024: 445; 464; 470). While interpreted and translated language differ in several respects, they also share more systematic similarities across multiple dimensions than might be assumed (Xu & Li 2024: 471). In dimensions associated with orality, interpreted speech patterns align more closely with native speech (Xu & Li 2024: 445), displaying fragmentation, simplification, and informal phrasing (Xu & Li 2024: 468; 470). At the same time, interpreted output also appears to resemble written translation, more specifically through the systematic underuse of stance-taking expressions (Xu & Li 2024: 468-469), which distinguishes both interpreted and translated speech from non-mediated speech.

Finally, the established insights offer a theoretical explanation for why interpreted speech often contains the temporal irregularities and lexical uncertainty that might challenge ASR tools. This variability manifests across several dimensions: temporal irregularities such as speech rate and pauses, disfluencies such as filled pauses, hesitations, self-repairs, lexical uncertainty, and syntactic or phonological inconsistencies (Zhao et al. 2023: 2–4; Bóna & Bakti 2020: 482, 497–502; Bosker et al. 2017: 174). Since non-standard phrasing might modify the acoustic signal in ways that are not usually reflected in training data, the machine's internal language models might struggle to accurately predict the next word in a sequence based on standard grammatical probabilities. Consequently, for this thesis, rather than being simple irregularities, disfluencies are considered markers of processing difficulty. Having established how these internal cognitive, linguistic, and psycho-emotional variables influence interpreted output, the next section examines the way contextual and linguistic factors across different speech types affect the performance of ASR systems.

2.5 Contextual, speaker-related, and linguistic constraints on ASR performance in institutional and non-native speech

The linguistic and contextual environment of any type of speech may present important factors to consider for the analysis of ASR outcomes. Specifically, given that the experiments conducted for this study are based on political discourse, it is necessary to examine how ASR architectures handle formal, terminology-dense input across varying levels of preparation. De Vos & Verberne (2020) compare ASR performance between prepared and spontaneous discourse in EU parliamentary committee meetings. They suggest that the short and largely scripted nature of plenary meetings may offer only a superficial basis for evaluation. Therefore, they selected committee meetings, which, in contrast, better reflect real-life scenarios. Additionally, there is a lack of official transcripts of such meetings, which is primarily due to the time and cost constraints of manual transcription (de Vos & Verberne 2020: 40). They also found that ASR systems frequently misrecognize proper names, specific terminology and institutional names, and these deviations should not only be seen as quantitative, measured in terms of word accuracy. Instead, they should be considered qualitative shortcomings that could lead to significant semantic distortions, making statements incomprehensible (de Vos & Verberne 2020: 42). However, it is noteworthy that their research was framed as a pilot study with a limited dataset, which might restrict the generalizability of their findings beyond the specific specialized context of EU politics.

Moving beyond the institutional setting, the physical profile of speakers may influence ASR performance as well. Zapata & Søeborg Kirkedal (2015) investigate ASR systems in dictation tasks within planned and spontaneous structures produced by speakers of various native languages and proficiency levels (Zapata & Søeborg Kirkedal 2015: 1; 205). Their findings suggest that the accuracy of ASR systems appears to be influenced not only by the ASR system used but also by speaker accent and profile (Zapata & Søeborg Kirkedal 2015: 205-207). Feng et al. (2024) expand on this by noting a measurable bias within ASR systems towards specific speaker groups, including non-native speakers and regional accent speakers. Importantly, they suggest that this bias should not fully be attributed to the speakers' pronunciation but appears to be a result of the ASR architecture itself. Hybrid DNN-HMM and end-to-end (E2E) models were found to perform differently across speaker profiles, suggesting that architecture-specific solutions may be necessary to improve inclusivity. Their findings reveal that ASR systems function best with input that is similar to native adult speakers with standard accents, and that any variation in terms of age, accent, and linguistic background results in increased recognition errors (Feng et al. 2024: 1).

Similar findings can be seen in a study conducted by Koenecke et al. (2020), where the error rate for Black individuals appeared to be almost twice as high as that of White individuals among the five most prominent ASR systems tested (Koenecke et al. 2020: 7684-7685). The authors did not trace the primary cause of errors to the complexity of the vocabulary, but rather "related to differences in pronunciation and prosody—including rhythm, pitch, syllable accenting, vowel duration, and lenition" (Koenecke et al. 2020: 7687). Importantly, they observed that the performance gap was localized in the acoustic models, indicating that the software struggled to recognize how words were pronounced, rather than what words were chosen. In contrast, the language models exhibited higher lower perplexity for the underrepresented group, meaning that it was easier to predict the speakers' sequences. Even though the language models also struggled with non-standard syntax, the ASR challenges were mostly triggered by acoustic failure. As can be seen from these documented disparities, they are a direct consequence of using homogeneous audio sets that do not have enough representation from diverse speaker populations (Koenecke et al. 2020: 7686-7687). In light of this, the authors recommend that future researcher strives to achieve equity by expanding their data set to other oppressed linguistic groups, particularly those with non-native and regional accents (Koenecke et al. 2020: 7688). Building on these insights, it is important to note that while speaker adaptation improves performance for some non-native speakers, it does not benefit all user groups equally, suggesting uneven system optimization

(Zapata & Søeborg Kirkedal 2015: 208). Even though the present study will not utilize optimized systems, such findings are important to consider when analyzing the ASR performance of interpreted speech, which frequently combines non-native accents, irregular prosody, and spontaneous reformulation under time pressure (Pöchhacker 2016, Gile 2009, Zhao et al. 2023).

Another important factor to consider in institutional speech is high terminology density. Zapata & Søeborg Kirkedal (2015) demonstrate that even small fluctuations in word accuracy may have disproportionate impacts on a transcript. Although their study was conducted aiming to improve medical dictation, it is consistent with the current research's focus on institutional, non-native speech. To limit the impact of the possible distortive factors, the authors replaced out-of-vocabulary (OOV) foreign words with generic English terms. However, they explicitly acknowledge a limitation of their study which is important to keep in mind for their results: the approach of the removal of OOV-terms does not authentically reflect real-life situations, as those normally include a high amount of foreign specialized and institution-specific terminology (Zapata & Søeborg Kirkedal 2015: 208).

Moreover, ASR performance, regardless of speech mode, may be influenced by the usage of non-lexical tokens (NLTs). Russell et al. (2024) argue that ASR performance is profoundly affected by conversational environments, noting that "NLTs carry meaning and should hence be transcribed" (Russell et al. 2024: 5). While the authors observe that commercial ASR systems mostly preserve the temporal location of many NLTs in their transcription, the system appeared to fail transcribing them accurately in terms of orthography (Russell et al. 2024: 1; 5). This failure suggests that there could be a bias in ASR, leaning towards representations of more written-like speech. If NLTs were to be transcribed inconsistently or removed entirely, this could lead to a less accurate representation of the speech or even a distortion of semantic meaning. Tran et al. (2023) confirm this risk, though they refer to non-lexical conversational sounds as NLCs. They claim that NLCs such as "Mm-hm" (Tran et al. 2023: 703) can convey vital information in conversations between clinical workers and patients, such as the confirmation of an allergy. In some cases of their study, ASR replaced tokens with "Hum-um" (Tran et al. 2023: 707) which could not be comprehensible for transcript readers and therefore lead to an information deficit. For clinical documentation purposes, the misrecognition or exclusion of NLCs could even mean putting patients health at risk (Tran et al. 2023: 710). One possible explanation for misrecognition is that many NLCs appear to sound similar (Tran et al. 2023: 703). However, it is noteworthy that their study utilized re-enacted recordings produced in a professional sound studio,

conducted with high quality equipment and native speakers (Tran et al. 2023: 705-707). Despite this idealized audio quality, they still recorded high NLT error rates. Additionally, both Tran et al. (2023) and Russell et al. (2024) utilized English, a high resource language. These conditions suggest that, in settings involving non-native speakers, underrepresented languages, lower quality equipment or uncontrollable factors such as noise interruptions, the misrecognition or omission of NLTs/NLCs would likely be even more present.

In sum, it can be assumed that ASR systems are optimized for written-style coherence and appear to exhibit higher error rates when exposed to institutional speech, where meaning is densely encoded, and small deviations may alter referential accuracy, making errors particularly consequential**.** Importantly, differences in ASR performance cannot be explained by the influence of a single factor but should rather be understood as context-sensitive phenomenon. The need for quality evaluation criteria that transcend simple word accuracy rates is discussed in the following chapter.

2.6 Quality assessment criteria for ASR comparison

The preceding chapters have illustrated that ASR performance varies systematically with speech type, linguistic structure, and interactional context. When it comes to measuring the quality of the resulting transcripts, the Word Error Rate (WER) appears to be the most commonly used measuring parameter, yet many ASR services do not specify their WER rate, instead presenting vague claims of qualitative accuracy across different settings (Kuhn et al. 2023: 2; Pentland et al. 2023: 672). Although Kuhn et al. (2023) acknowledge this issue, they do not hypothesize possible reasons or consequences for the lack of transparency. Text normalization procedures do not incorporate factors such as punctuation or capitalization, in other words, WER is seen as a necessary but insufficient transcription quality evaluation that needs supplementation with more microanalytic research strategies (Kuhn et al. 2023: 19-20). Consequently, recent research has increasingly questioned the assumption that ASR output can be adequately evaluated using a single aggregate metric such as WER. In a comparative evaluation of ASR systems, Russell et al. (2024) argue that although WER remains a useful way of measuring overall recognition accuracy, it only provides limited insight into transcript quality. For the purposes of analytical research, features such as the maintenance of turn structure and the faithful representation of non-lexical elements appear to be more significant indicators of quality than global word counts (Russell et al. 2024: 1; 12). Additionally, it is important to note that the misrecognition of names or institutions can carry high informational weight and therefore potentially change semantic meaning, which underscores the limitations

of using WER exclusively to assess ASR performance (de Vos & Verberne 2020: 42) and reinforces the need for a multi-parameter approach to ASR assessment in the present study's RQ2.

This need for a more qualitative approach is supported by the research of Davitti et al. (2024), who utilize a refined analysis for the interplay between speaker characteristics and system performance. They evaluate the use of ASR subtitles in broadcast contexts, aiming to provide a deeper understanding of how speech rate, accent, and fluency influence the ability of ASR tools to provide real-time transcription of speech produced by non-natives. They categorize ASR errors into content-based and form-based types (Davitti et al. 2024: 30), allowing for a deeper understanding of certain instances where content might be influenced by certain features, even though content is not altered. However, their research focuses mainly on the broader concept of speech within a broadcast scenario, which might have more factors influencing the overall intelligibility, compared to the speech types under investigation in the present thesis.

Moreover, regarding the processing of a final ASR transcript, there are automatic refinement programs, aiming to make transcripts more accessible to human readers. Liao et al. (2020) propose such a concept, namely ASR Post-Processing for Readability (APR), referring to the transformation of raw ASR output into more coherent text for human comprehension without altering the intended semantic meaning (Liao et al. 2020: 3). While categorizations such as content and form-based errors or post-processing strategies like APR will not be utilized in this thesis, understanding their purpose helps inform the suitable categories in RQ2. Instead, the present study will use the Normalized Error Rate (NER) per one thousand words, which aims to allow for a more transparent comparison across speech types while preserving the raw characteristics of ASR output. NER normalizes error frequency while maintaining sensitivity to speech-specific variability, therefore, it appears to be suitable for a comparative analysis across unequal speech samples. On this basis, the following sections define the six parameters utilized in the present research to identify specific error patterns **p**resented in corresponding NERs.

2.6.1 Fluency features

According to Bóna & Bakti (2020: 489), Fluency Features can be viewed as (filled) pauses, different kinds of repetitions, and broken words. The following proposed categories will be adopted for the current research:

a. Filled pauses: “Any sound or syllable which does not contribute to the meaning of the sentence“, such as “um” or “uh”;
b. Part-word repetitions: “A sound or syllable pronounced more than once with no additional meaning”;
c. Whole-word repetitions: “A word produced more than once with no additional meaning”;
d. Phrase repetitions: “More than one word produced more than once with no additional meaning”.

Furthermore, Liu et al. (2023) argue for a multimodal conceptualization of fluency that extends beyond purely auditory or acoustic signals. In this framework, speech fluency is a rather complex profile that incorporates text-related markers transcribed by the ASR system. The redundant repetition of lexical units are treated as indispensable components of the speaker's fluency profile (Liu et al. 2023: 2). Importantly, for the assessment of RQ2, it is important to note that ASR technology remains notably susceptible to non-standard pronunciation and various forms of speech disfluency, both of which frequently lead to more transcription errors (Prandi 2020: 15–16; Davitti et al. 2024: 3). For modern algorithmic architectures, standard fluency features like vocalized hesitations, verbal repetitions, and unfilled pauses pose major processing challenges (Fantinuoli 2017: 28). This technical difficulty may be increased even more when transcribing the output of trainee interpreters.

2.6.2 Syntactic Complexity

Ortega (2003: 492) characterizes Syntactic Complexity through “the range of forms that surface in language production and the degree of sophistication of such forms”. However, before attempting to define Syntactic Complexity, it appears useful to establish a basic understanding of complexity itself. According to Hulstijn & De Graaff (1994: 103), the “degree of complexity is contingent not so much on the number of forms in a paradigm, but rather on the number (and/or the type) of criteria to be applied in order to arrive at the correct form”. From this perspective, complexity refers to the density of the speaker’s mental decision-making process navigating linguistic rules, rather than the mere surface quantity of words produced.

This view can also be found in the work of Agmon et al. (2024), who studied sentence complexity by creating an automatic assessment system, tying it directly to syntactic structures (Agmon et al. 2024: 545). They found that simple utterance length becomes an

unreliable metric to evaluate ASR because the system's errors in word count and boundaries distort the data. When ASR utterance boundaries are inaccurate, every other measure of linguistic complexity appears to be negatively affected. They conclude that current ASR systems frequently fail to segment speech in a way that represents independent clauses, called T-units, accurately, which makes measuring sentence complexity in spoken discourse difficult (Agmon et al. 2024: 556).

Importantly, the spontaneous nature of interpreted speech can result in elliptical sentences that ASR systems might struggle to transcribe accurately (Fantinuoli 2017: 28). Some forms of non-interpreted speech, especially in informal settings, may present complex sentence structures that challenge ASR performance (Fantinuoli 2017: 29).

Additionally, Housen (2012: 23) further refines that Sentence Complexity can be considered a subcategory of Syntactic Complexity. Within this study, Syntactic Complexity is analyzed in terms of orthographic and syntactic structure. Specifically, it includes the hierarchical organization of words into phrases and larger units through a recursive set of linguistic rules (Agmon et al. 2024: 545). An error is coded into this category when the ASR system fragments features such as subordination, coordination, clause embedding, and non-linear sentence structures, resulting in the incorrect insertion of terminal boundaries that break apart a cohesive T-unit into disjointed or incomplete segments. These features increase the demands placed on both speaker and ASR, as they require the integration of multiple syntactic relationships within a single utterance. Additionally, this category includes orthographic inconsistencies in numerical rendering such as fluctuating between digits and words.

2.6.3 Speech errors

Drawing on the suggestions of Garrett (1975: 138), speech errors can be distinguished within the following categories:

a. "Addition", representing an unintentional insertion of an extra linguistic element that was not part of the intended message, occurring when two alternative words are both present in the output;
b. "General", which refers to the omission of a required linguistic unit, such as a prefix or a specific phoneme, from an intended word;
c. "Substitution", which is cosidered the replacement of a target word or sound with an alternative, frequently involving opposites or related concepts;

d. “Complex addition”; and e. “Complex deletion”, which involve the insertion or removal of larger, multi-segment portions of a word, representing a more significant disruption in word formation;
e. “Shift”, a specific unit (often a grammatical marker or affix) is moved from its intended position to another location within the utterance;
f. “Exchange”, where two discrete segments of the intended speech swap positions, providing evidence for the simultaneous processing of these elements;
g. “Fusion”, the blending of two competing lexical items, destined for the same functional role, into a single, hybrid word;
h. “Double whammy”, which is a dense error pattern where multiple types of slips occur simultaneously, resulting in a highly distorted output.

Similarly, Fromkin (1971: 30) categorizes most “segmental errors” into “substitution, transposition (metathesis), omission or addition”, explaining that “Simple anticipations result in a substitution of one sound in anticipation of a sound which occurs later in the utterance, with no other substitution occurring.”. This appears particularly relevant for the unplanned speech tasks. Specifically, as anxiety can lead to speech errors and negatively impact cognitive functions, this can affect the fluency and accuracy of speech production (s. Ch. 2.2). Consequently, the presence of Speech errors in both interpreted and non-interpreted speech are a key factor in assessing ASR performance, particularly when such errors stem from linguistic or cognitive stressors.

2.6.4 Phonological errors

Phonological errors can often be found in multilingual contexts, when the phonological information of one language triggers certain phonological associations in another language (Zhao et al. 2023: 3). Fromkin (1971: 30) establishes with practical examples that anticipations occur when a speaker substitutes a sound in anticipation of a sound that appears later in the utterance, whereas perseverance occurs when a sound from an earlier part of the utterance is carried over or persists into a later word. Zhao et al (2023: 2) provide similar explanations, and further elaborate that phonological errors typically manifest as mispronunciations in the target output through three primary mechanisms:

a. anticipation “(e.g., *leading list* instead of *reading list*)”;
b. preservation “(e.g., *a phonological fool* instead of *a phonological rule*)”;
c. phoneme exchange, where two distinct segments swap positions within an utterance “(e.g., *blake fruid* instead of *brake fluid*)”.

Fromkin (1971: 35) also demonstrates that errors can occur as a "single feature switch", where only a specific feature like nasality or voicing is changed while all other phonetic features remain intact. The present research will account for this utilizing the audio material, however, mispronunciations of standard forms are only counted as an error if the ASR and manual transcript do not match. Finally, any spelling deviations between ASR and manual transcript that are not considered to be more suitable in other error categories, are considered a phonological error.

2.6.5 Self-correction

One aspect that has been largely overlooked in previous studies is the role of self-correction during interpreting, which may further influence and possibly improve ASR performance. For this thesis, Self-corrections will be treated as a synonym for revision, which Bóna & Bakti (2020: 489) identify as "Instances when the speaker corrects an error". Furthermore, Fromkin (1971: 31) specifies that it is difficult to identify what the exact processes are when a "speaker corrects his error". The present thesis will adopt the proposed categories of Shen & Liang (2021: 768):

> (1) Repetition: the interpreter repeats one or more lexical items;
> (2) Restart: the interpreter restarts a new sentence before the completion of the previous one;
> (3) Replacement: the interpreter corrects phonological, lexical, grammatical and syntactic errors with immediate replacement;
> (4) Rephrasing: after the completion of a sentence, the interpreter rephrases it to bring it closer to the input in both form and meaning;
> (5) Delayed repair: the interpreter uses a different but improved version to interpret a word or a phrase once this word or phrase is said again by the original speaker.

Moreover, Shen & Liang (2021) discovered that trainee interpreters typically display a higher frequency of self-repairs and hesitations compared to established professionals (Shen & Liang 2021: 768). Because one experimental task of the present study specifically evaluates the output of a trainee interpreter, the analysis of Self-correction mechanisms serves as a critical evaluative parameter to analyze how ASR handles non-professional speech forms.

2.6.6 Punctuation errors

To understand the proposed error categories, this subchapter will include a brief explanation on the general topic of punctuation in the context of ASR systems. The coordination of punctuation in ASR output can be considered a task of structural reconstruction (Chen et al. 2015: 237). While technically modelled as a sequence labeling problem where the system predicts whether a token is followed by a specific mark or no mark at all, the practical reality

often appears to be one of high volatility (Agmon et al. 2024: 554-556). A possible explanation for this might be that punctuation marks often contain little acoustic weight (Chen et al. 2015: 238), making it difficult for automated systems to accurately identify suitable units when segmenting natural speech. Guo & Han (2024: 29) claim that this ASR-imposed punctuation process is often referred to as sentence boundary detection (SBD). For the purposes of the present study, a punctuation error will be categorized into the following types of deviation from the Gold Standard:

a. Insertion: The system inserts a sentence mark because of the incorrect identification of a sentence boundary (Guo & Han 2024: 36);
b. Deletion: The system overlooks a boundary produced by the speaker, leading to unwarranted merging (Guo & Han 2024: 36);
c. Substitution: The system changes a punctuation mark into an incorrect alternative (Chen et al. 2015: 238).

Having established these theoretical and evaluative foundations, the following chapter defines the research objective and the specific hypotheses guiding the empirical comparison of non-native interpreted and non-interpreted speech.

3 Research objective

After establishing that previous research has mostly been done on ASR's general performance in speech contexts (Alharbi et al. 2021; Zheng et al. 2023) as well as its cognitive implications for interpreters (Pisani & Fantinuoli 2021: 181), a significant gap remains regarding how these systems handle the linguistic product of interpretation itself. Given the growing utilization of AI tools in interpreting settings (s. Ch. 1), the present research aims to contribute to filling this gap by asking the following research questions: What are the key differences in ASR performance when transcribing non-native English speech produced in the modes of non-interpreted speech versus interpreted speech from German to English (RQ1)? Specifically, how might Fluency Features, Syntactic Complexity, Speech errors, Phonological errors, Self-correction and Punctuation errors influence the transcription (RQ2)?

Consistent with these objectives, the following hypotheses have been formulated based on the theoretical framework established in Chapter 2: First, E2 is expected to result in the highest accuracy (lowest overall NER). This assumption is based on the planned nature of the task, which provides a stable and syntactically coherent structure that more closely aligns with the written-style data on which ASR systems are typically trained. In contrast, the unplanned tasks (E1 and E3) are expected to exhibit higher overall NERs. While all three

tasks may be affected by non-native pronunciation and the presence of domain-specific or potentially unfamiliar institutional vocabulary, E1 and E3 introduce an additional layer of complexity through their unpredictable nature. In E1, errors are expected to arise primarily from real-time formulation processes, including hesitation phenomena and Self-corrections. These features are likely to result in elevated error rates in categories such as Fluency Features, Self-corrections, and Punctuation. In contrast, the challenges in E3 are expected to be more pronounced due to the cognitive demands of the simultaneous processing, reformulation, and production of speech (s. Ch. 2.1). Additionally, if the interpreter struggles with the terminology or is unfamiliar with the topic, they will likely rely more on notes. The increased cognitive load is likely to result in more frequent disfluencies such as filled pauses, and fragmentation, thereby posing greater challenges for the ASR system. As a result, E3 is expected to yield the highest overall NER among the three conditions.

Second, the ASR system is expected to exhibit a naturalization bias because of its inclination towards written style rather than spoken conventions. Considering the unpredictability of E1 and E3, both speakers in both tasks are expected to utilize NLTs, which ASR has been observed to transcribes unreliably (s. Ch. 2.2).

Third, the ASR system is expected to encounter difficulties in accurately identifying sentence boundaries, as participants are not instructed to explicitly verbalize punctuation markers. Instead, speech is produced in a natural manner, requiring the system to detect sentence boundaries based solely on prosodic and linguistic cues. Given the lack of explicit cues and the limited acoustic salience of punctuation (c. Ch. 2.6), all three tasks are expected to result in high error rates within Punctuation and Syntactic Complexity. Specifically in the unplanned tasks, likely containing more pauses or hesitations, these could be misinterpreted as sentence boundaries by the system.

The comparison of these tasks is expected to provide insights into how ASR systems perform across different speech conditions and levels of linguistic complexity. To address these hypotheses, the following chapter outlines the methodological framework, which adopts a sequential explanatory design combining quantitative error analysis with qualitative interpretation.

4 Methodology: Quantitative error analysis with qualitative interpretation

This Chapter outlines the planned empirical approach adopted in the present study. Building on the theoretical framework established in Chapter 2 and the hypotheses formulated in

Chapter 3, it describes the methodological procedures used to investigate ASR performance across the three experimental tasks. In particular, the chapter details the experimental design, the procedures for data collection and transcription, and the selection of participants and the ASR system.

4.1 Experimental design: the Spontaneous Speech, Read-aloud and Interpreting Task

This section outlines the analytical basis of the study and describes the three-step experimental setup to evaluate ASR performance across different levels of planning and mediation:

E1) Spontaneous speech task: A non-native English speaker with no formal linguistic training produces a spontaneous monologue in English that is simultaneously transcribed by the ASR system. Prior to starting, the participant receives a terminology checklist prepared by the researcher containing the following complex domain-specific terms: climate change, erosion of democracy, artificial intelligence, migrant smuggling, the Council of Europe, interinstitutional dialogue, values, freedom, security, and environmental protection. The participant is instructed to create a spontaneous unprepared monologue using this terminology, allowing for natural variation in syntax, fluency, content, and vocabulary.

E2) Read aloud task: The same participant (as in E1) reads a prepared English speech (Berset 2024) aloud, which is transcribed by the ASR system. The speech includes all terms from the list of E1, as well as other structured linguistic challenges, such as difficult-to-pronounce names (for example, Alain Berset). This task is designed to assess ASR performance with controlled and planned speech content.

E3) Interpreting task: The third task consists of live consecutive interpreting from German to English. The researcher reads the German version of the prepared speech (Berset 2024), and the participant delivers their consecutive interpretation into English, which is simultaneously transcribed by the ASR system.

Moreover, the selection of CI for the interpreting task in E3 aims to provide comparability with the spontaneous monologue produced in E1. For instance, the cognitive process of delivering a rendition in CI appears to resemble unmediated spontaneous speech more than in SI (s. Ch. 2.1). Additionally, CI removes the variable of simultaneous comprehension and rendering, which would otherwise introduce mechanical interference and time-pressure errors that would alter the interpreter's actual linguistic output. On the other hand, a simultaneous setup would have presented a practical recording constraint, as the researcher would not have been able to record the source delivery and the interpreted output

separately if they occurred at the same time. Consequently, E3 aims to provide a direct basis for analyzing how ASR systems handle the specific linguistic markers of interpreted speech when compared to the spontaneous (E1) and read-aloud (E2) baselines. However, it is important consider that while CI removes the overlap of tasks, it still presents its own unique cognitive demands that might impact the ASR performance. While this section established the experimental setup, allowing for a comparative analysis of ASR transcription performance across spontaneous, read, and interpreted speech, the following section will detail the analysis methods of the results.

4.2 Research design and analytical framework: quantitative error counting and qualitative interpretation

This section aims to provide transparency regarding the speech material utilized in the three-step experiment, while explaining how the specific, operationalized parameters elaborated in Chapter 2.6 will be practically applied to the data. The same speech material was utilized for all three experiments, though applied through different modalities: In E1, the terminology checklist provided to the participant was derived from the English version of the speech. In E2, the participant read the English version of the speech out aloud. Finally, in the consecutive interpreting task (E3), the researcher read the German version of the speech out aloud as the source for the interpreter's English rendition. The specific speech (Berset 2024) was chosen because the official website of the Council of Europe provides both English (Appendix SEN) and German (Appendix SGER) versions, providing the linguistic fidelity of the source material through a more reliable translation and mitigating the potential confounding variable of researcher-generated translations. While the chosen material did not contain excessive numerical data, a variable already extensively covered in previous ASR research, specific instances of numerals such as "75th anniversary" (Berset, 2024) were retained to maintain the texts as officially published. The German source segment selected for the interpreting task is approximately four to five minutes in duration. This length appears to be appropriate for evoking the sustained cognitive load and note-supported reconstruction processes that characterize CI in professional practice, aligning with the recommendation of Setton & Dawrant (2016) that CI should be tested via at least one long segment.

Furthermore, the current study is based on a sequential explanatory design (Kuckartz 2014: 162) aiming to not only comprehend the quantitative figures, but also their underlying causes. Precisely, for the present research, this means incorporating the potential explanations of the NERs that have been established for the ASR transcripts of all three experimental tasks.

The quantitative phase consisted of the systematic identification, counting, and categorization of ASR deviations from the Gold Standard based on the predefined typology of six parameters established in Chapter 2.6: Fluency Features; Syntactic Complexity; Speech errors; Phonological errors; Self-corrections; and Punctuation errors. The study maintains a distinction between inconsistencies in numerical rendering and capitalization. Within the category of Syntactic Complexity, the study specifically accounts for the inconsistent rendering of numerical data. Even though in the present research, an error is considered a deviation between Manual transcript and ASR transcript, there are two exceptions to be made: First, should the ASR system fluctuate between orthographic conventions in transcribing a number as a digit in one instance, and as a word in another, both instances are double-penalized even though one version may match the manual transcript. This decision is grounded in the orthographic inconsistency as well as the fundamental alternations in word count. However, to avoid redundant over-penalization, the study does not separately code the capitalization of words that immediately follow an algorithmic fragmentation, instead the instance will be displayed as one Syntactic Complexity error. Finally, while the coding framework in Chapter 2.6 establishes a detailed set of categories, not all of these were attested in the present dataset. This applies in particular to more complex categories such as delayed repairs, which were not found in the data.

Furthermore, to maintain consistency with the official source text, the Gold Standard adopts the lowercase orthography for “artificial intelligence” (Appendix SEN line 5; 6) as it appears in the original speech. Consequently, it is spelled in lowercase any time the term refers to the speech or the extracted terminology list. In the other parts of this thesis, it is spelled in capitals, as that would be considered the standard form.

Continuing to build directly on these quantitative findings, the second part of the analysis draws on principles of qualitative content analysis suggested by Kuckartz (2014: 109-113). “Qualitizing” (Kuckartz 2014: 88) seeks to interpret the quantitative data, taking the linguistic and structural evidence found within the experimental results into account. Precisely, for this research, this refers to the observational notes of the researcher, which were drafted in real time during each task. Their denaturalized version can be found at the end of Appendices MTE1, MTE2, and MTE3, aiming to provide insights into behavioral and procedural factors. Establishing the practical significance of the findings is important in small-scale T&I research where aggregate statistical significance is limited (Mellinger & Hanson 2016: 77) as mere statistical correlations can be considered reasonable explanations only when linked to the subjective or structural mechanisms underlying them (Kuckartz 2014:

78; Greene 2006: 93). Specifically, while a quantitative analysis might show that complex terminology correlates with higher error rates, only a qualitative investigation can reveal the causes as to why such a correlation occurs. This helps creating an intentional placement of both the quantitative identification of error patterns, as well as the interpretive framework, allowing for a basis to analyze the relation of ASR performance and the unique linguistic and communicative contexts of non-native interpreted and non-interpreted speech.

The ASR transcriptions of all three experiments can be found in Appendices E1, E2, and E3. These are compared to their corresponding manual transcripts MTE1, MTE2, and MTE3, serving as the Gold Standard adhering to the principles of denaturalized transcription (Bucholtz 2000: 1461). This means, all content, specifically NLTs, was transcribed based on the auditory perception of the researcher, aiming to maintain the surface form of all spoken utterances as they are encountered by the ASR system.

To provide transparency in the microanalytic process as well as easier error identification, all identified deviations between the ASR output, and the manual Gold Standard are highlighted in yellow in the Appendices. This visual marking includes both the primary errors contributing to the quantitative results as well as most of the speaker errors that were not counted as errors because they match the Gold Standard.

Having established the specific analytical procedures used to bridge the gap between these quantitative counts and their qualitative causes, the selection of the ASR tool and participants described in the next section further refine specifics regarding the three experiments.

4.3 Tool and participant selection

Building upon the analytical framework established above, this section describes the technical setup as well as the participant profiles recruited across the three experimental conditions. For all the pilot studies as well as the main experiments, a MacBook Air laptop with the newest update at that time was utilized. The tool selected for ASR transcription is Microsoft Word's Dictate function, based on the following methodological and practical considerations: first, it provides a live transcription feature within a Word document itself, unlike other ASR tools that often require uploading a pre-recorded audio file. Live transcription mimics the real-time processing demands that this study is trying to address, more closely. Secondly, the purpose of this study is to address transcription performance differences between non-interpreted and interpreted speech, rather than focusing on technical aspects that go beyond standard consumer hardware setup.

Additionally, it must be noted that there is no generic English language option to select from within the Word Dictate Function; instead, users must select from specific regional variations, which in this case included Indian, Canadian, Australian, British, and American English. This is why the language for the present study was set to English (USA) since this was assumed to provide the language variation that resembles most non-native accents.

Since variables such as accent, language proficiency, and familiarity with ASR tools can profoundly influence recognition error rate, the participant for E1 and E2 is a non-native English speaker, who had, at the time of the execution, no formal training in linguistic professions, aiming to ensure that speech production remains untrained and reflective of real-world non-expert conditions. In contrast, E3 requires a participant with advanced CI skills to evaluate how interpreted speech affects ASR performance. The task was therefore performed by a student from the University of Vienna, majoring in Conference Interpreting. The student thereby provides the necessary professional background for the interpreting task while introducing the additional complexity of a trainee-level interpreted output.

4.4 Execution of the experiments: Spontaneous Speech (E1), Read-Aloud (E2), and Consecutive Interpreting (E3)

This chapter details the execution phase of the experiments, including the time, date and place of recording. It starts with an overview of the pilot studies that were executed to assist in checking if the main experiments can be carried out the way they were planned and potentially pointing out any problems, followed with the description of the main three experiments.

4.4.1 Execution of the pilot studies

The pilot study of E1 was conducted on June 12th, 2025, in a domestic setting where specific measures were taken to eliminate potential background noise and external interruptions as much as possible. The participant is an acquaintance of the researcher who had no professional background in linguistics at that time. To start, the researcher explained the nature of the task: the participant must create a spontaneous monologue based on three terms (artificial intelligence, erosion of democracy, and climate change) which will be transcribed by an ASR system. Once the participant confirmed their readiness, the researcher explained that, to avoid interference with the transcription, they cannot talk any more from that point until the task is done. The Google Dictate Function was turned on, and the participant started

performing the task, however, the automatic punctuation system was turned off, so the ASR system started delivering the transcript in one non-stop text. The researcher did not to change any settings during the task, which took about 10 minutes total.

The pilot study of E2 was conducted immediately after the first pilot study, in the same environment, with the same participant. The researcher explained the second task, namely that the participant will now be asked to read a short excerpt of the speech from which the terms of the first task were extracted, which will again be transcribed by the ASR system. The researcher confirmed if the participant was ready, repeating that any interferences while the ASR system is transcribing her spoken output are to be avoided. The researcher enabled the automatic punctuation and started the Google Dictate function. The participant continued reading the selected part of the speech and gave a thumbs-up when she was finished. The researcher then pressed the stop button and confirmed the successful transcription. The task took about 8 minutes in total.

The pilot interpreting task (E3) was conducted on June 16th, 2025, at the Centre of Translation Studies, Vienna, in a quiet, lockable room. First, the researcher explained that the participant will be asked to do a consecutive interpretation on a small excerpt of a political speech from German into English. His interpretation would then be transcribed by an ASR system while the entire conversation would be recorded by the researcher's phone. After the participant's agreement, the researcher started the transcription and proceeded reading the selected excerpt from the speech (Berset, 2024). During that time, the participant took notes on his own papers. Then he proceeded to give the English interpretation, utilizing his notes. The process took about 15 minutes in total, including initial explanations.

4.4.2 Execution of the Spontaneous Speech Task (E1)

The first experiment was conducted on the 24th of November 2025, in a domestic setting with an acquaintance of the researcher. First, the task was explained, namely that the participant receives a list of political terminology to utilize for a spontaneous monologue. His output is then transcribed via Google Dictate function, and the whole session is simultaneously recorded via voice note on a phone. The researcher also informed the participant about the importance of minimizing interruptions during the task and asked him to refrain from touching the equipment or asking questions during the recording. The participant confirmed and was handed the terminology list on paper, which he read once. After checking the equipment for the correct language and the automatic punctuation settings, the researcher started the recording and transcription and aimed to minimize visual distractions by partially

closing the computer. The participant started his monologue, which lasted approximately four minutes. To confirm if the participant had finished, the researcher confirmed by giving a thumbs up before stopping the transcription as well as the phone recording. The details of the manual transcription process are elaborated in the following subchapter.

4.4.3 Execution of the Read-Aloud Task (E2)

The read-aloud task was conducted directly after E1, in the same location, with the same participant. The researcher begun by explaining the task: the participant digitally receives the prepared speech that the terminology from E1 was extracted from to read out aloud. The ASR system then transcribes the output while the phone records the entire session via voice note. After checking all settings, the researcher ran a brief test to see if the Dictate Function was working and enabled the transcription. The participant started reading the speech out loud from his phone, which lasted approximately three minutes, while the researcher noted observations and extra information. Upon finishing, the researcher stopped the ASR and phone recording and saved the Word document. As E1 and E2 were conducted directly after each other, the researcher was not able to start the manual transcription process before finishing both tasks. Upon completion, the researcher started listening to the voice notes via noise canceling headphones and manually transcribed everything the way it was heard from the recording. The process required multiple rounds of listening for E1 as well as E2, especially for segments that were ambiguous because of the speaker's articulation or because they were not loud enough. In the first rounds, the focus lied on merely transcribing content, followed by rounds of more thorough listening for NLTs and finally a last round to specifically check for deviations between manual and ASR transcript. For E1, the manual transcription process took approximately two hours, for E2 approximately one hour. Finally, the researcher added the observational notes taken during the task below the manual transcripts kept as bullet points.

4.4.4 Execution of the Interpreting Task (E3)

The interpreting task was conducted at the Centre for Translation Studies, Vienna, on June 16th, 2025. It took place in the same room that was utilized for the execution of the pilot study. The participant, a trainee interpreter majoring in conference interpreting, received instructions to consecutively interpret a German political speech into English. The participant asked for reassurance regarding the ability to use notes, which the researcher confirmed. Additionally, the participant inquired about unfamiliar terminology, therefore, the researcher

spelled out the name Alain Berset, which the participant wrote at the top of her paper. The participant brought a prepared binder of custom-cut papers to organize notes, as well as two pens. After confirming if the participant is ready and checking the equipment, the researcher read the speech excerpt aloud, which took approximately five minutes, while the interpreter took notes. After finishing, the participant started delivering her interpretation, which the Google Dictate Function transcribed. The participant proceeded to interpret the excerpt, which lasted approximately five minutes. The whole task, which took approximately fifteen minutes in total, was recorded via Smartphone. Immediately after completing E3, the researcher began the manual transcription process, resulting in approximately three hours, which is the longest overall time needed for manual transcription between the three tasks. The audios display many ambiguities in various instances, where the articulation of the interpreter is not immediately identifiable. Therefore, the manual transcription required multiple rounds of listening before proceeding, even though noise canceling headphones were used. Similarly to the production of the manual transcripts of E1 and E2, the first rounds of listening were used to transcribe content. After that, the focus shifted to NLTs and interpreting strategies and any other unclear instances of articulation. Finally, one last round of listening concerned the deviations between manual and ASR transcript. The qualitative notes taken during the interpretation, in form of bullet points, were included at the end of the manual transcript.

4.5 Procedures for error identification and counting

Following the detailed execution phase of the three experimental tasks, this chapter details the methods and procedures used to analyze the results. Importantly, the final manual transcripts (found in appendices MTE1, MTE2, and MTE3) prepared by the researcher serve as the Gold Standard for evaluating the raw ASR output (found in appendices E1, E2, and E3). After the execution of the experiments, the analytical process transitioned to a systematic, multi-stage procedure to identify and quantify every deviation that has been found between ASR and Gold Standard. These are displayed in comparative tables (Table 5.1, 5.2 and 5.3) and categorized into the six parameters established in Chapter 2.6. The comparative tables display the raw counts for every transcription discrepancy that has been found, mapped directly to their corresponding lines in both the ASR and manual transcripts. Where a single segment exhibits multiple errors in the same line, each point of divergence is recorded. Generally, errors of the same category are consolidated within a single row, however, some instances involving distinct error types are listed separately if it otherwise would have potentially created ambiguities in the tables. Because the three experimental tasks vary in length, raw

counts have been converted into NERs, calculated by dividing the raw error count by the total ASR word count and multiplying the resulting quotient by 1,000. This quantification allows for the identification of the performance hierarchy and the qualitative error profiles unique to each communicative mode. Finally, all NERs of each experiment are presented in an overview (s. Ch. 5.4).

Aiming for higher reliability, the identification process was conducted through a recursive analytical loop involving over ten rounds of iterative auditing. During these rounds, preliminary error counts were cross-referenced against original audio recordings and observational notes. These notes provided essential behavioral context, such as visible signs of speaker anxiety or task-specific hesitation. With these analytical procedures established, the following chapter presents the quantitative findings across the three experimental conditions to reveal the performance of ASR transcription between non-native interpreted versus interpreted speech and the unique error profiles generated by each communicative mode.

5 Quantitative results: ASR performance across interpreted and non-interpreted speech

This Chapter displays the results from the three experimental tasks: Experiment 1 (E1, Spontaneous Speech), Experiment 2 (E2, Read-Aloud), and Experiment 3 (E3, Interpreting). The analysis is based on the procedures established in the previous chapter 4.5.

For the purposes of this analysis, only errors that are attributable to ASR transcription were counted as such, or on other words, deviations from the Gold Standard. As a result, errors attributable to the speakers where the ASR transcriptions match the Gold Standard were not counted as errors. If they resulted in secondary structural or linguistic disruption in the ASR output, this is represented in the corresponding categories. The confirmed ASR output word counts used for normalization are: E1 = 456 words; E2 = 414 words; E3 = 381 words. The audio data for all tasks can be found in a Google Drive link.[1]

5.1: Results of Experiment 1 (E1): The Spontaneous Speech Task

In the spontaneous speech task, a non-native English speaker produced an unprepared monologue. This real-time generated output lasted approximately 5 minutes and 20 seconds

[1] https://drive.google.com/drive/folders/1BdSPOlR23hVOdjysuqSnh-eVQfSZXQML?usp=sharing

and displays a total of 456 words in the ASR output. In contrast, the Gold Standard consists of 455 words. The raw error count for E1 is 75, resulting in a total NER of 164.47.

Starting with the lowest error category, Self-correction, there are no instances of a mismatch between ASR transcription and manual transcript, resulting in a NER of 0.00. However, there are several instances where the speaker utilizes revision, though these match the manual transcript. They are nevertheless relevant for the interpretation of findings, which is why they are mentioned: First, the speaker performs the replacement "incidents. Incidences" (Appendix MTE1 line 10-11), which was transcribed the exact same way in the ASR version (Appendix E1 line 10-11). Second, ASR transcribed both tokens of the repetitions "is is" (Appendix MTE1 line 3-4; Appendix E1 line 3-4), and "our our" (Appendix MTE1 line 24, Appendix E1 line 24).

Continuing with Fluency Feature errors, displaying the second lowest error category (NER 4.39), the ASR system omitted one filled pause "Ehh" (Appendix MTE1 line 2) and one part-word repetition "s-" (Appendix MTE1 line 25). Nevertheless, there are numerous instances of errors produced by the speaker matching the Gold Standard such as the whole-word repetition "like like" (Appendix MTE1 line 9; Appendix E1 line 9); or the whole phrase repetition "people get people get" (Appendix MTE1 line 8; appendix E1 line 8). The repetition "this that that" (Appendix MTE1 line 24) was fragmented into "this. That that" (Appendix E1 line 24) and displayed in the Sentence Complexity error category, but the speakers repetition itself was transcribed accurately.

The third lowest error category (8.77 per 1,000 words) in E1 is phonological errors. Auditory analysis indicates that the speaker slurred his articulation and spoke rather quietly in the instances where errors occurred. Examples include the systems phonological misrecognition of "matter" (Appendix MTE1 line 22), instead transcribing "manner" (Appendix E1 line 22); as well as the transcription of "this side" (Appendix E1 line 16) instead of decide" (Appendix MTE1 line 16). Moreover, there was a phonological segmentation of the single unit "interinstitutional" (Appendix SE line 17), that the speaker pronounced as "intern-institutional" (Appendix MTE1 line 12), into the two single tokens "inter" and "institutional" (Appendix E1 line 12). Auditory analysis confirms that the speaker produced the non-standard form "intern-institutional" (Appendix MTE1 line 12), inserting an unwarranted letter followed by a distinct pause between the two units, coded as a perseverance.

Speech errors resulted in a NER of 19.74. Notably, the term "ChatGPT" was entirely deleted (Appendix MTE1 line 5; appendix E1 line 5). Lexical substitutions, such as the ASR

system turning “gonna” (Appendix MTE1 line 5) into “going to” (appendix E1 line 5), were found as well. Additionally, E1 displayed fluctuations in tense, such as transcribing “were” (appendix E1 line 9) instead of “are” (appendix MTE1 line 9); or “there was” (appendix E1 line 22) instead of “there is” (Appendix MTE1 line 22). However, the ASR system still transcribed the repetition of the substituted version: “There was there was” (appendix E1 line 22-23). In some instances, the ASR system changed words completely, causing a notable difference in ASR and manual transcription. In the first instance, it transcribed “How can I say” (Appendix MTE1 line 2) as “How can this say it?” (Appendix E1 line 2). Precisely, the ASR system performed a lexical substitution, replacing the first-person pronoun “I” with the demonstrative “this”, while simultaneously performing an addition by inserting the token “it” at the end of the phrase. Because of this overlap, this instance is classified as a double whammy. In another instance, the ASR system transcribed “alarm. But I” (appendix E1 line 18) instead of “or like they” (Appendix MTE1 line 17-18). Auditory analysis indicates that the speaker slurred his speech in this scenario.

Syntactic complexity errors exhibit the second highest error rate in E1, with a NER of 57.02. A limited number of errors appear to occur when the speaker uses filled pauses, such as in “because, umm, it affects” (Appendix MTE1 line 1-2) which ASR turned into “because. Um. It affects” (Appendix E1 line 1-2). More frequently, ASR fragmented segments of non-filled pauses. Auditory analysis indicates that this happened, for instance, when it transcribed “when they get” (appendix MTE1 line 8) as “when they. Get” (Appendix E1 line 8) with the speaker pausing between the words “they” and “get”.

Lastly, the highest error category is Punctuation errors, resulting in a NER of 74.56, mainly consisting of misplaced full stops in instances of fragmentation. In one specific instance, the ASR system inserted a question mark into the phrase “that's. Very hypocritical, no? Whatever.” (Appendix E1 line 4) while the Gold Standard displays “that's very hypocritical, no, whatever.” (Appendix MTE1, line 4). Auditory analysis of the source recording points to the speaker not raising his intonation typically associated with a question; rather, his tone remained flat and consistent as he appears to have paused before moving on to the next segment.

Moreover, it is noteworthy that there are instances where multiple errors have been identified in one sequence, for example the following: the Gold Standard displays “had to work and sleep at day, and work at night, but yeah” (Appendix MTE1 line 21), while ASR transcribed “had to work and sleep at day, at work at night, but. Yeah” (Appendix E1 line 21) fragmenting the sentence around the word “but”, inserting a full stop and capitalizing the next

word, as well as substituting “and” with “at” in a sequence where the speaker uses the word “at” twice. Consequently, this single instance is classified as a Syntactic Complexity error, a Speech error (substitution) and a Punctuation error (insertion).

Overall, E1 resulted in a high number of ASR errors affecting textual structure, particularly in relation to sentence segmentation. The two highest error categories, Syntactic Complexity and punctuation, were often connected, consisting mainly of misplaced full stops and sentence fragmentation, and the resulting inappropriate capitalization following automatic sentence breaks. Finally, instances such as the expansion of the contraction “gonna” (appendix MTE1 line 5; Appendix E1 line 5) the perseverance of “intern-institutional” (Appendix E1 line 12; Appendix MTE1 line 12); the insertion of the token “it” (appendix E1 line 2); the misrecognition of “decide” (Appendix MTE1 line 12) contributed to a difference in word count between ASR and manual transcript.

Table 5.1: ASR error analysis for Experiment 1 (Spontaneous Speech)

ASR error type	Specific error instance	ASR line	Manual transcript line	Raw count (C)	Normalized error rate
Fluency features				Total 2	4.39
	Omission of filled pause “Ehh”	2	2	1	
	Omission of “s-” (part-word repetition) in “they s- still”; ASR only transcribed “still”	25	25	1	
Syntactic complexity				Total 26	57.02
	“because, umm, it affects” fragmented into “because. Um. It affects”	1-2	1-2	1	
	“and me, personally” fragmented into “and. Me personally”	2	2	1	
	“don’t even think that” fragmented into “don't even. Think that”	3	3	1	

	“that’s very hypocritical” fragmented into “that's. Very hypocritical”	4	4	1	
	“I’m gonna skip that” fragmented into “I'm going to. Skip that”	5	5	1	
	“artificial intelligence is like ChatGPT and it's basically” fragmented into “artificial intelligence is like. And it's. Basically”	5-6	5-6	1	
	“like, it could” fragmented into “like. It could”	6	6	1	
	“smuggling is not a” fragmented into “smuggling is. Not a”	8	8	1	
	“because a lot of” fragmented into “because. A lot of”	8	8	1	
	“when they get” fragmented into “when they. Get”	8	8	1	
	“it is a political institution” fragmented into “it is a. Political. Institution.”	11-12	11-12	1	
	“what this is or what I could say about it” fragmented into “what the what is this or. World. I could say about it.”	13	13	1	
	“change over time and” fragmented into “change over time. And”	14	14	1	
	“to keep them” fragmented into “to keep. Them”	16	16	1	
	“keep them in your life or not.” fragmented into “keep. Them in	16	16	1	

	your life or not?" and changed into a question				
	"it's their freedom to decide" fragmented into "it's their freedom to. This side."	16	16	1	
	"who they gonna like" fragmented into "they. Gonna like"	17	17	1	
	"or like they" fragmented into "alarm. But I"	17	17-18	1	
	"don't see them anymore, but yeah. Umm." fragmented into "don't see them anymore. But yeah, umm."	18	18	1	
	ASR transcribes "10 to 15 years" as a number	20	19-20	1	
	"wasn't home in all those years. Because" fragmented into "wasn't at home. In all those years because"	20-21	20	1	
	"sleep at day and work at night, but yeah," fragmented into "sleep at day, at work at night, but. Yeah,"	21	21	1	
	"also a very important, um, matter" fragmented into "also a very. Important, um, manner"	22	22	1	
	"because there is, there is a lot" fragmented into "because. There was there was a lot" (phrase repetition)	22-23	22	1	
	"should, uhh, should not look away when they" fragmented	23	23	1	

	into “should. Uh should not look away. When they”				
	“that th- this that that” fragmented into “that do this. That that”	24	23-24	1	
	“and they feel guilty” fragmented into “and. They feel guilty”	25	25	1	
Speech errors				Total 9	19.74
	“How can this say it?” instead of “How can I say” (double whammy)	2	2	1	
	omission of “ChatGPT” (deletion)	5	5	1	
	“going to” instead of “gonna” (substitution)	5	5	1	
	“scarier” instead of “scary” (substitution)	7	7	1	
	“were” instead of “are” (substitution)	9	9	1	
	“their” instead of “your” (substitution)	17	17	1	
	“alarm. But I” instead of “or like they” (substitution)	17	17-18	1	
	“at” instead of “and” (substitution)	21	21	1	
	“there was” instead of “there is” (substitution)	22	22	1	
Phonological				Total 4	8.77
	“intern-institutional” (perseverance) separated into “inter institutional”	12	12	1	

	“this side” instead of “decide”	16-17	16	1	
	“well the thing” instead of “when I think”	18	18	1	
	“manner” instead of “matter”	22	22	1	
Self-correction				Total 0	0.00
Punctuation				Total 34	74.56
	Two misplaced full stops in “because. Um. It affects” (insertion)	1-2	1-2	2	
	Misplaced full stop in “and. Me personally” (insertion)	2	2	1	
	Misplaced full stop in “don't even. Think that” (insertion)	3	3	1	
	Misplaced full stop in “that's. Very hypocritical” (insertion)	4	4	1	
	Question mark instead of comma in “hypocritical, no? “ (substitution)	4	4	1	
	Misplaced full stop in “I'm going to. Skip that” (insertion)	5	5	1	
	Two misplaced full stops in “artificial intelligence is like. And it's. Basically” (insertion)	5-6	5-6	2	
	Misplaced full stop in “like. It could” (insertion)	6	6	1	
	Misplaced full stop in “smuggling is. Not a”	8	8	1	
	Misplaced full stop in “because. A lot of” (insertion)	8	8	1	
	Misplaced full stop in “when they. Get” (insertion)	8	8	1	

	Two misplaced full stops in “it is a. Political. Institution.” (insertion)	11-12	11-12	2	
	Two misplaced full stops in “what the what is this or. World. I could say about it.” (insertion)	13	13	2	
	Misplaced full stop in “change over time. And” (insertion)	14	14	1	
	Misplaced full stop in “to keep. Them” (insertion)	16	16	1	
	Question mark instead of full stop (substitution)	16	16	1	
	Two misplaced full stops in “it's their freedom to. This side.” (insertion)	16-17	16	2	
	Misplaced full stop in “they. Gonna like” (insertion)	17	17	1	
	Misplaced full stop in “alarm. But I” (insertion)	17	17-18	1	
	Two misplaced full stops in “don't see them anymore. But yeah, umm.” (substitution)	18	18	2	
	Misplaced full stop in “wasn't at home. In all those years because” (insertion)	20-21	20	1	
	Misplaced full stop in “sleep at day, at work at night, but. Yeah,” (insertion)	21	21	1	
	Misplaced full stop in “also a very. Important, um, manner” (insertion)	22	22	1	

	Misplaced full stop in “because. There was there was a lot” (insertion)	22-23	22	1	
	Two misplaced full stops in “should. Uh should not look away. When they” (insertion)	23	23	2	
	Misplaced full stop in “that do this. That that” (insertion)	24	23-24	1	
	Misplaced full stop in “and. They feel guilty” (insertion)	25	25	1	
Total errors				75	164.47

5.2: Results of Experiment 2 (E2): The Read-Aloud Task

In E2, the same participant as in E1, read a prepared political speech (Berset 2024) aloud, which includes all the terms from the list given to the participant in E1. The reading lastex approximately 3 minutes and 25 seconds. The ASR output as well as the Gold Standard display 414 words. A total of 27 ASR errors were identified, corresponding to a normalized error rate of 65.22 errors per 1,000 words.

First, Fluency Feature errors are limited to one instance only, displaying the lowest error category in E2, with a NER of 2.42. The instance recorded in this category is the unwarranted insertion of the conjunction “and” (Appendix E2 line 5), which does not correspond to any lexical item or NLT produced by the speaker.

The second lowest error categories are Phonological and Self-correction, both displaying a raw count of 3 errors and a NER of 7.25. Starting with Phonological errors, the following instances have been identified: the misrecognition of the name “Alain Berset” (Appendix MTE2 line 7; Appendix SEN line 7; 21), transcribed as “Ellen Perset” (Appendix E2 line 7); the fluctuation in the regional spelling variant “centre” (Appendix E2 line 23); and “we’ll” (Appendix E2 line 25) instead of “I will” (Appendix MTE2 line 26).

Following with Self-corrections, in the first instance, ASR deleted the speaker’s mispronunciation “citn” (Appendix MTE2 line 11) and only transcribed the replacement “citizens” (Appendix E2 line 11). Similarly, the token “pri” (Appendix MTE2 line 11) was omitted, while only the replacement “priorities” (Appendix E2 line 11) was transcribed. Moreover, one of the speaker’s self-corrections, even though it is not a mismatch between ASR transcript and Gold Standard, is being displayed for the analysis of results. According to

auditory analysis and observational notes, the speaker first appears to have struggled reading, therefore articulating the initial version “he had” (Appendix MTE2 line 19), before correcting “he added” (Appendix MTE2 line 19). The entire sequence “he had, he added” (Appendix E2 line 19) was transcribed by the ASR system, even though it fragmented the phrase beforehand “formed by. Our values” (Appendix E2 line19) by inserting a full stop.

Punctuation and Syntactic Complexity errors display a NER of 12.08. Specifically, Punctuation errors and corresponding sentence fragmentation arose mostly during the speaker’s pauses. For instance, the misplaced full stop in “formed by. Our values” (Appendix E2 line 19) caused sentence fragmentation and was therefore counted as a syntactic complexity error as well. Auditory analysis suggests that in this instance, the speaker hesitated and briefly paused between the words “by” and “our” (Appendix MTE2 line 19). Fragmentation and a following misplaced full stop indicating the end of a sentence, can be observed in the sequence “Dialogue is a paramount. Importance” (Appendix E2 line 16). Here, the speaker paused between the words “paramount” and “importance” (Appendix MTE2 line 16). Additionally, the system deleted a full stop in “Importance A dialogue” (Appendix E2 line 16), however, it capitalized the beginning of the next sentence. In another instance, the ASR system substituted a full stop with a comma, as seen in “to name but a few,” (Appendix E2 line 5). Based on auditory analysis, the speaker paused and decreased his pitch after the word “few” (Appendix MTE2 line 5) which could be interpreted as the end of a sentence.

Speech errors display the most frequent error category in E2, with 10 errors in total, contributing to a NER of 24.14. Examples include the inconsistent capitalization of institutional titles such as “Environmental Protection” (Appendix E2 line 5-6) or “Artificial Intelligence” (Appendix E2 line 6-7), which was capitalized in one instance, but not in another (Appendix E2 line 5). Additionally, the tense of the verb “build” (Appendix MTE2 line 4) was changed and replaced with “built” (Appendix E2 line 4). Finally, it is worth noting that the abbreviation “PACE” (Appendix E2 line 9) was transcribed in capitals only, matching the Gold Standard (Appendix MTE2 line 9).

Table 5.2: ASR error analysis for Experiment 2 (Read-aloud)

Error type	Specific error instance	ASR transcript line	Manual transcript line	Raw count (C)	Normalized error count

Fluency				Total 1	2.42
	Insertion of “and”	5	5	1	
Syntactic complexity				Total 5	12.08
	Transcribed “75” as a number	10	10	1	
	Transcribed “three” as a word	11	11	1	
	“Dialogue is a paramount importance. A dialogue between” fragmented into “Dialogue is a paramount. Importance A dialogue between”	16	16	1	
	“formed by our values” fragmented into “formed by. Our values”	19	19	1	
	“security that democracy brings” fragmented into “security. That democracy brings.”	24	24-25	1	
Speech				Total 10	24.14
	“the” instead of “a” (substitution)	2	2	1	
	“factor” instead of “fact” (substitution)	3	3	1	
	“build” instead of “built” (substitution)	4	4	1	
	Capitalized “Environmental Protection”	5-6	5-6	1	

	Inconsistent capitalization of “artificial intelligence” (capitalized in line 6-7, but not line 5)	5-7	5-7	1	
	Did not capitalize “secretary” (instance 1 and 2) or “general”	8	8	3	
	Capitalized “Heads” in “Heads of state”, but not “state” or “government”	21	21	2	
Phonological				Total 3	7.25
	“Ellen Perset” instead of “Alain Berset”	7	7	1	
	“centre” (fluctuation in regional spelling variant)	23	24	1	
	“we'll” instead of “I will”	25	26	1	
Self-correction				Total 3	7.25
	Deletion of the speaker’s token “citn” and only transcribed “citizens” (replacement)	11	11	1	
	Deletion of the speaker’s token “pri” and only transcribed “priorities” (replacement)	11	11	1	
	Deletion of the speaker’s token “gover” before “government” (replacement)	21	21	1	
Punctuation				Total 5	12.08

	Comma instead of full top in “to name but a few” (substitution)	5	5-6	1	
	Misplaced full stop in “paramount. Importance” (insertion)	16	16	1	
	Missing full stop after “Importance” (deletion)	16	16	1	
	Misplaced full stop in “formed by. Our values” (insertion)	19	19	1	
	Misplaced full stop in “freedom and security. That democracy brings” (insertion)	24-25	24-25	1	
Total errors				27	65.22

5.3: Results of Experiment 3 (E3): The Interpreting Task

In E3, a trainee interpreter performed a consecutive interpreting task from German to English, based on the same speech used for E2. The reading of the German version by the researcher lasted approximately 4 minutes and 30 seconds, while the interpreters output resulted in approximately 4 minutes and 10 seconds. The ASR output resulted in 381 words, while the Gold Standard displays 407. A total of 69 ASR errors has been identified, resulting in a NER of 181.10.

Self-correction errors present the lowest error category with a raw count of two and a NER of 5.25. While approximately eight total instances of the speaker utilizing Self-corrections have been identified, two of them have been counted as ASR errors. For instance, the reformulation “when he was, uhh, starting his, uhh, f-uhh-period” (Appendix MTE3 line 10) that the ASR system fragmented into “his. Period” (Appendix E3 line 9) omitting all filled pauses. In this case, the interpreter corrects the aborted articulatory attempt “f-uhh-” with the immediate substitution “period of” (Appendix MTE3 line 10). In another sequence, the ASR system performs a substitution by transcribing only the speaker’s corrected version “a” (Appendix E3 line 7), excluding the corrected word “an” (Appendix MTE3 line 8). Other

instances of Self-correction made by the speaker but not counted as an ASR error include the following examples: the rephrasing of “such as, uhmm, helping, uhmm, for safer, making migration safe” (Appendix MTE3 line 5) which ASR transcribed as “such as. Helping for safer making migration safe” (Appendix E3 line 4-5); the replacement of “these, uhh, take” with “start taking to over” (Appendix MTE3 line 12), which ASR transcribed as “these. Take start taking to over” (Appendix E3 line 11); the rephrasing “put my work and efforts into the, uhh, work and efforts into the” (Appendix E3 line 23) which ASR transcribed as “put my work and efforts into the work and efforts into the” (Appendix E3 line 21-22). In these instances, the system frequently transcribes both the initial version and the subsequent correction. This results in multiple duplicated word sequences, such as “central powers are powers are centralizing” (Appendix E3 line 10) or “put my work and efforts into the work and efforts into the” (Appendix E3 line 21-22), where the system deletes the filled pauses. These omissions are displayed in the Fluency Feature error category, whereas the fragmentations can be found in the Syntactic Complexity category.

The second lowest error category, Phonological errors, resulted in a total of 6 identified errors and a corresponding NER of 15.75. Phonological errors include the misrecognition of the proper name “Berset” (Appendix MTE3 line 24), which has been transcribed as “birth set” (Appendix E3 line 23); as well as the regional spelling variant “center” (Appendix E3 line 19) instead of “centre” (Appendix MTE3 line 20). Additionally, the ASR version displays the word “OK” (Appendix E3 line 1), deviating from the Gold Standard displaying “okey” (Appendix MTE3 line 1).

Following with Speech errors and an identified raw count of 7, the result is a NER of 18.37. One noteworthy error example is Appendix MTE3 line 20), Another instance worth noting is the lexical omission of the core lexical item “period” (Appendix MTE3 line 20) combined with a Punctuation error. While the speaker utters this term as part of the phrase “a goal of his, uhh, active period, that every human” (Appendix MTE3 line 20) the ASR system renders this literally as a full stop (Appendix E3 line 18). This instance resulted in the transcription “a goal of his. Active. That every human” (Appendix E3 line 18), categorized as a Speech error and a Syntactic Complexity error due to the resulting syntactic fragmentation of the clause. Additionally, the ASR system inconsistently capitalized the phrase “all 46 member States” (Appendix E3, line 15). While it correctly identified the numerical data, it deviates from the manual Gold Standard by localizing the capitalization to the token “States” while leaving the associated descriptor “member” in lowercase and transcribing “46” as a number instead of a word.

An additional deviation was observed in the rendering of the institutional title "Secretary General of the Council of Europe" (Appendix SEN line 8; 9). The interpreter produced the phrase "the General Secretary" (Appendix MTE3 line 9), resulting in a non-standard word order in English, so the ASR system transcribed "general" (Appendix E3 line 8), in lowercase, keeping the syntactic structure of the speaker. However, there was another instance where "general secretary" was kept in all lowercase (Appendix E3 line 16).

Punctuation errors (41.99 NER) and Syntactic Complexity errors (31.50 NER) are frequently connected to the interpreter's use of pauses. The ASR system frequently treats these pauses as sentence boundaries, resulting in fragmentation, misplaced capitalization, and incomplete clauses. In one instance, a multi-layered error was identified where the ASR system fragments "more regular. Summits." (Appendix E3 line 17) instead of "more regular [su-hh], summits," (Appendix MTE3 line 19) displayed in the Gold Standard. This instance is coded across three distinct error categories: First, the aborted articulatory attempt at the target word "summits" (Appendix MTE3 line 19) is classified a Fluency error. Second, the system omits the NLT "[su-hh]", classified a part-word repetition. Second, auditory analysis reveals that this repetition is followed by a distinct drop in pitch, which is considered a continuation pause requiring a comma (Appendix MTE3 line 19). This deviates from the ASR version exhibiting a full stop and is therefore considered the first punctuation error (substitution). Additionally, ASR inserted a second full stop after the word "Summits." (Appendix E3 line 17), classified as an insertion, resulting two punctuation errors in total for this instance. Finally, a Syntactic Complexity error was identified, because the ASR system inserts a terminal boundary into a cohesive noun phrase, fragmenting a single intended T-unit into two grammatically incomplete segments, misrepresenting the speaker's syntactic structure. In another error-dense instance, the ASR system fragments an introductory clause and deletes a filled pause while retaining a redundant lexical repetition. The speaker's utterance is as follows: "where the powers are diverging, and central powers are, uhh, powers are centralizing," (Appendix MTE3 line 11-12), whereas the ASR output only captures "where. The powers are diverging and Central Powers are powers are centralizing" (Appendix E3 line 10). The filled pause "uhh", classified as a Fluency error, was omitted, whereas "Central Powers", classified a Speech error, was incorrectly capitalized. A similar pattern was observed in the following case: the Gold Standard shows: "said in the, uhh, the planetary meeting" (Appendix MTE3 line 9-10) while the ASR transcription displays "said in the the planetary meeting" (Appendix E3 line 9-10), keeping the speaker's lexical whole-word repetition, while removing the interjection that separates them. However, in another instance,

the conjunction “and” as well as the filled pause “uhh” in “And the, and, uhh, it doesn't matter” (Appendix MTE3 line 4) was omitted, resulting in the ASR version “and the IT. It doesn't matter” (Appendix E3 line 4).

The highest error category in E3 is Fluency Features. The ASR system omitted a total of 26 hesitation sounds such as “uhh” or “uhmm” (Appendix MTE3). These omissions are frequently followed by sentence breaks or syntactic distortion, increasing the overall impact of the error. Notably, there is one single instance of the filled pause “um” (Appendix MTE3 line 1; Appendix E3 line 1) which the ASR system did not omit, matching the Gold Standard.

Table 5.3: ASR error analysis for Experiment 3 (Interpreting Task)

ASR error type	Specific error instance	ASR line	Manual transcript line	Raw count (C)	Normalized error rate
Fluency features	Total			Total 26	68.24
	Omission of filled pause “uhh” before “something”	2	2	1	
	Omission of filled pause “ehh” before “in”	2	2	1	
	Omission of filled pause “uhh” after “in”	2	2	1	
	Omission of filled pause “uhh” before “our”	3	4	1	
	Omission of filled pause “uhh” after “and”	3-4	4	1	
	Omission of filled pause “uhh” after “it’s”	3-4	4	1	
	Omission of filled pause “uhmm” (instance 1) before “helping”	4	5	1	
	Omission of filled pause “uhmm” after “helping” (instance 2)	5	5	1	

	Omission of filled pause “uhmm” (instance 3)	5	5	1	
	Omission of filled pause “uhh”	7	8	1	
	Omission of filled pause “uhmm”	7	8	1	
	Omission of filled pause „uhh“ between repetition of “the”	8	9	1	
	Omission of filled pause between “in” repetition (instance 1)	9	10	1	
	Omission of filled pause “uhh” after “was” (instance 2)	9	10	1	
	Omission of filled pause “uhh” after „his“ (instance 3)	9	10	1	
	Omission of filled pause “uhh” before “powers”	10	11	1	
	Omission of filled pause “uhh” before “take”	11	12	1	
	Omission of filled pause “uhmm” before “help”	12	13	1	
	Omission of filled pause “uhh” before “we’ve been”	13	14	1	
	Omission of filled pause “uhh” before “which”	13-14	15	1	
	Omission of filled pause “uhmm” (instance 1) before “This is what”	16	17	1	
	Omission of filled pause “uhmm” (instance 2) after “This is what”	16	18	1	
	Omission of “[su-hh]” in “summits” (part-word repetition)	17	19	1	

	Omission of filled pause “uhh” before “active”	18	20	1	
	Omission of filled pause “uhh” before “work”	21	23	1	
	Omission of filled pause “uhh” before “these”	22	24	1	
Syntactic complexity				Total 12	31.50
	“we haven't seen before, ehh, in, uhh, our society. And I like to call it the perfect storm.” fragmented into “we haven't seen before. Our society and I like to call it the perfect storm”	2-3	2-3	1	
	“our strength, uhh, our agility, our reaction and our experience. And” fragmented into “our strength. Or agility or reaction and our experience and”	3-4	3-4	1	
	“experience. And the, and, uhh, it doesn't matter” fragmented into “experience and the IT. It doesn't matter”	4	4	1	
	“it's, uhh, about topics such as, uhmm, helping, uhmm, for safer, making migration safe. Uhmm” fragmented into “it’s about topics such as. Helping for safer making migration safe”	4-5	4-5	1	
	Unwarranted full stops and the subsequent capitalization of “In” and “Period”	9	10	1	

	"And these, uhh, take, start taking it over and in times" fragmented into "and these. Take start taking to over and. In times"	10-11	12	1	
	"values and, uhmm, help truth. And" fragmented into "values and. Help truth and"	12	13	1	
	Writes "46" as a number instead of a word	13	15	1	
	Insertion of two full stops around "active" fragmenting the clause into disjointed segments	18	19	1	
	Transcription of "Uhmm. This is what, uhmm, the general secretary said." as a question "This is what? The general secretary said."	16	17-18	1	
	"make more regular [su-hh], summits" fragmented into "make more regular. Summits." (part-word repetition)	17	19	1	
	"It is a goal of his active period, that every human" fragmented into "That is a goal of his. Active. That every human"	18	19-20	1	
Speech error				Total 7	18.37
	Omission of "and"	4	4	1	
	Insertion of "IT"	4	4	1	
	Did not capitalize "general"	8	9	1	
	Capitalized "Central Powers"	10	11	1	

	Capitalized “States”, but not “member”	13	15	1	
	“That” instead of “it” (substitution)	18	20	1	
	Transcription of a full stop instead of the word “period”	18	20	1	
Phonological error				Total 6	15.75
	“OK” instead of “okey”	1	1	1	
	“or” instead of “our” (instance 1 and 2)	3	4	2	
	“They’re” instead of “where”	18	19	1	
	“That” instead of “It”	18	20	1	
	“center” instead of “centre” (regional spelling variant)	19	20	1	
	“birth set” instead of “Berset”	23	24	1	
Self-correction				Total 2	5.25
	Speaker corrects “an” to “a” (replacement); ASR only transcribes replaced version “a”	7	8	1	
	Speaker uses 3 filler sounds “uhh, f-uhh-“ before “period of” (replacement of an aborted attempt), ASR fragments “his. Period”	9	10	1	
Punctuation error				Total 16	41.99
	Full stop instead of comma in “strength. Or agility” (insertion)	3	3		
	Misplaced full stop (instance 1) in “and the IT. It doesn't matter”	4	4		

	Misplaced full stop (instance 2) in “about topics such as. Helping for safer” (insertion)	4	5		
	Missing full stop before “When” (deletion)	5	5		
	2 misplaced full stops in “In. When he was starting his. Period” (insertion)	9	10	2	
	Misplaced full stop in “a world where. The powers are” (insertion)	10	11		
	2 misplaced full stops in “these. Take start taking to over and. In times” (insertion)	11	12	2	
	Misplaced full stop in “our values and. Help truth” (insertion)	12	13		
	Misplaced full stop in “which are. All proud” (insertion)	14	15		
	Wrong question mark in “This is what? The general secretary said” (insertion)	16	17-18		
	2 misplaced full stops in “more regular. Summits.” (substitution and insertion)	17	19	2	
	2 misplaced full stops in “goal of his. Active. That every” (insertion)	18	20	2	
Total errors				69	181.10

5.4: A comparative overview of the results of all three experiments

This chapter offers a comparative overview of the findings that appear relevant for the research questions of the present study. The most relevant discovery in addressing RQ1 is that

interpreted speech (E3) appears to be the most error-prone task between the three, displaying a NER of 181.10, that is almost three times as high as the one calculated for E2 (65.22 NER). E3 surpasses the unmediated spontaneous monologue in E1 (164.47 NER) as well, though the difference in the total error rate is smaller than compared to E2.

E2 displays the lowest NERs across 4 categories (Fluency, Phonological, Syntactic Complexity, and Punctuation errors), however, it reached the highest number in Speech errors across all three tasks (24.14 NER). E1 and E3 both show higher overall error rates, but with distinct profiles. E1 displays the highest Punctuation error rate (74.56 NER), as well as the highest Syntactic Complexity rate (NER 57.02) across all three tasks. In contrast, E3 shows the highest NER for Fluency Features (68.24), which is more than ten times higher than in E1 or E2, as well as the overall highest error category within E3 itself. Phonological errors, with the highest NER across all tasks (15.75), peaked in E3 as well. Even though this is the least frequent error category in E3, it resulted in a higher NER than in both other tasks.

Table 5.4: A comparative overview of the ASR error analysis of E1, E2 and E3

	E1 (Spontaneous) Rate (per 1k words)	E2 (Read-Aloud) Rate (per 1k words)	E3 (Interpreting) Rate (per 1k words)
Fluency features	4.39	2.42	68.24
Phonological errors	8.77	7.25	15.75
Speech errors	19.74	24.14	18.37
Syntactic complexity	57.02	12.08	31.50
Self-correction	0.00	7.25	5.25
Punctuation errors	74.56	12.08	41.99
Total errors	164.47	65.22	181.10

Taken together, the results suggest that, out of the experiments selected for the present study, ASR performance exhibits the highest error rate for interpreted speech, slightly exceeding the error rate observed for non-interpreted spontaneous speech. In contrast, planned read-aloud speech appears to present the least challenging input for ASR systems between the ones selected for this study.

Additionally, a comparison of word counts in manual and ASR transcripts reveals noteworthy differences. E2 reveals the exact same word count across both versions, while the unplanned tasks do not. The ASR version of E1, consisting of 456 words, is longer than the

Gold Standard with 455 words, while the ASR transcript with 381 words in E3 is notably shorter than the corresponding Gold Standard of 407 words. Importantly, the data suggests that interpreted speech (E3) and spontaneous monologue (E1) are not merely different in the volume of errors, but in their qualitative error profiles, which will be discussed in the next chapter.

6 Discussion of findings in ASR performance

While the figures of the previous chapter established the numerical data, they do not explain the possible causes behind the observed ASR performance. To produce a holistic account of how human speech performance interacts with machine recognition and answer the established research questions, the following chapter will analyze the numerical findings and aim to find possible explanations for the data across the three experimental conditions.

6.1 Interpretation of key findings and performance hierarchy in interpreted versus non-interpreted speech

The primary focus of this investigation is the comparative assessment of ASR performance across three distinct modes of non-native oral production. In alignment with the initial hypothesis, interpreted speech (E3) yields the overall highest NER of 181.10, significantly exceeding the planned read-aloud task with the lowest NER of 65.22. This establishes a performance hierarchy wherein the read-aloud task (E2) maintains maximal accuracy, while unplanned tasks (E1 and E3) appear to have triggered significantly higher recognition error rates, indicating that ASR performance varies systematically depending on speech production mode and degree of planning.

The classification of E2 being as planned versus E1 and E3 as unplanned forms of speech (Dufour et al. 2014) represent, aside from mediation versus non-mediation, a primary task divergence that could be a possible explanation for the observed NER hierarchy. Specifically, as established in Chapter 2.4.1, pre-planning allows for greater predictability and a more stable acoustic signal, thereby facilitating ASR processing. Consequently, the high error rates found in E1 and E3 can be explained by previous predictions and observations on specific factors of unplanned discourse, such as repairs, false starts, and irregular prosodic patterns challenging ASR performance (Dufour et al. 2014; Bazillon et al. 2008; Fantinuoli 2017; Sun 2023).

Additionally, the overall low error rates in E2 could be explained by its condition of maximal predictability and reduced cognitive load, suggesting a predictive power of structured input. This reinforces the assumption that ASR performance increases in non-spontaneous, more fluent speech (Fantinuoli 2017; Bazillon et al. 2008; Sun 2023). Despite being delivered by a non-native speaker and containing specialized terminology, the consistency of the acoustic signal, largely free from pauses, self-repairs, and syntactic restructuring, allowed the ASR system to process the input efficiently. However, these error rates should not be interpreted as reflecting ASR performance in isolation. Since the NERs only capture deviations from the manual transcript, speaker-produced disfluencies that were accurately transcribed by the system are not represented as errors, however, they play an important role in the analysis of results, which is why they have been displayed in Chapter 5 as well. The lower error rate in E2 may therefore be attributed to the relative absence of such disfluencies in the input itself, whereas the higher error rates in E1 and E3 could partly result from a greater presence of disfluencies such as filled pauses, repetitions and other features unknown to the system, introducing more opportunities for deviations between the ASR version and the Gold Standard.

Importantly, the high error rates in the unplanned tasks occurred despite participants exhibiting speaker characteristics generally favorable to ASR systems. All participants in this study are young adults producing non-native English, with, based on the researcher's auditory assessment, a standard accent. With ASR systems performing best for speakers whose characteristics match those in the training data (Feng et al. 2024), it can be assumed that the key differences in ASR performance observed in this study were not triggered by age, demographic or regional-accent biases. Instead, they appear to be qualitatively shaped by the task-specific transcription challenges, specifically providing a possible explanation for the highest error rate of E3 (191.60 NER).

Moreover, when analyzing the identified error instances displayed in the tables (s. Ch. 5), it is important to relate the numbers to the speaker-produced errors that were accurately transcribed as well. For instance, within Self-correction errors (NER 0.00 in E1; NER 7.25 in E2; NER 5.25 in E3), speakers produced numerous instances that match the ASR transcript. In E1, there are two accurately transcribed instances of Self-correction, where the speaker's clear articulation and the lack of NLTs likely contributed to the match between ASR and manual transcript. The same circumstances can be applied for the one instance of a Self-correction error in E1 (Appendix E2 line 19) that was accurately transcribed, with 3 replacements of part-word repetitions that are counted as errors. They were likely omitted by

the system because they were followed by the corrected version immediately. In contrast, most Self-correction error instances in E3 include the omission a filled pause, resulting in a total of eight Self-Correction instances made by the speaker, of which only two are counted as an error. Similarly, in E1, the Fluency Feature “Ehh” (Appendix E1 line 2) was deleted, while “Um” (Appendix E1 line 1) and “uh” (Appendix E1 line 2) was transcribed. It is noteworthy that “Ehh” is immediately followed by “erosion”, containing the same starting vowel, which is likely why the system treated it as a redundant part-word repetition, deleting it in favor of readability.

Moreover, phonological errors must be interpreted in relation to cognitive and psychological load. The most phonological errors across all tasks were identified in E3 (NER 15.75) even though the same ASR system processed a phonologically comparable but cognitively less demanding input in E2. The planned nature of E2 confirms that non-native accent alone does not account for elevated phonological error rates. Qualitative observations revealed that the interpreter’s anxiety and lack of confidence about English as her B-language likely created a high-stakes psychological environment and therefore affected pronunciation stability during the task. This anxiety was manifested even prior to the onset of speech production, as the participant was observed making “sounds of despair” (Appendix MTE3) while listening to the task instructions.

Beyond the task-specific findings, challenges with domain-specific terminology, proper names, and non-native pronunciation, materialized consistently across tasks. The deletion of “ChatGPT” (Appendix MTE1 line 5) and misrecognition of “Alain Berset” (Appendix E2 line 7; Appendix E3 line 8)) confirm prior findings that general-purpose ASR struggles with institutional vocabulary, especially OOV items (Zapata & Søeborg Kirkedal 2015: 208). Specifically, the omission of “ChatGPT” (Appendix MTE1 line 5) makes the ASR version “Artificial intelligence is like. And it's.” (Appendix E1 line 5) incomprehensible, as the full stop replaces the term of comparison that the word “like” refers to. Moreover, the ASR system inconsistently capitalized the institutional title “Secretary General” (Appendix SEN line 8; 9). In E3, it capitalized “Secretary”, but not “general” (Appendix E3 line 8), while in E2, both words are displayed in lowercase in one instance (Appendix E2 line 8) and in upper case in another (Appendix E2 line 9). In the instance where it capitalized “Secretary” in E3, the broader context of the utterance “the general Secretary of the Council of Europe” (Appendix E3 line 8), points to a possible explanation on why ASR recognized it as an institutional title, which is the presence of the institutional title “Council of Europe” (Appendix MTE3 line 9) in the same sentence. Similarly, in E2, the article “The” in

“The Secretary General” could have led the system to capitalize both. In contrast, the lowercase writing in the instance “in his in his new capacity of secretary secretary general” (Appendix E3 line 8) could be explained by the combination of the speaker’s repetition, as well as the lack of an article and the missing connection to the institution “Council of Europe”. Additionally, the orthographic fluctuation of the British version of **“**centre” in E2 (appendix E3 line 23) versus the American version “center” (Appendix E3 line 19) align with prior observations that ASR tools frequently transcribe the same word differently when processing speech that diverges from their training norms (s. Ch. 2.3). Notably, the system utilized the version “centre” in E2 (Appendix E2 line 23), despite the language setting English (United States) was selected for all experiments.

Finally, the differences in ASR processing across conditions are reflected in the word counts of the transcripts. E2 shows identical word counts between the manual and ASR-transcripts, even though there is an insertion (Appendix E2 line 5), which could be explained by the system transcribing “we'll” (Appendix E2 line 25), instead of “I will” (Appendix MTE2 line 26). While “I will” is counted as 2 words, “we'll” is counted as one, which could have made up for the insertion of “and”, resulting the restoration of the original word count of the speech that was being read aloud. On the other hand, both E1 and E3 exhibit noticeable discrepancies. Specifically, E1 deviates in one word only, which could be explained by the following instances: While the system's naturalization bias omitted the filled pause “Ehh” (Appendix MTE1 line 2) and the term “ChatGPT” (Appendix MTE1 line 5), it simultaneously performed the unwarranted lexical addition, of “it” (Appendix MTE1 line 2; Appendix E1 line 2). Additionally, the ASR system attempted to normalize the token “inter institutional” (Appendix E1 line 12) by separating it into two words, likely caused by the speakers perseverance “intern-institutional” (Appendix MTE1 line 12). Similarly, it turned “decide” (Appendix MTE1 line 16), counted as one word in the Gold Standard into “this side” (Appendix E1 line 16-17), counted as 2 in the ASR version. In contrast, the ASR transcript of E3 is 26 words shorter than the Gold Standard, likely caused by the frequent omissions of filled pauses and fragmentation (s. Ch. 5.3).

This indicates that in the more controlled speech condition (E2), the ASR system remains largely faithful to the linear structure of the source utterance. However, errors still occurred, indicating that pre-planned speech alone does not fully mitigate ASR limitations when challenging items are involved. Therefore, ASR performance does not appear to be strictly task-dependent, but rather structural, which becomes particularly evident when considering the principle of naturalized transcription. In contrast, the deviations observed in

E1 and particularly in E3 indicate that the system does not merely transcribe but actively reshapes the spoken input. These discrepancies can largely be attributed to the omission of disfluencies such as filled pauses, the compression of repair sequences, and occasional duplication effects. As such, differences in word count provide further evidence that ASR performance is not only a matter of recognition accuracy but also of structural transformation, reinforcing the view that the system functions as an active co-producer rather than a neutral transcription tool.

6.2 The structural fragmentation of E1 versus adaptive interference in E3

This subchapter contrasts the observed errors between the two unplanned tasks, presenting potential explanations for the erratic structural fragmentation of the unmediated monologue and the machine's interference with the non-standard features of the interpreter.

As established in chapter 2.4.1, speaking is already a complex task that requires cognitive effort (Levelt 1998: 1). Therefore, it can be assumed that spontaneous content generation involves even more cognitive effort than planned speech, however the differences between interpreted and non-interpreted speech must be considered. In E1, cognitive load could be attributed to spontaneous content generation, whereas in CI, there appear to be additional demands of bilingual reformulation, note-processing, and time pressure.

The total error rates for the two unplanned tasks (164.47 NER for E1; 181.10 NER for E3), which are significantly less apart than any of them compared to E2 (NER 65.22), were likely caused by different underlying error mechanisms. Specifically, the overall highest proportional failure in Punctuation (74.56 NER) and Syntactic Complexity (57.02 NER) in E1 could be explained by the speaker's unmediated, irregular prosody, resulting from cognitive burden of real-time content generation. An instance of structural instability can be seen in the speaker's sequence "And Artificial intelligence is like ChatGPT and it's basically taking over everything, like, it could take over everything or most of it" (Appendix MTE1 line 5-7). The system fragmented it into the disjointed segments "artificial intelligence is like. And it's. Basically taking over everything like. It could take over everything or most of it" (Appendix E1 line 5-7). The speaker uses multiple embedded T-units, resulting in a quantitatively long sentence with multiple commas, which the system does not represent accurately. The sentence structure as well as auditory analysis suggest that the speaker paused around the word "like", pointing to the process of his mental decision-making, which could have additionally influenced the systems misrecognition.

Additionally, qualitative observations confirm that the participant appeared overwhelmed (Appendix MTE1) by the sudden introduction of specialized political terms and expressed significant anxiety regarding his linguistic proficiency. This state of uncertainty likely induced the frequent, non-standard acoustic gaps that the ASR system frequently misinterpreted as terminal sentence boundaries, resulting in syntactic fragmentation.

In contrast, most observed errors in the interpreting task (E3) likely occurred because it was challenging for the system to process certain the non-conventional features specific to this type of speech. For instance, the interpreter made use of numerous filled pauses where the system frequently appears to have inserted full stops as grammatical endpoints, fragmenting otherwise continuous clauses. This can be seen in segments such as “show our strength, uhh, our agility, our reaction and our experience.” (Appendix MTE3 line 4) that was transcribed as “how our strength. Or agility or reaction and our experience” (Appendix E3 line 3-4) by the ASR system. The interpreter’s filled pause “uhh” likely contributed to the fragmentation as it disrupts standard syntax. As observed in Table 5.3, the interpreter frequently utilizes similar filled pauses, which, as auditory analysis suggests, was the case when she could not find the right word and needed time to think, resulting in a total of 26 Fluency errors. Additionally, the repetition of “our” was not articulated clearly, which is what likely caused the system to substitute it with “or”.

When the interpreter performed self-repair, the ASR system did not suppress the redundant information, resulting in sequential transcription of both erroneous and corrected segments. For instance, “Central Powers are powers are centralizing” (Appendix E3 line 10) and repeated clause restarts illustrate how ASR processes acoustic input literally without semantic filtering in some cases. According to auditory analysis, in these instances of reformulation, the articulation was loud and clear. However, in another identified instance, the conjunction “and” as well as the filled pause “uhh” in “And the, and, uhh, it doesn't matter” (Appendix MTE3 line 4) was omitted, resulting in the ASR version “and the IT. It doesn't matter” (Appendix E3 line 4). This discrepancy can be attributed to acoustic ambiguity caused by the interpreter’s disfluent and hesitant speech production. The system likely expected a noun after the preceding words “and the”, which likely resulted in the capitalized “IT”.

The ASR system’s limited accommodation of these features might have led to the highest rates of Fluency Feature errors (68.24) which is particularly relevant for RQ1, as it suggests that interpreting does not merely reproduce the challenges of spontaneous speech but intensifies them through the interaction of multiple linguistic and cognitive pressures. The combination of non-native speech production, domain-specific terminology, time pressure,

and reformulation leads to a constellation of vulnerabilities that exceed those observed in unmediated spontaneous discourse.

However, similar cases can also be observed in E1, as seen in Examples such as “is is,” (Appendix MTE1, line 3-4) or “people get people get,” (Appendix MTE1, line 8), where the system transcribes both the initial and the corrected version. In other cases, the ASR system deletes an erroneous form and retains only the corrected version, transcribing “a” (Appendix E3 line 7) while omitting the initial form “an” (Appendix MTE1 line 8), suggesting an inconsistency in suppression mechanisms.

Moreover, even though zero Self-correction errors have been identified for E1, the system indeed produced multiple accurate transcriptions of repetitions such as “people get people get” (Appendix MTE1 line 8) or “our our” (Appendix MTE1 line 24). This could have occurred because, as auditory analysis indicates, the repetitions that the system did not omit, appear to be acoustically clear, following standard prosodic patterns. During another self-correction process, the speaker appears uncertain whether the correct plural form of incident is “incidents” (Appendix MTE1 line 10; Appendix E1 line 10-11), likely repeating a second token to fix a perceived error in the first. Therefore, this instance can be considered as a functional repair rather than a redundant fluency marker, though ASR transcribed both versions literally. Auditory analysis further points to the speaker’s fluctuations in voice and pitch in this scenario, specifically, as perceived by the researcher, it went down after the speaker first said “incidents” (Appendix MTE1 line 10), sounding like the end of a sentence. The participant appears to have decreased his intonation again after the second instance “Incidences” (Appendix MTE1 line 10). The ASR system did not only transcribe the phonological instances the way they appear in the Gold Standard, but the punctuation markers as well. However, in another instance, the ASR system treated the phrase “that's. Very hypocritical, no?” (Appendix E1, line 4) as a rhetorical question, inserting a question mark, even though auditory analysis suggests that the speaker did not raise his intonation indicating a question. Instead, he appears to have paused to search for an alternative lexical choice to the word “hypocritical” (Appendix MTE1 line 4) before moving on to the next segment, while he appears to have said “no” to himself, indicating that his word choice was wrong. This algorithmic insertion resulted in a terminal sentence boundary that was likely not intended in the speaker’s intended discourse. These results raise the question of the ASR system detects sentence boundaries from syntax or imposes them based on prosodic cues such as falling pitch. The little acoustic weight that punctuation marks hold might be a possible explanation on why the system treats gaps or drops in intonation as sentence boundaries

Furthermore, the contrast between E1 and E3 must be viewed considering the different speech production demands and speaker backgrounds. The interpreter in E3 had several years of formal linguistic training in both German and English, whereas the speaker in E1 reported limited exposure to English and lower confidence in language use, which was ultimately confirmed by auditory analysis as well. However, the interpreter did, at the time of the experiment, not have any experience in professional settings, and was still learning the efficient use of note-taking techniques and other strategies typically used for CI. This difference in linguistic background likely contributed to the error rates observed in E1 and E3 as well and should be considered when interpreting the results. Ultimately, E3 exhibiting the overall highest error rates suggests that trainee interpreter output poses an even greater challenge for ASR than spontaneous non-interpreted speech.

6.3 The naturalization bias of the ASR system

The findings of this study show that, as predicted, the ASR system frequently appears to have altered the transcription process to follow the principle of naturalized transcription (Bucholtz 2000: 1461). Therefore, this chapter aims to find explanations and demonstrate the instances where the system appears to prioritize a written-style language over a literal phonetic representation of the spoken input. The analysis of the results points to this naturalization bias not only in the interpreting task, but in all three ASR transcriptions.

First, a possible explanation for the suppressed hesitation sounds, filled pauses and anything the system is likely not trained to process, might be the classification of content as either Primary or Secondary Information (Gile 2009: 59). Additionally, when such acoustic cues occurred at structurally sensitive points, misclassification frequently led to fragmentation. The resulting error rates in E1, specifically Punctuation with a NER of 74.56 and Syntactic Complexity with a NER of 57.02; as well as Punctuation with 21.99 and Syntactic Complexity with a NER of 31.50 in E3, suggest that when human cognitive effort becomes audible, the ASR system treated it as noise rather than Primary Information. However, the Punctuation and Syntactic Complexity errors in E1 must be viewed in light of the low Fluency Feature error rate (4.39). On the one hand, the ASR system appears to have tried achieving naturalization by simply deleting some acoustic markers of hesitation (Appendix MTE1line 2; 25). A possible explanation for this is that ASR systems frequently struggle to divide suitable T-units when segmenting speech (Agmon et al. 2024: 556). On the other hand, lexical substitutions such as rendering the abbreviation “gonna” as “going to” (Appendix MTE1 line 5; Appendix E1 line 5) and grammatical substitutions, such as “there

is" becoming "there was" (Appendix MTE1 line 22; Appendix E1 line 22) illustrate how the system actively reshapes the spoken input.

Moreover, the ASR system substituted a comma with a question mark in the sequence "that's. Very hypocritical, no?" (Appendix E1 line 4). As established in Chapter 6.2, the speaker's token "no" (Appendix MTE3 line 4) appears to refer to his lexical choice. The fact that ended his sentence with the word "whatever" (Appendix MTE3 line 4), as well as qualitative observations noting that he appeared to be disappointed about not finding the lexical choice he would have preferred, imply that he might have thought he will not find a better lexical alternative at that point, leading him to move on with his monologue. The speaker's utterance "that's very hypocritical, no, whatever" could be seen as a deviation from standard written style syntax, which is likely why the ASR system prematurely inserted a question mark before the sentence was finished. ASR systems predict boundaries based on linguistic probability (Chen et al. 2015: 237), specifically, in this instance, the system seems to have prioritized the high statistical probability of "no" appearing as a tag question, thereby hallucinating a terminal boundary that contradicts the acoustic evidence of the speaker's intention.

Qualitative observations further suggest that specific pronunciation issues led to transcription deviations. The omission of the article "a" in "a lot of people" (Appendix E1, line 8; Appendix MTE1, line 8) likely occurred because the speaker pronounced it very quietly, indicating that the detection of unstressed function words, was challenging for the ASR system. The speaker's quiet rendering could be associated with the non-native speaker insecurity observed during the task. Similarly, the grammatical substitution of "are" with "were" (Appendix E1, line 9; Appendix MTE1, line 9) likely resulted from the speaker swallowing the word rather than articulating it in what could be considered a more native-like manner. These instances also highlight that phonological precision alone did not guarantee transcription accuracy in non-native contexts.

In contrast, the Fluency error rate in E3 (68.24 NER), which is more than ten times higher than in E1 or E2, mostly consists of the deletion of filled pauses such as "uhh" or "uhmm" (Appendix MTE3). This can be seen in the sequence "of his, uhh, active period, that" (Appendix MTE3 line 20), where the ASR system did not only delete the token "uhh", but also misinterpreted the core lexical item "period" (Appendix MTE3 line 20) as a literal formatting command, outputting a full stop instead (Appendix E3 line 18). This incident could have been triggered by the hesitation sound "uhh" (Appendix MTE3 line 20), and, as auditory analysis suggests, the speaker's pauses and unclear articulation around the word

“active” (Appendix MTE3 line 20). Combined with the distinct form of interpreting that the system likely was not trained on, this scenario might have led to the lack of taking the surrounding syntactic context into account. Finally, the system appears to have prioritized the probabilistic weight of a formatting command over the word’s lexical meaning, creating a syntactically impossible terminal boundary after an adjective. In another instance, the system deleted the aborted articulatory attempt “[su-hh]” (Appendix MTE3 line 19) before the word “summits”. Although the speaker produced a part-word repetition, likely triggered by high cognitive load, the ASR system did not attempt to transcribe this non-standard segment. Instead, it appears to have suppressed the sound entirely to prioritize written-style content.

However, while the ASR system frequently omitted disfluencies such as filled pauses in the interpreting task, there is only one Fluency error in E2, namely the insertion of the conjunction “and” in the segment “issues such as combating migrant smuggling and Environmental Protection” (Appendix E2 line 5). It was likely caused because the ASR system took the introductory phrase “such as” as an expectation for a coordinator between listed terms. Consequently, the internal language model seems to have prioritized the probability of a formal written list over the literal acoustic signal. Auditory analysis confirms that the speaker produced a minor acoustic gap at the time of the inserted word, suggesting that when ASR encountered at a highly predictable signal, its internal language model may have come up with the insertion as a strategy to create a transition and therefore make the syntax resemble a formal written language.

Finally, this study identified a structural bias of the technology itself, indicating that it does not merely transcribe speech but actively edits it in line with internal language-model expectations, and its difficulty with non-standard speech.

6.4 Cognitive saturation in CI and its influence on ASR transcription

This discussion analyses the correlation between the cognitive strain in E3, illustrating how the specifics of the trainee interpreter's performance appear to have influenced the acoustic signal. First, it is worth mentioning that in some instances, the interpreter seems to have relied on source-language syntax, aligning with previous observations that high cognitive load in CI can reduce the capacity for syntactic reorganization, leading to increased instances of interference from the source language (Zhao et al. 2023: 2-4). This can be observed by an example where the interpreter appears to have struggled to abandon the sentence structure of the original text, formulating “in a world where the powers are diverging and central powers are, uhh, powers are centralizing, and these, uhh, take, start taking to over” (Appendix E3 line

10-12)", closely mimicing the German structure "In einer Welt, die Zentrifugalkräften ausgesetzt ist, in einer Welt, in der divergierende Kräfte derzeit häufig die Oberhand zu gewinnen scheinen" (Appendix SEN line 16-17). Here, the German version of "to" was translated word for word, even though the English version "taking over" would not require an infinitive marker. Moreover, in the sequence "the General Secretary of the Council of Europe said" (Appendix MTE3 line 9), the interpreter produced a structurally non-target-like formulation by rendering the German title "Generalsekretär" (Appendix SGER line 8;11) as "General Secretary" (Appendix MTE3 line 9), instead of the conventional English form "Secretary General" (Appendix SEN line 8; 9). The linear transfer of source-language word order could be explained by the cognitive saturation of the interpreter, who may have, because of real-time content generation constraints, prioritized rapid lexical retrieval over target-language restructuring.

The high error rates observed in E3 (s. Ch. 5.4) could also partially be grounded in the training stage of the participant. The disproportionately high frequency of Fluency Feature omissions (68.24 NER) and eight instances of Self-correction (s. Ch. 6.5 for an elaboration on why only two of these have been coded as errors) suggests that the trainee was navigating the complex isolation of key semantic markers and partitioning the source discourse into functional segments to facilitate note-supported reconstruction**.** Nevertheless, it is important to consider that the professional interpreting should be an ongoing learning process (s. Ch. 2.1). Moreover, the reliance on handwritten notes during the task may have further contributed to reformulation and self-repair, as attention was split between the reconstruction of meaning from handwritten notes and target-language production, which can be viewed as non-automatic operations that approach the interpreter's processing capacity limits (s. Ch. 2.1). As indicated by qualitative observations, the participant relied on a previously prepared binder of A5 custom-cut papers for her notes. This organizational strategy appears to have supported the interpreter's cognitive load management, allowing her to structure the source information into manageable chunks and anticipate potential reformulations. However, when she could not read her notes properly, it resulted in confusion, unclear articulation and repetitions such as "in the, uhh, the planetary meeting" (Appendix E3 line 9-10). Throughout the whole task, she appears to have utilized many symbols and arrows instead of full words, which suggests that she has practiced specific note-taking techniques.

Taken together, this chapter illustrated that the elevated error rates in E3 cannot be attributed to isolated ASR recognition failures but rather emerge from the interaction between the interpreter's cognitive load, the structural properties of interpreted speech, and the

system's limitations in processing non-linear input. In particular, the reliance on source-language syntax, the fragmentation of discourse into chunked, note-based segments, and the increased presence of self-repair and hesitation phenomena collectively appear to have shaped the acoustic signal.

6.5 Error classification rationale: Delineating ASR-deviations from the Gold Standard

Documenting the human behaviors in interpreted as well as non-interpreted speech through machine recognition presented a significant analytical challenge. Even though parameters and analysis strategies have been predefined, the identification and categorization of errors is nevertheless inevitably based on subjective decisions by the **r**esearcher. Since many previous studies lack transparency in the regard of coding logic (Polio 1997: 129) and the present study might display instances that appear ambiguous, this chapter aims to make the decisions in error counting as explicit as possible.

Starting with E1, the ASR system inconsistently capitalized the sequence "Heads of state and government" (Appendix E1 line 21). "Heads of State" and "Government" are capitalized in the Gold Standard in accordance with the original speech (Appendix SEN line 21) where the phrase appears to function as a standardized institutional designation used by organizations like the Council of Europe. The deviations of lowercase versions in the ASR versions are counted as errors, whereas the matching capitalized version "Heads" was not.

The production of "intern-institutional" (Appendix MTE1 line 12) is categorized as a phonological perseverance because a phonetic feature from the preceding lexical unit "institution" (Appendix MTE1 line 12) appears to have persisted into a subsequent production unit (Fromkin 1971: 30). The nasal consonant "n" likely remained active in the articulatory plan, causing the speaker to erroneously carry that sound over into the prefix "inter-".

In another instance, the system appears to have fluctuated between regional standards. While the ASR system transcribed the British version "centre" in E2 (appendix E3 line 23), it applied its American configuration in E3 (Appendix E3 line 19) by outputting "center". The Gold Standard displays the British version "centre" (Appendix MTE3 line 20), because this version is used in the original speech (Appendix SEN line 23). Because of the inconsistency, both versions are counted as an error, even though one version matches the Gold Standard.

Moreover, a single recognition failure frequently triggered multiple types of disruption, forcing a decision on whether to double-penalize the system. For instance, in E2, the misplaced full stop in the phrase "formed by. Our values" was coded as both a punctuation error and a Syntactic complexity error. This dual categorization is grounded in

the findings of Agmon et al. (2024), who conclude that when ASR utterance boundaries are inaccurate, every other measure of linguistic complexity is negatively affected. Because these misplaced full stops represent a failure in algorithmic SBD while simultaneously causing fragmentation of the discourse, recording both the cause and the structural consequence was essential to capture the full impact on transcript utility.

The maintenance of a consistent coding logic was further challenged when the system produced the sequence: “Dialogue is a paramount. Importance A dialogue” error (Appendix E2 line 16), indicating an internal desynchronization within the ASR architecture. The system did not transcribe a full stop after “Importance” (a Punctuation error) yet nevertheless triggered the capitalization of the following word “A”. A possible explanation is that the system's language model treated this as a sentence start without executing the corresponding visual cue. Finally, this instance includes orthographic inconsistency and structural failure, though it was excluded from the error count to avoid redundant over-penalization, treating the capitalization as a direct mechanical consequence of the segmentation failure already recorded.

As displayed in Chapter 5.3, only two of approximately eight instances of Self-correction were counted as errors. In the two coded instances, the system was perceived to fail representing the speaker’s Self-correction accurately. This can be seen in the sequence “when he was, uhh, starting his, uhh, f-uhh-period” (Appendix MTE3 line 10) that resulted in the ASR version “starting his. Period” (Appendix E3 line 9), losing the representation of the speaker’s replacement of an aborted attempt. In the other instance, the speaker replaces “an” with “a” (Appendix MTE3 line 8) and ASR only transcribes corrected version “a” (Appendix MTE3 line 8). The reason why this is counted as an error is that the system appears to have recognized that the speaker was performing an act of Self-correction, and actively omitted the wrongly perceived initial token, therefore deviating from the Gold Standard. In the other instances that were not counted as Self-correction errors, the aim was to avoid over-penalizing, as the corresponding omissions of filled pauses are already coded as a Fluency error, whereas the fragmentation is counted in the Syntactic Complexity category.

Moreover, the interpreter produced an interjection perceived phonetically as “okey” Appendix MTE3 line 1) at the beginning of her delivery, likely intending to signal that she was ready. The ASR system transcribed this as the standard capital “OK” (appendix E3 line 1), which appears to be the systems naturalized form (s. Ch. 2.3), likely resulting from its preference for formal written conventions. Because it actively erased the phonetic shape of

the spoken utterance, which deviates from the Gold Standard this instance is counted as an error.

Notably, there are numerous identified instances where the study does not over-penalize formatting byproducts that result from a primary algorithmic failure. This can be seen in the segment “We’ve had, uhh, we’ve been” (Appendix MTE3 line 14), which was transcribed as “We've had we've been” (Appendix E3 line 13). While the Gold Standard includes commas to set off the interpreter’s filled pause, the ASR system’s omission of these marks was not recorded as a separate punctuation error. Because the suppression of the NLT “uhh” was already counted as a Fluency Feature error, counting the missing commas would have over-represented the error frequency without adding analytical depth. Instead, this instance serves as a microanalytic example of how the machine’s naturalization bias removes the audible marker of cognitive effort, automatically dissolving the syntactic boundaries associated with it.

In another instance, the omission of “[su-hh]” (Appendix MTE3 line 19) in “more regular. Summits.” (Appendix E3 line 17) is classified as a Fluency Feature error because the system likely considered the speaker’s anticipation as “Secondary Information” (Gile 2009). This likely misled the system’s internal language model, which misinterpreted the resulting acoustic gap as a terminal boundary, triggering the unwarranted full stop and fragmentation observed in the ASR transcript.

Collectively, the systematic identification of errors turned out to be a highly iterative process, necessitating a recursive analytical loop between the manual transcripts and ASR outputs. Following the initial categorization and the preliminary calculation NERs, subsequent microanalytic reviews frequently uncovered more phonological or structural deviations that were initially overlooked by the researcher. This required a comprehensive revision of the error tables and a recalculation of the total NERs in over ten separate instances to ensure statistical fidelity. Having established these specific coding challenges, the final section addresses the broader methodological considerations and limitations that shaped the study's design as well as recommendations for future research.

6.6 Limitations and future research suggestions for ASR in interpreting contexts

Experimental designs in Translation and Interpreting Studies often aim to or are limited to generalize findings to a limited professional population and must therefore report research in a transparent way, with analytically rich rather than aggregate data and metrics (Mellinger & Hanson 2016: 137). Despite aiming to provide maximal possible transparency given limited

resources, the study is subject to several limitations. First, the small sample of the present research, consisting of one person per experiment, restricts generalizability.

Furthermore, the speech experiments relied on an internal laptop microphone in a domestic setting, mirroring authentic, non-laboratory environments. However, it is possible that the documented error rates in this research were exacerbated by the hardware's limited capacity to capture the nuances of the non-native speech signal. As Alharbi et al. (2021) and Koenecke et al. (2020) suggest, ASR performance is often sensitive to recording quality, and mediocre microphones can significantly reduce the accuracy of the acoustic model's ability to parse rhythm and pitch.

Another limitation was found in the live monitoring of the experiments. It became apparent that complete observational neutrality is difficult to maintain when the researcher is present during task performance. The live ASR transcription was considered very distracting, and it appeared tempting to not display facial expressions of surprise when transcription behavior was unexpected, which is why the desktop was finally partially closed. Rather than attempting to eliminate the influence of the researcher entirely, observable behavioral cues were noted selectively when they appeared relevant for understanding specific error patterns.

Moreover, an important observation emerged during the pilot study, when automatic punctuation was disabled, initially displaying the ASR output as a single, uninterrupted block of text. The decision to enable automatic punctuation for the main experiment, while necessary for readability and identification of Punctuation and Syntactic Complexity errors, introduced a critical confounding variable. The resulting high punctuation NERs could be a direct consequence of the ASR system's structural imposition interacting with the speaker's numerous filled pauses.

On another note, while some dictation workflows may require speakers to deliberately vocalize formatting commands (Zapata & Søeborg Kirkedal 2015), the participants in this research were not asked to provide such explicit cues. Instead, the experimental design requires them to maintain a natural delivery style, which could present a significant variable in system performance, as it forces the ASR architecture to algorithmically infer boundaries from prosodic and linguistic cues alone.

Moreover, it should be noted that E3 was conducted under relatively controlled conditions. The task was a simulation, and the interpreter was not exposed to the psychological pressure of a real conference setting in which listeners rely entirely on the interpretation. In addition, the researcher delivered the prepared source text with the intention of speaking slowly and clearly. However, because this assessment is based solely on the

researcher's own perception and was not independently verified, it must be treated as a methodological limitation rather than an objective claim. This suggests that error rates in non-simulated interpreting settings, where speech may be faster and less controlled, could be even higher. These observations point to the relevance of studying more complex interpreting modes, such as relay interpreting, where the input itself is already an interpretation and may introduce additional layers of error.

It must be acknowledged that the production of the manual transcription as well as the error analysis of this study inevitably involve an element of researcher judgment, since neither was validated externally[2]. Specifically, the systematic identification of errors and the subsequent mathematical calculations for NERs were not verified by an independent second rater. While independent verification should be a standard practice to mitigate subjective bias and ensure reliability (Polio 1997: 130; Sun 2023; Zhao et al. 2023), the resource constraints of this project necessitated a single-rater approach. To account for these limitations and provide a basis for the replicability of future studies, Chapter 4.5 explains the transcription process of the Gold Standard, while Chapter 6.5 elaborates decisions in error coding. Additionally, all audio files of the experiments can be accessed openly through the link provided.

Finally, these limitations can be considered an inspiration for several directions of future research rather than to shortcomings of the present study alone. Future research may expand participant samples and systematically compare multiple ASR systems to improve the generalizability of the findings, as the resources of this study only allowed for the examination of one ASR system, suggesting that the findings cannot be assumed to apply equally to all. Another consideration would be the systematic application of APR (outlined in Chapter 2.5) for the transcription of interpreted output, to determine whether automated grammatical repair can mitigate punctuation and Syntactic complexity errors generated by interpreter chunking.

Another relevant direction for future research concerns variation in interpreter experience**.** Future studies could compare the ASR-processed output of trainee and professional practicing interpreters with experience. This would allow researchers to examine whether high Self-correction instances decrease as interpreters gain strategic control. At the

[2] Disclaimer: The researcher hereby acknowledges the utilization of AI for basic grammar checks and formulation support as English is not their first language.

same time, it must be acknowledged that interpreter competence develops gradually and requires ongoing practice, which complicates attempts to model interpreter output as a stable and uniform input type for ASR. Moreover, comparative studies across different interpreting modes, such as simultaneous or relay interpreting, could contribute to determining whether the challenges identified in the present study are specific to CI or characteristic of interpreted speech more generally.

Another area of exploration for future studies could be the application of the framework of content-based versus form-based error analysis (s. Ch. 2.6). While individual omissions of NLTs may carry limited semantic weight, their cumulative effect may be more significant, particularly in non-native speech. The ASR system appears to prioritize clearly articulated lexical items over grammatical function words, suggesting an internal hierarchy that favors acoustic salience over grammatical completeness. Future research could therefore investigate if or how the omission of NLTs affects meaning transfer and functional adequacy, rather than focusing solely on phonological deviation. Finally, after acknowledging identified limitations and suggesting areas for future research, the final chapter will present a conclusion of the main findings of this study.

7 Conclusion

This thesis investigated ASR performance when transcribing non-native English speech across different production modes, specifically comparing unmediated spontaneous speech (E1), controlled read-aloud speech (E2) and CI from German to English (E3). In doing so, it addressed two central research questions asking what key differences between can be found within the ASR transcriptions of these speech modes and how they are influenced by Fluency Features, Syntactic Complexity, Speech errors, Phonological errors, Self-correction and Punctuation errors.

The findings confirm the hypothesis that the read-aloud task yields the lowest error rates (65.22 NER), reflecting its planned, more text-like nature, which aligns closest with the type of data ASR systems are typically trained on between the three tasks. In contrast, both non-interpreted spontaneous speech (NER 164.47) and CI (NER 181.10) resulted in significantly higher error rates, though for different underlying reasons. E1 is characterized by structural instability, leading to the highest error rate in Punctuation (NER 74.56). The CI task exhibits the highest overall NERs in two of six categories, yet the disproportionately high Fluency Feature error rate (NER 68.24) contributed to the overall highest NER across all

tasks (compared to a NER 4.39 in E1; 2.42 in E2). This error profile appears to be dominated by omissions of filled pauses such as “uhh” (Appendix MTE3 line 2), except for one deleted part-word repetition “[su-hh]” (Appendix MTE3 line 19), suggesting that the cognitive effort of simultaneous reconstruction of notes, target language formulation and time pressure destabilized the internal language model.

By prioritizing readability and written-style coherence, the system appears to have suppressed numerous NLTs and non-standard formulations across tasks, reshaping the transcription toward a naturalized version. This is particularly evident where it turned the informal abbreviation “gonna” (Appendix MTE1 line 5) into “going to” (Appendix E1 line 5). In some instances, the system disregarded the surrounding syntactic context, for instance, when substituting the lexical item “period” (Appendix MTE3 line 20) for a full stop, treating it as a formatting command. These observations are crucial for RQ1 as this naturalization bias actively reshaped the output in all tasks.

However, all error rates should be viewed in the light of how many instances of potential ASR errors are produced from the speakers, and how many of those were captured accurately. For instance, the NER of 0.00 Self-correction in E1 does not mean that there are no instances of Self-correction made by the speaker, it only displays that the identified instances of the ASR version match the Gold Standard. The speaker, in fact, attempted two instances of Self-correction, however, the clear articulation in these instances likely led to no ASR errors. In E2, the three penalized errors are exclusively deletions of part-word repetitions such as “pri-” before “priorities” (Appendix MTE2 line 11), where the ASR system likely transcribed the corrected version only because it immediately follows the percieved redundant part-word. However, clearly articulated full-word Self-corrections such as “he had, he added” (Appendix E2 line 19) were accurately captured.

Moreover, significant differences in word counts throughout the tasks have been identified. E2 displays the same word count for the ASR version as well as the Gold Standard, even though the ASR version displays an insertion and the separation of a token from 2 to 1 word. In contrast, the word count of the unplanned tasks shows discrepancies: E1 deviates only minimally, which can be attributed to two omissions and an addition. By comparison, E3 exhibits a substantially larger divergence, reflecting the cumulative impact of deletions of filled pauses, reformulations, and structural fragmentation.

Overall, the findings of this study show that, as hypothesized, the ASR system normalized or omitted disfluencies in spontaneous and interpreted speech. The overall high error rates across all tasks in Punctuation and Syntactic Complexity confirm the hypothesis

that the system misinterprets filled pauses as sentence boundaries. Importantly, the specific Fluency and Self-correction instances in E3 highlight the need to consider interpreting output as a distinct and complex speech type that should not be treated as equivalent to either spontaneous non-mediated or planned speech. Finally, the present study carries implications for both research and practice, as it shows that ASR performance is not solely determined by acoustic input but also influenced by the linguistic and cognitive properties of speech production.

9 Bibliography

Agmon, Galit; Pradhan, Sameer; Ash, Sharon; Nevler, Naomi; Liberman, Mark; Grossman, Murray & Cho, Sunghye (2024). Automated Measures of Syntactic Complexity in Natural Speech Production: Older and Younger Adults as a Case Study. *Journal of Speech, Language, and Hearing Research* 67 (2), 545–561. DOI: 10.1044/2023_JSLHR-23-00009.

Alharbi, Sadeen; Alrazgan, Muna; Alrashed, Alanoud; Alnomasi, Turkiayh; Almojel, Raghad; Alharbi, Rimah; Alharbi, Saja; Alturki, Sahar; Alshehri, Fatimah; & Almojil, Maha (2021). *Automatic Speech Recognition: Systematic Literature Review.* IEEE Access, 9, 131858–131876. DOI: 10.1109/ACCESS.2021.3112535.

Baumgarten, Stefan & Bourgadel, Carole (2024). Digitalisation, neo-Taylorism and translation in the 2020s. *Perspectives* 32 (3), 508–523. DOI: 10.1080/0907676X.2023.2285844.

Bazillon, Thierry; Esteve, Yannick & Luzzati, Daniel (2008). Manual vs assisted transcription of prepared and spontaneous speech. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2008/pdf/277_paper.pdf

Berset, Alain (2024)**.** *Secretary General stresses the Council of Europe's crucial role during the Assembly's autumn plenary session* [Speech]. Council of Europe. Retrieved January, 2025, from https://www.coe.int/de/web/portal/full-news/-/asset_publisher/y5xQt7QdunzT/content/id/272920028?_com_liferay_asset_publisher_web_portlet_AssetPublisherPortlet_INSTANCE_y5xQt7QdunzT_languageId=en_GB#p_com_liferay_asset_publisher_web_portlet_AssetPublisherPortlet_INSTANCE_y5xQt7QdunzT

Bóna, Judit & Bakti, Mária (2020). The effect of cognitive load on temporal and disfluency patterns of speech: Evidence from consecutive interpreting and sight translation. *Target. International Journal of Translation Studies* 32 (3), 482–506. DOI: 10.1075/target.19041.bon.

Bosker, Hans Rutger; Reinisch, Eva & Sjerps, Matthias J. (2017). Cognitive load makes speech sound fast, but does not modulate acoustic context effects. *Journal of Memory and Language*, *94*, 166–176. DOI: 10.1016/j.jml.2016.12.002.

Bucholtz, Mary (2000). The politics of transcription. *Journal of Pragmatics*, *32*(10), 1439–1465. DOI: 10.1016/S0378-2166(99)00094-6.

Chen, Jinying; Cao, Huaigu & Natarajan, Premkumar (2015). Integrating natural language processing with image document analysis: what we learned from two real-world applications. *International Journal on Document Analysis and Recognition (IJDAR)* 18 (3), 235–247. DOI: 10.1007/s10032-015-0247-x.

Davidson, Christina (2009). Transcription: Imperatives for Qualitative Research. *International Journal of Qualitative Methods* 8 (2), 35–52. DOI: 10.1177/160940690900800206.

Davitti, Elena; Sandrelli, Annalisa; Korybski, Tomasz; Zou, Yuan; Orasan, Constantin; Braun, Sabine (2024). Using ASR tools to produce automatic subtitles for TV broadcasting: A cross-linguistic comparative analysis. *Journal of Audiovisual Translation* 7 (2), 1-35. DOI: https://doi.org/10.47476/jat.v7i2.2024.305

De Vos, Hugo & Verberne, Suzan (2020). Challenges of Applying Automatic Speech Recognition for Transcribing EU Parliament Committee Meetings: A Pilot Study. In: Fišer, Darja, Eskevich, Maria, & de Jong, Franciska (Eds.), *Proceedings of the Second ParlaCLARIN Workshop* (pp. 40–43). Marseille, France: European Language Resources Association. https://aclanthology.org/2020.parlaclarin-1.8/.

Dufour, Richard; Estève, Yannick & Deléglise, Paul (2014). Characterizing and detecting spontaneous speech: Application to speaker role recognition. *Speech Communication*, *56*, 1–18. DOI: 10.1016/j.specom.2013.07.007.

Fantinuoli, Claudio (2017). Speech recognition in the interpreter workstation. *Proceedings of the Translating and the Computer 39*, 25–34. DOI: 10.1075/target.19166.def.

Feng, Siyuan; Halpern, Bence Mark; Kudina, Olya; Scharenborg, Odette (2024). Towards inclusive automatic speech recognition. ComputerSpeechandLanguage, 84, Article 101567. DOI: https://doi.org/10.1016/j.csl.2023.101567.

Fu, Xue-Yong; Chen, Cheng; Laskar, Md Tahmid Rahman; Bhushan, Shashi & Corston-Oliver, Simon (2021). Improving Punctuation Restoration for Speech Transcripts via External Data. In: Xu, Wei, Ritter, Alan, Baldwin, Tim, & Rahimi, Afshin (Eds.), *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)* (pp. 168–174). Online: Association for Computational Linguistics. DOI: 10.18653/v1/2021.wnut-1.19.

Gile, Daniel (2009). Basic concepts and models for interpreter and translator training. In: *Basic Concepts and Models for Interpreter and Translator Training*. John Benjamins Publishing Company.

Garrett, Merrill F. (1975). The Analysis of Sentence Production. In: *Psychology of Learning and Motivation*. Elsevier. Vol. 9. 133–177. DOI: 10.1016/S0079-7421(08)60270-4.

Guo, Meng & Han, Lili (2024). From manual to machine: Evaluating automated ear–voice span measurement in simultaneous interpreting. *Interpreting. International Journal of Research and Practice in Interpreting*,*26* (1), 24–54. DOI: https://doi.org/10.1075/intp.00100.guo

Greene, Jennifer C. (2006). Toward a Methodology of Mixed Methods Social Inquiry. *Re search in the Schools*, *13*(1), 93-99.

Housen, Alex (2012). *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins Publishing Company.

Hulstijn, Jan H. & Rick de Graaff (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review*, *11*, 97–112.

Karakasidis, Georgios; Kurimo, Mikko; Bell, Peter & Grósz, Tamás (2024). *Comparison and analysis of new curriculum criteria for end-to-end ASR.* Speech Communication, 163, Article 103113. Elsevier B.V. DOI: https://doi.org/10.1016/j.specom.2024.103113

Koenecke, Allison; Nam, Andrew; Lake, Emily; Nudell, Joe; Quartey, Minnie; Mengesha, Zion; Toups, Connor; Rickford, John R.; Jurafsky, Dan & Goel, Sharad (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117 (14), 7684–7689. DOI: 10.1073/pnas.1915768117.

Kuckartz, Udo (2014). *Mixed Methods: Methodologie, Forschungsdesigns und Analyseverfah ren*. Wiesbaden: Springer VS.

Kuhn, Korbinian; Kersken, Verena; Reuter, Benedikt; Egger, Niklas & Zimmermann, Gottfried (2023). Measuring the Accuracy of Automatic Speech Recognition

Solutions. *ACM Transactions on Accessible Computing* 16 (4), 1–23. DOI: https://doi.org/10.1145/3636513

Levelt, Willem J. M. (1998). *Speaking: from intention to articulation*. Cambridge, Mass.: MIT Pr. 1. MIT Press paperback ed. 1993, 5. printing.

Liao, Junwei; Eskimez, Sefik Emre; Lu, Liyang; Shi, Yu; Gong, Ming; Shou, Linjun; Qu, Hong & Zeng, Michael (2020). Improving Readability for Automatic Speech Recognition Transcription. arXiv. DOI: 10.48550/arXiv.2004.04438.

Liu, Jiajun; Wumaier, Aishan; Fan, Cong & Guo, Shen (2023). Automatic Fluency Assessment Method for Spontaneous Speech without Reference Text. *Electronics* 12 (8), 1775. DOI: 10.3390/electronics12081775.

Malik, Mishaim; Malik, Muhammad Kamran; Mehmood, Khawar & Makhdoom, Imran (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications* 80 (6), 9411–9457. DOI: 10.1007/s11042-020-10073-7.

McMillan, Corey T. & Corley, Martin (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition* 117 (3), 243–260. DOI: 10.1016/j.cognition.2010.08.019.

Mellinger, Christopher D. & Hanson, Thomas A. (2016) *Quantitative research methods in translation and interpreting studies.* London: Routledge. https://doi-org.uaccess.univie.ac.at/10.4324/9781315647845

Ortega, Lourdes. (2003). Syntactic complexity measures and their relationship to L2 profi ciency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24 (4), 492–518. DOI: 10.1093/applin/24.4.492

Pentland, Steven J.; Fuller, Christie M.; Spitzley, Lee A. & Twitchell, Douglas P. (2023). Does accuracy matter? Methodological considerations when using automated speech-to-text for social science research. *International Journal of Social Research Methodology* 26 (6), 661–677. DOI: https://doi.org/10.1080/13645579.2022.2087849.

Pisani, Elisabetta & Fantinuoli, Claudio (2021). Measuring the impact of automatic speech recognition on number rendition in simultaneous interpreting. In: Wang, Caiwen & Zheng, Binghan (eds.) *Empirical Studies of Translation and Interpreting*, New York: Routledge, 181-197. DOI: 10.4324/9781003017400-14.

Pöchhacker, Franz (2016). Introducing Interpreting Studies (2nd ed.). Routledge. https://doi-org.uaccess.univie.ac.at/10.4324/9781315649573

Point, Sébastien & Baruch, Yehuda (2023). (Re)thinking transcription strategies: Current challenges and future research directions. *Scandinavian Journal of Management* 39 (2), 101272. DOI: https://doi.org/10.1016/j.scaman.2023.101272.

Polio, Charlene G. (1997). Measures of Linguistic Accuracy in Second Language Writing Research. *Language Learning*,*47* (1), 101–143. DOI: 10.1111/0023-8333.31997003.

Prandi, Bianca (2020). The use of CAI tools in interpreters' training: a pilot study. In: Proceed ings of the Translating and the Computer 37 Conference, London. http://www.intralinea.org/specials/article/2512

Russell, Sam O'Connor; Gessinger, Iona; Krason, Anna; Vigliocco, Gabriella & Harte, Naomi (2024). What automatic speech recognition can and cannot do for conversational speech transcription. *Research Methods in Applied Linguistics* 3 (3), 100163. DOI: 10.1016/j.rmal.2024.100163.

Setton, Robin (1999). *Simultaneous Interpretation*. Amsterdam: John Benjamins. DOI: 10.1075/btl.28.

Setton, Robin & Dawrant, Andrew (2016). *Conference interpreting: a trainer's guide*. Amsterdam: John Benjamins.

Shen, Mingxia & Liang, Junying (2021). Self-repair in consecutive interpreting: Similarities and differences between professional interpreters and student interpreters. *Perspectives: Studies in Translatology* 29 (5), 761–777. DOI: 10.1080/0907676X.2019.1701052.

Sun, Weina (2023). The impact of automatic speech recognition technology on second lan guage pronunciation and speaking skills of EFL learners: a mixed methods investigation. *Frontiers in Psychology* 14, 1210187. DOI: 10.3389/fpsyg.2023.1210187.

Tardel, Anke (2020). Effort in semi-automatized subtitling processes: Speech recognition and experience during transcription. *Journal of Audiovisual Translation* 3 (1), 79–102. https://doi.org/10.47476/jat.v3i2.2020.131.

Tammasrisawat, Pannapat & Rangponsumrit, Nunghatai (2023). The Use of ASR-CAI Tools and their Impact on Interpreters' Performance during Simultaneous Interpretation. *New Voices in Translation Studies* 28. DOI: 10.14456/NVTS.2023.27.

Tran, Brian D.; Latif, Kareem; Reynolds, Tera L.; Park, Jihyun; Elston Lafata, Jennifer; Tai-Seale, Ming & Zheng, Kai (2023). "Mm-hm," "Uh-uh": are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology? *Journal of*

the American Medical Informatics Association 30 (4), 703–711. DOI: 10.1093/jamia/ocad001.

Wang, Caiwen, & Zheng, Binghan (Eds.) (2021). *Empirical studies of translation and inter preting: The post-structuralist approach*. Abingdon, New York: Routledge. https://doi.org/10.4324/9781003017400

Wang, Xiaoman & Fantinuoli, Claudio (2024). *Exploring the Correlation between Human and Machine Evaluation of Simultaneous Speech Translation*. DOI: 10.48550/arxiv.2406.10091.

Xu, Cui & Li, Dechao (2024). More *spoken* or more *translated*?: Exploring the known unknowns of simultaneous interpreting from a multidimensional analysis perspective. *Target. International Journal of Translation Studies* 36 (3), 445–480. DOI: https://doi.org/10.1075/target.22028.x

Zapata, Julián & Søeborg Kirkedal, Andreas (2015). *Assessing the Performance of Automatic Speech Recognition Systems When Used by Native and Non-Native Speakers of Three Major Languages in Dictation Workflows*. In: Megyesi, Beáta (Ed.) *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. Vilnius, Lithuania: Linköping University Electronic Press, Sweden. 201–210. https://aclanthology.org/W15-1825/.

Zhao, Nan; Cai, Zhenguang G. & Dong, Yanping (2023). Speech errors in consecutive interpreting: Effects of language proficiency, working memory, and anxiety. *PLOS ONE* 18 (10), e0292718. DOI: 10.1371/journal.pone.0292718.

Zheng, Wei-Zhong; Han, Ji-Yan; Cheng, Hsiu-Lien; Chu, Wei-Chung; Chen, Ko-Chiang & Lai, Ying-Hui (2023). Comparing the performance of classic voice-driven assistive systems for dysarthric speech. Biomedical Signal Processing and Control,81, 104447. DOI: 10.1016/j.bspc.2022.104447.

10 Appendices

Appendix PS1 (ASR transcript of E1 Pilot study)

Today I had a talk with my friend about climate change and then he told me some things about artificial intelligence and the erosion of democracy and it was it's these are really intense topics and don't know much about them and yes it's really difficult for me to find words right now but yes I am glad I can participate paid in this project and yeah I'm really curious about what will be to come out of this and realizing right now my English is not really good thank you

Researcher notes:

- P seems very nervous and mentions her English skills are very rusty

Appendix MTPS1 (manual transcription of E1 Pilot study)

Today I had a talk with my friend about climate change and then he told me some things about artificial intelligence and the erosion of democracy. And it was, uhh, these are really intense topics and I don't know much about them. And yes, it's really difficult for me to find words right now, but yes, I am glad I can participate in this project. And yeah, I'm really curious about what will be to come out of this. And I'm realizing right now, my English is not really good. Thank you.

Researcher notes:

- participant immediately expresses concern not knowing a lot about the terms provided
- researcher clarifies that the quality of the output itself is not what will be measured, because its only a pilot study to account for any potential problems that may arise during the actual experiment
- Participant agrees to proceed but is very worried because she didn't speak proper English for a while and has mainly been exposed to the language through social media

Appendix PS2 (ASR transcript of E2 Pilot study)

The Secretary General. And phases the crucial role of the Council of Europe, as it celebrated its 75th anniversary in guaranteeing a future of peace, stability, prosperity, security and dignity for Europe's citizens. He also highlighted the three priorities of his mandate, support for Ukraine's action to. Through vitalized democracy and upholding the unity of the European family in all its diversity, in a world that is plagued by centrifugal forces, where the divergent movements seem to be gaining the upper hand at the moment, we must work our way back to stronger convergence and to achieve that convergence. You have to stick to our truth and values. Dialogue is of paramount importance. A dialogue between the Council of Europe's 46 members, which are just described as importance proud equals. It is an inter institutional dialogue, both formal and informal. With the Council of Europe and it's a dialogue that respects diversity but is also deeply anchored in the foundations formed by our values, he added.

Researcher notes:

- P expresses relief that she must only read now
- seems very insecure, almost ashamed about her English

Appendix MTPS2 (manual transcription of E2 Pilot study)

The Secretary General emphases the crucial role of the Council of Europe, as it celebrated its seventy-fifth anniversary in guaranteeing a future of peace, stability, prosperity, security and dignity for Europe's citizens. He also highlighted the three priorities of his mandate, support for Ukraine's action to- through vitalized democracy and upholding the unity of the European family in all its diversity, in a world that is plagued by centrifugal forces, where the divergent movements seem to be gaining the upper hand at the moment, we must work our way back to stronger convergence and to achieve that convergence you have to stick to our truth and values. Dialogue is of paramount importance. A dialogue between the Council of Europe's forty-six members, which are just described as importance proud equals. It is an interinstitutional dialogue, both formal and informal. With the Council of Europe and it's a dialogue that respects diversity but is also deeply anchored in the foundations formed by our values, he added.

Appendix PS3 (ASR transcript of E3 Pilot study)

OK, Hmm. Watch whether you face global challenges and it's. Most important to. Pursuit together in our in our world with a lot of convergency we have to remain on the course of the request that we had before and it's important to continue our dialogue between 46 member states of our Union and Council was. To play a major role in this election.

Appendix MTPS3 (manual transcription of E3 Pilot study)

Mhm. Umm. In our world, where we face global challenges and it's very, its mostly important to pursue it together. In our in our world with a lot of convergency we have to remain on the course of the request that we have before. And it's important to continue our dialogue between 46 member states of our union and Council wants to play a major role in this direction.

Researcher notes:

- P expresses concern, since English is his C-language and he dislikes CI
- first working language is Ukrainian and his second German
- heavy Ukrainian accent
- after finishing, participant asks if this was it, expresses that the source speech was complex

Appendix E1 (experimental step 1): ASR transcription of Spontaneous Speech Task

Climate change is a topic that I think about. More often than others, probably because. Um. It affects everyone and. Me personally, I'm looking for my uh. How can this say it? Erosion footprint, no CO2 footprint I guess, and a lot of other people don't even. Think that climate change is is real or exists? So yeah, that's. Very hypocritical, no? Whatever. Um, I don't really know what erosion of democracy means, so I'm going to. Skip that. And artificial intelligence is like. And it's. Basically taking over everything like. It could take over everything or most of it. Most of the things I guess. And that's a bit scarier, but yeah. That's the change in the world. Umm. I think migrant smuggling is. Not a good thing because. A lot of people get people get hurt when they. Get smuggled like like there were a lot of dead people or missing people. A lot of humans don't have another choice but to to flee. But it's it's tragic. I mean, there are often tragic incidents. Incidences. Um. I also don't really know anything about the Council of Europe except that it is a. Political. Institution. And that's basically all I know. Um, into the inter institutional dialogue is also something that I have no idea what the what is this or. World. I could say about it. So yeah. Um, every human has values and they can also change over time. And most of the time the humans around you also change because of your values because some people don't change with you and you have to decide if you want to keep. Them in your life or not? So it's their freedom to. This side. Who they want to, uh, have in their life and, uh, who they. Gonna like cancel the friendship alarm. But I don't see them anymore. But yeah, umm. Yeah, well, the thing about security, the first thing that comes to my mind is my dad, because he worked as a security guard in bars for like 10 to 15 years. And I also remember the the times he wasn't at home. In all those years because he had to work and sleep at day, at work at night, but. Yeah, he's not doing that anymore. And I think Environmental Protection is also a very. Important, um, manner because. There was there was a lot to protect, and more people should. Uh should not look away. When they see stuff that do this. That that destroys our our environment. But a lot of people do because they don't want to see it and. They feel guilty, but they still don't act on their guilt. But yeah. That's that I guess.

Appendix MTE1 (manual transcription of E1)

Climate change is a topic that I think about more often than others, probably because, um, it affects everyone and me personally, I'm looking for my uh. How can this say? Ehh, erosion footprint, no CO2 footprint I guess, and a lot of other people don't even think that climate change is is real or exists, so, yeah, that's very hypocritical, no, whatever. Um, I don't really know what erosion of democracy means, so I'm gonna skip that. And artificial intelligence is like ChatGPT and it's basically taking over everything, like, it could take over everything or most of it. Most of the things I guess. And that's a bit scary, but yeah. That's the change in the world. Umm. I think migrant smuggling is not a good thing, because a lot of people get people get hurt when they get smuggled, like like there are a lot of dead people or missing people. A lot of humans don't have another choice but to to flee, but it's it's tragic. I mean, there are often tragic incidents. Incidences. Um. I also don't really know anything about the Council of Europe except that it is a political institution and that's basically all I know. Um, intern-institutional dialogue is also something that I have no idea what that, what is this or what I could say about it, so, yeah. Um, every human has values and they can also change over time and most of the time the humans around you also change because of your values because some people don't change with you and you have to decide if you want to keep them in your life or not. So it's their freedom to decide who they want to b-, uh, have in your life and, uh, who they gonna like cancel the friendship or like they don't see them anymore, but yeah. Umm. Yeah, when, I think about security, the first thing that comes to my mind is my Dad, because he worked as a security guard in bars for like ten to fifteen years and I also remember the the times he wasn't at home in all those years because he had to work and sleep at day, and work at night, but yeah, he's not doing that anymore. And I think environmental protection is also a very important, um, matter because there is, there is a lot to protect, and more people should, uhh, should not look away when they see stuff that this that that destroys our our environment. But a lot of people do because they don't want to see it and they feel guilty, but they s- still don't act on their guilt, but yeah. That's that I guess.

Researcher notes:

- Participant and researcher discussed the basic frame before starting
- P looks a little startled, researcher elaborates that it does not have to be a perfect monologue, and it would absolutely be acceptable if there were mistakes
- P expresses concern about speaking English

- P looks at terminology before starting the task and shakes his head and laughs because he does not know a lot of the words provided
- P struggles to find the right word for what he then says as “hypocritical”, seems confused and uncomfortable, then shakes head and ends with “whatever”, seems disappointed that he did not find the word he was thinking of
- After task 1, the participant does not want to take a break but instead continue right away with E2

Appendix E2 (experimental step 2): ASR transcription of Read-aloud Task

The combined effects of these wars. Together with climate change and the erosion of democracy, constitution and unprecedented challenge for our societies, the challenge I think we could call the perfect storm in the factor of the challenge that. The Council of Europe has already shown its solidity, its great experience, its agility and its capacity to respond. Build on issues such as combating migrant smuggling and Environmental Protection, or artificial intelligence, to name but a few, the Council of Europe has often shown the way. Our new Convention on Artificial Intelligence that was recently opened for signature is a perfect example. Said Ellen Perset, speaking for the first time in his in his new capacity of secretary secretary general at the autumn plenary session of the PACE. The Secretary General emphasizes the crucial role of the Council of Europe, as it celebrated its 75th anniversary, in guaranteeing a future of peace, stability, prosperity, security and dignity for European citizens. He also highlighted the three priorities of his mandate, support for Ukraine's action to revitalize democracy and upholding the unity of the European family in all its diversity. In the world that is plugged by centrifugal forces, where divergent movements seem to be gaining the upper hand at the moment, we must work our way back to stronger convergence. And to achieve that convergence, we have to stick our truth to our truth and values. Dialogue is a paramount. Importance A dialogue between the Council of Europe's 47 members, which I just described as proud equals. It is an interinstitutional dialogue, both formal and informal, within the Council of Europe, and it is a dialogue that respects diversity but is also deeply anchored in the foundations formed by. Our values he had, he added. He also mentioned an action plan for promoting democracy in order to revitalize the democratic process, as well as the possibility of organizing a summit of Heads of state and government on a regular basis. And then Berset concluded by saying during my terms of office, we'll work with passion and determination to continue building a Europe in which every human being is at the centre of our efforts, with the ultimate aim that everyone may live in the freedom and security. That democracy brings. I am very much looking forward to what we can accomplish together, and we'll put all my energy into working for that mission and for our shared values into serving the Council of Europe.

Researcher notes:

- rushed into the task because he was in a hurry
- seems relieved that it's over and asked if everything worked out

Appendix MTE2 (manual transcription of E2)

The combined effects of these wars, together with climate change and the erosion of democracy, constitution and unprecedented challenge for our societies, a challenge I think we could call the perfect storm in the fact of the challenge that the Council of Europe has already shown its solidity, its great experience, its agility and its capacity to respond. Built on issues such as combating migrant smuggling, environmental protection, or artificial intelligence, to name but a few. The Council of Europe has often shown the way. Our new Convention on artificial intelligence that was recently opened for signature is a perfect example. Said Ellen Berset, speaking for the first time in his in his new capacity of Secretary Secretary General at the autumn plenary session of the PACE. The secretary general emphasizes the crucial role of the Council of Europe, as it celebrated its seventy-fifth anniversary, in guaranteeing a future of peace, stability, prosperity, security and dignity for Europe's citn-citizens. He also highlighted the three pri-priorities of his mandate, support for Ukraine action to revitalize democracy and upholding the unity of the European family in all its diversity. In the world that is plugged by centrifugal forces, where divergent movements seem to be gaining the upper hand at the moment, we must work our way back to stronger convergence. And to achieve that convergence, we have to stick our truth to our truth and values. Dialogue is a paramount importance. A dialogue between the Council of Europe's forty-seven members, which I just described as proud equals. It is an interinstitutional dialogue, both formal and informal, within the Council of Europe. And it is a dialogue that respects diversity but is also deeply anchored in the foundations formed by our values he had, he added. He also mentioned an action plan for promoting democracy in order to revitalize the democratic process, as well as the possibility of organizing a summit of Heads of State and Gover-Government on a regular basis. And then Berset concluded by saying during my terms of office, I will work with passion and determination to continue building a Europe in which every human being is at the centre of our efforts, with the ultimate aim that everyone may live in the freedom and security that democracy brings. I am very much looking forward to what we can accomplish together, and I will put all my energy into working for that mission and for our shared values into serving the Council of Europe.

Researcher notes:

- P wants to start right away after E1 no break, expresses relief about completing the task, sighs and says it took longer than anticipated
- P misreads "he added", looks closer at the text, then corrects to "added", looks confused, but continues

Appendix E3 (experimental step 3): ASR transcription of the Interpreting Task

Yes. OK, um. The combination of the effects of war, climate change and the erosion of democracy is something of which we haven't seen before. Our society and I like to call it the perfect storm. We, as the Council of Europe, have been able to show our strength. Or agility or reaction and our experience and the IT. It doesn't matter if it's about topics such as. Helping for safer making migration safe When it comes to climate change or environmental safety or artificial intelligence, we've always we as the Council of Europe have always been a leading the way, showing the way. Our new agreement about AI is a great example for this. And we are going into the right direction. This is what the general Secretary of the Council of Europe said in the the planetary meeting in autumn. In. When he was starting his. Period of in his function. We are living in a world where. The powers are diverging and Central Powers are powers are centralizing, and these. Take start taking to over and. In times like these, we need more convergence so that we can increase our values and. Help truth and what's important to improve the situation is communication and dialogue. We've had we've been talking with like with all 46 member States and which are. All proud and equal partners of these dialogues. These dialogues are interinstitutional and not only formal but also informal, and they are all about our diversity and about our fundamental birth of democracy. This is what? The general secretary said. Our action plans are trying to strengthen democracy, and he's also planning on trying to make more regular. Summits. They're all the Heads of State can meet. That is a goal of his. Active. That every human is in the center of the actions of the Council of Europe and that every human being can experience the safety and freedom which democracy has to offer. And I'm looking forward and I'm glad to what we can achieve in this time and I'm happy that I can put my work and efforts into the work and efforts into the Council of Europe. And I'm looking forward to this. This is what these were the closing words of birth set.

Comment: The following part of the original speech was left out: Der Generalsekretär unterstrich die entscheidende Rolle des Europarates – der sein 75-jähriges Bestehen feiert –, indem er den europäischen Bürgerinnen und Bürgern eine Zukunft in Frieden, Stabilität, Wohlstand, Sicherheit und Würde garantiert. Es wurden auch die drei Prioritäten seiner Amtszeit hervorgehoben: Unterstützung für die Ukraine, Maßnahmen zur Neubelebung der Demokratie und Förderung der Einheit der europäischen Familie in all ihrer Vielfalt.

Appendix MTE3 (manual transcription of E3)

Yes, okey. Umm. The combination of the effects of war, climate change and the erosion of democracy is, uhh, something of which we haven't seen before, ehh, in, uhh, our society. And I like to call it the perfect storm. We, as the Council of Europe, have been able to show our strength, uhh, our agility, our reaction and our experience. And the, and, uhh, it doesn't matter if it's, uhh, about topics such as, uhmm, helping, uhmm, for safer, making migration safe, uhmm, when it comes to climate change or environmental safety or artificial intelligence. We've always, we as the Council of Europe have always been, leading the way, showing the way. Our new agreement, uhh, about AI is an-a great example for this. And, uhmm, we are going into the right direction. This is what the General Secretary of the Council of Europe said in the, uhh, the planetary meeting in autumn, uhh, in, when he was, uhh, starting his, uhh, f-uhh-period, of in his function. We are living in a world where the powers are diverging and central powers are, uhh, powers are centralizing, and these, uhh, take, start taking to over and in times like these, we need more convergence so that we can increase our values and, uhmm, help truth and what's important to improve the situation is communication and dialogue. We've had, uhh, we've been talking with like, with all forty-six member states and, uhh, which are all proud and equal partners of these dialogues. These dialogues are interinstitutional and not only formal but also informal, and they are all about our diversity and about our fundamental birth of democracy. Uhmm. This is what, uhmm, the General Secretary said. Our action plans are trying to strengthen democracy, and he's also planning on trying to make more regular [su-hh], summits where all the Heads of State can meet. It is a goal of his, uhh, active period, that every human is in the centre of the actions of the Council of Europe and that every human being can experience the safety and freedom which democracy has to offer. And I'm looking forward and I'm glad to what we can achieve in this time. And I'm happy that I can put my work and efforts into the, uhh, work and efforts into the Council of Europe. And I'm looking forward to this. This is what, uhh, these were the closing words of Berset.

Researcher notes:

- seems visibly nervous and makes sounds of despair before even starting, Researcher provided some difficult terms beforehand (Erosion der Demokratie)
- P asks how to spell the name “Berset” before starting, because not sure
- has a major in conference interpreting, German as an A and English as a B-language
- often displays pronunciation issues during interpreting, swallowing the ends of words, though good note strategy with many arrows and symbols, fewer full words

Appendix SGER (speech: German version)

„Die kombinierten Auswirkungen der Kriege, zusammen mit dem Klimawandel und der demokratischen Erosion, sind eine beispiellose Herausforderung für unsere Gesellschaften. Eine Herausforderung, die ich als *the perfect storm* zu beschreiben versucht habe. Angesichts dieser Herausforderung hat der Europarat bereits seine Stärke, seine große Erfahrung, seine Agilität und seine Reaktionsfähigkeit unter Beweis gestellt. Ob es um die Bekämpfung der Schleusung von Migranten, den Umweltschutz oder die künstliche Intelligenz geht – der Europarat hat oft den Weg gewiesen. Unser neues Übereinkommen über künstliche Intelligenz, das vor Kurzem zur Zeichnung aufgelegt wurde, ist ein perfektes Beispiel dafür“, erklärte der Generalsekretär des Europarates, Alain Berset, als er zum ersten Mal in seiner neuen Funktion bei der Herbstplenarsitzung der Parlamentarischen Versammlung des Europarates eine Rede hielt. Der Generalsekretär unterstrich die entscheidende Rolle des Europarates – der sein 75-jähriges Bestehen feiert –, indem er den europäischen Bürgerinnen und Bürgern eine Zukunft in Frieden, Stabilität, Wohlstand, Sicherheit und Würde garantiert. Es wurden auch die drei Prioritäten seiner Amtszeit hervorgehoben: Unterstützung für die Ukraine, Maßnahmen zur Neubelebung der Demokratie und Förderung der Einheit der europäischen Familie in all ihrer Vielfalt. „In einer Welt, die Zentrifugalkräften ausgesetzt ist, in einer Welt, in der divergierende Kräfte derzeit häufig die Oberhand zu gewinnen scheinen, muss wieder mehr Konvergenz erreicht werden. Und um mehr Konvergenz zu erreichen, müssen wir den Kurs der Wahrheit und der Werte beibehalten. Der Dialog ist von größter Bedeutung. Ein Dialog zwischen den 46 Mitgliedsstaaten des Europarates, die ich zuvor als stolz und gleich beschrieben habe. Es ist ein interinstitutioneller Dialog innerhalb des Europarates, sowohl formell als auch informell. Und es ist ein Dialog, der die Vielfalt achtet, aber auch grundlegend im Fundament unserer Werte verankert ist“, erklärte er. Ein Aktionsplan zur Förderung der Demokratie, der darauf abzielt, die Kraft des demokratischen Prozesses neu zu beleben, kam ebenfalls zur Sprache sowie die Möglichkeit, regelmäßig ein Gipfeltreffen der Staats- und Regierungschefinnen und -chefs zu veranstalten. „Während meiner Amtszeit werde ich mit Leidenschaft und Entschlossenheit daran arbeiten, weiterhin ein Europa aufzubauen, in dem jeder Mensch im Zentrum unseres Handelns steht, wobei das letztliche Ziel darin besteht, dass jede und jeder in der Freiheit und Sicherheit leben kann, welche die Demokratie bietet. Ich freue mich auf das, was wir gemeinsam erreichen werden, und ich beabsichtige, meine ganze Energie in den Dienst unseres Auftrags und unserer gemeinsamen Werte, in den Dienst des Europarates zu stellen“, so Berset abschließend. (Berset 2024)

Appendix SEN (speech: English version)

"The combined effects of these wars, together with climate change and the erosion of democracy, constitute an unprecedented challenge for our societies, a challenge I think we could call the *perfect storm*. In the face of that challenge, the Council of Europe has already shown its solidity, its great experience, its agility and its capacity to respond. Be it on issues such as combating migrant smuggling, environmental protection or artificial intelligence to name but a few**,** the Council of Europe has often shown the way. Our new Convention on artificial intelligence that was recently opened for signature is a perfect example," said Alain Berset speaking for the first time in his new capacity of Secretary General at the autumn plenary session of the PACE . The Secretary General emphasised the crucial role of the Council of Europe as it celebrated its 75th anniversary, in guaranteeing a future of peace, stability, prosperity, security and dignity for Europe's citizens. He also highlighted the three priorities of his mandate: support for Ukraine, action to revitalise democracy and upholding the unity of the European family in all its diversity. "In a world that is plagued by centrifugal forces, where divergent movements seem to be gaining the upper hand at the moment, we must work our way back to stronger convergence. And to achieve that convergence, we have to stick to our truth and values. Dialogue is of paramount importance. A dialogue between the Council of Europe's 46 members, which I just described as proud equals. It is an interinstitutional dialogue, both formal and informal, within the Council of Europe. And it is a dialogue that respects diversity but is also deeply anchored in the foundations formed by our values," he added. He also mentioned an action plan for promoting democracy in order to revitalise the democratic process, as well as the possibility of organising a Summit of Heads of State and Government on a regular basis. Alain Berset concluded by saying "During my term of office, I will work with passion and determination to continue building a Europe in which every human being is at the centre of our efforts, with the ultimate aim that everyone may live in the freedom and security that democracy brings. I am very much looking forward to what we can accomplish together and I will put all my energy into working for that mission and for our shared values, into serving the Council of Europe." (Berset 2024)