Speaking like Machines:
Examining the Effects of AI-influenced Language on Interpreting Students

Final Masters Degree Thesis by Ezra W. Enns

Faculty of Social Sciences and Humanities

Department of Translation and Interpreting and Multilingual Communication

Máster Universitario en Interpretación de Conferencias

Comillas Pontifical University

Director of Masters Degree Thesis: Alessio Ghirlanda

May 22, 2026

Abstract

The overwhelming amount of text written by generative AI tools has caused shifts in human language in certain contexts. As human language incorporates typical Large Language Model language features, interpreters may need to adapt to new difficulties. In this paper, previous studies from the fields of linguistics provided a framework for understanding what effects AI is having on language, and an overview of previous approaches to assessing interpreting quality guided the design of an experiment to test interpreting students' ability to react to these language changes. In a mixed methods experiment, Masters in Conference Interpreting students from Comillas Pontifical University interpreted speeches, some of which contained AI-consistent features, and provided responses in a questionnaire. Together with rubric-based evaluations of the interpretations, the quality of the interpretations was compared to the source material difficulty. The results showed that AI-consistent features in a speech corresponded to lower quality of interpretation in dense speeches, but higher quality in less dense speeches. Future studies would be able to further investigate the mechanisms that make these AI-consistent features have a negative impact on interpretation.

Keywords: Large Language Models, simultaneous interpreting, informational density, difficulty, readability, source material

La abrumadora cantidad de texto escrito por herramientas de IA generativa ha provocado cambios en el lenguaje humano en determinados contextos. Mientras el lenguaje humano incorpora características de lenguaje típicas de modelos extensos de lenguaje, puede que los intérpretes necesiten adaptarse a dificultades nuevas. En este ensayo, los estudios previos del ámbito de la lingüística proporcionaron un marco para entender qué efectos la IA está teniendo en el lenguaje, y una visión general de los enfoques anteriores para evaluar la calidad de la interpretación sirvió de guía para el diseño de un experimento con el fin de comprobar la capacidad de alumnos de interpretación para reaccionar ante estos cambios lingüísticos. En un experimento de métodos mixtos, alumnos del Máster en Interpretación de Conferencias de la Universidad Pontificia de Comillas interpretaron discursos, algunos de los cuales contenían características típicas de la IA, y respondieron a un cuestionario. Junto con evaluaciones de las interpretaciones basadas en rúbricas, se comparó la calidad de las interpretaciones con la

dificultad del material de fuente. Los resultados demostraron que la presencia de características típicas de la IA en un discurso correspondía con una menor calidad en discursos más densos, pero con una mayor calidad en discursos menos densos. Estudios futuros podrían continuar la investigación de los mecanismos que hacen que estas características típicas de la IA tengan un impacto negativo en la interpretación.

Palabras claves: Modelos Extensos de Lenguaje, interpretación simultánea, densidad de información, dificultad, legibilidad, material de fuente

Contents

Speaking like Machines:
Examining the Effects of AI-influenced Language on Interpreting Students

Digital technology in 2026 is defined by its relationship to the fast-growing Artificial Intelligence industry. New models are released increasingly frequently (LLM Stats 2026), and scholarship on AI is hurrying to keep up. Interpreters are paying close attention to these developments. Some are concerned about the use of AI tools for interpreting assistance, and the potential loss of skills that reliance on technology like Computer Assisted Interpreting (CAI) can cause for interpreting students (Yang 2025; Guo et al. 2025; Reinhart et al. 2025). Others have suggested that upskilling as interpreters will become increasingly important in order to adapt to a private market looking to cut costs through the use of AI tools (Fantinuoli 2023; Giustini 2024; Guo et al. 2025). This paper also examines the crossroads between interpreting studies and technology but is more focused on the growing use of AI tools in the greater context of society, where the spoken material interpreted by interpreters originates. The prevalence of generative AI in society means that interpreters may run up against AI-generated content whether they like it or not, be it in their training materials or in the speech read out loud by a conference delegate. As AI-consistent language use becomes more widespread, it is important for interpreting students to not only understand how to use the technology, but also to be prepared for the ways that it might affect language in general. Given the emerging changes to the English language, it will be important to discuss the features of language that are most difficult for interpreting students to manage. The present study examined the effects of AI-consistent language use on interpreting students' performance.

Literature Review

AI-generated content is becoming more prevalent, expanding into contexts where until recently it had not been common (Sun et al. 2025). Certain words and phrases get called out for sounding like AI, even in the British Parliament (Pimlico Journal 2025). People are noticing changes to language use since the development of Large Language Models (LLMs), but they may be unable to prevent it. Before discussing the impact that these language changes might have on interpreters, and specifically on interpreters in training, a brief overview of the nature of AI-generated content and its recent rise to ubiquity will aid in justifying the relevance of the present study. Then, a review of the particular ways that LLMs use language differently from

humans follows, with particular attention paid to lexical choice and syntax-level structures. These differences have impacts on communication, which naturally implies an intersection with interpreting. Linguistic features on multiple levels can contribute to the difficulty of interpreting a given speech. The words and syntax that LLMs use might therefore have effects on source material difficulty, and these effects will be explored in this study. Finally, the widely debated subject of measuring interpreter performance will be explored. With all of these pieces in place, I will propose a method for observing the effect that AI-consistent language has on student interpreting performance.

What is AI-generated Content?

For the purposes of this paper, the most salient kind of Artificial Intelligence is the Large Language Model type. These models have exploded in popularity since the release of ChatGPT in November 2022 (AI Trends 2026), and now content generated by these language models can be found in every corner of the internet, from journal-published research paper abstracts to viral recipes on TikTok. LLMs are algorithms designed for Natural Language Processing (NLP) that create predictions based on how often linguistic structures appear together in large amounts of Natural Language material (McDonough 2026). LLMs can be trained on many kinds of material, but the most powerful ones are trained on huge parts of the internet, with datasets including Wikipedia and Common Crawl. Part of the training process before the LLM is released publicly always includes human-supervised output evaluation to determine the suitability of the output (Wang 2024). AI-generated content is any text, image, audio, or video created by these AI models. It can be used in isolation or as a supplement to human-produced content, or to revise human-produced content.

AI Language Use

LLMs as a class use language differently than the average human user, and these differences are constantly morphing (Merrill 2025). Famously, the word “delve” began showing up in an inordinate number of scientific abstracts (Juzek and Ward; Kobak et al.; Kousha 2025), with its usage seeing a significant drop shortly after it made waves in academia (Geng and Trotta). Companies that want to boost engagement with AI-assisted content encourage their writers to avoid the current AI-associated words (e.g. Pugalia 2026), and students and

professionals alike try to avoid the words that could see their work flagged by AI detectors which are also steadily improving in effectiveness (Shawky 2025). Lexical peculiarities are often called out for marking AI-generated text, which makes sense from a public communication standpoint—people without a background in linguistics can understand the importance of word choice more easily than that of deeper linguistic structures—but the overuse of specific words is only one feature of LLM-generated text. Syntax- and discourse-level patterns also differ between humans and LLMs.

While LLMs all work on the same principles, each one has its own idiosyncratic language tendencies. Sun et al. (2025a) found that each model they studied used links and qualifiers in different proportions, demonstrating that LLMs can be distinguished from each other by this metric. Many of the other studies referenced in this paper also distinguished between models in their own analyses. While there are many overlapping tendencies among LLMs, it bears repeating that each model behaves somewhat differently in the parameters detailed below, and some precision in the original studies is lost by generalizing between them. Given the variation between individual LLMs, a text generated by a single model might not incorporate all of the features that the cited studies have shown to be consistent with AI-generated language. The effects of individual AI models on natural human speech have not been studied as much as the effects on written text, but conference interpreters often encounter speeches that have been written out beforehand, and these AI-consistent features have already appeared in high-level political settings, as previously mentioned.

AI Lexicon

LLMs do use some words much more frequently than humans, though this varies between models and the style of writing. There are lists of these words on sites that claim to detect AI-generated text (e.g. Grammarly 2025) but they unfortunately fail to provide links to quantitative measurements of these disproportionalities. Other research has indicated that humans who frequently use or read AI-generated text can detect it themselves with high accuracy, often considering specific “AI words” to be a significant tell (Russell et al. 2025). Some researchers (Reinhart et al. 2025, Juzek and Ward 2025, Geng et al. 2025, Yakura et al. 2025, Anderson et al. 2025) generated lists of these words from corpora, sometimes restricted to particular genres or styles. For their respective corpora, Juzek and Ward (2025) focused on

scientific abstracts; Geng et al. (2025) used papers and presentations from machine learning conferences; and Yakura et al. (2025) retrieved words from the transcripts of academic YouTube videos and conversational podcast episodes.

Generative AI models use words that occupy a high register, often sounding academic in contexts where this register is not appropriate. Reinhart et al. (2025), using a generalized corpus of internet texts spanning numerous styles, found that as a possible consequence of overrepresented genres in LLM training material, namely "academic writing, newspapers, or television scripts" (p. 3), LLMs tend to use words describing complex relationships between concepts, such as *tapestry* and *intricate*. According to Reinhart et al., “these words together may signal a preference for grandiose, if hollow, summative sentences” (p. 4). If these words indeed lack clarity or substantial meaning, it could indicate a possible source of difficulty for interpreters. There is also some contention around the possibility that these particular words are overrepresented in LLMs because of the influence of regional dialects of English on the training of LLMs (Hern 2024; Juzek and Ward 2024). Interpreters are, of course, expected to be able to understand a wide range of dialects, but even so, if the words from one dialect appear in a linguistic context where it has never appeared before, an interpreter may be caught by surprise.

Morphosyntax

Larger linguistic structures such as morphology and syntax also differ between LLMs and humans, and these may have a greater effect on interpreters than word choice, because interpreters must take each word and sentence as part of a whole rather than an element to be translated on its own (Yuan 2022), ha Subordinate clauses are more common in AI-generated content (Zamaraeva 2025), and Gile (2009) indicated that such embedded structures increase cognitive load. However, there is comparatively little research on these structures in AI-generated content compared to that on words (Georgiou 2025; Kujur 2025), and they may differ significantly between models.

Reinhart et al. (2025) found that LLMs tend to use a noun-heavy style in English, due in part to the liberal use of "that” clauses as subjects as well as present participle clauses. Georgiou (2025) corroborated this finding in academic exam writing samples, also finding an increased number of conjuncts, adjectival modifiers, and direct objects in AI-generated text. Interestingly, when it comes to Part of Speech analysis, Muñoz-Ortiz (2024) seems to have found precisely the

opposite, at least in news-based text, where humans displayed a preference for noun and adjective content words. Kujur (2025) found both a preference for adjectives and verbs in AI-generated essays. However, since both Muñoz-Ortiz and Kujur restricted their studies to the genre of news text and academic writing, respectively, whereas the corpus in Reinhart et al. (2025) was composed of a wide variety of genres, their results carried more weight in the design of the present study.

On a higher level of syntax, LLMs use certain discourse-level constructions consistently more than humans, highlighted by Russell et al. (2025). Famously, the rule of three, which has long been generally accepted as a strong speechwriting strategy, now sounds suspiciously like AI because of how overused it has become in AI-generated content. The structure "not only ___, but also ___" is another example. The key to understanding AI-influenced linguistic shifts is understanding that GenAI uses the same English as humans, but it uses fewer creative choices, leading to a less diverse version of English, and that version then becomes reinforced (Zamaraeva et al. 2025). The same argumentative formulas appear again and again, leading the annotators in the Russel et al. experiment to consider sentence structure to be the second clearest indicator of AI-generated text, behind vocabulary (2025).

Perhaps the most subtle of AI language patterns takes place on a stylistics level. After a qualitative analysis of ten AI-generated and ten human-written essays, Kujur (2025) used the word “uniformity” to summarize generative AI’s writing style. AI-generated text rarely deviates from a stable register, formulaic reasoning, and a neutral, impersonal tone. In contrast, human writing uses slight variations in style that create interest and communicate the voice of the speaker. While this aspect may be the hardest to measure quantitatively, it may have a profound impact on interpretation. Interpreters depend on their understanding of the source text to provide high quality interpretation, and the idiosyncrasies of human language become points of interest that an interpreter can easily grasp amid what might be an otherwise indistinctive speech. If style, tone, and personality are reduced to monotony, interpreters have less to connect to personally and must rely more heavily on other competencies. On the other hand, the consistency of AI-generated language might rather reduce the cognitive load of interpreters who no longer need to react to as many unexpected twists as occur in human speech.

Impact on Discourse and Communication

AI is being used to create endless pages of text that appears on the internet, sometimes overtly and other times covertly, material which can vary in both veracity and aesthetic quality (Liang et al. 2024; Yakura et al. 2025; Cava et al. 2025; Anderson and Niu 2025; Hanley and Durumeric 2024). Up to half of new webpages could be written exclusively by AI tools, and it's highly likely that other webpages include AI tools in some stage of the otherwise human-driven writing process (Paredes et al. 2024). Even though a much larger percentage of pages that are displayed by search engines, and therefore interacted with, remain human-produced (Smith et al. 2024), the internet is becoming populated by AI-generated posts, comments, and webpages (Cava et al. 2025; Sun et al. 2025b). Regardless of whether people use generative AI for public communications or not (Pimlico Journal 2025), there is a real fear of the consequences that doing so would have on human interaction. By sheer amount of exposure to AI-generated content, humans are witnessing a shift in the demographics of their textual environment.

Words that LLMs use disproportionately compared to humans are showing up in natural human speech as well as in text, at a rate that suggests that AI language use is affecting the rate of English lexical change. Geng et al. (2024) found signs that suggest that LLM use has an impact on the words used in academic oral papers, but the effects go beyond academia. There is evidence showing that spoken content (like YouTube videos) has been significantly impacted by the accessibility of LLMs (Anderson and Niu 2025; Yakura et al. 2025). They can be used for script writing, which means that viewers of YouTube videos can hear AI-generated content being spoken with human (or non-human) voices. Yakura et al. empirically showed the extent to which these videos contribute to an AI-influenced English language community. It is no longer necessary for humans to interact directly with LLMs for their language patterns to mark everyday language use. Though these effects were felt first in academic and scientific environments, they are now becoming embedded in human communication by way of education and business contexts. It is natural to assume that the formal settings in which conference interpreters most often work will not be the last to be affected by this ongoing shift.

The effect of linguistic features on source material difficulty

How might the features consistent with AI-generated content affect how difficult a speech is to interpret? Interpreters, of course, often exist in the same linguistic communities as

one or both participants of an interpreting encounter, and therefore will be already familiar with everyday language use in that language. It is possible that interpreters are already adapting their own skills unconsciously to the changing probabilities of certain words and formulations in that language. In fact, interpreters may be further along than your average person—many see AI as an endless supply of training material (Goldsmith 2025) and so will be more familiar with the patterns in AI-generated text. On the other hand, it's possible that AI features are currently causing unique problems for interpreters in training. This could be explained by the concepts of cognitive load and the effort model (Gile 2009). Any number of interferences to the listening and production efforts could be attributed to the particular language of LLMs, on whatever morphosyntactic or discourse level we wish to examine. If those linguistic choices are unusual or the interpreter does not expect them, the cognitive load required to interpret increases. The present study was designed to observe whether there is a difference in the performance of interpreting students as a result of AI-consistent language features, and whether it created problems for the interpreters.

Controlling an experiment for the effect of language use on interpreting quality requires the source material to maintain a comparable level of difficulty, but it is difficult to maintain consistency across several specific linguistic features that contribute to the overall interpreting situation. Any aspect of the source material that has the potential to demand heavier cognitive resources from an interpreter increases the theoretical difficulty of the speech. Yuan (2022) ascribed greater interpreting difficulty to larger logical structures over word-level and sentence-level structures, and claimed that a speech's structure should be the primary factor in determining its difficulty level, especially for interpreting students. The words used by LLMs are easy to quantify in a corpus-based analysis but potentially have a smaller impact on the actual interpretation compared to syntax and discourse-level structure. Interpreters may strategically diverge from the lexicosyntax structure of the original (Dam 2001), and this could be due to several factors other than the difficulty of those particular words. Moser-Mercer (1996) recommended the use of propositional analysis or difficulty ratings as reliable ways to measure source material difficulty, but expert ratings of difficulty can be unreliable, particularly if the experts focus on different aspects of the text in their evaluations (Liu and Chiu 2009). Other research has advocated for readability as a metric for source material difficulty (Kuang and Zheng 2023; Liu and Chiu 2009), because it agglomerates multilevel linguistic features such as

word length, sentence length, and syntactic structure. The Flesch Reading Ease tool (among others) can be used to calculate the readability of a text. Combining this tool with the subjective judgement of the interpreters, speeches utilizing the features above can be measured against speeches without these features.

The level of difficulty of an interpreted speech is related to several factors in conjunction with one another including delivery speed, subject matter, and informational density (Moser-Mercer 1996; Gile 2009; Liu and Chiu 2009). The last of these perhaps needs the most clarification. Muñoz-Ortiz (2024) found that in the style of news-based text, LLMs' sentence lengths are more uniform than those written by humans, tending towards sentences of 10-20 words, with fewer outliers than human writers. Interestingly for interpreting studies, the length of a sentence does not necessarily correspond with density of information, which is a much more important metric to track for the purpose of determining source material difficulty. As such, Moser-Mercer (1996) proposed using either propositional analysis or expert ratings, which can reliably track information density, to systematically determine source material difficulty.

Measuring Interpreting Performance

Here it is pertinent to discuss the measurement of interpreting quality. Before one can compare interpreting performance of control speeches with speeches containing AI-consistent language, one must establish a benchmark for measuring interpreting performance. Historically, measuring quality in interpreting has been highly contested, as it consists of the interplay between the levels of morphosyntax, semantics and pragmatics, and discourse. Lee (2015) provided an overview of some of the many systems that have been developed by researchers since the mid-1980s for this purpose. Many methods for measuring interpreting performance have been developed, and studies of interpreting quality lack unanimously agreed-upon criteria to measure quantitatively. Each researcher has needed to decide for themselves which of the numerous facets of interpreting were relevant and which were not, and whether to use holistic or analytic scales for their evaluations (pp. 231-236). Holistic evaluation methods are best suited for comparative studies between two groups of interpreters, and ideally a large number of raters, because these types of studies do not need to identify specific language features that cause interpreters to have difficulty maintaining quality (Ding 2017). An analytic approach provides the researcher with specific qualitative information about relevant indicators of quality, which

makes the approach complementary to interpreter self-evaluation. Analytic scales divide the evaluation of interpretation across various criteria, which may be weighted according to their importance.

Lexical items can of course be compared between original and interpretation, often through transcription, but doing so prioritizes the words used over the act of communicating. Pöchhacker (2002) said that textual features only tell part of the story, and that interpreting quality is also the result of "complex psychocommunicative relationships and effects" (p. 420). In between the text and the understanding of it is the content's meaning, which is the priority of meaning-based interpreting (Dam 2001). Lexical similarity between original and interpretation on its own is not a sign of good interpreting, either in the estimation of external evaluators or the interpreters themselves (Chiaro and Nocella 2004). In fact, the form of the original often needs to be transformed to maintain the intent in a target language. As an analytic tool, something akin to propositional analysis can be a way for researchers to measure the amount of information retained in an interpretation without basing the measurement on one-to-one lexical correspondence. However, in order for propositional analysis to yield useful results, the division of propositions and the acceptability of their interpretations should come from the agreement of multiple experts, rather than a single evaluator.

The most complete picture of interpreting quality arises from the analysis of several levels of granularity in concert (Bartłomiejczyk 2007). Interpreting quality is the result of complex processes that mutually influence each other, as much as researchers would like to separate them. Thus, the ideal measurement of interpreting performance, as stated by Pöchhacker (2002), combines "corpus-based observation, survey research (interviews), participant observation and documentary analysis so as to ensure a holistic view on quality also at the levels of intended effect and successful interaction" (p.420). Corpus-based observation can be used to detect errors in substitution and lexical content, a function also filled by propositional analysis, but these functions can also be included in rubric-based qualitative evaluations, albeit with the loss of quantifiable lexical correspondence. Limited resources may also make it impossible to include analysis from all of the multitude of perspectives that are involved in an interpreting event, which Pöchhacker calls the "constellation of interactants" (p. 412). However, surveys can be collected from the participating interpreters on their perceptions of performance and contributing factors, and outside observers (like the researcher) can make their evaluation based

on their more complete knowledge of the material. Given the nature of the present experiment, which limits the number of participants in the first place, these two perspectives allow for an insightful comparison of interpreting quality perceptions.

Rubric Design

The optimal design of error scales for interpreting evaluations is another contested territory in research centered around interpreting. When listing the types of errors that evaluators will look for, there should be a clear distinction between error types, without overlapping (Moser-Mercer 1996). Naturally, these error types correspond to the factors that are prioritized by the evaluator depending on their own purposes. The criteria that are important for a professional accreditation test are different from those that interpreting instructors consider important in formative tests for their students. For this study, it was important to find an assessment rubric that was designed and used for a similar purpose, that is, for measuring the effects of differences in the source texts on the interpretations.

The assessment rubric developed by Khorami and Modarresi (2019) was consulted because it somewhat resembled the experimental design of the present study, namely in that it was designed for the evaluation of interpreting students (albeit for consecutive interpreting). However, they included the factor of personality traits, including items like “[h]aving personal aptitude for interpreting”. These personality traits are difficult to measure without extensive training in interpreting assessment and furthermore have little bearing on the measurement of performance in an experimental setting involving interpreting students. This rubric generally paid attention to the underlying competencies used in interpreting rather than the resulting interpretation, thus requiring evaluators to make inferences. The current study was oriented more around product than process.

Lee (2008) and Lee (2015) developed very similar assessment rubrics for assessing consecutive interpreting students based on their interpretations. The rubric proposed by Lee (2015) was ultimately chosen because it required the least number of modifications to fit the present experimental design. Ultimately, the clearly stratified criteria for the interpretations themselves present in the rubric by Lee (2015) made it the basis for the rubric used in the present study. The criteria “Finishing interpretation within the time limit” was dropped because it didn't apply to simultaneous interpreting. Additionally, in the interests of consistency with the

interpreter questionnaires, the original scale from 0 to 4 was changed to a scale from 1 to 5, with 1 representing “No characteristics present.” The rubric with these modifications can be found in Appendix F.

Addressing the weight of each measured component, another important factor in designing such scales, the scale in Lee (2015) weighted Content, Form, and Delivery according to a 2:1:1 ratio. Lee (2008), on the other hand, placed equal weight on accuracy and target language quality, which roughly correspond to Content and Form, respectively. Although Ding (2017) points out that delivery is often left out of wholistic assessments of interpreting performance because it typically does not vary much between interpreters who share the same level of language expertise, it was considered in this study because dysfluencies might indicate higher levels of cognitive load (Gile 2009). Lee (2015) also provided a justification for placing more weight on Content than any other component. Two of the criteria in that category (“no opposite meanings” and “accurate rendition of main ideas”) were the only criteria in the rubric that received a mean valuation above “very important” from interpreting instructor respondents.

Methodology

The experiment was designed in two parts: the elaboration of the source material for interpreting, and the evaluation of the interpretations themselves. The material consisted of four speeches, two of which had been altered to contain various AI-consistent language features. The modifications to the speeches were not intended to replicate a hypothetical text generated by AI, but rather to represent a speech written by a human who has been influenced by linguistic shifts tending towards the same patterns that LLMs exhibit. The evaluation of interpretations consisted of a mixed-methods design of questionnaire-based self-evaluation and rubric-based assessment to capture potential differences between the perceived source material difficulty and actual interpreting performance.

Sampling of Participants and Recruitment

Participants were sampled by convenience from the Masters in Conference Interpreting at Comillas Pontifical University in Madrid. All nine of the participants had been studying conference interpreting at the Masters level for six months (October 2025 through March 2026). The interpreters' previous experience with interpreting before the start of the program varied.

They had been studying and practicing simultaneous interpretation in the program since December 2025, having studied consecutive from October to December. All of the interpreters had Spanish as an A language and English as a B or C language. The fairly equivalent experience levels of the interpreters served to reduce the difficulty of controlling variables in a small dataset. The researcher who evaluated the interpretations and facilitated the experiment was also a student of the program. He had English as an A language and Spanish as a B language, and had been studying for the same amount of time as the participating interpreting students.

Ethics

The interpreters gave their consent beforehand for their recordings and transcriptions of the recordings to be used for the purposes of the study. Their personal information was protected under EU General Data Processing Regulation 2016/679.

Materials

The material consisted of four speeches on two topics related to current affairs, topics with which all the interpreters had a similar degree of familiarity. One speech was on nuclear energy and current trends in policy (denoted as N1), and the other was on social media bans for teenagers (T1). The speeches were all written by the researcher to ensure consistency of style. One original speech was written for each topic, and then a second version was adapted from the first incorporating AI-consistent language features, resulting in four total (these altered versions were called N2 and T2, respectively). Specialized terminology was avoided, given that the interpreters would not be given time to research the topic beforehand. Written speeches read out loud are often difficult for interpreters because written text tends to be much more lexically dense and syntactically complex (Gile 2009, p. 181). To avoid these pitfalls, the speeches were first delivered with only an outline, and a transcription was made of the resulting natural speaking of the researcher. The speeches were written and edited to be approximately four minutes long when delivered at a speed of 120 words per minute.

Before adding certain AI-consistent features to the speeches, they were put through a readability test to ensure a baseline comparability between the two. Thus, in theory, any AI features added in the following steps would be the sole determiners of increased difficulty. Liu and Chiu (2009) also used readability scores, noting however that their dependence on word and

sentence length means they are not optimized to operate on spoken language, nor do they take syntax or style into account as contributing factors of difficulty. However, Liu and Chiu (2009) do suggest that readability might be a better indicator of source material difficulty for simultaneous interpreting than consecutive (p. 259), and Liu et al. (2004) found a correlation between difficulty according to readability tests and simultaneous interpreting performance.

A new version of each speech was made containing the lexical and morphosyntactic patterns which are commonly found in AI-generated text as discussed in this paper. First a selection of some commonly used AI words was incorporated based on several lists that was compiled in Appendix A. From the list of words created by Reinhart et al. (2025), several were removed such as *ebook*, *bam*, *paperback*, and *bananas*, because they did not fit with the topics of the speeches (2025). Some words that were present on multiple lists (*align*, *boast*, *comprehending*, *crucial*, *delve*, *significant*, *surpass*, *underscore*) were included only once in the master list. Multiple inflections of the same lexical root (e.g. *comprehend* and *comprehensive*) were likewise reduced to a single entry. After adding "AI words", several sentences were modified to include more adjuncts, and implicit rhetorical structures were made more explicit. References to the speaker were removed to create a more neutral tone. The noun-heavy style of GenAI was emulated by rearranging some sentences to use "that" clauses and present participles as subjects. The rule of three and the structure "not only ___, but also ___" were also included in these altered versions. The new versions of the speeches were read aloud and recorded by the researcher, ensuring consistent voice quality and speed for all four. The alterations are highlighted in Appendix B.

These recordings were then transcribed and divided into units of meaning by the researcher, and this permitted a measurement of the difficulty level of the speeches by way of informational density. As noted by Liu and Chiu (2009), the proportion of the number of propositions to the total number of words offers a way to measure the relative informational density of each speech. Each proposition was labeled as either critical or secondary, as in Liu et al. (2004)..Propositions with AI-consistent features in the altered versions were bolded. Rather than separating propositions by their syntactic function, such as predicates and modifications (Ding 2017), this method emphasized the importance of prioritizing meaning over form in simultaneous interpreting.

Interpretation

Prior to the task, the participants were informed that their interpretations would be used as data in a study of interpreting quality and AI, though the specific purpose of the study and the specific assessment criteria were not revealed to them. The participants took part individually in the study as was convenient to their schedules, and they were told not to discuss the content of the speeches with other participants. Some of them had interpreted earlier in the day, while others participated in the experiment before attending their first classes of the day. The following measures were also taken to control the experiment for quality of interpreting, which was the dependent variable:

The speeches were video recorded ahead of time to ensure proper and consistent speed of delivery (approx. 120 words per minute), in addition to consistent audio quality. The recordings showed the speaker seated at a desk and speaking directly to the camera.

The mode of interpretation in this experiment was simultaneous. In simultaneous interpreting, interpreters are forced to make rapid choices, and they are not privy to the overall shape or direction of the speech before delivering their interpretation. This prevents interpreters from using the greater context of the speech to compensate for momentary difficulties, of which AI-consistent language may be a possible cause.

During the task, each participant was given one speech with AI-consistent features and one without, such that they would interpret speeches on both topics. The order of the speeches was randomly determined. Before each speech, the participant was given the topic and a few important terms that would appear in the speech. They were given time to find equivalents of these terms and discuss them with the researcher. The participant was then asked to watch and listen, recording the audio of their simultaneous interpretation. After delivering their interpretation, they were given one minute to fill out the short questionnaire for the speech. To ensure consistent performance between renditions, they were then allowed up to five minutes to rest before repeating this process with the second speech. Finally, the participants were given an open invitation to share their general impressions of the experiment in their own words in the questionnaire.

Evaluation

Interpreter self-evaluation

The interpreters were asked to fill out a short questionnaire after each interpretation (see Appendix E). They could rate their perception of the speech difficulty and indicate the features of the source material that contributed to their rating. They could also indicate whether they felt that the speeches in this experiment were more or less difficult than the typical practice speeches used in their classes. In the second section of the questionnaire, the interpreters gave their assessment of the quality of their interpretations based on Content, Form, and Delivery. In contrast to the rubric used by the researcher, however, the interpreters did not have an itemized breakdown of the criteria in each category. Instead, they had to indicate their score on a scale of 1 to 5, with 1 indicating very unsatisfactory performance, and 5 indicating very satisfactory performance.

Rubric-based Evaluation

The researcher took on the role of evaluator in the experiment. He made an initial evaluation while the participant delivered the interpretation using a rubric based on Lee's (2015) (see Appendix F). He made a second evaluation of each speech two weeks later while listening to the audio recordings of the interpretations. The rubric used for this evaluation was the same in both the first and second evaluations. The researcher followed the same order of interpreters and speeches for the second evaluation.

Results

The researcher collected the questionnaires, the audio recordings, and the rubric evaluation results for each of the eighteen interpretations. The data was compiled in Microsoft Excel, and the same software was used to create visual representations of the source material difficulty and the interpreters' performances. The perceived difficulty of the speeches was compared to the readability and informational density to evaluate the difficulty of the source material from several perspectives. The self scores were likewise compared to the rubric-based evaluations. To this effect, the interpreters' ratings of Content were weighted, just as the Content score in the rubric based on Lee (2015) was weighted. After compiling these results, the

researcher was able to discuss how language influenced by generative AI may have affected the quality of interpreting students' renditions.

Source material difficulty

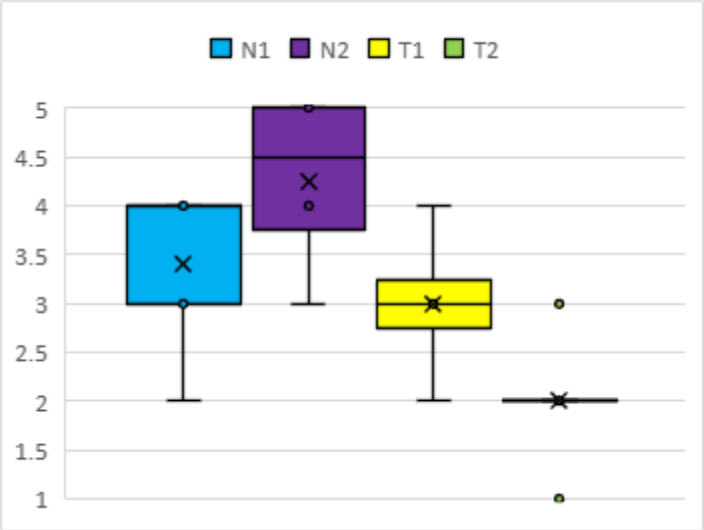
Perceived difficulty

The participating interpreting students rated the difficulty of each of their two speeches immediately after delivering the interpretation. The questionnaire gave them a scale from 1 to 5, with 1 indicating a very easy speech and 5 indicating a very difficult speech. Two of the speeches were interpreted by five different interpreters (N1, T2), and the other two speeches were interpreted by four (N2, T1). The difficulty ratings were grouped by speech into a box and whisker plot as shown in Figure C1. An *x* marked the mean rating, and a horizontal line marked the median. Only T2 had outliers, since three of the five ratings put the difficulty at 2.

The difficulty scores as reported by the interpreters marked a difference between the four speeches. The participants rated the difficulty of the unaltered speeches between 2 and 4 on the scale, inclusive. However, three of the five participants who rated N1 rated it at difficulty 4, whereas two of the four participants who rated T1 rated it at difficulty 3 (see Table C1 for each interpreter's response). The mean and median difficulty ratings for N1 were therefore higher than those of T1. The difference in perceived difficulty for the speeches containing AI-consistent features was much greater. All four participants who interpreted T1 and N2 rated N2 as more difficult than T1. Conversely, all five of the participants who interpreted N1 and T2, regardless of the order in which they interpreted the speeches, rated T2 as more difficult than N1. It can therefore be inferred that the difficulty of the source material from the perspective of the interpreting students, did not correspond to the presence of AI-consistent language. These added features appeared to have a negative effect on the perceived difficulty of N2 but a positive effect on that of T2.

Figure C1

Perceived Difficulty by Speech



After providing their numerical difficulty rating, the participants were asked to report the characteristics of the speech that contributed to their previous response in the questionnaire. They then indicated whether the speech was easier, of similar difficulty, or more difficult than the typical speeches that they practice with in their classes (Table C1). The participants who interpreted N2 described it as having high density (two participants), a complex topic, a fast speed, and technical vocabulary. The two participants who gave N2 a difficulty rating of 5 also reported it to be more difficult than in-class speeches. These were the only participants who considered any of the speeches they interpreted to be more difficult than the speeches they practiced with in class. All the participants who interpreted N1 considered it to be of similar difficulty to in-class speeches. Conversely, only T1 and T2 were considered easier than in-class speeches. T2 was reported to have a slow speed (three participants), accessible vocabulary (three participants), and clear structure (three participants), and three of the five participants considered it easier than in-class speeches, while the other two considered it to be similar.

Readability

The readability scores of the speeches showed similarities across various metrics (see Appendix D). Many of the readability tests, including the Flesch-Kincaid formula, Gunning Fog index, SMOG index, Coleman Liau index, and Automated Readability Index, were based on

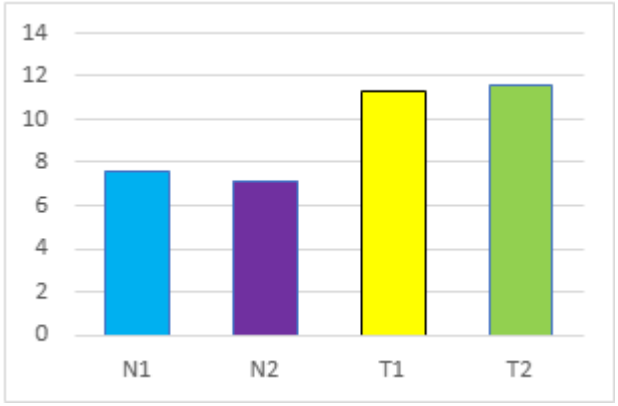
average sentence length, word length, and number of polysyllabic words. Others, like the Dale-Chall formula, measure proportion of easy words according to a predetermined list. All the readability scores indicated a higher difficulty level for the speeches with AI features than for their respective originals. In this respect, the readability did not always match the observed difficulty according to the participants. In particular, the participants rated T2 as easier on average than the version without AI-consistent features, which according to readability scores was less readable than T1.

Informational density

Informational density was determined by the researcher in terms of words per proposition (see Figure C2). N1 and N2 were found to be more information dense than T1 and T2, which corroborated the participants' own observations. The difference in informational density between N1 and T1 was greater than the difference between each original speech and their altered versions. In fact, the difference in informational density between original and modified speech was only 0.5 words per proposition between N1 and N2, and only 0.3 words per proposition between T2 and T1. Nevertheless, the variation that was observed between versions of the same topic marginally followed the pattern observed in the difficulty according to the participants. N2 was more difficult than N1, but T2 was easier than T1. The speeches that incorporated AI-consistent features (that is, N2 and T2), were not necessarily more informationally dense than the speeches without AI-consistent features. This measurement is consistent with both the interpreters' determination and the readability scores.

Figure C2

Words per Proposition



Interpreting performance

The combined results of the rubric-based evaluations were compared to the participant self scores. The perception of interpretation quality differed between the participants and the evaluator, and the difference varied by speech. Figure C3 plots the combined score averages of according to the participants themselves, the first rubric-based score, and the second. For every speech, the average self score was lower than either of the rubric-based scores, The difference was greater in N1 and N2 than in T1 and T2. The average scores of T1 were nearly identical to the average scores of T2 (for the self score, 13.75 and 13.80, respectively; for the first rubric-based evaluation, 15.75 and 15.6; and for the second, 15.25 and 15.8).

Figure C3

Average Interpreter Performance by Speech

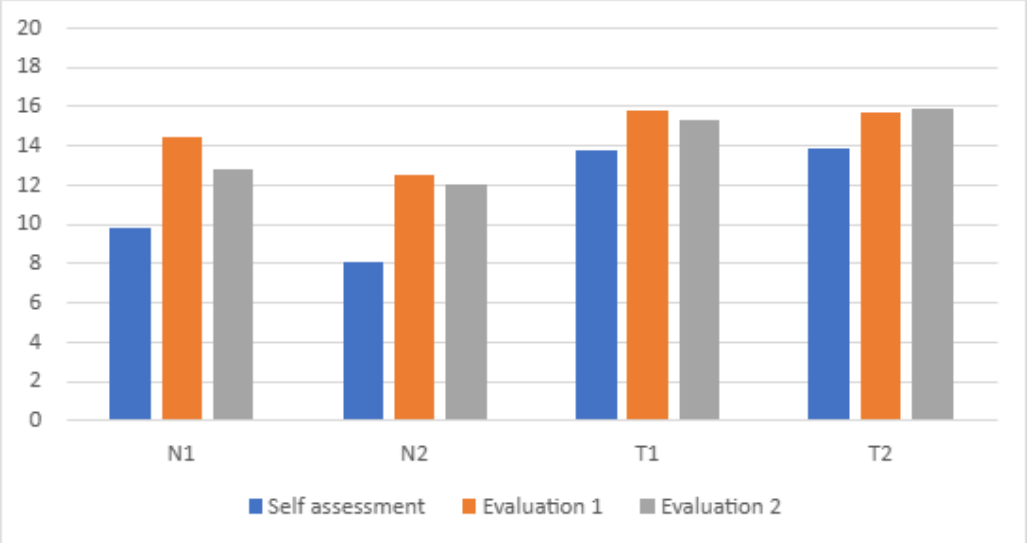
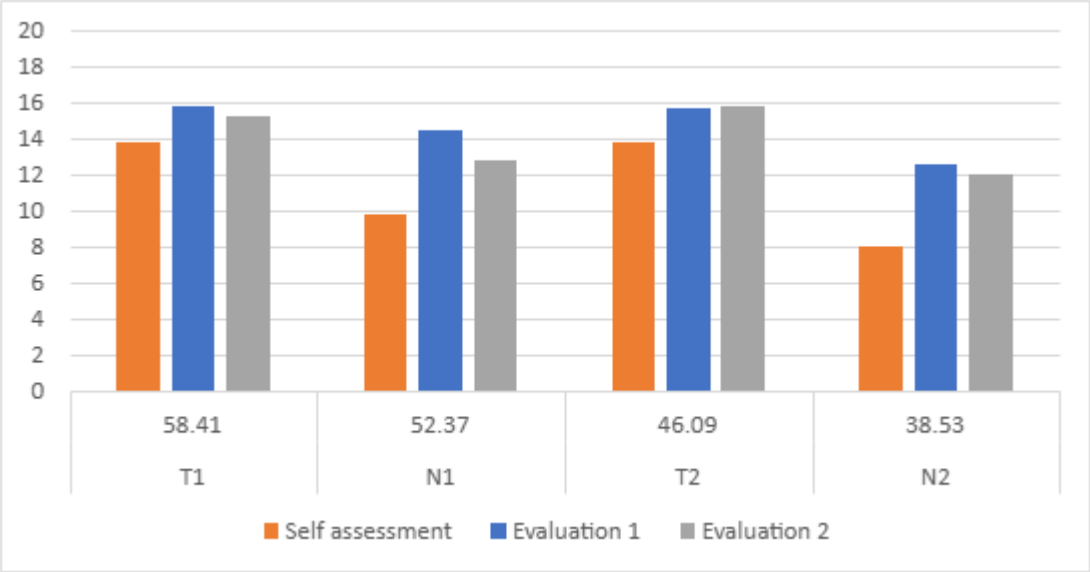


Figure C4 shows the speeches rearranged in order of ascending difficulty according to their Flesch-Kincaid readability scores. T2 was measured to be less readable than T1 and N1, but interpreting students performed just as well on T2 as on T1, and better than on N1, on average. That is, more readable speeches did not always correspond to higher interpreting quality. Rather, lower informational density in words per proposition, and lower difficulty ratings of the participants themselves corresponded to the highest quality of interpretation.

Figure C4

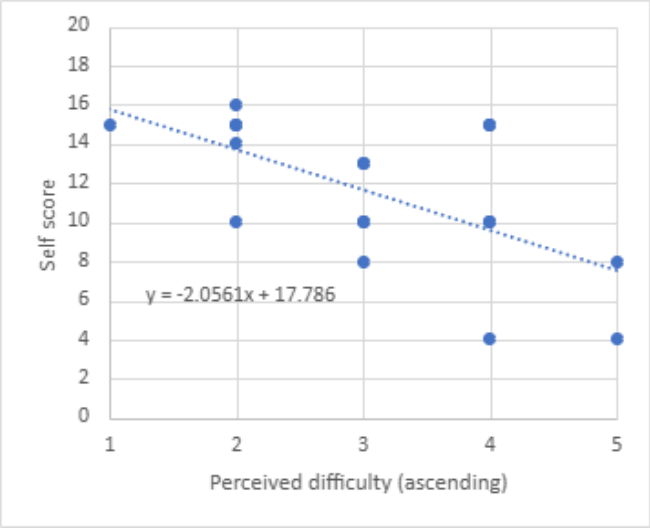
Interpreter Performance by Flesch-Kincaid Score



Naturally, the interpreters tended to consider their performance to reflect their perception of the source material difficulty, as they had reported in the same questionnaire. The downward slope of the trendline in Figure C5 shows that higher perceived difficulty correlated to comparatively lower scores. There appears to be an exception to this rule in the single high score given to a difficulty 4 speech. Interpreter 3 rated T1 as having a difficulty of 4 but gave a high rating of 15 to their performance. This was not an exception to the trend, however, because the same interpreter gave a difficulty rating of 5 to N2, and reported a self score of 8. Higher perceived difficulty was reflected by lower perceived quality of interpretation in every case.

Figure C5

Self Score as Function of Perceived Difficulty



Figures C6 and C7 in the appendix show that the rubric-based scores reflected the perceived difficulty level to a lesser extent. As shown in Figure C3 and Table C2, the participants tended to give themselves lower scores than the external evaluator. In fact, there were only three instances where a participant gave themself a higher score than the evaluator did. On average, the evaluator gave slightly lower scores in the second round of evaluations than in the first. The average scores for N1/N2 differed much more than the average scores for T1/T2, which were virtually identical. There was also a greater disparity between the self-score and the evaluation in N1/N2 than in T1/T2. Figures C8 and C9 show these differences organized by interpreter and speech. The bars extending into negative y-values show the three occasions of the evaluator scoring the participant lower than the participant rated themselves.

Figure C8

Difference Between Self and Average Rubric Evaluation – N1, T2

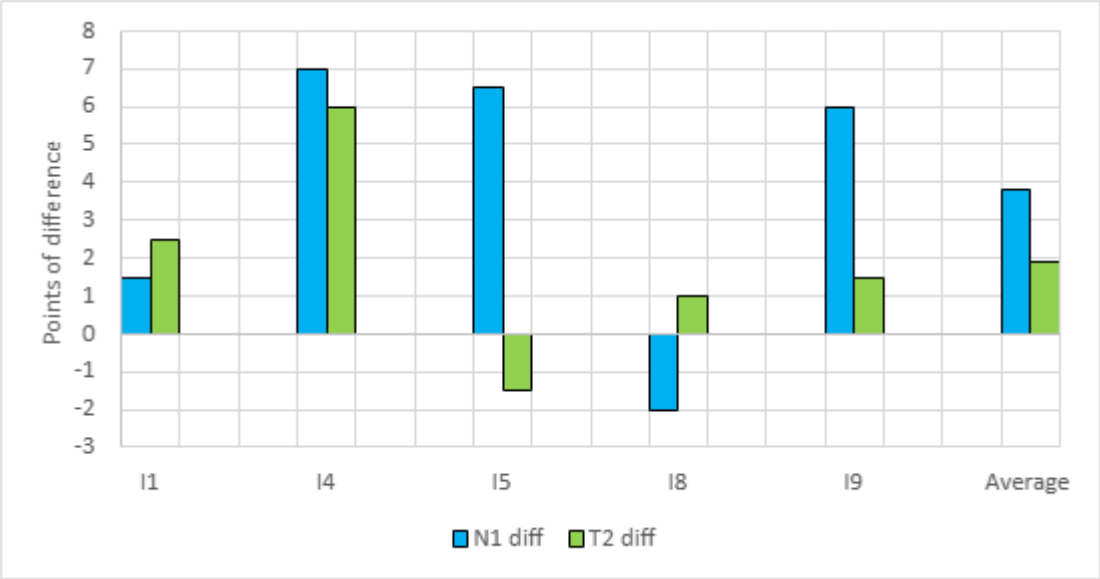
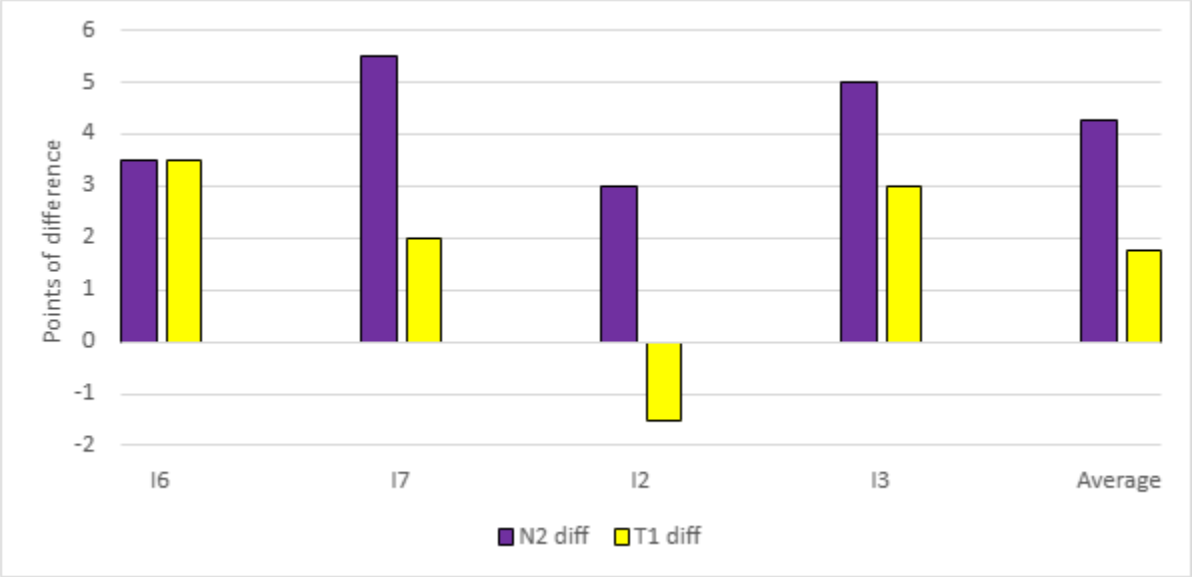


Figure C9

Difference Between Self and Average Rubric Evaluation – N2, T1

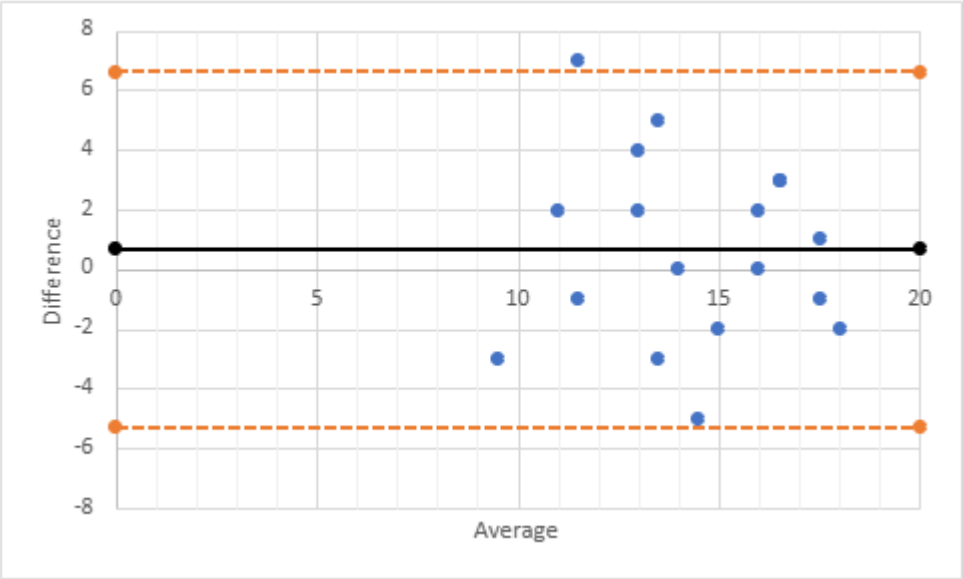


The two rubric-based evaluations appeared to be more consistent with each other compared to the self score, but they also gave conflicting reports of the quality of interpretation. A Bland-Altman plot of the two sets of results is shown in Figure C10, with each point showing

the average between the two evaluations of each speech on the x-axis and the difference between the first and the second scores on the y-axis. The range of differences and the standard deviation were high, indicating a low level of agreement between ratings. Additionally, the coefficient of variation (CV) was calculated for each of the three sets of scores individually, to show the spread of scores. The first rubric-based evaluation CV was 18.60%, and the second was higher at 21.16%, indicating that there was more variation in the evaluator's second score. The participants' self score CV was higher at 32.08%, which indicates that the interpreters gave themselves a wider range of scores than the evaluator.

Figure C10

Bland-Altman Plot Comparing Rubric Evaluations



Discussion

The most striking result of the experiment arose from the difference between the apparent difficulty of the two speeches with AI-consistent features. The results suggest that the difficulty of a speech for interpreting students does not depend solely on these characteristics. Interpreting students were able to perform well on T2 despite the presence of the lexical, grammatical, and structural features often found in AI-generated content. The interpreters performed more poorly when the speeches were denser as in N1 and N2, and it appears they could recognize that density of information as they interpreted, because they noted that a high density of ideas led them to

give that speech a higher difficulty rating. They also reported higher speed for these speeches that had a higher informational density, even though all four speeches were delivered at the same number of words per minute. The impression of speed could be a sign of higher cognitive load on the interpreter when the source material was very dense, and the opposite could be said for speeches with lower density.

Readability scores failed to accurately determine source material difficulty for the interpreting students in this study. The readability scores of N1 and T1 were similar (see Appendix D). They were designed this way to ensure that AI-consistent features were the only significant change between versions of the speeches. Nevertheless, if one assumes that interpreters will on average perform better on easier speeches, and vice versa, the results may indicate that readability cannot be the sole metric to set a baseline difficulty level for simultaneous interpretation at this level. The difficulty ratings that the participants attributed to the speeches were more indicative of the resulting interpreting quality, the average perceived difficulty of N1 being slightly higher than that of T1. However, given that seven out of the nine participants considered all of the speeches to be either of similar difficulty or easier than speeches they typically practice with in their classes, future research using interpreter ratings of difficulty will need to employ more finely tuned instruments to allow the participants to indicate the features of the speech that cause them difficulty. The measurement in this study that most closely predicted the quality of interpretation was informational density, determined by dividing the speech into propositions as discussed previously. Readability, which often incorporates sentence length in its calculation, does not incorporate sentence structure, and subordinate clauses can greatly increase the complexity of a sentence without altering its readability score. In the case of the speeches used in this study, it may be inferred that readability alone was insufficient for translating linguistic complexity to interpreting difficulty.

The fact that one speech became easier after AI-consistent features were incorporated while the other became more difficult suggests that other variables may have had a greater impact on the interpreting task. Some of these AI-consistent features might even have aided the participants rather than cause problems (future research could be directed towards examining the impact of GenAI's more overt logical structure on interpretation quality). The key to understanding the difference in difficulty between T2 and N2 is therefore best explained by the intersection of informational density and these AI-consistent features in concert. N1 was already

denser than T1 from the beginning. The density of N1 appeared to have a negative effect on interpretation quality compared to T1, as reported by the participants and reflected in the quality of interpretation. When AI-consistent features were incorporated into the second version (N2), interpreter performance dropped further still.

There was an observed difference between the average rubric-based evaluation and the score that the participating interpreting students gave themselves. The participants tended to give themselves lower scores than the evaluator on more difficult speeches. The difference between self score and rubric-based evaluation could have been a result of the different measuring systems used in each type of assessment. Given that the rubric contained several items in each category (Content, Form, and Delivery), a score of 1 for any given category was very rare, because the evaluator could almost always identify one or more criteria present in the interpretation for each category. The participants of the study, on the other hand, merely rated their performance in each category on the basis of their own satisfaction, rather than by the presence of discrete items indicating quality, which could be a possible reason for lower scores. It is also possible that the interpreters were more critical of their interpretations. In the rubric-based assessments, the participants scored just over 4 out of 5 on average in the Form category, whereas the self scores for the same category put them just under 3 out of 5 on average. There was greater discrepancy between the two scoring methods in N1/N2 than in T1/T2, which could have arisen from negativity bias on the part of the participants, since N1 and N2 were the speeches that they considered to be the most difficult.

Limitations

The parameters of the source speeches that were measured by readability scales were insufficient for the definitive determination of source material difficulty. By comparing the readability scores with the judgement of the interpreting students, it was determined that the speeches were not uniform in their difficulty level. More precise measurement of source material difficulty could potentially increase the clarity of the results of the experiment. To this purpose, the speeches could be submitted to several interpreting professionals who could give them difficulty ratings. A comprehensive method for separating speeches into propositions as in those used by Bovair and Kieras (2017) or Ding (2017) could also increase the accuracy of the difficulty scores. If the source material difficulty is measured with sufficient confidence (by

interpreting professionals, for example), the experiment design could be scaled to increase the statistical significance of the results.

A substantial limitation of this study was the lack of access to large numbers of interpreting students and professionals. Stronger data would be gathered with a larger number of participating interpreting students, and several assessors selected from professional interpreters with English and Spanish in different language combinations. The effects of GenAI on English speeches may be perceived differently by interpreters with English as an A language. A further limitation related to the lack of access to a sufficient number of assessors was the lack of validation for the rating scale used by the researcher to evaluate the interpretations. With more resources and time available, the rubric could be validated by having professional interpreters and/or interpreting instructors give ratings to various component parts, which would increase the applicability of the resulting scores.

The lack of expert evaluators also increased the subjectivity of the results. The researcher (a Masters student himself) was the only evaluator to assess the quality of the interpretations. Though student interpreters have also been raters in previous studies, of the six characteristics of an ideal assessor, the evaluator in this study only possessed two: excellent command of the language pairs, and history of study in an interpreting program (Ding 2017, p. 41). The evaluator was also the author of the speeches, which meant he was not only familiar with the material, but also knew the intent behind the words of the speech, which could have introduced a bias in the scoring. Additionally, the evaluator conducted the second rubric-based assessment by listening to audio recordings of the interpretations two weeks after the first evaluation. There was a substantial difference in score between the two assessments, which may have been caused by these factors. Reliability of the evaluation could be improved in future studies by selecting multiple evaluators from a group of professional interpreters or interpreting instructors, ensuring that they are trained with the rubric, and conducting as many evaluations as possible for each interpretation.

The interpreting students' self score provided an interesting comparison with the external evaluation, but the much simpler design of the questionnaire did not allow for a direct comparison between the two quality ratings. Accuracy in the measurement of interpreting quality could be improved by changing the phrasing of the questions to be more specific, or even subdividing the questions on Content, Form, and Delivery to include the criteria present in the

rubric. The 5-point scale in the questionnaire could then be more defined, with 1 indicating serious and frequent errors and 5 indicating minor and infrequent errors. The evaluation rubric could be similarly improved by making the 5-point scale more closely match that of the questionnaire. For example, a score of 1 could be expanded to indicate "very few or no characteristics present", and a score of 5 could indicate "nearly all or all characteristics present". These modifications to the continuum scales would align with Angelelli's recommendations for assessment rubrics (2009).

In the present study, the evaluator made assessments of the presence or absence of various criteria over the entirety of the interpreting performance, but difficulty and informational density fluctuate over the course of a speech. The most difficult components of a speech could be identified if the researcher were to obtain time-bound data points (Gieshoff and Albl-Mikasa 2024). Future research could use detailed propositional analysis to not only measure informational density more precisely, but also to trace the fluctuations in informational density over time in the source material, rather than simply taking an average of the entire text. Additionally, certain linguistic features might have localized effects on interpreters' cognitive load (Gile 2009), and a propositional analysis of the interpretations could be used to identify the most problematic passages for interpreters. In this way, one could obtain more concrete information on the effect of AI-consistent features on interpreters in the moment of processing them, and potentially even identify the most difficult features of AI-consistent language to interpret.

Conclusion

Artificial Intelligence is being used to create vast amounts of texts in various registers and genres using Large Language Models. Detecting this content is becoming increasingly important, but it is difficult to identify features which unerringly reveal AI. There is research that suggests that since the advent of LLMs, human language use has shifted to incorporate AI-consistent language features on multiple levels. As these patterns become widespread in natural language, especially among academic and scientific communities, they have the potential to affect the work of interpreters. The overrepresentation of certain "AI words" and high informational density, among other features at various linguistic levels, in speeches written by humans, has the potential to make interpreting more challenging. While interpreting quality is

the result of numerous interdependent factors that come from context, it can be measured with high confidence provided a variety of methods are utilized, including error scales and participant questionnaires. The present study explored the possible effects that the language changes caused by LLMs could have on student interpreters.

The researcher wrote two speeches on different topics, ensuring a similar readability score between the two before creating a second version of each incorporating several linguistic features which have been shown to be overrepresented in AI-generated texts, including specific words, embedded clauses and noun-heavy syntax. The intention was to emulate the style of someone who through exposure to AI-generated text had incorporated some of its language patterns into their own writing and speaking. Interpreting students from the Masters of Conference Interpreting at Comillas Pontifical University interpreted two of these speeches each, one without AI-consistent features and one with them. Their impressions of the interpretations were compared to the researcher's evaluations of their performance. The speeches were also analyzed for difficulty using interpreter feedback, readability scores, and by dividing the speech into units of meaning.

The results indicated that AI-consistent features may have an impact on interpreting quality, but they may be less of a determining factor than other characteristics that increase the difficulty of interpreting a speech, such as informational density. However, in speeches with high informational density, it could be argued that these AI-consistent features exacerbate the difficulty of interpreting them. Readability did not correlate with the quality of interpretation precisely because it does not measure informational density. The participants tended to give themselves lower scores than the researcher, but in general the quality of interpretation was higher in low-density speeches than in high-density speeches. The combination of high informational density and AI-consistent features resulted in the poorest interpreting quality of the four speeches, suggesting a possible compounding effect.

For better or for worse, conference interpreters at the highest levels are regularly required to interpret dense material. Conference interpreting students are encouraged to set their sights on the world's largest international bodies for their future careers. In the European Union, it's typical for statements, especially those delivered to the plenary, to be written carefully and efficiently. While succinct, they also have a high density of information. These highly structured conference settings where everything is well prepared ahead of time could become increasingly

difficult if the influence of LLMs on language extends into political discourse (Pimlico Journal 2025). Interpreting students of today will need to begin preparing for these new difficulties if they want to be able to capably respond to this kind of material. Just as it has been demonstrated that people who read large amounts of AI-generated text develop instincts to distinguish it from human-written text, it may be possible for interpreting students to prepare for the changing linguistic landscape by familiarizing themselves with the features of LLM output that are being incorporated into human speech. While this could mean listening to AI-generated output and using it in practice sessions, it might be more effective to practice interpreting the speech of people who use LLMs on a regular basis, or speeches from the academic and scientific environments that have been shown to be most affected up until now. If interpreting students are made aware of these language changes and practice responding to them, they will be more able to adapt to the changes in difficulty that each context demands, even as those demands change in the era of Artificial Intelligence.

References

AI Trends (2026). Llm-stats.com. Accessed February 10, 2026, from https://llm-stats.com/ai-trends

Anderson, B., Galpin, R., & Juzek, T. S. (2025). Model misalignment and language change: Traces of AI-associated language in unscripted spoken English. http://arxiv.org/abs/2508.00238

Anderson, T., & Niu, S. (2025). Making AI-enhanced Videos: Analyzing generative AI use cases in YouTube content creation. Proceedings of the extended abstracts of the CHI Conference on Human Factors in Computing Systems, 1–7. https://doi.org/10.1145/3706599.3719991

Angelelli, C. (2009). Using a rubric to assess translation ability: Defining the construct. In C. V. Angelelli, & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting Studies.* John Benjamins Publishing. https://doi.org/10.1075/ata.xiv

Bartłomiejczyk, M. (2007). Interpreting quality as perceived by trainee interpreters: Self-evaluation. *The Interpreter and Translator Trainer*, *1*(2), 247–267. https://doi.org/10.1080/1750399X.2007.10798760

Bovair, S., & Kieras, D. E. (1985). A guide to propositional analysis for research on technical prose. In B. K. Britton & J. B. Black (Eds.), *Understanding expository text* (1st ed.), 315–362. Routledge. https://doi.org/10.4324/9781315099958-12

Cava, L. L., Aiello, L. M., & Tagarelli, A. (2025). Machines in the crowd? Measuring the footprint of machine-generated text on Reddit (No. arXiv:2510.07226). arXiv. https://doi.org/10.48550/arXiv.2510.07226

Chiaro, D., & Nocella, G. (2004). Interpreters' perception of linguistic and non-linguistic factors affecting quality: A survey through the World Wide Web. *Meta*, *49*(2), 278–293. https://doi.org/10.7202/009351ar

Dam, H. V. (2001). On the option between form-based and meaning-based interpreting: The effect of source text difficulty on lexical target text form in simultaneous interpreting. In: *The Interpreters' Newsletter*, *11*, 27–55. EUT Edizioni Università di Trieste.

Fantinuoli, C. (2023). Towards AI-enhanced computer-assisted interpreting. In G. Corpas Pastor & B. Defrancq (Eds.). *IVITRA Research in Linguistics and Literature*, *37*, 46–71. John Benjamins Publishing Company. https://doi.org/10.1075/ivitra.37.03fan

Geng, M., & Trotta, R. (2025). Human-LLM coevolution: Evidence from academic writing. https://doi.org/10.48550/arXiv.2502.09606

Geng, M., Chen, C., Wu, Y., Wan, Y., Zhou, P., & Chen, D. (2025). The impact of large language models in academia: From writing to speaking. https://doi.org/10.48550/arXiv.2409.13686

Georgiou, G. P. (2025, November 13). Differentiating between human-written and AI-generated texts using linguistic features automatically extracted from an online computational tool. https://doi.org/10.48550/arXiv.2407.03646

Gieshoff, A. C., & Albl-Mikasa, M. (2024). Interpreting accuracy revisited: A refined approach to interpreting performance analysis. *Perspectives*, *32*(2), 210–228. https://doi.org/10.1080/0907676X.2022.2088296

Gile, D. (2009). *Basic Concepts and Models for Interpreter and Translator Training*. Benjamins Translation Library. https://doi.org/10.1075/btl.8

Giustini, D., & Dastyar, V. (2024). Critical AI literacy for interpreting in the age of AI. *Interpreting and Society*, *4*(2), 196–213. https://doi.org/10.1177/27523810241247259

Goldsmith, J. (2025, June 6). How to use AI to generate interpreting practice speeches and boost your skills. *Techforword*. Accessed February 10, 2026, from www.techforword.com/blog/ai-interpreting-practice-speeches

Grammarly. (2025, April 9). Decoding AI language: Common words and phrases in AI-generated content. Retrieved February 10, 2026, from https://www.grammarly.com/blog/ai/common-ai-words/

Guo, M., Xie, Y., Han, L., Lei, V. L. C., & Li, D. (2025). Bridging traditional and AI-assisted simultaneous interpreting: Empirical insights for curriculum design. *The Interpreter and Translator Trainer*, *19*(3–4), 425–443. https://doi.org/10.1080/1750399X.2025.2533007

Hanley, H. W. A., & Durumeric, Z. (2024). Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*, 542–556. https://doi.org/10.1609/icwsm.v18i1.31333

Hern, A. (2024, April 16). TechScape: How cheap, outsourced labour in Africa is shaping AI English. *The Guardian*. Retrieved February 10, 2026, from

https://www.theguardian.com/technology/2024/apr/16/techscape-ai-gadgest-humane-ai-pin-chatgpt

Juzek, T. S., & Ward, Z. B. (2024, December 16). Why does ChatGPT "delve" so much? Exploring the sources of lexical overrepresentation in large language models. https://doi.org/10.48550/arXiv.2412.11385

Khorami Nia, F., & Modarresi, G. (2019). A Rasch-based validation of the evaluation rubric for consecutive interpreting performance. *Sendebar*, *30*, 221–244. https://doi.org/10.30827/sendebar.v30i0.8512

Kousha, K. & Thelwall, M. (2025). How much are LLMs changing the language of academic papers after ChatGPT? A multi-database and full text analysis. https://doi.org/10.48550/arXiv.2509.09596

Kujur, A. (2025, November 2025). A comparative analysis of AI-generated and human-written text: Linguistic patterns, detection accuracy, and implications for modern communication. http://dx.doi.org/10.2139/ssrn.5833302

Lee, J. (2008). Rating scales for interpreting performance assessment. *The Interpreter and Translator Trainer*, *2*(2), 165–184. https://doi.org/10.1080/1750399X.2008.10798772

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews (No. arXiv:2403.07183). arXiv. https://doi.org/10.48550/arXiv.2403.07183

Liu, M., & Chiu, Y.-H. (2009). Assessing source material difficulty for consecutive interpreting: Quantifiable measures and holistic judgment. *Interpreting*, *11*(2), 244–266. https://doi.org/10.1075/intp.11.2.07liu

McDonough, M. (2026, January 6). Large language model. *Encyclopedia Britannica*. Retrieved February 10, 2026, from https://www.britannica.com/topic/large-language-model

Merrill, J.B. Chen, S.Y., & Kumer, E. (2025, November 13). How to detect text from ChatGPT? Look for these emojis and other tells. *Washington Post*. Retrieved February 10, 2026, from https://www.washingtonpost.com/technology/interactive/2025/how-detect-chatgpt-em-dash/

Moser-Mercer, B. (1996) Quality in interpreting: Some methodological issues. *The Interpreters' Newsletter*, *7*, 43–55. Trieste, Edizioni LINT. http://hdl.handle.net/10077/8990

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, *57*(10), 265. https://doi.org/10.1007/s10462-024-10903-2

Paredes, J.L., Smith, E., Druck, G., Benson, B. (2024, May 8). More articles are now created by Ai than humans. *Graphite*. Retrieved February 10, 2026, from graphite.io/five-percent/more-articles-are-now-created-by-ai-than-humans.

Pimlico Journal. (2025, September 1). MPs are almost certainly using chatgpt to generate commons speeches. Retrieved February 11, 2026, from https://www.pimlicojournal.co.uk/p/mps-are-almost-certainly-using-chatgpt

Pöchhacker, F. (2002). Quality assessment in conference and community interpreting. *Meta*, *46*(2), 410–425. https://doi.org/10.7202/003847ar

Pugalia, R. (2026, January 5). List of 300+ AI words, phrases and sentences to avoid (2026). *Content Beta*. Retrieved February 10, 2026, from https://www.contentbeta.com/blog/list-of-words-overused-by-ai/

Readability Checker – Readability Calculator. (n.d.). Retrieved April 29, 2026, from https://charactercalculator.com/readability-checker/

Reinhart, A., Markey, B., Laudenbach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, *122*(8), e2422455122. https://doi.org/10.1073/pnas.2422455122

Russell, J., Karpinska, M., & Iyyer, M. (2025). People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text (arXiv:2501.15654). arXiv. https://doi.org/10.48550/arXiv.2501.15654

Shawky, N. (2025, October 6). Common AI words to avoid if you want to bypass AI detectors. *GPTHuman*. Retrieved February 10, 2026, from https://gpthuman.ai/common-ai-words-to-avoid-if-you-want-to-bypass-ai-detectors/

Smith, E., Druck, G., Benson, B. (2024, May 8). How does AI-generated content perform in search and answer engines? *Graphite*. Retrieved February 10, 2026, from https://graphite.io/five-percent/ai-content-in-search-and-llms.

Sun, M., Yin, Y., Xu, Z., Kolter, J. Z., & Liu, Z. (2025). Idiosyncrasies in Large Language Models (No. arXiv:2502.12150). arXiv. https://doi.org/10.48550/arXiv.2502.12150

Sun, Z., Zhang, Z., Shen, X., Zhang, Z., Liu, Y., Backes, M., Zhang, Y., & He, X. (2025). Are we in the AI-generated text world already? Quantifying and monitoring AIGT on social media. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics.* https://doi.org/10.48550/arXiv.2412.18148

Wang, Z., Chu, Z., Doan, T. V., Ni, S., Yang, M., & Zhang, W. (2024). History, development, and principles of Large Language Models: An introductory survey. (arXiv:2402.06853). arXiv. https://doi.org/10.48550/arXiv.2402.06853)

WebFX. (n.d.). *Readability Test.* Retrieved April 29, 2026, from https://www.webfx.com/tools/read-able/

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., Soraperra, I., & Rahwan, I. (2025). Empirical evidence of Large Language Model's influence on human spoken communication. arXiv. https://doi.org/10.48550/arXiv.2409.01754

Yang Jiaming. (2025). Interpreting teaching in AI era: Opportunities and challenges. *Sino-US English Teaching*, *22*(5). https://doi.org/10.17265/1539-8072/2025.05.002

Zamaraeva, O., Flickinger, D., Bond, F., Gómez-Rodríguez, C. (2025). Comparing LLM-generated and human-authored news text using formal syntactic theory. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, *1*, 9041–9060. Association for Computational Linguistics. 10.18653/v1/2025.acl-long.443

Appendix A

Compiled AI Words

Word	Source
additionally	Geng et al. (2025)
advancements	Juzek and Ward (2024)
align	Juzek and Ward (2024) Anderson et al. (2025)
amidst	Reinhart et al. (2025)
boast	Juzek and Ward (2024) - Focal word Anderson et al. (2025)
cacophony	Reinhart et al. (2025)
capabilities	Geng et al. (2025)
camaraderie	Reinhart et al. (2025)
comprehending	Juzek and Ward (2024) - focal word Geng et al. (2025) (comprehensive)
continuation	Reinhart et al. (2025)
crucial	Yakura et al. (2025) Geng et al. (2025)
delve	Yakura et al. (2025) Juzek and Ward (2024)
effectively	Geng et al. (2025)
emphasizing	Juzek and Ward (2024)
enhance	Geng et al. (2025)
fleeting	Reinhart et al. (2025)
garnered	Juzek and Ward (2024)
goodnight	Reinhart et al. (2025)
grapple	Reinhart et al. (2025)
groundbreaking	Juzek and Ward (2024)
ignite	Reinhart et al. (2025)
intricate	Juzek and Ward (2024)

	Reinhart et al. (2025)
necessity	Yakura et al. (2025)
palpable	Reinhart et al. (2025)
pang	Reinhart et al. (2025)
policymaker	Reinhart et al. (2025)
prioritize	Reinhart et al. (2025)
realm	Juzek and Ward (2024)
reminder	Reinhart et al. (2025)
shoutout	Reinhart et al. (2025)
showcase	Juzek and Ward (2024)
significant	Geng et al. (2025) Geng et al. (2025) (significantly) Anderson et al. (2025)
solace	Reinhart et al. (2025)
strategically	Anderson et al. (2025)
surpass	Juzek and Ward (2024) - focal word Anderson et al. (2025)
swiftly	Yakura et al. (2025)
tapestry	Reinhart et al. (2025)
underscore	Yakura et al. (2025) Juzek and Ward (2024) - focal word Reinhart et al. (2025)
unease	Reinhart et al. (2025)
unravel	Reinhart et al. (2025)
unspoken	Reinhart et al. (2025)
valuable	Geng et al. (2025)
vibrant	Reinhart et al. (2025)

Appendix B

Speech Transcripts

N1 full text (original version)

In September of last year, the European Court of Justice ruled that nuclear and natural gas-based electricity can still count as sustainable for the purposes of investing. The reason for this is that some countries do not have the means to transition quickly away from fossil fuels. Thanks to this decision they can still contribute to transitioning away from fossil fuels even though nuclear is not technically a green sustainable source of energy. Slovakia and the Czech Republic are 2 European countries that depend on nuclear power for the majority of their energy. With this court decision in place, they can continue adding nuclear energy capacity and still be working towards a green future as the EU understands it. However, though nuclear power may be better than fossil fuels in terms of emissions, it still causes damage to the environment, particularly through the mining of uranium.

This is not the only example of governments putting softer limits on what can be considered green power.

In the United States, Donald Trump has signed several executive orders boosting coal, keeping coal plants producing electricity for longer. These orders allow coal plants to bypass restrictions on certain chemical emissions for 2 years. This means that they can continue emitting chemicals like mercury, arsenic, and benzene. Donald Trump is branding it as "Clean Coal", but I think it's another example of setting the bar lower for clean energy.

So why is the bar lowering? Why is the definition of sustainable changing? The main reason is an increase in the amount of power needed and projected to be needed in the near future. Data centres put a huge strain on power grids, and we can only see that demand going up. As a consequence of laxer policies, the investment that might have been destined for purely sustainable sources of energy like wind and solar now gets diluted. When you add other sources of energy into the mix, like nuclear for example, companies that want to invest in "green energy" can happily invest in these less sustainable sources. This means that rich countries that have the means to develop true green energy slow down their progress towards it by splitting their attention.

What about other countries that can't afford to invest heavily in solar power or wind power? Maybe nuclear could be a stepping stone to net zero emissions. It's possible, but

uncertain. This option that the EU is offering its countries won't be offered to just anyone, because of the danger of nuclear weapons. That is, the nuclear powers don't want any new players to worry about.

There's a big question here about power dynamics of the energy sector. Powerful countries in the global North are allowed to use nuclear power and call it green, but other countries are being blocked from using nuclear as a first step. It's starting to look like the countries with nuclear weapons also want to be the countries with the electricity of the future. Thank you.

N1 with propositions (Critical; Secondary)

// In September // of last year //, the European Court of Justice ruled that nuclear // and natural gas-based // electricity can still count as sustainable // for the purposes of investing. // The reason for this is that some countries do not have the means to transition quickly away from fossil fuels. // Thanks to this decision // they can still contribute to transitioning away from fossil fuels // even though nuclear is not technically a green sustainable source of energy. // Slovakia and the Czech Republic are 2 European countries that depend on nuclear power // for the majority of their energy. // With this court decision in place, // they can continue adding nuclear energy capacity // and still be working towards a green future // as the EU understands it. // However, // though nuclear power may be better than fossil fuels in terms of emissions, // it still causes damage to the environment, // particularly through the mining of uranium. //

// This is not the only example // of governments putting softer limits on what can be considered green power. //

// In the United States, // Donald Trump has signed several executive orders // boosting coal, // keeping coal plants producing electricity for longer. // These orders allow coal plants to bypass restrictions // on certain chemical emissions // for 2 years. // This means that they can continue emitting chemicals // like mercury, arsenic, and benzene. // Donald Trump is branding it as "Clean Coal", but // I think it's another example of setting the bar lower for clean energy.

// So why is the bar lowering? // Why is the definition of sustainable changing? // The main reason is an increase in the amount of power needed // and projected to be needed in the near future. // Data centres put a huge strain on power grids, // and we can only see that demand going up. // As a consequence of laxer policies, // the investment that might have been destined

for purely sustainable sources of energy // like wind and solar // now gets diluted. // When you add other sources of energy into the mix, // like nuclear for example, // companies that want to invest in "green energy" can happily invest in these less sustainable sources. // This means that rich countries // that have the means to develop true green energy // slow down their progress towards it // by splitting their attention. //

// What about other countries that can't afford // to invest heavily in solar power or wind power? // Maybe nuclear could be a stepping stone to net zero emissions. // It's possible, but uncertain. // This option that the EU is offering its countries won't be offered to just anyone, // because of the danger of nuclear weapons. // That is, the nuclear powers don't want any new players to worry about. //

// There's a big question here about power dynamics of the energy sector. // Powerful countries // in the global North // are allowed to use nuclear power and call it green, // but other countries are being blocked from using nuclear // as a first step. // It's starting to look like the countries with nuclear weapons also want to be the countries with the electricity of the future. // Thank you. //

Critical propositions: 33

Secondary propositions: 32

TOTAL propositions: 65

= 1 proposition / ~7.6 words

N2 (altered version) full text

[Text in bold shows modifications incorporating AI-consistent language features]

In September of last year, the European Court of Justice ruled that, **at least for the purposes of investing,** nuclear and natural gas-based electricity can still **be considered** sustainable. Thanks to this decision, **those countries which do not have the means to transition quickly away from fossil fuels can still begin the transition**, even though **technically speaking**, **nuclear energy is not an entirely renewable energy source. Three European countries that serve as examples are Slovakia, the Czech Republic, and France**, where nuclear power has **surpassed** all other sources in electricity production. With this court decision in place, they can continue adding nuclear energy capacity and still **maintain their alignment to** sustainability **as defined by the** EU.

However, **although** nuclear power may be better than fossil fuels in terms of emissions, it still causes damage to the environment, particularly through the **mining of uranium**, **the release of heated water into ecosystems, and the generation of radioactive waste.**

Regardless of the environmental impact, **there are other examples** of **policymakers** putting softer limits on what can be considered green power.

A second example comes from the United States. **Last year**, Donald Trump signed several executive orders boosting coal. **Not only do these orders** keep coal plants producing electricity for longer, **but they also** allow coal plants to bypass restrictions on certain chemical emissions for 2 years. This means that they can continue emitting chemicals like mercury, arsenic, and benzene. Donald Trump **underscores** that this is "Clean Coal", but this is another example of setting the bar lower for clean energy.

That the bar is lowering is clear, but why is this, **and** why **does there appear to be a change in the definition** of sustainable? One reason **is an increase in the demand for energy, both currently and projected for the near future**. Data centres put a **significant** strain on power grids, and **that demand is only** going up. As a consequence of laxer policies, **some** investment that might have been destined for purely sustainable sources of energy like **wind, solar, and hydro** now gets **diverted**. **Adding** nuclear power **and even coal to the set of green sources** allows companies that want to invest in "green energy" to **prioritize these** less sustainable sources, **meaning** that rich countries that have the **capability** to develop true green energy slow down their progress towards it by splitting their attention.

With that said, there are countries **that lack the economic capacity** to invest heavily in **solar, wind, or hydro** power. For them, **investing** in nuclear power could be a **crucial** stepping stone towards net zero emissions. **However,** this possibility is uncertain, **because the option provided to EU member countries is not available to all**. **The possibility of the development of nuclear weapons poses a danger that the world's nuclear powers don't want to invite by allowing new players**.

In conclusion, the intricate power dynamics of the energy sector are giving rise to many uncertainties. **Not only are** powerful countries in the global North allowed to **strategically** use nuclear power and call it green, **but they are also blocking** other countries from using nuclear as a first step. It's starting to look like the countries **controlling** nuclear weapons also want to be the countries **controlling** the electricity of the future.

N2 (altered version) with propositions (Critical; Secondary)

[Text in bold shows modifications incorporating AI-consistent language features]

In September // of last year, // the European Court of Justice ruled that, // **at least for the purposes of investing,** // nuclear // and natural gas-based // electricity can still **be considered** sustainable. // Thanks to this decision, // **those countries** // **which do not have the means to transition quickly away from fossil fuels** // **can still begin the transition**, // even though // **technically speaking**, // **nuclear energy is not an entirely renewable energy source. // Three European countries that serve as examples** // **are Slovakia, the Czech Republic, and France**, // where nuclear power has **surpassed** all other sources in electricity production. // With this court decision in place, // they can continue adding nuclear energy capacity // and still **maintain their alignment to** sustainability // **as defined by the** EU. // However, // **although** nuclear power may be better than fossil fuels in terms of emissions, // it still causes damage to the environment, // particularly through the **mining of uranium**, // **the release of heated water into ecosystems, // and the generation of radioactive waste.** //

Regardless of the environmental impact, // **there are other examples** // of **policymakers** putting softer limits on what can be considered green power. // //**A second example comes from** the United States. // **Last year**, // Donald Trump signed several executive orders boosting coal. // **Not only do these orders** keep coal plants producing electricity for longer, // **but they also** allow coal plants to bypass restrictions // on certain chemical emissions // for 2 years. // This means that they can continue emitting chemicals like mercury, arsenic, and benzene. // Donald Trump **underscores** that this is "Clean Coal", // but this is another example of setting the bar lower for clean energy. //

// **That the bar is lowering is clear**, // but why is this, // **and** why **does there appear to be a change in the definition** of sustainable? // One reason **is an increase in the demand for energy,** // **both currently and projected for the near future**. // Data centres put a **significant** strain on power grids, // and **that demand is only** going up. // As a consequence of laxer policies, // **some** investment that might have been destined for purely sustainable sources of energy // like **wind, solar, and hydro** // now gets **diverted**. // **Adding** nuclear power // **and even coal** // **to the set of green sources** // allows companies that want to invest in "green

energy" to **prioritize these** less sustainable sources, // **meaning** that rich countries // that have the **capability** to develop true green energy // slow down their progress towards it // by splitting their attention. //

// **With that said,** // there are countries **that lack the economic capacity** // to invest heavily in **solar, wind, or hydro** power. // For them, // **investing** in nuclear power could be a **crucial** stepping stone // towards net zero emissions. // **However**, // this possibility is uncertain, // **because the option provided to EU member countries is not available to all**. // **The possibility of the development of nuclear weapons poses a danger** // **that the world's nuclear powers don't want to invite** // **by allowing new players**. //

// In conclusion, // the intricate power dynamics of the energy sector are giving rise to many uncertainties. // **Not only are** powerful countries in the global North allowed to **strategically** use nuclear power and call it green, // **but they are also blocking** other countries from using nuclear as a first step. // It's starting to look like the countries **controlling** nuclear weapons also want to be the countries **controlling** the electricity of the future. //

Critical propositions: 42

Secondary propositions: 34

TOTAL propositions: 76

= 1 proposition / ~7.1 words

T1 (unaltered version) full text

Teenagers are being banned from social media in many countries around the world. The Spanish government is currently waiting for parliamentary approval for a ban of everyone under 16 years old, and many other European countries are doing the same. But if you look at the world map of all the countries that are putting these bans in place, the continent of Africa is very empty. So far only Nigeria, Egypt, and South Africa have started discussing putting a ban in place for teenagers on social media.

This tells us a lot. Africa is the youngest continent in the world. According to the United Nations, approximately 60% of the African population is under 25 years old. You might think that African countries would be more likely to have bans on teenage social media users, because there are so many of them. But the reality is that Africa as a continent has far less Internet access than the rest of the world. This is a well-known phenomenon known as the digital gap between

Africa and the rest of the world. I wonder: could the bans on social media in other countries have the unintended consequence of reducing this digital gap? Perhaps we could see a future where companies invest more money into infrastructure and connectivity in Africa because that's where they can find more users among youth. Perhaps this could lead to widespread digital literacy in Africa.

We could also see a dark side of all of this, of course. But the Internet and social media have been powerful ways for youth to organize and build democracy in Africa recently. An example of this is in the Gen Z protests in Kenya in 2024. Young people used social media to educate each other on the politics of the day and on a controversial bill that was being debated in parliament. Social media is a powerful organizing tool for activists all over the world, and young people don't want to be left out.

Before I close, it's important to point out that social media is a double-edged sword. It can be used as a way to isolate people from each other, but it can also be a way to build community between people that otherwise wouldn't interact. The internet is often a tool of exploitation, but it can also be a tool of solidarity. I think that the problems young people are facing go much deeper than a simple addiction to their phones. And I'm not sure if these bans are going to solve those problems.

Only time will tell what this all means for the new generation. Depending on which countries and how many of them put bans in place, we could see some very different trends begin to emerge among young people. If these bans become widespread, it'll certainly be interesting to see how Gen Z and Gen Alpha compare in 10 or 20 years.
Thank you.

T1 (unaltered version) with propositions (Critical; Secondary)

Teenagers are being banned from social media // in many countries // around the world. // The Spanish government is currently waiting for parliamentary approval // for a ban of everyone under 16 years old, // and many other European countries are doing the same. // But if you look at the world map of all the countries that are putting these bans in place, // the continent of Africa is very empty.// So far only Nigeria, Egypt, and South Africa have started discussing putting a ban in place // for teenagers on social media. //

// This tells us a lot. // Africa is the youngest continent in the world. // According to the United Nations, approximately 60% of the African population is under 25 years old. // You might think that African countries would be more likely to have bans on teenage social media users, // because there are so many of them. // But the reality is that Africa as a continent has far less Internet access than the rest of the world. // This is a well-known phenomenon // known as the digital gap between Africa and the rest of the world. // I wonder: // could the bans on social media in other countries have the unintended consequence of reducing this digital gap? // Perhaps we could see a future where companies invest more money into infrastructure and connectivity in Africa // because that's where they can find more users among youth. // Perhaps this could lead to widespread digital literacy in Africa. //

// We could also see a dark side of all of this, of course. // But the Internet and social media have been powerful ways for youth to organize and build democracy in Africa recently. // An example of this is in the Gen Z protests in Kenya in 2024. // Young people used social media to educate each other on the politics of the day // and on a controversial bill that was being debated in parliament. // Social media is a powerful organizing tool for activists all over the world, // and young people don't want to be left out. //

Before I close, // it's important to point out that social media is a double-edged sword.// It can be used as a way to isolate people from each other, // but it can also be a way to build community between people that otherwise wouldn't interact. // The internet is often a tool of exploitation, // but it can also be a tool of solidarity. // I think that the problems young people are facing go much deeper than a simple addiction to their phones. // And I'm not sure if these bans are going to solve those problems. //

// Only time will tell what this all means for the new generation. // Depending on which countries and how many of them put bans in place, // we could see some very different trends begin to emerge among young people. // If these bans become widespread, it'll certainly be interesting to see how Gen Z and Gen Alpha compare in 10 or 20 years. //

Thank you. //

Critical propositions: 22

Secondary propositions: 21

TOTAL propositions: 43

= 1 proposition / ~11.3 words

T2 (altered version) full text

[Text in bold shows modifications incorporating AI-consistent language features]

In many countries around the world, policies banning teenagers from social media are being adopted. In Spain, **policymakers** are currently waiting for parliamentary approval for a ban of everyone under 16 years old, and many other European countries are doing the same. **However**, **a glance at a** world map **showing** all the countries that are putting these bans in place reveals that the continent of Africa is not following suit. So far, only Nigeria, Egypt, and South Africa have started discussing putting a **comprehensive** ban in place for teenagers on social media.

This difference is worth paying attention to. As a continent, Africa **boasts** the youngest population in the world, **with approximately 60% of the population under 25 years old, according to the United Nations**. **As a result, it would be easy to** think that African countries would be more likely to have bans on **the multitude of** teenage social media users. **However,** the reality is that **the continent of** Africa has **significantly** less internet access than the rest of the world. **That this digital gap between Africa and the rest of the world exists is well-documented**. **Could** the bans on social media in other countries have the unintended consequence of reducing this digital gap? **There might be** a future where, **to take advantage of larger user bases among youth in Africa,** companies **prioritize** investment in **infrastructure, greater connectivity, and digital competencies**, **leading** to more widespread digital literacy **in the region**.

Of course, there is a dark side to all of this, but the **internet, social media, and other digital tools** have been **crucial** for **young people** to organize and build democracy in Africa recently. An example of this is in the Gen Z protests in Kenya in 2024. **Not only** did young people use social media **strategically** to educate each other on the politics of the day, **but they also** raised awareness on a controversial bill that was being debated in parliament. **Speaking more broadly**, social media is a **valuable** organizing tool for activists all over the world, and young people don't want to be left out.

At the end of the day, it's important to **underscore** that social media is a double-edged sword. It can be used **to** isolate people from each other, but it can also be an **effective** way to build community between people that otherwise wouldn't interact. The internet is often a tool of

exploitation, but it can also be a tool of solidarity. The problems young people are facing go much deeper than **phone addiction, doomscrolling, and exposure to mature content**. I**t is unclear whether** these bans are truly a **groundbreaking** solution to these **intricate** problems. **In conclusion**, only time will tell what **consequence these bans will bring** for the new generation. It will depend on which countries, and how many of them, put bans in place. If they become widespread, they **will not only have an impact on trends among young people but also have an interesting effect on how** Gen Z and Gen Alpha compare in 10 or 20 years. Thank you.

T2 (altered version) with propositions (Critical; Secondary)

[Text in bold shows modifications incorporating AI-consistent language features]

// **In many countries** // **around the world,** // **policies banning teenagers from social media are being adopted.** // **In Spain, policymakers** are currently waiting for parliamentary approval for a ban // of everyone under 16 years old, // and many other European countries are doing the same. // **However, a glance at a** world map **showing** all the countries that are putting these bans in place // reveals that the continent of Africa is not following suit. // So far, only Nigeria, Egypt, and South Africa have started discussing putting a **comprehensive** ban // in place for teenagers on social media. //

// **This difference is worth paying attention to**. // As a continent, Africa **boasts** the youngest population in the world, // **with approximately 60% of the population under 25 years old, according to the United Nations**. // **As a result,** // **it would be easy to** think that African countries would be more likely to have bans // on **the multitude of** teenage social media users.// **However,** the reality is that **the continent of** Africa has **significantly** less internet access than the rest of the world. // **That this digital gap between Africa and the rest of the world exists is well-documented**. // **Could** the bans on social media in other countries have the unintended consequence of reducing this digital gap? // **There might be** a future // where, **to take advantage of larger user bases among youth in Africa,** // companies **prioritize** investment in **infrastructure, greater connectivity, and digital competencies**, // **leading** to more widespread digital literacy **in the region**. //

// **Of course, there is** a dark side to all of this, // but the **internet, social media, and other digital tools** have been **crucial** for **young people** to organize and build democracy in Africa

recently. // An example of this is in the Gen Z protests in Kenya in 2024. // **Not only** did young people use social media **strategically** to educate each other on the politics of the day, // **but they also** raised awareness on a controversial bill that was being debated in parliament. // **Speaking more broadly**, social media is a **valuable** organizing tool for activists all over the world, // and young people don't want to be left out. //

At the end of the day, // it's important to **underscore** that social media is a double-edged sword. // It can be used **to** isolate people from each other, // but it can also be an **effective** way to build community between people that otherwise wouldn't interact. // The internet is often a tool of exploitation, // but it can also be a tool of solidarity. // The problems young people are facing go much deeper than **phone addiction, doomscrolling, and exposure to mature content**. // **It is unclear whether** these bans are truly a **groundbreaking** solution to these **intricate** problems. //

In conclusion, // only time will tell what **consequence these bans will bring** for the new generation. // It will depend on which countries, and how many of them, put bans in place. // If they become widespread, they **will not only have an impact on trends among young people** // **but also have an interesting effect on how** Gen Z and Gen Alpha compare in 10 or 20 years. //

Thank you. //

Critical propositions: 23

Secondary propositions: 21

TOTAL propositions: 44

= 1 proposition / ~11.6 words

Appendix C

Tables and Graphs

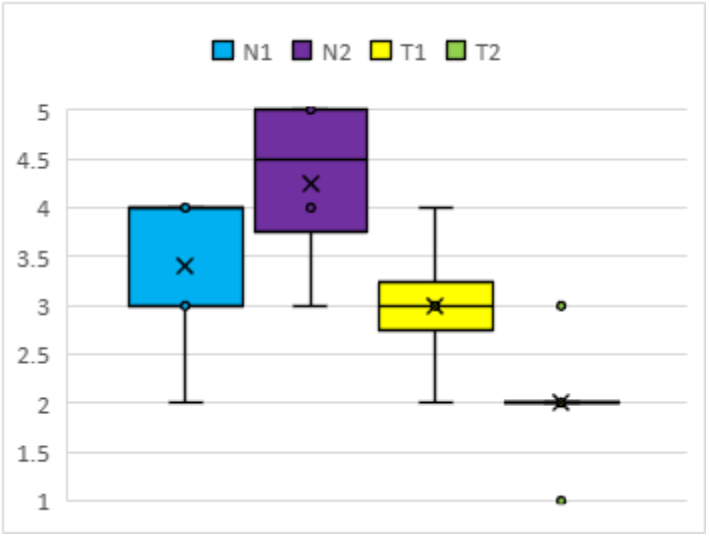


Figure C1. *Perceived Difficulty by Speech*

Interpreter (speech)	Perceived difficulty (1=very easy; 5=very difficult)	Reported characteristics contributing to difficulty	Compared to speeches from class
1a (N1)	2	Well structured sentences, accessible vocabulary, well read	Similar
1b (T2)	1	Slow, general knowledge	Easier
2a (N2)	4	Dense structure, normal speed, surprising topic	Similar
2b (T1)	3		Similar
3a (T1)	4	Speed, dense structure	Similar
3b (N2)	5	Speed, dense vocabulary, complex topic	Harder
4a (T2)	3	Few empty ideas, clear speech, accessible clear speed	Similar

Interpreter (speech)	Perceived difficulty (1=very easy; 5=very difficult)	Reported characteristics contributing to difficulty	Compared to speeches from class
4b (N1)	4	Complex and specialized ideas, good speed and speaking, easy to follow	Similar
5a (N1)	4	requires close concentration, not slow	Similar
5b (T2)	2	Clear argument, current topic, well organized, simple vocabulary	Easier
6a (N2)	3	Technical vocabulary	Similar
6b (T1)	2	Simple vocabulary, clear structure, visual	Easier
7a (T1)	3	Low volume	Easier
7b (N2)	5	High density of terms	Harder
8a (T2)	2	Clear, structured, easy to follow logic, known vocabulary	Similar
8b (N1)	4	High density of ideas	Similar
9a (N1)	3	Speed, vocabulary	Similar
9b (T2)	2	Logical structure, slower speed	Easier

Table C1. *Interpreter Ratings of Source Material Difficulty*

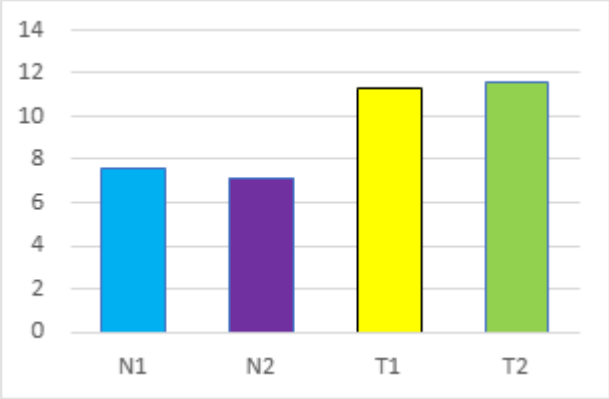


Figure C2. *Words per Proposition*

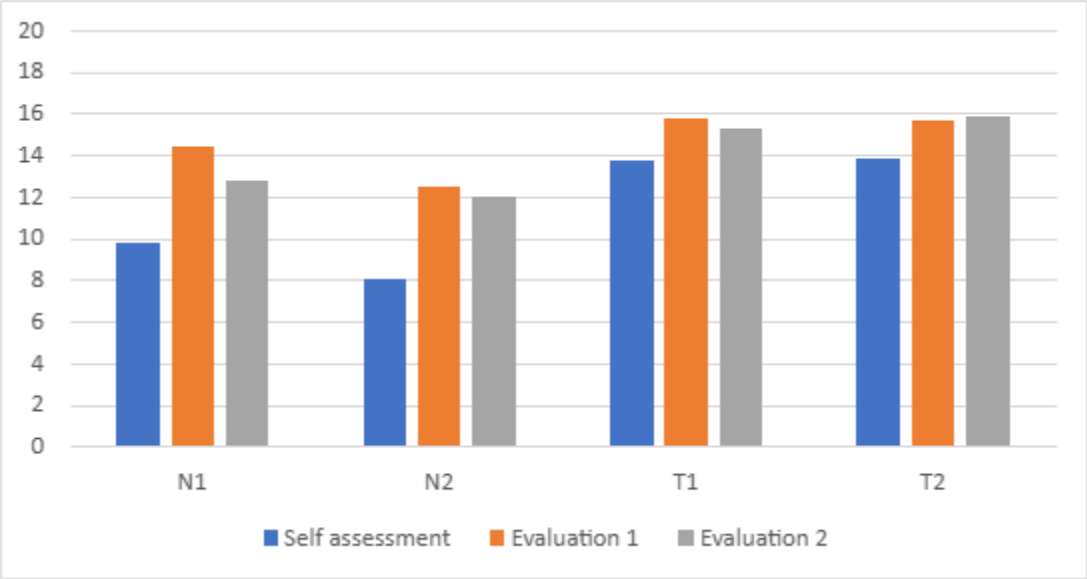


Figure C3. *Average Interpreter Performance by Speech*

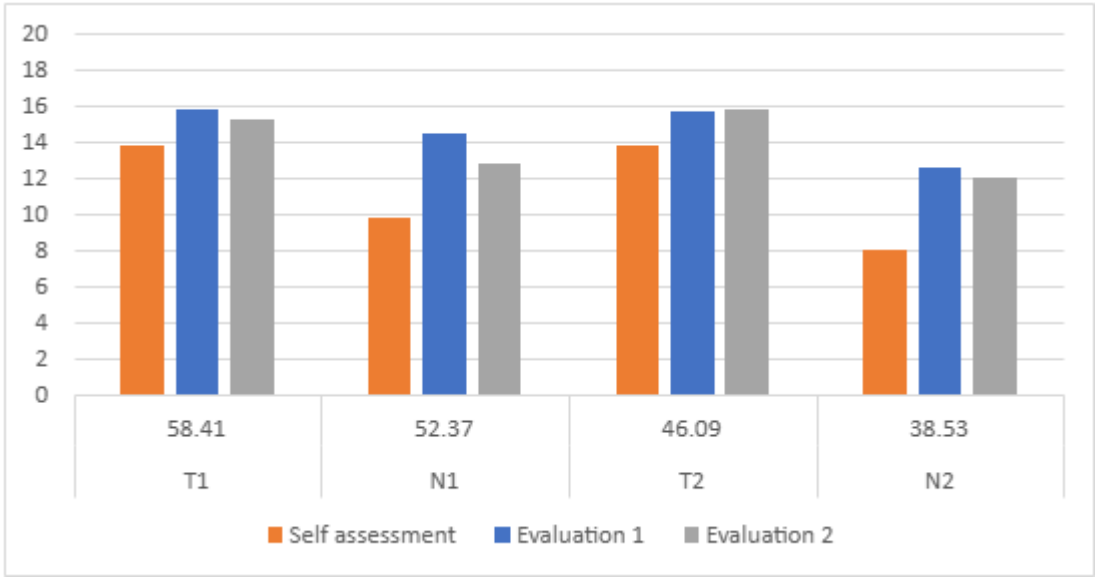


Figure C4. *Average Interpreter Performance by Flesch-Kincaid Score*

In C5, C6, and C7, each data point represents the interpretation of one speech. Some points do not appear because they overlap precisely with the data of another interpretation.

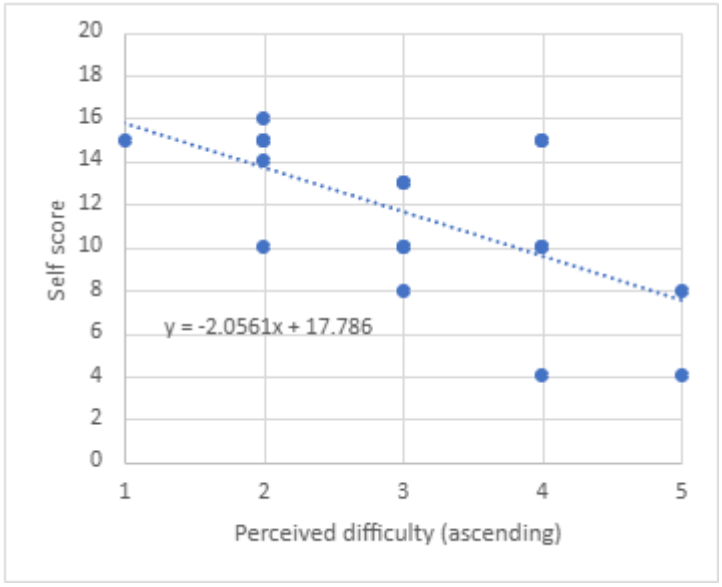


Figure C5. *Self Score as Function of Perceived Difficulty*

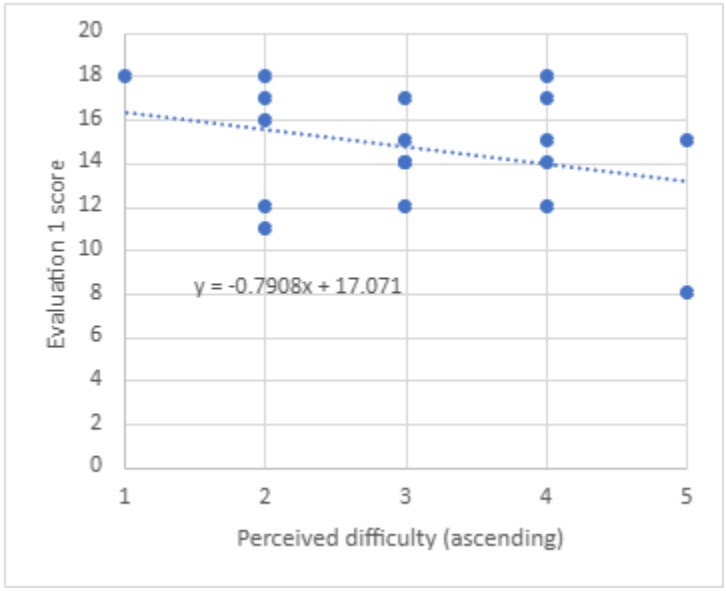


Figure C6. *Rubric Evaluation 1 as Function of Perceived Difficulty*

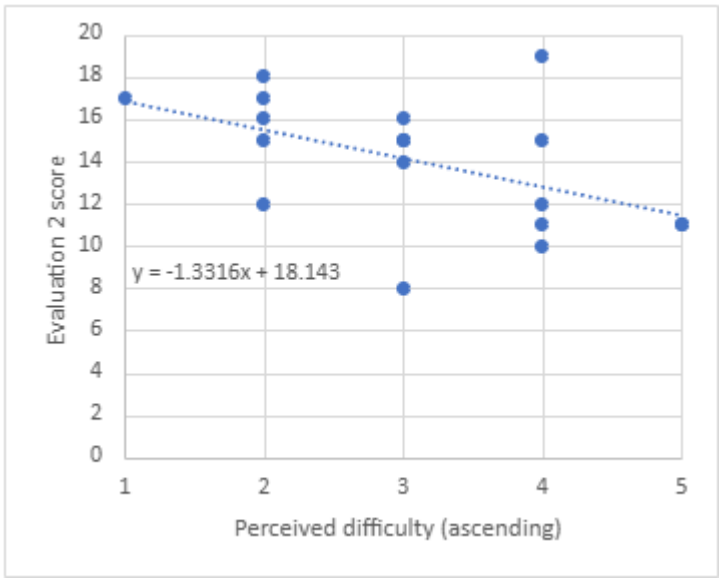


Figure C7. *Rubric Evaluation 2 as Function of Perceived Difficulty*

Interpreter	Method	N1	N2	T1	T2
	Self	10			15
I1	Eval 1	11			18
	Eval 2	12			17
	Self		10	13	
I2	Eval 1		15	15	
	Eval 2		11	8	
	Self		8	15	
I3	Eval 1		15	17	
	Eval 2		11	19	
	Self	4			8
I4	Eval 1	12			14
	Eval 2	10			14
	Self	10			16
I5	Eval 1	18			12
	Eval 2	15			17
	Self		10	14	
I6	Eval 1		12	17	
	Eval 2		15	18	
	Self		4	13	
I7	Eval 1		8	14	
	Eval 2		11	16	
	Self	15			15
I8	Eval 1	14			16
	Eval 2	12			16
	Self	10			15
I9	Eval 1	17			18
	Eval 2	15			15

Table C2. *Comparison of Self Score and Rubric-Based Evaluations by Interpreter and Speech*

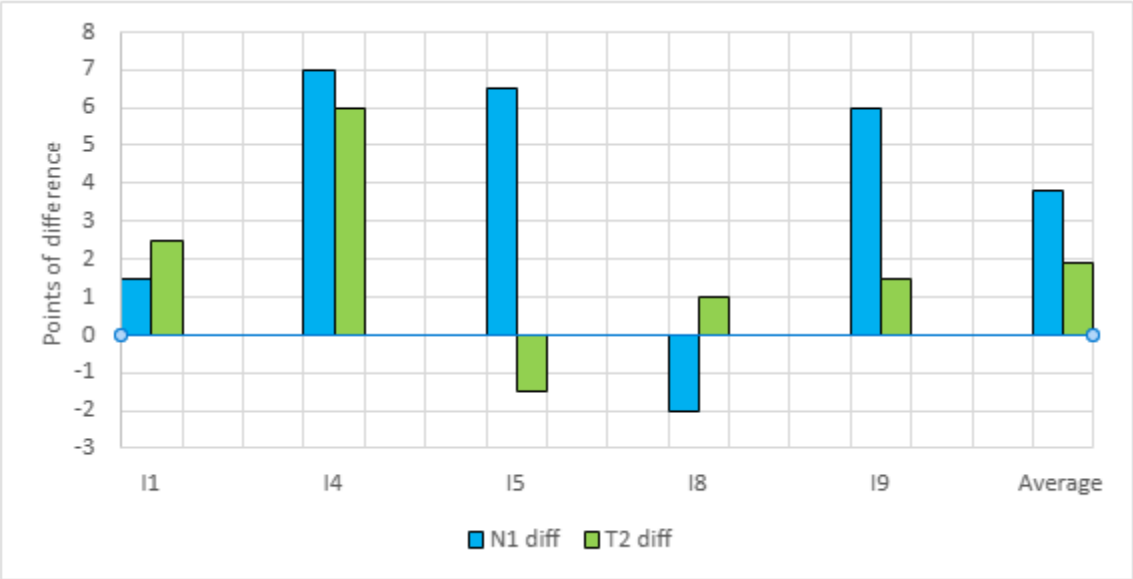


Figure C8. *Difference Between Self and Average Rubric Evaluation – N1, T2*

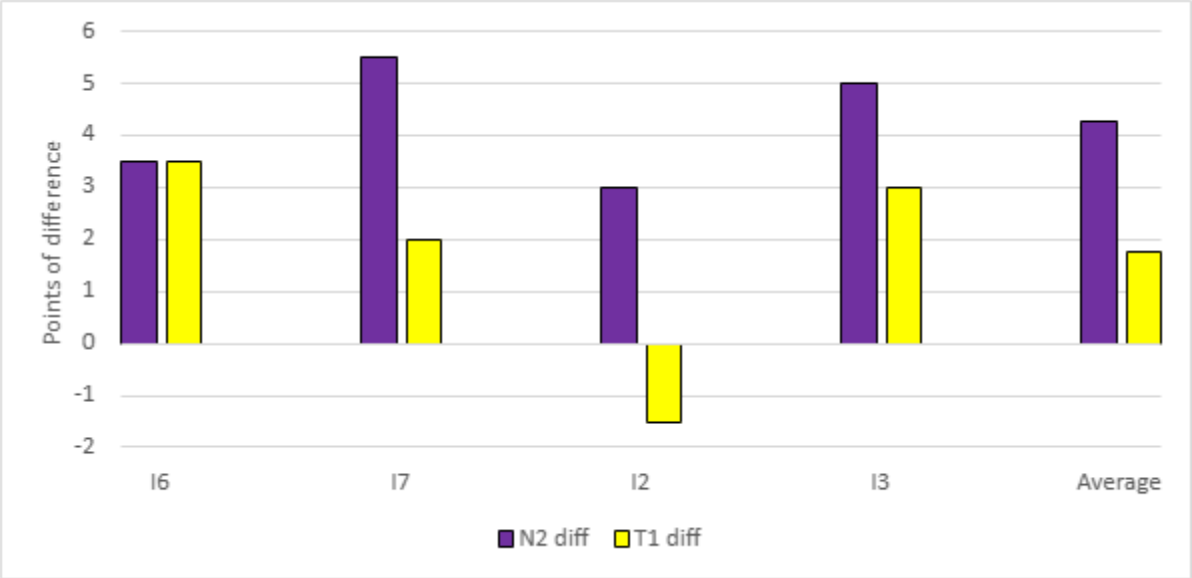


Figure C9. *Difference Between Self and Average Rubric Evaluation – N2, T1*

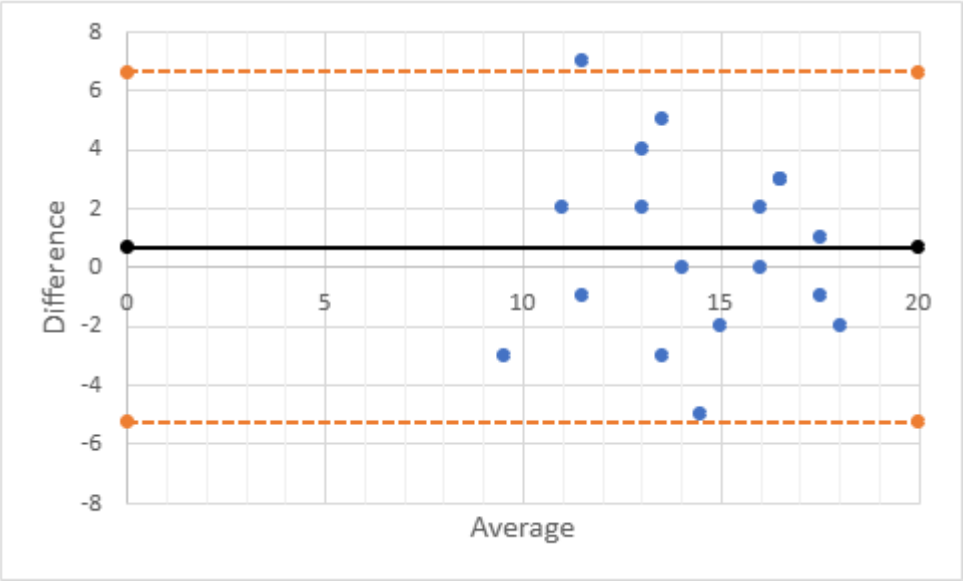


Figure C10. *Bland-Altman Plot Comparing Rubric Evaluations*

Average difference between evaluation scores: 0.7

Standard deviation: 3.0

Upper limit: 6.62

Lower limit: 5.29

Appendix D

Readability Scores Comparison

Readability Test	N1 Result	N2 Result	T1 Result	T2 Result
Flesch-Kincaid	52.37 (Grade 10-12) – "Fairly difficult to read"	38.53 (College) – "Difficult to read"	58.41 (Grade 10-12) – "Fairly difficult to read"	46.09 (College) – "Difficult to read"
Gunning Fog	13.65 (College freshman) – "Difficult to read"	17.03 (College graduate) – "Very difficult to read"	13.70 (College freshman) – "Difficult to read"	16.42 (College senior) – "Difficult to read"
SMOG	12.83 (Grade 7) – "Plain English"	15.22 (Undergraduate) – "Difficult to read"	12.90 (Grade 7) – "Plain English"	14.85 (Undergraduate) – "Difficult to read"
Automated Readability Index (ARI)	16.06 (College student) – "Age 18–22"	19.74 (College student) – "Age 18-22"	13.82 (Grade 12) – "Age 17–18"	16.94 (College student) – "Age 18-22"
Dale-Chall	6.10 (Grade 7-8) – "Fairly easy to read"	6.97 (Grade 7-8) – "Fairly easy to read"	5.21 (Grade 5-6) – "Easy to read"	6.32 (Grade 7-8) – "Fairly easy to read"

Table D1. *Character calculator (Readability Checker n.d.)*

Metric	N1 Result	N2 Result	T1 Result	T2 Result
Flesch-Kincaid Reading Ease	54.1	40.3	57.9	46.1
Flesch-Kincaid Grade Level	10.3	13.5	9.7	12.1
Gunning Fog Score	13.1	16.4	12.4	15.5
SMOG Index	9.7	12.1	9.9	11.8
Coleman Liau Index	12.8	14.3	10.3	12.1
Automated Readability Index	10.8	14.4	8.5	11.5
Sentences	26	23	27	24
Word count	485	542	488	510

Metric	N1 Result	N2 Result	T1 Result	T2 Result
Complex Words	73	101	78	100
Percentage Complex Words	15.05%	18.63%	15.98%	19.61%
Average Words per Sentence	18.65	23.57	18.07	21.25
Average Syllables per Word	1.58	1.69	1.54	1.65
Hard Sentences	11	8	12	11
Very Hard Sentences	3	9	2	5
Passive Voice Sentences	5	3	5	5
Adverbs Used	8	11	7	12

Table D2. *WebFX Readability Tool (WebFX n.d.)*

Appendix E

Participant Questionnaire

Para los fines de un estudio de caso sobre interpretación simultánea, responda a las siguientes preguntas. Su respuesta será anónima y se utilizará únicamente por parte de la persona investigadora

Instrucciones para intérpretes:

1. **Interprete el primer discurso**
2. **Inmediatamente después de haber terminado, rellene el primer cuestionario**
3. **Una vez completado el cuestionario, puede tomar hasta 5 minutos de descanso**
4. **Interprete el segundo discurso y rellenar el segundo cuestionario**
5. **Ha terminado el experimento. Puede añadir comentarios adicionales sobre los discursos o la experiencia en general en este apartado.**

Comentarios adicionales:

DISCURSO 1

Material				
¿Qué grado de dificultad le atribuiría al discurso? (1 = muy fácil, 5 = muy difícil)				
1	**2**	**3**	**4**	**5**
¿Qué características del discurso contribuyeron a su respuesta?				

¿Cómo compararía la dificultad del discurso con los discursos practicados en clase?		
Más fácil	**Más o menos igual**	**Más difícil**

Interpretación				
¿Cómo evaluaría su interpretación según los siguientes criterios? (1 = muy insatisfactoria, 5 = muy satisfactoria)				
Contenido (exactitud de las ideas principales, coherencia lógica, exhaustividad de la información, etc.)				
1	**2**	**3**	**4**	**5**
Forma (naturalidad de la producción en la lengua de llegada, terminología, interferencias de la lengua de partida, frases incompletas, etc.)				
1	**2**	**3**	**4**	**5**
Ejecución (fluidez de la locución, seguridad, muletillas y pausas, gestión del tiempo, etc.)				
1	**2**	**3**	**4**	**5**

DISCURSO 2

Material				
¿Qué grado de dificultad le atribuiría al discurso? (1 = muy fácil, 5 = muy difícil)				
1	**2**	**3**	**4**	**5**

¿Qué características del discurso contribuyeron a su respuesta?
¿Cómo compararía la dificultad del discurso con los discursos practicados en clase?
Más fácil **Más o menos igual** **Más difícil**

Interpretación				
¿Cómo evaluaría su interpretación según los siguientes criterios? (1 = muy insatisfactoria, 5 = muy satisfactoria)				
Contenido (exactitud de las ideas principales, coherencia lógica, exhaustividad de la información, etc.)				
1	**2**	**3**	**4**	**5**
Forma (naturalidad de la producción en la lengua de llegada, terminología, interferencias de la lengua de partida, frases incompletas, etc.)				
1	**2**	**3**	**4**	**5**
Ejecución (fluidez de la locución, seguridad, muletillas y pausas, gestión del tiempo, etc.)				
1	**2**	**3**	**4**	**5**

Appendix F

Evaluation Rubric

The rubric used in this study was modified from the design in Lee (2015).

Assessment categories and criteria (checklist for rating) Note. Criteria in each category are arranged in order of importance.	Level (1-5)	LEVELS OF EFFECTIVENESS 5 Complete *ALL* characteristics present 4 Extensive *MOST* characteristics present 3 Moderate *SOME* characteristics present 2 Limited *FEW* characteristics present 1 Zero *NO* characteristics present COMMENTS (FEEDBACK) Assessor's comments for additional feedback to assessee
1. CONTENT (7 criteria) ___ *No opposite meanings ___ *Accurate rendition of main ideas ___ No unjustified change in meaning ___ Logical cohesion ___ High level of completeness of information (except numbers and names) ___ Accurate rendition of numbers and name ___ No unjustified additions		
2. FORM (7 criteria) ___ No incomplete sentences ___ Natural/idiomatic target-language expres ___ Unambiguous and clear diction ___ Appropriate register and speech level ___ Little source-language interference ___ Correct terminology ___ Grammatical correctness		
3. DELIVERY (6 criteria) ___ Fluency of delivery (general concept/impr ___ No significant repairs or backtracking ___ Impression of confidence ___ Few fillers, hesitations and pauses ___ Lively intonation and stress ___ No slips of the tongue		
Total	/15	Total = (CONTENT score×2) + (FORM score×1) + (DELIVERY score×1)

Note. The two asterisked components ("no opposite meanings" and "accurate rendition of main ideas") may be weighted within the category concerned. Only these two criteria were rated above 4.0 (= "very important") in the survey.