International Journal of Language, Translation and Intercultural Communication

Vol 11 (2026)

The Sustainability of Interpreting as a Profession in the Era of Artificial Intelligence

Simultaneous Interpreting and AI Text-Mediated Output

Anastasios Ioannidis

doi: 10.12681/ijltic.44104

Copyright © 2026, ANASTASIOS IOANNIDIS

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0.

To cite this article:

Ioannidis, A. (2026). Simultaneous Interpreting and AI Text-Mediated Output: A Comparative Analysis of English–Greek Institutional Discourse. *International Journal of Language, Translation and Intercultural Communication*, *11*, 86–99. https://doi.org/10.12681/ijltic.44104

https://epublishing.ekt.gr | **e-Publisher:** EKT | **Downloaded at:** 23/05/2026 15:34:53

Simultaneous Interpreting and AI Text-Mediated Output: A Comparative Analysis of English–Greek Institutional Discourse

Anastasios Ioannidis
Ionian University
ioanndis@ionio.gr

Abstract

This study investigates how closely AI-based text-mediated output approximates professional human simultaneous interpreting with respect to target-language output characteristics in a high-stakes institutional context. Using three plenary speeches from the European Parliament, it compares official Greek simultaneous interpretation with AI-generated Greek output produced via a neural machine translation system. The analysis combines proposition-level error annotation with qualitative discourse-oriented analysis. While the AI output displays fewer overall errors, it shows systematic limitations in pragmatic appropriateness, evaluative intensity, terminology, and stylistic naturalness. Human interpreting, by contrast, demonstrates greater sensitivity to contextual cues and audience-oriented reformulation, despite strategic information compression.

Keywords: *artificial intelligence, simultaneous interpreting, speech translation, communicative adequacy*

1 Introduction

Technological innovation has long played a role in shaping the practice of interpreting, but the pace and scope of digital change in recent years have transformed the landscape more profoundly than in any previous period, contributing to what several scholars describe as a technological turn in the interpreting profession (Fantinuoli, 2019). Within this rapidly evolving landscape, recent advances in artificial intelligence (AI)–driven language mediation in spoken discourse have enabled the deployment of Machine Interpreting (MI) systems in an expanding range of communicative environments, including business meetings, healthcare, education, and conference settings (Horváth, 2021: 181).

In practice, these MI systems are implemented through either cascaded architectures or end-to-end models. The former combine automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (ASR → MT → TTS) and remain the dominant configuration, largely due to the relative maturity and robustness of each component (Horváth, 2021: 176; Fantinuoli, 2022: 518-519). At the same time, advances in neural sequence-to-sequence modelling have facilitated the development of end-to-end deep learning architectures, in which an encoder processes incoming speech and a decoder generates target-language output directly from speech input (Cao, 2024: 732; Horváth, 2021: 175). Despite their architectural differences, both configurations are capable of producing fluent and lexically detailed output. As a result, MI systems are increasingly marketed as viable real-time solutions, with industry reports pointing to accelerating commercialisation and adoption (Nimdzi, 2019).

This rapid expansion has, in turn, intensified scholarly and professional debate over whether AI-based interpreting systems might eventually replace, rather than merely support, human interpreters. Against this backdrop, the present study examines how closely current AI-based language mediation (operationalised here through a text-mediated neural MT proxy) approximates professional human simultaneous interpreting in a real institutional context, focusing on target-language output rather than on temporal processing dynamics. By analysing three short plenary speeches from the European Parliament together with their official human interpretation and an AI-generated version, the study investigates quantitative error patterns and qualitative differences in communicative and discourse-level behaviour.

2 Theoretical background

Despite growing scholarly interest in MI, direct comparative studies between human interpreting and AI-generated interpreting remain scarce. One major reason concerns the methodological gap in evaluating the performance of automatic interpreting. MI systems are primarily evaluated using computational metrics such as BLEU, METEOR, TER, or embeddings-based similarity measures (Lu & Han, 2023). While these metrics provide a quick indication of system performance, they are largely designed to capture surface-level similarity and lexical overlap and therefore fail to account for the communicative impact of interpreting output as situated discourse (Sperber et al., 2025). As a result, a divide persists between the quantitative evaluation paradigms of computational linguistics and the qualitative, communicative frameworks traditionally employed in Interpreting Studies – a divide that recent research has increasingly sought to bridge through human-centred evaluation in ecologically valid tasks.

A foundational contribution in this line of research is the communicative evaluation framework proposed by Fantinuoli and Prandi (2021). The authors examine real-time speech translation systems from a communicative perspective by comparing a speech-to-text–based speech translation system with professional simultaneous interpreters, adopting a user-oriented framework based on Tiselius's (2009) intelligibility and informativeness scales. This approach enables the assessment of machine output in relation to human interpreting performance as perceived by listeners, focusing on ease of comprehension and information uptake rather than formal correspondence to the source text or process-level features of simultaneity.

A related strand of research adopts corpus-based approaches, including quantitative metrics and manual analytical methods, to compare human and machine interpreting output. Tan and Gao (2025) analyse outputs produced by three professional human interpreters alongside two commercial AI-based speech translation tools using linguistic indices related to lexis, syntax, readability, and textual cohesion. Their analysis highlights systematic differences in how human and machine systems present information, with particular attention to context-awareness and audience-oriented processing strategies. Similarly, Liu and Liang (2024) and Lee and Cha (2023) compare human interpreting output with MT-generated target texts, drawing respectively on parallel Chinese–English interpreting corpora and Korean–English spoken interview data. Together, these studies reveal consistent differences across lexical, syntactic, and discourse-related levels through multidimensional linguistic indices and a manual error analysis respectively. Their findings show that MT exhibits limitations in handling spoken discourse features, pragmatic intent, and discourse-level meaning compared with human interpretation.

Complementary evidence is also provided by small-scale task-based studies. Cao (2024), for instance, examines human and machine interpreting in a pilot consecutive-mode study based on a controlled task in which interpreting students and an ASR–MT-based commercial system interpret the same source speeches. Relying on human evaluation of communicative adequacy rather than on automatic metrics, the study aligns with human-centred approaches to interpreting assessment and illustrates the feasibility of small-scale, manually evaluated designs grounded in MT-generated output.

Taken together, the studies reviewed above reveal a consistent pattern across languages, interpreting modes, and evaluation designs. While AI-based systems tend to perform strongly in terms of surface-level linguistic processing, such as lexical accuracy and information density, human interpreters consistently demonstrate advantages in communicative mediation. These advantages manifest in higher intelligibility and readability, more coherent discourse organisation, more effective management of contextual and pragmatic constraints, and heightened sensitivity to audience-oriented reformulation.

3 Methodology

In alignment with prior comparative research reviewed above, this study adopts a small-scale, mixed-methods comparative design combining quantitative error annotation with qualitative, discourse-oriented analysis. The empirical material analysed here consists of excerpts from a plenary session of the European Parliament, an institutional setting characterised by high information density, specialised terminology, rhetorical complexity, and strict quality norms governing professional interpreting. These features make it particularly suitable for investigating the limits of AI-based speech mediation in high-stakes contexts. The availability of high-quality, publicly accessible audiovisual data with parallel interpretation channels further enables transparent and ethically sound data collection (Bernardini et al., 2018: 22).

Within this institutional setting, three short excerpts were selected from the same European Parliament plenary session held on 24 November 2025.[1] Each excerpt was drawn from a different debate segment and characterised by distinct institutional functions, thematic domains, and rhetorical demands. Consistent with corpus-based interpreting research (Russo, 2019), the selection was guided by thematic diversity, clear audio quality in both the source and interpretation channels, moderate to high terminological density, and natural variation in speaking rate, accent, and rhetorical style. More specifically, the selected excerpts were the following:

- Excerpt 1: From the debate *Existence of a clear risk of a serious breach by Hungary of the values on which the Union is founded*, delivered by Tineke Strik (rapporteur), with a duration of approximately 5 minutes.
- Excerpt 2: *From the debate Enhancing police cooperation in relation to the prevention, detection and investigation of migrant smuggling and trafficking in human beings; enhancing Europol's support to preventing and combating such crimes*, delivered by Jeroen Lenaers (rapporteur), lasting approximately 4.5 minutes.
- Excerpt 3: From the debate marking the *30th anniversary of the Barcelona Process and the new Pact for the Mediterranean*, delivered by Dubravka Šuica (Member of the European Commission), with a duration of approximately 7.5 minutes.

For each excerpt, three parallel datasets were compiled. First, the English source speeches were transcribed using an automatic speech recognition (ASR) tool (TurboScribe). Although official transcripts are available on the EP website, ASR transcription was deliberately chosen in order to align the source data with the input conditions typically faced by AI-based speech mediation

[1] https://www.europarl.europa.eu/plenary/en/debate-details.html?date=20251124&detailBy=date

pipelines. Disfluencies characteristic of spontaneous speech were retained, and ASR errors were not corrected.

Second, the official Greek simultaneous interpretation produced by a professional EP interpreter was extracted from the Greek interpretation channel of the EP recordings and transcribed following the same procedure. In this case, obvious transcription errors were corrected to ensure that the transcript accurately reflected the interpreter's actual output, thereby establishing a reliable benchmark for comparison.

Third, an AI-generated Greek output was produced using DeepL, a commercial neural machine translation system, applied to the English ASR transcripts. Although DeepL is not designed for automatic simultaneous interpreting platforms, it is employed here as a high-quality text-mediated proxy for AI-based interpreting output. This approach reflects a best-case scenario for the translation component of cascaded speech-to-speech interpreting pipelines, in which ASR output is passed to a text-to-text MT system. In this sense, the use of DeepL in the present study does not aim to model a specific commercial simultaneous interpreting platform nor to assume functional equivalence with human simultaneous interpreting; rather, it enables the examination of target-language output characteristics under unconstrained informational conditions, with latency, incremental decoding, and other time-based constraints excluded.

All three source texts were subsequently segmented into proposition-level analytical units, defined as minimal meaning-bearing segments that can be independently evaluated. Segmentation was based on functional coherence and communicative unity rather than on sentence or clause boundaries. On this basis, the dataset – comprising a combined duration of approximately 17 minutes and a total of 2,152 words in the English source texts – was segmented into 57 proposition-level units across the three excerpts (18, 25, and 14 units respectively).

These units were then analysed using a predefined error taxonomy comprising five categories: (1) omissions; (2) mistranslations and semantic shifts; (3) terminology errors; (4) cohesion and coherence issues; and (5) pragmatic appropriateness issues. These categories are grounded in communicative and discourse-oriented dimensions repeatedly identified in the literature discussed earlier as central to interpreting quality (Fantinuoli & Prandi, 2021; Tan & Gao, 2025; Liu & Liang, 2024; Sperber et al., 2025), thereby functioning as an applied bridge between the human-centred evaluation frameworks and the empirical comparison of human and AI-generated output. Consistent with this focus, the taxonomy excludes variables related to temporal performance, such as latency.

With regard to the annotation procedure, a single-annotator design was adopted given the small-scale and exploratory nature of the study. Both the human and AI outputs were then manually annotated by the author following a consistent and explicitly defined analytical protocol based on the predefined error taxonomy described above. To minimise post-hoc interpretive bias, the protocol was established prior to analysis and applied uniformly across datasets. Accordingly, all annotation decisions were grounded in functional criteria related to communicative impact rather than surface-level deviation from the source text.

Each identified error was further classified as minor or major based on its assessed communicative impact, as determined by its effect on the intended communicative function of the segment. Errors were considered major when they affected core propositional meaning, institutional or legal references, evaluative stance, or pragmatic intent – dimensions consistently highlighted in the literature as crucial for communicative effectiveness in professional interpreting

(Fantinuoli & Prandi, 2021; Horváth, 2021). Minor errors involved surface-level inaccuracies, stylistic infelicities, or omissions that did not substantially compromise discourse coherence or audience comprehension.

4 Results

Drawing on the annotated material described above, this section presents the study's results in two subsections, addressing quantitative and qualitative analyses respectively.

4.1 Quantitative results

The quantitative analysis reveals that, across the dataset, a total of 49 errors were identified in the human interpreting output, compared to 28 errors in the AI-generated output (Table 1). When normalised by source word count, this corresponds to 22.8 errors per 1,000 words for HI and 13.0 errors per 1,000 words for the AI system. At the aggregate level, the human output thus exhibits a higher overall number of annotated errors. However, error counts alone do not capture differences in the functional role or communicative impact of individual error types, an issue examined in detail in the qualitative analysis below.

Table 1: Distribution of errors by category (aggregated counts)

Error Type	**Human Interpreting (HI)**	**AI Output (AI)**
Omissions	31	0
Mistranslations / Semantic Shifts	8	10
Terminology Errors	2	5
Cohesion & Coherence	7	2
Pragmatic appropriateness	0	11
Total	**49**	**28**

The human interpreter's errors are dominated by omissions, which account for well over half of all errors. These are followed by cohesion- and coherence-related issues and a smaller number of mistranslations or semantic shifts. By contrast, the AI output shows no omissions across the analysed excerpts. Instead, its errors cluster in categories related to semantic accuracy, terminology, and pragmatic appropriateness. No errors were coded under the pragmatic category in the human output within the present taxonomy, as audience-oriented adaptations (e.g., intensification, colloquialisation, segmentation) were treated as strategic reformulation rather than as pragmatic errors. Accordingly, the absence of omissions in the AI output results from the classification of evaluative attenuation, metaphor loss, and reductions in rhetorical force under semantic shift or pragmatic appropriateness categories rather than as omissions.

In addition to distributional differences, the two systems also diverge in the severity profile of their errors. Across the three excerpts, the human interpreter produced four major errors, consisting primarily of omissions that resulted in information loss. The remaining HI errors were assessed as minor. The AI system produced three major errors, involving one major terminology error and two major pragmatic inadequacies. When normalised by source word count, this corresponds to 1.9 major errors per 1,000 words for HI and 1.4 for the AI system. While the overall rate of major errors is comparable, their qualitative distribution differs: major HI errors primarily affect information completeness, whereas major AI errors affect institutional reference and pragmatic stance.

Finally, the analysis examined whether AI errors could be attributed to upstream automatic speech recognition (ASR) errors. Cascaded pipelines are structurally vulnerable to error propagation, as recognition errors introduced at the ASR stage may propagate downstream into translation and synthesis. In the present dataset, two AI errors – both in the first excerpt – were clearly traceable to ASR misrecognition, accounting for 7.1% of all AI errors. Although limited in number, these cases illustrate the structural vulnerability of cascaded pipelines to error propagation across processing stages.

4.2 Qualitative analysis

The qualitative analysis examines selected segments from a discourse-oriented perspective, comparing human and AI output in terms of pragmatic intent, evaluative force, terminology, and stylistic naturalness across the three excerpts analysed.

4.2.1 Pragmatic Adequacy and Communicative Intent

The first aspect examined concerns the rendering of the speaker's pragmatic stance and illocutionary intent. In the opening segment of the first speech on breaches of EU values in Hungary, the speaker contrasts her disappointment with the Council against her appreciation of the Commission's presence (Example 1). Through the paraphrastic formulation *αλλά δυστυχώς δεν μπορώ να το πω / but unfortunately, I cannot say it*, the human interpreter reinforces the implication of unmet expectations, conveying a subtly ironic yet clearly critical tone in a manner that is idiomatic and natural in Greek. Similarly, the addition of *που είναι εδώ τουλάχιστον / for being here at least* strengthens the rhetorical contrast between an absent or underperforming Council and a present, supportive Commission. The AI output, by contrast, remains more literal and structurally faithful, but fails to capture the underlying evaluative nuance of the utterance. Although the propositional content is preserved, the unmodulated rendering does not convey the speaker's implicit reproach nor re-establish the intended contrast between the Council and the Commission. In addition, the AI's direct address to the Commissioner as an individual (*κύριε Επίτροπε*) lacks the institutional framing found in the human rendition, where the interpreter opts to thank the institution collectively (*Ευρωπαϊκή Επιτροπή*), thereby reinforcing the rhetorical opposition encoded in the source speech.

Example 1

Original	**Human output**	**AI output**
Thank you, Chair, dear Chair. I wish I could have said thank you, Council, but I'm sorry. But thank you very much, Commissioner, for being here.	Ευχαριστώ, κύριε Πρόεδρε. Θα ήθελα να πω ευχαριστώ και στο Συμβούλιο, αλλά δυστυχώς δεν μπορώ να το πω. Ευχαριστώ την Ευρωπαϊκή Επιτροπή που είναι εδώ τουλάχιστον.	Σας ευχαριστώ, κύριε Πρόεδρε, αγαπητέ Πρόεδρε. Μακάρι να μπορούσα να ευχαριστήσω το Συμβούλιο, αλλά λυπάμαι. Σας ευχαριστώ πολύ, κύριε Επίτροπε, που είστε εδώ.

A second example of pragmatic divergence arises in the handling of gender agreement in the same speech. In Example 2, the human interpreter renders the clause accurately, correctly applying feminine grammatical agreement in line with the gender of the speaker. In Greek, self-referential adjectives must agree with the speaker's gender, reflecting pragmatic sensitivity to speaker-related contextual cues. The AI output, by contrast, fails to make this adjustment and defaults to the masculine form. While this does not impede propositional comprehension, it disrupts interpersonal alignment and diminishes communicative naturalness.

Example 2

Original	Human output	AI output
So, I am proud to present this report…	Χαίρομαι και είμαι περήφανη που σας παρουσιάζω αυτή την έκθεση…	Είμαι λοιπόν περήφανος που παρουσιάζω την παρούσα έκθεση …

Pragmatic adaptation is also evident in the rendering of address forms in the closing appeal of the first speech (Example 3). Here, the human interpreter avoids vocative address altogether and reformulates the appeal as a second-person plural directive, thereby preserving the illocutionary force of the utterance while conforming to Greek discourse norms. In Greek, metonymic references – using the name of an institution to refer to its members – are common in declarative contexts, but are pragmatically infelicitous in vocative form. The AI output, by contrast, retains the vocative *Λοιπόν, Συμβούλιο,… / Well, Council,…*, which, although grammatically well-formed and semantically transparent, is pragmatically unnatural in Greek, as it treats an impersonal institution as a direct addressee.

Example 3

Original	Human output	AI output
The Council itself concluded that more Member-States want to see action. Well then, Council, take it.	Το Συμβούλιο το ίδιο συμπέρανε ότι πιο πολλά κράτη-μέλη ζητούν να αναληφθεί δράση. Ήρθε η ώρα να λάβετε δράση, λοιπόν.	Το ίδιο το Συμβούλιο κατέληξε στο συμπέρασμα ότι περισσότερα κράτη μέλη επιθυμούν να δουν δράση. Λοιπόν, Συμβούλιο, αναλάβετε δράση.

A related pragmatic divergence appears in the third speech on the new Pact for the Mediterranean, where the speaker responds to circulating disinformation on the Erasmus Plus programme (Example 4). While semantically adequate, the verb *να αναφερθώ / to refer to* in the AI rendition reframes the utterance as a neutral act of reference, weakening the corrective and confrontational illocutionary force implied in this context. By contrast, the human interpreter's reformulation (*να διασκεδάσω κάποιες φήμες / to dispel some rumours*) explicitly signals an act of clarification and refutation, more effectively preserving the speaker's communicative intent.

Example 4

Original	Human output	AI output
And let me briefly address the disinformation on Erasmus Plus, which has emerged here in the corridors.	Και επειδή υπάρχει κάποια παραπληροφόρηση για το Εράσμους Συν, θέλω να διασκεδάσω κάποιες φήμες που ακούγονται στους διαδρόμους.	Και επιτρέψτε μου να αναφερθώ εν συντομία στην παραπληροφόρηση σχετικά με το Erasmus Plus, η οποία έχει εμφανιστεί εδώ στους διαδρόμους.

4.2.2 Evaluative Intensity and Rhetorical Attenuation

A further systematic divergence between human and AI output concerns the degree to which evaluative intensity and rhetorical emphasis are maintained or attenuated in the target output. A first illustration comes from the speech on risks to EU values in Hungary (Example 5). The source text encodes strong evaluative intensity through the metaphorical adjective *dizzying*, which conveys both quantitative excess and affective overload. The human interpreter renders this evaluative force through an idiomatic intensification (*σε βαθμό που μας φέρνει πονοκέφαλο / to the point of giving us a headache*) that is not lexically present in the source text but effectively mirrors its rhetorical impact. The AI output, by contrast, remains closer to the surface structure of the English utterance, omitting the metaphorical component altogether. While the resulting formulation is informationally accurate, the evaluative intensity of the source is attenuated, yielding a flatter and less affectively charged utterance that weakens the speaker's stance.

Example 5

Original	**Human output**	**AI output**
In this report, we point out a dizzying array of new, even more extreme violations.	Σε αυτή την έκθεση βλέπουμε ακόμα περισσότερες, ακόμα πιο ακραίες παραβιάσεις σε βαθμό που μας φέρνει πονοκέφαλο.	Στην παρούσα έκθεση επισημαίνουμε μια σειρά από νέες, ακόμη πιο ακραίες παραβιάσεις.

A similar pattern appears in Example 6, where the speaker criticises the Commission's failure to act decisively. The human interpreter selects a formulation that preserves not only the factual content but also the evaluative implication of *failed to*. The verb *δεν κατάφερε να / did not manage to* conveys non-achievement against expectation and implicitly assigns responsibility, thereby maintaining the critical stance of the source. The AI output, by contrast, produces a factually accurate but evaluatively weaker rendering. By framing the event as simple non-action, it neutralises the notion of institutional failure embedded in the source text. As a result, the propositional meaning is preserved, but the rhetorical criticism is attenuated, constituting a minor semantic shift.

Example 6

Original	**Human output**	**AI output**
The Commission failed to take decisive and effective action at crucial moments.	Η Ευρωπαϊκή Επιτροπή δεν κατάφερε να αναλάβει δράση αποτελεσματική στις κρίσιμες στιγμές.	Η Επιτροπή δεν έλαβε αποφασιστικά και αποτελεσματικά μέτρα σε κρίσιμες στιγμές.

A similar pattern of evaluative attenuation appears in the third speech on the new Pact for the Mediterranean, in a segment addressing circulating disinformation about the Erasmus Plus programme (Example 7). In the source text, evaluative intensity is conveyed through categorical refutation, followed by a metadiscursive clarification that frames the explanation as a response to circulating claims. The human interpreter reinforces this rhetorical trajectory through segmentation and intensification: the dismissal is delivered in short, emphatic units (*Και αυτό δεν είναι αλήθεια. Πολύ απλά. / And this is not true. Quite simply),* while the subsequent clarification is expanded with causal motivation (*επειδή ακούγονται διάφορα / because various things are being said*), explicitly anchoring the explanation in a context of misinformation. The AI output, by contrast, remains semantically accurate but rhetorically attenuated, with reduced evaluative force and no explicit anchoring in the surrounding misinformation.

Example 7

Original	**Human output**	**AI output**
Some are arguing that Erasmus Plus promotes migration from the south. This simply is not true. […] It has been functioning for some time now to the benefit of all societies.	Κάποιοι λένε ότι το Εράσμους Συν προωθεί τη μετανάστευση από τον Νότο. Και αυτό δεν είναι αλήθεια. Πολύ απλά. […] Λειτουργεί εδώ και κάποιο καιρό προς	Κάποιοι υποστηρίζουν ότι το Erasmus Plus προωθεί τη μετανάστευση από τον Νότο. Αυτό απλά δεν είναι αλήθεια. […] Λειτουργεί εδώ και αρκετό καιρό προς όφελος όλων των

Just to clarify this Erasmus Plus story.	όφελος όλων των κοινωνιών. Αυτό το λέω για να διευκρινίσω, επειδή ακούγονται διάφορα για το Εράσμους Συν.	κοινωνιών. Απλώς για να ξεκαθαρίσω αυτό το θέμα με το Erasmus Plus.

4.2.3 Terminology and Institutional References

A further area of qualitative analysis concerns the handling of terminology by the human interpreter and the AI system. In the first speech on breaches of EU values in Hungary, when the source text refers to the Court of Justice of the European Union (CJEU), the human interpreter renders the reference as *Ευρωπαϊκό Δικαστήριο / European Court.* Although this does not reproduce the official designation (*Court of Justice of the European Union*), it reflects a conventional simplification strategy in spoken Greek. The AI output, by contrast, renders CJEU as *Γενική Ελεγκτική Υπηρεσία / General Audit Office*, an institutionally unrelated body. This error can be traced to upstream ASR misrecognition propagated by the MT system, resulting in a severe distortion of institutional reference. A similar divergence appears in the rendering of the concept of the rule of law. The human interpreter consistently uses the established Greek EU term *κράτος δικαίου*, thereby preserving the legal and normative specificity of the concept. The AI system, however, produces *αρχή χρηστής διακυβέρνησης*, a notion corresponding to *good governance* rather than *rule of law*, leading to a conceptual mismatch that alters the legal framing of the argument.

Proper names and culturally embedded references further illustrate the contrast between human inferencing and AI error propagation. In the same excerpt, the source mentions that *the organiser of the Pride in Pécs is currently facing a criminal investigation*. An ASR misrecognition renders the city name *Pécs* as *Page*, an error subsequently propagated in the AI output, resulting in a mistranslation of a concrete geographic referent. The human interpreter, by contrast, correctly produces *Πεκς*, drawing on contextual and world knowledge. In the same segment, the human interpreter retains the English term *Pride* in the Greek output, in line with established usage in Greek public discourse (e.g. *Athens Pride*). The AI system instead opts for the literal rendering *παρελάσεις υπερηφάνειας / pride parades*, which, while semantically transparent, is atypical in Greek oral discourse.

In the second speech on enhancing Europol's role, the source text consistently refers to *smugglers* and *smuggling*, terms which in EU legal and policy discourse denote the facilitation of irregular migration by criminal networks. The human interpreter systematically employs the established Greek equivalents *διακινητές* and *διακίνηση*, thereby preserving the legal framing of the issue. The AI system, by contrast, displays terminological inconsistency, alternating between migration-related terms and renderings such as *λαθρέμποροι* and *λαθρεμπόριο*, which in Greek are conventionally associated with the illicit trade of goods rather than with offences against persons or migration-related crimes. This instability results in a domain shift that reframes the issue as commercial contraband, weakening the legal specificity of the source discourse.

Terminological divergence is also evident in the third speech on the new Pact for the Mediterranean. While the human interpreter consistently uses the institutionally established term *Σύμφωνο για τη Μεσόγειο*, the AI output alternates between *σύμφωνο* and *συμφωνία*. Although both renderings are semantically intelligible, the latter reduces the Pact to a generic agreement, thereby weakening its institutional specificity. Occasional cases of terminological under-specification are also observed in the human output. For instance, *High Representative* is rendered simply as *εκπρόσωπος / representative*, obscuring the formal institutional role, while *economic footprint* is rendered as *οικολογικό / environmental*, introducing a semantic shift from economic presence to environmental impact. Unlike the AI-related errors discussed above, such cases are plausibly attributable to momentary processing pressure and do not indicate systematic terminological instability.

4.2.4 Fluency and Stylistic Naturalness

A final area of qualitative comparison concerns fluency and stylistic naturalness in the interpreting outputs. In Example 8, drawn from the speech on breaches of EU values in Hungary, the human interpreter produces minor surface-level errors characteristic of spontaneous simultaneous production, such as subject–verb agreement errors or non-standard adverbial endings. These errors do not impede comprehension. On the contrary, the interpreter's lexical choices and syntactic strategies – most notably the use of shorter clauses, less formal vocabulary, and idiomatic collocations – align closely with spoken Greek and facilitate processing for the recipient.

Example 8

Original	**Human output**	**AI output**
Corruption remains widespread and anti-corruption bodies are being obstructed. School curriculums are under increasing ideological influence and teachers' right to strike is curtailed.	Η διαφθορά παραμένει διάχυτη και τα όργανα κατά της διαφθοράς παρεμποδίζεται το έργο τους. Τα σχολικά προγράμματα ελέγχονται ιδεολογικά όλο και περισσότερα. Το δικαίωμα στην απεργία των καθηγητών περιορίζεται.	Η διαφθορά παραμένει ευρέως διαδεδομένη και οι φορείς καταπολέμησης της διαφθοράς παρεμποδίζονται. Τα σχολικά προγράμματα σπουδών υφίστανται αυξανόμενη ιδεολογική επιρροή και το δικαίωμα των εκπαιδευτικών να απεργούν περιορίζεται.

A similar pattern is observed in a subsequent segment of the same speech describing media repression (Example 9). Here, the human interpreter restructures a dense English sentence into a sequence of shorter main clauses, segmenting the information into cognitively manageable units. Although several elements are simplified or omitted in this process – the adverb *relentlessly* is not rendered, *intimidating investigations* is condensed into the more general *τις απειλεί*, and *civil society organisations* is rendered through a metonymic generalization – the resulting output reduces cognitive load and supports fluency in spoken discourse.

Example 9

Original	**Human output**	**AI output**
The media environment is in the hands of pro-government propaganda, attacking critical voices relentlessly. The government subjects independent media and civil society organisations to smear campaigns and intimidating investigations.	Το γενικότερο περιβάλλον των μίντια είναι στα χέρια της φιλοκυβερνητικής προπαγάνδας. Και οι κριτικές φωνές δέχονται επιθέσεις. Η κυβέρνηση δεν αφήνει τα ανεξάρτητα μέσα και την κοινότητα πολιτών να προχωρήσουν. Τις συκοφαντεί, τις απειλεί.	Το περιβάλλον των μέσων ενημέρωσης βρίσκεται στα χέρια της φιλοκυβερνητικής προπαγάνδας, η οποία επιτίθεται αμείλικτα στις κριτικές φωνές. Η κυβέρνηση υποβάλλει τα ανεξάρτητα μέσα ενημέρωσης και τις οργανώσεις της κοινωνίας των πολιτών σε εκστρατείες δυσφήμισης και εκφοβιστικές έρευνες.

Similarly, in Example 10 from the speech on enhancing police cooperation in relation to migrant smuggling, the human interpreter produces a highly compressed and colloquial reformulation.

While this compression entails significant information loss – references to mandate extension, human trafficking, and causal justification are omitted – it simultaneously yields an output that is fluent and stylistically aligned with spontaneous spoken Greek.

Example 10

Original	**Human output**	**AI output**
Europol's mandate now also covers the digital dimension of migrant smuggling and human trafficking, including activities conducted on social media platforms. This is essential because much of the work of these criminal networks these days takes place online.	Θα ενισχυθούν επίσης οι δραστηριότητες στα social media, γιατί εκεί πολλά πράγματα συμβαίνουν σήμερα online.	Η εντολή της Ευρωπόλ καλύπτει πλέον και την ψηφιακή διάσταση της παράνομης διακίνησης μεταναστών και της εμπορίας ανθρώπων, συμπεριλαμβανομένων των δραστηριοτήτων που διεξάγονται σε πλατφόρμες κοινωνικών μέσων. Αυτό είναι ουσιαστικής σημασίας, διότι μεγάλο μέρος της δραστηριότητας αυτών των εγκληματικών δικτύων πραγματοποιείται σήμερα στο διαδίκτυο.

These collocational patterns are consistently observed in the AI output. This is evident, for instance, in the rendering of expressions such as *is currently facing a criminal investigation* (in references to the organiser of the Pride in Pécs in Section 4.2.1) and *which has emerged here in the corridors* (in references to disinformation surrounding the Erasmus Plus programme in Section 4.2.3). The human interpreter selects idiomatic and discourse-appropriate alternatives (e.g. *διώκεται* and *φήμες που ακούγονται στους διαδρόμους*), whereas the AI system produces semantically transparent but collocationally unnatural formulations in Greek (e.g. *αντιμετωπίζει ποινική έρευνα* and *παραπληροφόρηση που έχει εμφανιστεί*).

Across these examples, the human interpreting prioritises fluency, idiomaticity, and discourse optimisation, even at the cost of selective information loss. By contrast, the AI output preserves the full informational load of the source text and is often more complete at the propositional level. However, its reliance on literal, text-oriented rendering results in stylistically heavy and collocationally infelicitous formulations that are less compatible with spoken Greek, producing a perceptible translationese effect.

Conclusion

This study examined how closely current AI-based language mediation approximates professional human simultaneous interpreting in a high-stakes institutional setting, focusing on target-language output. Using three short plenary excerpts from the European Parliament and comparing the official Greek simultaneous interpretation with an AI-generated Greek version, the analysis combined proposition-level error annotation with a discourse-oriented qualitative analysis. In doing so, the study moved beyond surface-level accuracy metrics to assess communicative adequacy in terms of pragmatics, rhetorical stance, terminology, and stylistic naturalness, thereby prioritising discourse-level analysis over statistical generalisation.

At the quantitative level, the human interpreting output contained more annotated errors overall (49 vs. 28), yielding a higher error density. When interpreted alongside the qualitative findings, however, this difference does not indicate inferior communicative performance by the human interpreter. Rather, it reflects different operating logics. The human output was dominated by low-impact omissions and, to a lesser extent, mistranslation- and cohesion-related issues, which are consistent with coping strategies in simultaneous interpreting aimed at preserving fluency and discourse coherence under real-time and cognitive constraints. By contrast, the AI output produced no omissions, preserving the informational load of the source more fully. Yet its errors clustered in categories with greater potential communicative risk: pragmatic errors, mistranslations or semantic shifts, terminology errors, as well as cohesion and coherence problems.

Severity profiles were broadly comparable (four major errors in HI; three in AI), but again the nature of high-impact errors differed in ways that matter for institutional communication. Human major errors primarily involved information loss through omission. AI major errors, by contrast, tended to involve institutional reference and pragmatic stance, including cases where incorrect institutional naming or register choices could misrepresent authority, responsibility, or evaluative intent. A subset of AI errors was traceable to upstream ASR misrecognitions, illustrating a structural vulnerability of cascaded pipelines whereby early-stage recognition errors can propagate into consequential distortions.

Systematic differences emerge in how human and AI-mediated output functions at the level of discourse and interaction. The human output prioritises discourse optimisation over informational completeness, particularly in dense argumentative stretches. Even where omissions entail partial information loss, they are frequently offset by strategies that support comprehension, including syntactic restructuring, idiomatic phrasing, and the redistribution of information into cognitively manageable units. Human interpreting consistently excels in areas not captured by "completeness" measures, such as sensitivity to rhetorical contrast, pragmatically appropriate address strategies, and speaker-related contextual cues. As a result, even compressed renditions often preserve illocutionary force and interactional positioning more effectively than AI-generated output.

By contrast, the AI-generated output prioritises information density and lexical coverage, often appearing more accurate by retaining content omitted by the interpreter. Yet the analysis identifies systematic limitations that are particularly consequential in spoken discourse. The AI output frequently displays pragmatic infelicity in Greek and limited contextual inference, resulting in over-literal renditions that are grammatically correct but pragmatically marked and socially or interactionally misaligned. In addition, evaluative intensity is often attenuated: metaphorical or rhetorically charged source expressions are rendered in flatter formulations, thereby weakening the speaker's stance. Similar limitations are evident in terminology and style, including occasional domain drift, susceptibility to error propagation from misrecognised input, and a reliance on semantically transparent but collocationally heavy constructions, which result in a persistent "translationese" effect.

Taken together, the findings indicate that divergences between human and AI output cannot be reduced to a simple opposition between accuracy and error frequency. Rather, they reflect two fundamentally different modes of communicative processing. Human interpreting operates as adaptive mediation shaped by cognitive, contextual, and interactional constraints, involving strategic compression, pragmatic modulation, and dynamic redistribution of meaning to maintain coherence and audience accessibility in real time. AI-mediated output, by contrast, tends to function primarily as structurally faithful reproduction, prioritising lexical equivalence and informational completeness while lacking the capacity for reliable pragmatic modulation, register adaptation, and institutional framing.

This distinction aligns with the conceptualisation of human interpreting within Interpreting Studies as situation-embedded communication unfolding under cognitive and contextual constraints, as opposed to automated speech translation, which operates primarily as a unidirectional process of linguistic transcoding (Horváth, 2021; Sperber et al., 2025). While AI-based tools may offer useful support in routine or low-risk communicative scenarios, they consistently fall short in cultural sensitivity and register adaptability and struggle with complex discourse, cultural references, nuance, and emotionally laden communication (Cela Gutiérrez, 2025). These findings collectively point to the need for evaluation frameworks that move beyond automatic metrics and capture discourse-level organisation, pragmatic adequacy, and communicative effectiveness – dimensions that can be operationalised even in small-scale, manually annotated analyses, as applied in the present study.

However, several limitations constrain the generalisability of the findings. The dataset is small and restricted to a single institutional setting and language pair; the AI output was generated via a text-mediated proxy rather than a fully integrated simultaneous interpreting system, and it is not subject to latency constraints; and annotation was conducted by a single researcher. Future research should therefore extend the analysis to larger and more diverse corpora, compare multiple AI-based interpreting systems under comparable conditions, integrate temporal performance measures, and combine expert analysis with listener-based evaluation in order to triangulate communicative impact from multiple perspectives. Against this background, and on the basis of the present findings, current AI-based language mediation should at this stage be understood primarily as a complementary resource rather than a substitute for professional human interpreting in high-stakes institutional settings (see also Peeters et al., 2025: 20).

References

Bernardini, S., Ferraresi, A., Russo, M., Collard, C. & Defrancq, B. (2018). Building interpreting and intermodal corpora: a how-to for a formidable task. In M. Russo, C. Bendazzoli & B. Defrancq (eds) *Making way in corpus-based interpreting studies* (pp. 21-42). Singapore: Springer.

Cao, Y. (2024). Comparative analysis of machine interpreting and human interpreting: Insights into consecutive interpreting teaching. In M. F. b. Sedon, I. A. Khan, M. C. Birkök & K. S. Chan (eds), *Proceedings of the 2024 International Conference on Social Sciences and Humanities* (pp. 730-741). Paris: Atlantis Press.

Cela Gutiérrez, C. (2025). AI-assisted Interpreting: Valuable Tool for Professional Interpreters or Job Displacement? *Traduction et Langues, 24 (1),* 263-281.

Fantinuoli C. (2022). Conference interpreting and new technologies. In Michaela Albl-Mikasa and Elisabet Tiselius (eds), *Routledge Handbook of Conference Interpreting* (pp. 508-522). New York: Routledge.

Fantinuoli, C. (2019). The technological turn in interpreting: The challenges that lie ahead. In W. Baur & F. Mayer (eds) *Proceedings of the Conference Übersetzen und Dolmetschen 4.0 – Neue Wege im digitalen Zeitalter*. (pp. 334-354). Berlin: BDÜ Fachverlag.

Fantinuoli, C. and Prandi, B. (2021). Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. In M. Federico, A. Waibel, M. R. Costa-jussà, J. Niehues, S. Stuker, & E. Salesky (eds), *Proceedings of the 18th International Conference on Spoken Language Translation* (pp. 245-254). Stroudsburg: Association for Computational Linguistics.

Horváth, I. (2021). Speech translation vs. Interpreting. Language Studies and Modern Humanities, 3(2), 174–187. Berlin: Russian State Pedagogical University Herzen.

Lee, J.-H. & Cha, K.-W. (2023). Human interpretation and machine translations based upon interviews with director Joon-ho Bong. *Korean Journal of English Language and Linguistics*, 23, 204-219.

Liu, Y. & Liang, J. (2024). Multidimensional comparison of Chinese–English interpreting outputs from human and machine: Implications for interpreting education in the machine-translation age. *Linguistics and Education*, 80, 101273, 1-13.

Lu, X.-L. & Han, C. (2023). Automatic assessment of spoken-language interpreting based on machine-translation evaluation metrics: A multi-scenario exploratory study. *Interpreting: International Journal of Research and Practice in Interpreting*, 25(1), 109-143. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Nimdzi (2019). The Nimdzi Interpreting Index. [Online]. Available at: https://www.nimdzi.com/the-2019-nimdzi-interpreting-index/ (accessed 02.01.2026).

Peeters, K., Daems, J., Plieseis, C., Rivas Ginel, M. I., Şahin, M. (2025). *AI for Translation and Interpreting. A Roadmap for Users and Policy Makers*. Conseil Européen pour les Langues / European Language Council.

Russo, M. (2019). Corpus-based studies in conference interpreting. *Slovo.ru: Baltijskij accent*, 10(1), 87-100.

Sperber, M., de Seyssel, M., Bao, J., & Paulik, M. (2025). Toward Machine Interpreting: Lessons from Human Interpreting Studies. In C. Christodoulopoulos, T. Chakraborty, C. Rose & V. Peng (eds), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 23349–23364). Suzhou, China: Association for Computational Linguistics.

Tan, Y. L. & Gao, L. (2025). Pointing to context from a Relevance Theory perspective: A comparative study of human and machine interpreting. *Mediterranean Journal of Social Sciences*, 16(3), 1-14.

Tiselius, E. (2009). Revisiting Carroll's scales. In C. V. Angelelli & H. E. Jacobson (eds), *Testing and Assessment in Translation and Interpreting Studies* (pp. 95-121). Amsterdam/Philadelphia: John Benjamins Publishing Company.