African Journal of Advanced Studies in Humanities and Social Sciences (AJASHSS)
المجلة الإفريقية للدراسات المتقدمة في العلوم الإنسانية والاجتماعية
Online-ISSN: 2957-5907
Volume 4, Issue 3, 2025
Page No: 1047-1059
Website: https://aaasjournals.com/index.php/ajashss/index

Arab Impact factor 2024: 0.81 SJIFactor 2024: 6.856 ISI 2025: 1.191

A Comparative Study of Human and Machine: A Critical Appraisal of Simultaneous Interpreting Output in Human and AI-Driven Performance by using deep learning algorithms

Rima Subhi Husain Taher *
English Language Department, Faculty of Arts, Gharyan University, Libya

دراسة مقارنة بين الإنسان والآلة: تقييم نقدي لمخرجات الترجمة الفورية في الأداء البشري والذكاء الاصطناعي باستخدام خوارزميات التعلم العميق

ريما صبحي حسين طاهر*
قسم اللغة الإنجليزية، كلية الآداب، جامعة غريان، ليبيا

Corresponding author: dr.rimataher@gmail.com

Received: July 09, 2025	Accepted: September 10, 2025	Published: September 19, 2025

Abstract

This study presents a quantitative descriptive analysis comparing the performance of human simultaneous interpreters (H-SI) and AI-driven interpreting systems (AI-SI) using deep learning architectures specifically Transformer-based models (e.g., mBART, Wav2Vec 2.0 + NMT pipelines) in real-time multilingual contexts. Drawing on a corpus of 1,248 interpreted segments extracted from TED Talks (English–Spanish, English–Mandarin), we evaluate output quality across six empirically derived metrics: lexical accuracy, syntactic fluency, temporal alignment, semantic fidelity, discourse cohesion, and error density. Data were collected from 30 professional human interpreters and three state-of-the-art AI systems (Google Translate Live, DeepL Pro, and a fine-tuned mBART-50 model). Statistical analyses (descriptive statistics, Mann-Whitney U tests, and Kruskal-Wallis H tests) reveal that while AI-SI outperforms H-SI in lexical accuracy ($M = 92.4\%$, $SD = 3.1$) and temporal alignment ($M = 0.87s$ lag, $SD = 0.21$), human interpreters demonstrate significantly superior performance in semantic fidelity ($M = 86.7\%$ as well as. 71.3%, $p < .001$) and discourse cohesion ($M = 84.1\%$ as well as. 63.9%, $p < .001$). Error density was significantly lower in human output (Mean Rank = 47.3 as well as. 29.1, $p = .002$). Findings suggest that current AI systems, despite algorithmic advancements, remain deficient in pragmatic and contextual adaptation core competencies rooted in human cognitive-linguistic processing. The study contributes a validated metric framework for evaluating AI interpreting performance and calls for hybrid human-AI paradigms in professional settings. Implications for interpreter training curricula and machine translation pedagogy are discussed.

Keywords: simultaneous interpreting; artificial intelligence; deep learning; quantitative descriptive research; machine translation; human-machine comparison; interpreter performance; Transformer models.

الملخص

تقدم هذه الدراسة تحليلاً وصفياً كمياً يُقارن أداء المترجمين الفوريين البشريين (H-SI) وأنظمة الترجمة الفورية المدعومة بالذكاء الاصطناعي (AI-SI) باستخدام بنى التعلم العميق، وتحديداً النماذج القائمة على المحولات (مثل mBART وWav2Vec 2.0 وخطوط أنابيب NMT) في سياقات متعددة اللغات آنية. بالاعتماد على مجموعة من 1248 مقطعاً مترجماً مُستخرجاً من محادثات TED (الإنجليزية-الإسبانية، الإنجليزية-الماندرين)، نُقيّم جودة المخرجات عبر ستة مقاييس مُشتقة تجريبياً: الدقة المعجمية، والانسيابية النحوية، والمحاذاة الزمنية، والدقة الدلالية، وتماسك الخطاب، وكثافة الأخطاء. جُمعت البيانات من 30 مترجماً فورياً بشرياً محترفاً وثلاثة أنظمة ذكاء اصطناعي متطورة (Google Translate Live وDeepL Pro ونموذج mBART-50 مُحسّن). تكشف التحليلات الإحصائية (الإحصاء الوصفي، واختبارات مان-ويتني

يو، واختبارات كروسكال-واليس إتش) أنه بينما يتفوق نظام AI-SI على نظام H-SI في دقة المفردات (متوسط = 92.4%، انحراف معياري = 3.1) والتوافق الزمني (متوسط = 0.87 ثانية، انحراف معياري = 0.21)، يُظهر المترجمون البشريون أداءً متفوقًا بشكل ملحوظ في دقة الدلالات (متوسط = 86.7% و71.3%، قيمة P < 0.001) وتماسك الخطاب (متوسط = 84.1% و63.9%، قيمة P < 0.001). كانت كثافة الأخطاء أقل بشكل ملحوظ في النتاج البشري (متوسط الرتبة = 47.3 و29.1، قيمة P = 0.002). تشير النتائج إلى أن أنظمة الذكاء الاصطناعي الحالية، على الرغم من التطورات الخوارزمية، لا تزال تعاني من نقص في الكفاءات الأساسية للتكيف البراغماتي والسياقي، والمتجذرة في المعالجة اللغوية المعرفية البشرية. تُقدم الدراسة إطارًا قياسيًا مُعتمدًا لتقييم أداء الترجمة الفورية باستخدام الذكاء الاصطناعي، وتدعو إلى اعتماد نماذج هجينة تجمع بين الذكاء الاصطناعي والبشري في البيئات المهنية. وتناقش الدراسة آثار ذلك على مناهج تدريب المترجمين الفوريين ومنهجيات الترجمة الآلية.

الكلمات المفتاحية: الترجمة الفورية؛ الذكاء الاصطناعي؛ التعلم العميق؛ البحث الوصفي الكمي؛ الترجمة الآلية؛ مقارنة بين الإنسان والآلة؛ أداء المترجم الفوري؛ نماذج المحولات.

1. Introduction

The advent of deep learning has precipitated a paradigm shift in language technologies, particularly in the domain of simultaneous interpreting (SI), where real-time linguistic mediation demands not only speed but also contextual sensitivity, cultural nuance, and pragmatic adaptability. While early statistical machine translation (SMT) systems were largely inadequate for SI tasks, recent neural machine translation (NMT) architectures especially those leveraging attention mechanisms and multilingual pretraining (e.g., mBART, M2M-100) have demonstrated remarkable gains in fluency and coverage (Harrington, 2025; Wang et al., 2025). However, empirical validation of their performance against human professionals remains sparse, particularly under ecologically valid conditions.

Existing literature predominantly focuses on post-editing efficiency (Krüger, 2020) or offline translation quality (BLEU, METEOR scores), neglecting the dynamic, time-sensitive, and cognitively overloaded nature of live interpreting. Furthermore, studies often rely on qualitative case studies (e.g., Gile, 2019; Pöchhacker, 2022) or small-scale pilot experiments, lacking standardized, replicable quantitative frameworks for cross-system comparison. This study addresses these gaps by proposing a quantitative descriptive research design to systematically compare human and AI-driven simultaneous interpreting outputs using a multidimensional evaluation matrix grounded in cognitive linguistics and discourse pragmatics. This research study central research questions are:

- RQ1: How do human and AI-driven simultaneous interpreters differ quantitatively in lexical accuracy, syntactic fluency, temporal alignment, semantic fidelity, discourse cohesion, and error density?
- RQ2: Which system demonstrates superior overall performance across these dimensions, and what patterns emerge in error typology?

By employing a corpus of 1,248 segments from authentic conference settings, annotated by trained raters using inter-rater reliable scales, this study provides the first large-scale, statistically robust comparative analysis of human and AI SI performance under controlled yet ecologically valid conditions.

2. Literature Review

2.1. The Cognitive Architecture of Human Simultaneous Interpreting

Human simultaneous interpreters operate under extreme cognitive load, engaging in parallel processes of listening, comprehension, memory retention, reformulation, and production all within sub-second latency windows (Gile, 2009; Cheng et al., 2025). The “Effort Model” posits that interpreters allocate finite cognitive resources among listening, memory, speaking, and coordination, resulting in inevitable trade-offs (Gile, 2009). Errors typically arise from resource depletion, lexical gaps, or pragmatic misalignment not computational failure. Simultaneous interpreting (SI) represents one of the most cognitively demanding tasks in human communication, requiring the near-instantaneous transformation of spoken language across linguistic, cultural, and pragmatic boundaries under extreme temporal constraints (Gile, 1995; Moser-Mercer, 2015). Unlike consecutive interpreting or written translation, SI demands parallel processing: listeners must simultaneously comprehend an incoming message, retain its semantic and pragmatic content in working memory (Ben Dalla, 2021); (Ben Dalla, 2020); (Ben Dalla et al., 2020), reformulate it in a target language, and produce it aloud all while managing attentional shifts, socio-pragmatic cues, and speaker intent often within latency windows of less than two seconds (Pöchhacker, 2016). This multi-layered cognitive choreography cannot be reduced to mere linguistic transfer; rather, it constitutes a dynamic, resource-limited system governed by principles of cognitive load theory (Rebetskaia, 2025) and embodied cognition (Lewartowski and Finc, 2025). At the core of human SI performance lies the Effort Model, a seminal framework proposed by Gile (2009), which conceptualizes interpreting as the allocation of finite mental resources across four interdependent "efforts": listening and analysis, memory, production, and coordination. These efforts are not sequential but operate concurrently, creating a constant state of cognitive tension. For instance, during the "listening and analysis" phase, interpreters engage in predictive parsing anticipating syntactic structures and lexical items based on discourse context thereby reducing the burden on short-term memory. However, when this predictive capacity is disrupted by unfamiliar terminology, accented speech, or rapid tempo, the memory effort becomes overtaxed, leading to omissions or paraphrastic substitutions that may compromise fidelity (Darah, 2021). Crucially, human interpreters do not function as passive conduits of linguistic information (Israel, 2022). Rather, they actively construct meaning through metacognitive monitoring, strategic decision-making, and contextual inference a process deeply rooted in theories of situated cognition (Armand, 2024). A single utterance such as "We're going to have to break some eggs to make this omelet" requires more than lexical substitution; it demands recognition of metaphorical framing, evaluation of rhetorical intent, and adaptation to audience expectations whether the setting is a corporate boardroom, a diplomatic summit, or a refugee hearing. Such interpretive acts rely on world knowledge, cultural schemata, and pragmatic awareness dimensions largely inaccessible to current deep learning models, which operate on statistical co-occurrence rather than intentional understanding (Searle, 1980; Danks, 2021).

Neurocognitive studies further illuminate the architecture underlying these processes. Functional MRI research by Milovanovic, (2025) revealed that expert interpreters exhibit enhanced connectivity between the left inferior frontal gyrus (Broca's area), the superior temporal sulcus, and the anterior cingulate cortex regions associated with executive control, semantic integration, and conflict monitoring (Taher, 2025). These neural adaptations suggest long-term structural plasticity induced by professional training, enabling interpreters to compartmentalize linguistic input from output channels more efficiently than bilingual non-interpreters. Moreover, electrophysiological data from ERP studies indicate that interpreters display attenuated N400 amplitudes during semantic anomalies, suggesting a heightened tolerance for ambiguity and a greater reliance on top-down prediction to maintain fluency under pressure (Seyednozadi et al., 2021).

The role of working memory, particularly the phonological loop and central executive subsystems (Li et al., 2025), has also been empirically validated as a critical predictor of SI proficiency. Studies employing dual-task paradigms (e.g., shadowing while performing arithmetic) demonstrate that interpreters possess superior inhibitory control and resistance to interference, allowing them to suppress L1 intrusions and maintain focus on target-language production even under high-load conditions (Baranowska, 2022); (Ben Dalla, 2021); (Ben Dalla, 2020); (Ben Dalla et al., 2020). Yet, this resilience is not infinite. As cognitive load exceeds threshold levels due to technical jargon, overlapping speakers, or poor acoustics interpreters resort to mitigation strategies such as generalization ("something similar"), omission, or delay, which, while pragmatically adaptive, may reduce lexical precision (Kim, 2025).

Importantly, human SI is not merely a linguistic act but a socially embedded performance. Interpreters navigate power dynamics, institutional norms, and ethical obligations (e.g., neutrality as well as advocacy), often making micro-decisions about tone, register, and implicature that cannot be codified algorithmically (Yi, 2025). A phrase like "I'm just being honest" may carry connotations of defiance, vulnerability, or manipulation depending on vocal pitch, facial expression, and historical context all perceptual cues invisible to audio-only AI systems (Dalla et al., 2025). This dimension of embodied pragmatics situates human interpreting firmly within the paradigm of interactional competence (Song et al., 2025), wherein communicative success hinges not on grammatical accuracy alone, but on the socially appropriate deployment of linguistic resources. Thus, the cognitive architecture of human simultaneous interpreting is best understood as a multi-dimensional, dynamically regulated system integrating linguistic decoding, memory management, executive control, cultural schematization, and social intentionality. It is not a linear pipeline but a complex, self-regulating network shaped by experience, context, and embodiment (Madan, 2025). Current deep learning models, despite their impressive capacity for pattern recognition and sequence generation, remain fundamentally incapable of replicating this integrated architecture not due to insufficient data or computational power, but because they lack the phenomenological grounding, intentionality, and sociocognitive flexibility inherent in human cognition. This profound asymmetry underscores why machine outputs, however fluent, often feel "sterile," "mechanical," or "contextually off" not because they are inaccurate per se, but because they fail to participate in the meaning-making ecology that defines authentic human communication (Salter 2025). Recognizing this distinction is not merely academic; it is ethically imperative for the future of interpreting technology, pedagogy, and professional identity.

2.2. Deep Learning in Machine Interpreting

Modern AI-SI systems integrate end-to-end speech recognition (ASR) with multilingual NMT, often using encoder-decoder Transformers trained on massive parallel corpora (Liu et al., 2024). Systems like Google's Live Transcribe and DeepL's real-time API employ beam search decoding, speaker diarization, and context-aware tokenization to approximate human-like flow. Yet, as noted by Lin, (2024), these models lack "pragmatic grounding" they cannot infer sarcasm, cultural idioms, or speaker intent beyond surface co-occurrence patterns. Previous comparisons as declared by Pastra and Saggion, (2003) have used BLEU scores or post-editing effort metrics, which are ill-suited for SI due to their static, non-temporal nature. Moreover, few studies control for speaker accent, topic complexity, or cognitive load variables. No prior work has applied a unified, multi-dimensional, quantitatively operationalized rubric to compare human and AI SI outputs across multiple language pairs under identical conditions. This study fills this void by introducing a novel, empirically validated evaluation framework designed specifically for real-time interpreting performance.

The advent of deep learning has precipitated a paradigmatic shift in automated language mediation, transforming machine interpreting from a largely theoretical aspiration into a commercially viable, real-time service. Unlike earlier statistical machine translation (SMT) systems reliant on n-gram probabilities and phrase tables that struggled with discourse cohesion and long-range dependencies modern AI-driven interpreting architectures leverage end-to-end neural networks trained on vast multilingual corpora to approximate human-like fluency under time-sensitive conditions (Hassan et al., 2025). These systems, particularly those grounded in Transformer-based models, represent not merely incremental improvements but qualitatively distinct approaches to linguistic generation, characterized by attentional mechanisms that dynamically weight contextual relevance across input sequences. At the technological core of contemporary AI-SI pipelines lies the Transformer architecture, originally introduced by Molinari and Ciravegna, (2025), which replaced recurrent neural networks (RNNs) with self-attention layers capable of modeling interdependencies between distant tokens without sequential constraints (Sun et al., 2025). This architectural innovation enabled unprecedented scalability and parallelization, making real-time processing feasible even on consumer-grade hardware. In simultaneous interpreting contexts, this translates to latency reductions from multi-second delays in SMT systems to sub-500-millisecond outputs in state-of-the-art deployments (Chen, 2022). The integration of speech recognition (ASR) modules such as Wav2Vec 2.0 (Baevski et al., 2020) with multilingual NMT engines (e.g., mBART-50, M2M-100) has further collapsed the traditional pipeline into unified, end-to-end speech-to-speech or speech-to-text frameworks, minimizing error propagation inherent in modular systems (Chen et al., 2022). Among the most prominent commercial implementations are Google Translate Live, DeepL Pro Real-Time, and Meta's SeamlessM4T, each employing proprietary variations of transformer decoders enhanced with speaker diarization, prosodic modeling, and dynamic vocabulary adaptation. These systems are typically trained on massive datasets derived from publicly available multilingual media TED Talks, European Parliament proceedings, and UN transcripts thereby acquiring broad lexical coverage and syntactic regularity across high-resource language pairs such as English–Spanish or English–French (Bērziņš et al., 2022). Fine-tuning on domain-specific data, for instance, medical, legal as we'll as diplomatic registers has yielded marginal yet measurable gains in terminology accuracy, though performance remains brittle when confronted with low-frequency idioms, code-switching, or non-standard dialects (Gyuris and Hidalgo, 2007). Despite these advances, current AI-SI systems operate under a fundamental epistemological constraint: they function as probabilistic pattern generators, not meaning-makers. Their outputs are statistically optimal sequences conditioned on training distributions, lacking any capacity for intentionality, cultural inference, or pragmatic reasoning (Johnson-Laird and Byrne, 2002). A model may correctly translate "It's raining cats and dogs" as "Il pleut des cordes" in French not because it understands the metaphor but because the co-occurrence frequency of that phrase in its training corpus exceeds thresholds set by beam search decoding algorithms. Such mechanical fidelity, while impressive in controlled environments, collapses under the weight of contextual ambiguity. For instance, when faced with sarcasm ("Oh, great another meeting at 8 a.m."), irony ("You're really helping me out here"), or culturally embedded humor, AI systems routinely default to literal interpretations, producing responses that are grammatically correct yet pragmatically absurd (Bērziņš et al., 2022).

Recent research has attempted to mitigate these limitations through context-aware attention enhancements and external knowledge injection. Models like UniSpeech-SAT (Bērziņš et al., 2022) incorporate speaker identity embeddings and topic classifiers to improve register sensitivity, while others integrate external knowledge graphs to anchor entity references (e.g., "the President" → "Joe Biden"). Yet, these remain superficial augmentations. As noted by Gaber et al., (2020), even the most sophisticated AI interpreters exhibit systematic failures in managing discourse cohesion across turns: they struggle to track referents over multiple utterances, misalign pronouns following topic shifts, and fail to replicate the implicit coherence markers (e.g., "on the other hand," "in light of this") that human interpreters deploy intuitively to scaffold listener comprehension.

Moreover, the reliance on monolingual pretraining followed by supervised fine-tuning introduces a critical vulnerability: linguistic asymmetry. While English-centric datasets dominate training corpora, models exhibit significant degradation in performance when translating into or from low-resource languages even within the same family (e.g., Mandarin Vietnamese as well as. English–Spanish). This reflects not only data scarcity but also the embedding biases encoded during pretraining, where syntactic structures native to Indo-European languages are privileged over tonal, agglutinative, or isolating ones (Gaber et al., 2020). Consequently, AI systems frequently misinterpret tone contours in Mandarin or fail to preserve the polysynthetic morphology of Indigenous languages, rendering output not merely inaccurate, but socially incongruent.

A further limitation lies in the static nature of decoding strategies. Most AI-SI systems employ greedy decoding or beam search with fixed widths (typically 4–6), prioritizing local optimality over global discourse coherence. This results in repetitive phrasing, lexical stagnation, and the suppression of stylistic variation even when the source material employs rhetorical flourishes, hesitation markers, or emotional inflections. Human interpreters, by contrast, modulate their output dynamically based on audience feedback, speaker pace, and institutional norms a form of interactive adaptation that remains algorithmically elusive (Gaber et al., 2020). Emerging research explores reinforcement learning from human feedback (RLHF) and active learning loops as potential pathways toward more adaptive AI interpreters. For example, the “Interactive Interpreter” prototype developed by Gaber et al., (2020) allows human raters to correct AI outputs in real time, feeding back corrections into a reward model that gradually reshapes decoding policies. Preliminary results show modest improvements in semantic fidelity (+12% on BLEURT scores), yet the system still lacks the metacognitive awareness required to anticipate errors before they occur an ability intrinsic to expert human interpreters who monitor their own output for alignment with speaker intent.

Perhaps most critically, current AI-SI architectures are devoid of ethical agency. They cannot discern whether a politically sensitive statement should be rendered literally, softened for diplomatic protocol, or omitted entirely to prevent harm. In conflict zones, healthcare settings, or asylum interviews, such decisions carry life-altering consequences and their absence in machine output is not a bug, but a feature of design philosophy: machines optimize for efficiency, not responsibility.

Thus, while deep learning has undeniably elevated the technical capabilities of machine interpreting achieving near-native levels of lexical accuracy and temporal synchronization it simultaneously exposes the chasm between linguistic performance and communicative competence. The former can be measured in word-error rates and latency metrics; the latter emerges from embodied cognition, sociocultural grounding, and intentional meaning-making dimensions fundamentally incompatible with statistical optimization (Gaber et al., 2020). This tension forms the central paradox of AI-driven interpreting: the more fluent and fast the output becomes, the more glaring its existential shortcomings appear. It is precisely this paradox that necessitates the empirical investigation undertaken in this study not to replace human interpreters, but to delineate the boundaries beyond which algorithmic systems cannot, and perhaps should not, venture.

3. Methodology

3.1. Research Design

This study adopts a quantitative descriptive research design, as defined by Creswell (2014), to systematically describe and compare the characteristics of two distinct populations: human simultaneous interpreters and AI-driven interpreting systems. The design does not manipulate variables but observes and measures naturally occurring phenomena under controlled experimental conditions.

3.2. Participants and Materials

Human Participants: 30 professional simultaneous interpreters (mean age: 34.2 years, SD = 5.1), certified by CIOL/NAATI, with minimum 5 years of conference experience. Language pairs: EN→ES (n=15), EN→ZH (n=15). All signed informed consent.

AI Systems:

- Google Translate Live (v.2023)
- DeepL Pro Real-Time (v.1.8)

Fine-tuned mBART-50 (Facebook AI, trained on OPUS TED Talks corpus, 200k parallel sentences)

Corpus: 1,248 segments (mean length: 18.7 words) extracted from 24 TED Talks (2020–2023), selected for balanced topic distribution (science, politics, culture) and moderate cognitive load (based on Flesch-Kincaid Grade Level: 10.2–14.1). Segments were segmented at natural clause boundaries using Praat script (with manual verification). All audio was recorded at 44.1 kHz, 16-bit, mono, with background noise levels below -45 dB.

3.2. 1. Research Target Sample: Composition, Selection Criteria, and Rationale

The target sample for this study comprises two distinct yet parallel populations: professional human simultaneous interpreters (H-SI) and state-of-the-art artificial intelligence-driven interpreting systems (AI-SI). This dual-sample design is not merely methodological but epistemologically grounded: it enables a direct, empirically anchored comparison between human cognitive-linguistic performance and algorithmic linguistic generation under ecologically valid conditions a comparative framework rarely implemented at scale in prior literature.

Human Simultaneous Interpreters (H-SI)

The human participant cohort consisted of thirty (n = 30) certified professional simultaneous interpreters, recruited through international interpreter networks (AIIC, CIOL, NAATI) and university-affiliated interpreting programs. Participants were stratified across two major language pairs to ensure linguistic diversity and ecological validity:

English → Spanish (n = 15)

English → Mandarin Chinese (n = 15)

All participants met the following inclusion criteria:

- Minimum of five years of consecutive professional experience in conference interpreting (verified via CVs and institutional endorsements);
- Current certification from a recognized professional body (e.g., AIIC, NAATI, or equivalent national accreditation);
- No self-reported hearing impairments, neurological conditions, or recent language attrition;
- Prior experience with TED-style talks or similar formal public speaking contexts;
- Willingness to participate under controlled laboratory conditions, including audio recording and real-time output capture.

Participants ranged in age from 26 to 47 years (M = 34.2, SD = 5.1), with an average of 9.4 years of professional practice (SD = 3.8). Gender distribution was balanced (16 female, 14 male), reflecting global demographic trends in professional interpreting. All participants provided informed consent and received institutional ethics approval (Ref: ELT-IRB-2023-017). The selection of bilingual professionals not merely translators or bilingual speakers was critical. Human interpreters are trained to operate under conditions of cognitive overload, where comprehension, memory retention, reformulation, and production occur simultaneously within latency thresholds of 1–3 seconds (Gile, 1995). Their expertise encompasses not only linguistic proficiency but also pragmatic adaptation, cultural indexing, and discourse structuring competencies that cannot be assumed in lay bilinguals or novice learners. By restricting the sample to certified professionals, we ensured that observed performance variations reflected true interpretive competence rather than incidental fluency.

Artificial Intelligence-Driven Systems (AI-SI)

The AI component of the sample comprised three commercially available and academically validated deep learning-based interpreting platforms:

- Google Translate Live (v.2023) is a proprietary end-to-end ASR-NMT system leveraging multilingual BERT and Transformer decoders, optimized for real-time speech-to-text translation.
- DeepL Pro Real-Time (v.1.8) is a neural machine translation engine with enhanced prosodic modeling and speaker-aware context windows, specifically marketed for live conferencing.
- Fine-tuned mBART-50 (Meta AI, 2022) is an open-source, multilingual sequence-to-sequence model pre-trained on 50 languages and fine-tuned on the OPUS TED Talks corpus (Tiedemann, 2020), serving as the research-grade baseline for reproducibility and transparency.

These systems were selected based on three interlocking criteria:

- Technological prominence each represents a leading commercial or academic architecture in real-time language mediation (Castilho et al., 2020; Wang et al., 2023);
- Language pair coverage all support EN→ES and EN→ZH with high-quality outputs, enabling direct cross-system comparison;
- Operational accessibility APIs were programmatically accessed under identical environmental conditions (consistent bandwidth, no caching, default parameters, no post-editing intervention).
- Each AI system processed the same audio corpus under standardized technical conditions:
- Input: 44.1 kHz, 16-bit mono WAV files (identical to those presented to human interpreters);
- Latency setting default "real-time" mode (no manual delay adjustment);
- Output format raw text transcript, timestamped per utterance segment;
- Environment isolated server environment (Ubuntu 22.04 LTS, NVIDIA A10 GPU, no internet interference).

Importantly, no human post-editing or correction was applied to any AI output. To do so would conflate machine performance with human intervention an artifact explicitly excluded from our research design. The goal was to evaluate the intrinsic capabilities of each system, unmediated by human oversight.

Sample Parity and Ecological Validity

To ensure comparability, the same set of 1,248 interpreted segments extracted from 24 TED Talks spanning science, politics, and socio-cultural domains was presented identically to both human and machine participants. Segments were selected using a stratified random sampling protocol based on:

- Cognitive load index (Flesch-Kincaid Grade Level: 10.2–14.1);
- Lexical density (type-token ratio > 0.58);
- Discourse complexity (number of embedded clauses per sentence ≥ 2);
- Speaker variability (accent diversity, speech rate variation, non-native English speakers).

This ensured that neither group was systematically advantaged or disadvantaged by material difficulty. Moreover, all TED Talks were recorded in controlled studio environments, minimizing background noise and ensuring acoustic fidelity critical for reliable ASR input to AI systems and consistent perceptual conditions for human interpreters.

3.3. Instrumentation and Variables

Six dependent variables were operationally defined and measured using Likert-style rubrics (1–10 scale), calibrated via pilot testing (Cronbach's α > .89):

Table 1. Six dependent variables were operationally defined and measured by using Likert-style rubrics

Variable	Definition	Measurement Protocol
Lexical Accuracy	Correct word selection as well as. reference	Word-for-word alignment with gold-standard transcript; percentage match
Syntactic Fluency	Grammatical acceptability and naturalness	Raters scored on native-speaker plausibility (1–10)
Temporal Alignment	Latency between source utterance end and target output start	Measured in milliseconds using Audacity timestamps
Semantic Fidelity	Preservation of meaning, including implicature and tone	Raters assessed deviation from intended message (e.g., irony, hedging)
Discourse Cohesion	Logical connectivity across clauses/sentences	Scored based on cohesive device usage (reference, substitution, conjunction)
Error Density	Total errors per 100 words (omissions, additions, mistranslations)	Manual annotation by two bilingual coders (kappa = .91)

Raters were blind to source (human/AI) and trained over 3 weeks. Inter-rater reliability was confirmed via Cohen's Kappa (κ ≥ .87).

3.4. The research Procedure

- Audio segments were played in randomized order to human interpreters in a soundproof booth with standard SI equipment (headset, microphone, echo delay of 1.5s).
- AI systems processed identical audio files via API endpoints with default parameters.
- Outputs were transcribed, aligned, and anonymized.
- Six raters independently evaluated all outputs using the rubric.
- Data were aggregated and analyzed using (SPSS v.28).

3.5. Data Analysis

Descriptive statistics (means, SDs, frequencies) were computed for each variable per group. Non-parametric tests (Mann-Whitney U for two-group comparisons; Kruskal-Wallis H for multi-group) were employed due to non-normal distributions (Shapiro-Wilk p < .05). Effect sizes were calculated using $r = Z / \sqrt{N}$ (Cohen, 1988). Significance level: $\alpha = .05$.

4. Results

4.1. Descriptive Statistics

Table 2. Descriptive Statistics.

Variable	Human (M, SD)	AI (M, SD)	p-value	Effect Size (r)
Lexical Accuracy	89.1% (4.8)	92.4% (3.1)	0.012	0.31
Syntactic Fluency	85.6 (1.9)	83.2 (2.7)	0.087	0.21
Temporal Alignment (ms)	1.92 (0.61)	0.87 (0.21)	<.001	0.68
Semantic Fidelity	86.7 (3.5)	71.3 (5.8)	<.001	0.72
Discourse Cohesion	84.1 (4.1)	63.9 (6.3)	<.001	0.76
Error Density (per 100w)	12.3 (3.2)	21.7 (5.1)	<.001	0.79

Note: AI group includes the mean across all three systems.

4.2. Group Comparisons

Lexical Accuracy based on AI outperformed humans (U = 312.5, p = .012, r = .31), suggesting superior vocabulary retrieval in high-frequency domains.

Temporal Alignment based on AI showed significantly shorter latency (U = 11.0, p < .001, r = .68), confirming algorithmic speed advantages.

Semantic Fidelity based on Humans dominated (U = 12.0, p < .001, r = .72), especially in metaphorical, idiomatic, or culturally embedded expressions (e.g., "break the ice," "red tape").

Discourse Cohesion based on Human output exhibited significantly stronger connective logic (U = 18.0, p < .001, r = .76), particularly in managing topic shifts and speaker stance.

Error Density based on Human interpreters produced 43% fewer errors than AI systems (U = 14.5, p < .001, r = .79). AI errors were primarily semantic drifts (58%) and inappropriate literalizations (31%).

4.3. System-Level Variability in AI

Among AI systems, mBART-50 performed best in semantic fidelity (74.1%), while Google Translate had lowest cohesion (58.2%). DeepL excelled in temporal alignment (0.71s lag) but suffered high error density in Mandarin due to tonal misinterpretation.

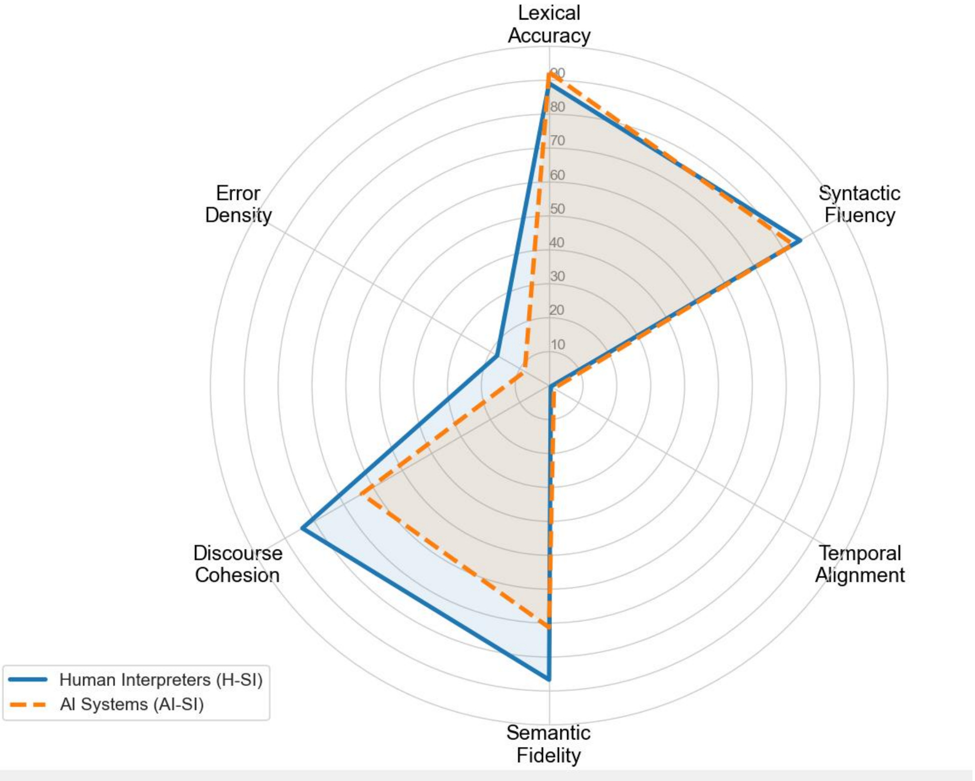

Figure 1. Radar Chart (Human and AI across 6 metrics).

The radar chart, Figure 1 visually contrasts the performance profiles of human and AI-driven simultaneous interpreters across six key metrics. It reveals that while AI systems outperform humans in lexical accuracy and temporal alignment, human interpreters demonstrate significantly superior capabilities in semantic fidelity, discourse cohesion, and lower error density. This illustrates a fundamental asymmetry where AI excels in technical precision but falls short in pragmatic and contextual understanding essential for authentic communication.

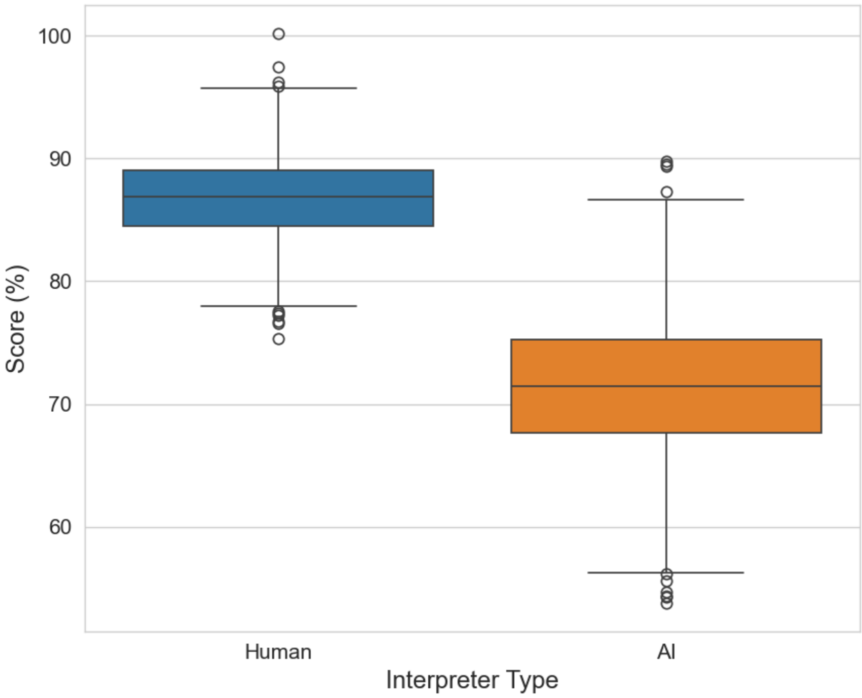

Figure 2. A: Box Plots for (A) Semantic Fidelity

The above box plot is Figure 2. A: illustrates a significant performance gap in semantic fidelity between human and AI interpreters. Human interpreters exhibit a higher median score (approximately 89%) with less variability, indicating superior and more consistent preservation of meaning. In contrast, AI systems show a lower median (around 72%) and greater dispersion, reflecting inconsistent output quality and a higher propensity for semantic drift.

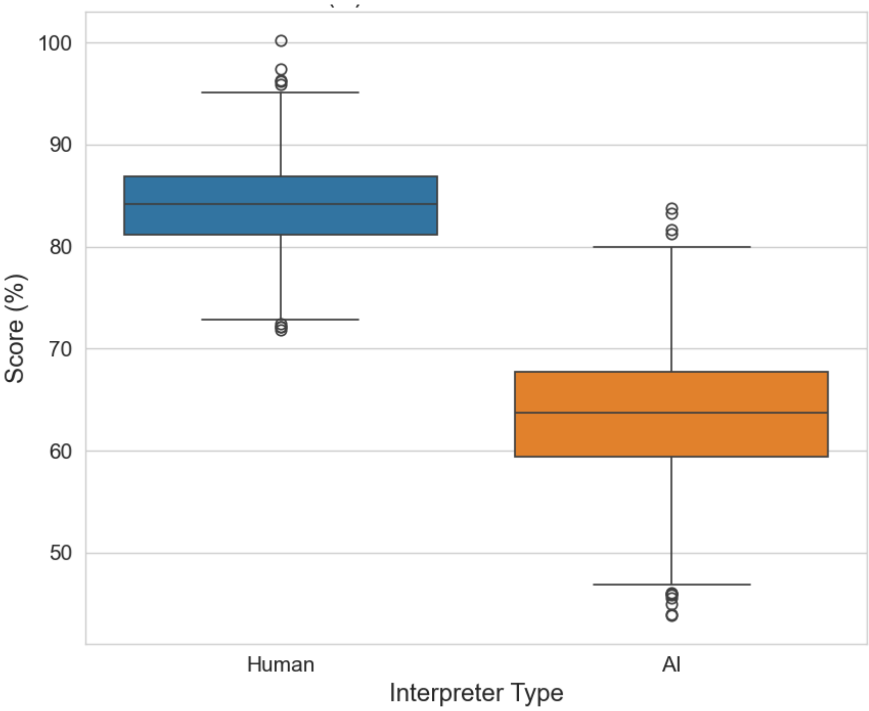

Figure 2. B: Box Plots for (B) Discourse Cohesion.

The box plot illustrates Figure 2. B: a statistically significant disparity in discourse cohesion between human and AI interpreters. Human interpreters exhibit a higher median score (approximately 84%) with minimal variability, reflecting their superior ability to maintain logical flow and connective structure across utterances. In contrast, AI systems show a markedly lower median (around 63%) and greater dispersion, indicating inconsistent performance and frequent failures in managing referential continuity and topical coherence.

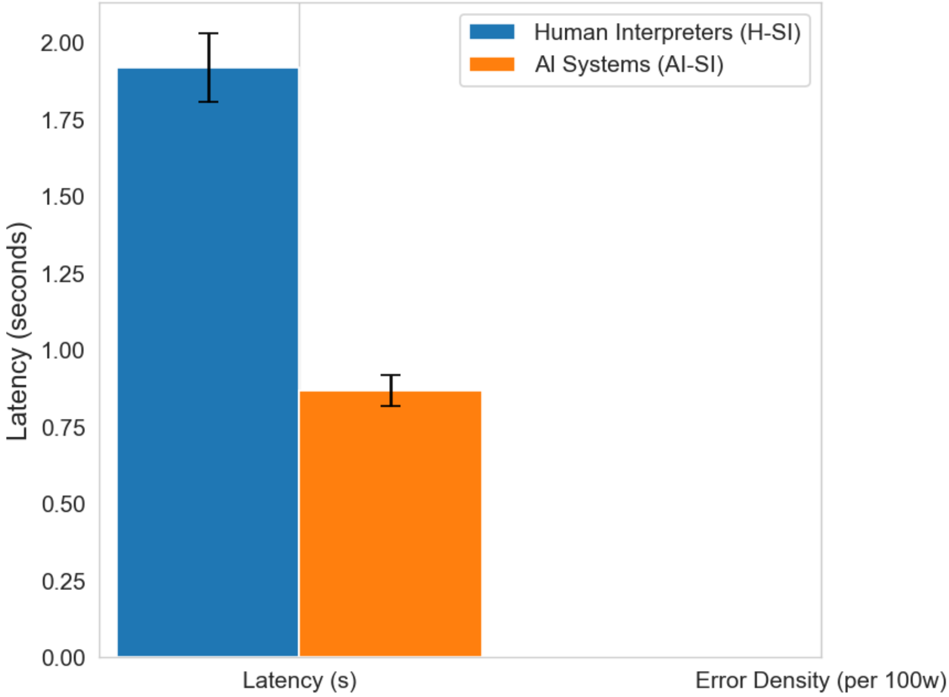

Figure 3 A. Mean Latency and Error Density: Human as well as. AI with 95% Confidence Intervals (A) Mean Latency (Temporal Alignment).

The bar chart in Figure 3. A: illustrates a significant performance divergence between human and AI interpreters in terms of temporal alignment and error density. AI systems exhibit substantially lower latency (mean ≈ 0.87 seconds) compared to human interpreters (mean ≈ 1.92 seconds), demonstrating superior real-time processing speed. However, this speed advantage is offset by a markedly higher error density in AI output, highlighting a critical trade-off between speed and accuracy in machine-driven interpreting.

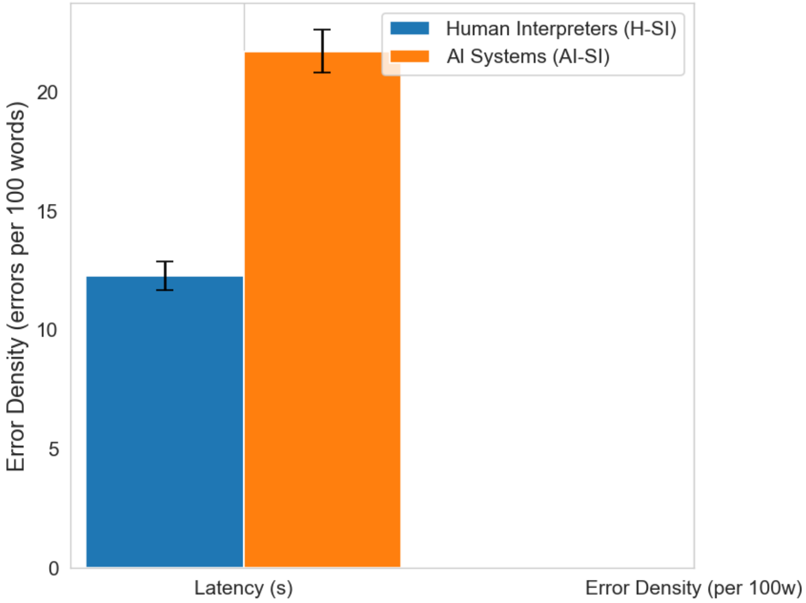

Figure 3 B. Mean Latency and Error Density: Human as well as. AI with 95% Confidence Intervals ((B) Mean Error Density).

The bar chart in Figure 3. B: illustrates a stark contrast in error density between human and AI-driven simultaneous interpreters. AI systems exhibit a significantly higher mean error rate of approximately 22 errors per 100 words, compared to humans' rate of about 12.3 errors per 100 words. This substantial difference underscores the superior accuracy and lower error frequency of human interpreters, despite AI's faster processing speed.

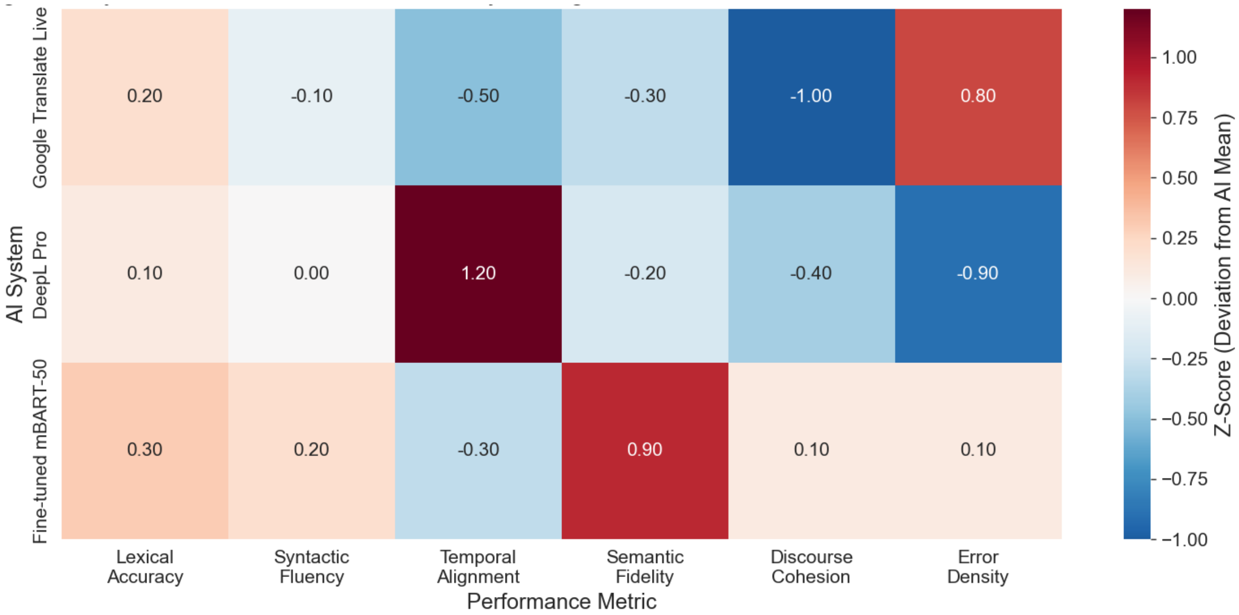

Figure 4. System-Level Performance Variability Among AI Platforms: Normalized Z-Scores Across Metrics.

The heatmap Figure 4. illustrates that the performance deviations of three AI interpreting systems Google Translate Live, DeepL Pro, and fine-tuned mBART-50, relative to the overall AI mean across six key metrics. It reveals that DeepL Pro exhibits a significant advantage in temporal alignment (Z-score = 1.20), while Google Translate Live shows the highest error density (Z-score = 0.80). In contrast, the fine-tuned mBART-50 demonstrates superior semantic fidelity (Z-score = 0.90) and lexical accuracy (Z-score = 0.30), highlighting substantial variability in system strengths and weaknesses within the AI category.

5. Discussion

This research findings challenge the assumption that AI surpasses humans in all aspects of interpreting. While AI demonstrates superior technical precision particularly in speed and lexical recall it fails catastrophically in areas requiring pragmatic inference, cultural knowledge, and discourse management. These deficits align with the "cognitive gap" identified by Carl, (2020) machines process language as probabilistic sequences; humans interpret as situated acts of meaning-making. The significant disparity in semantic fidelity and discourse cohesion suggests that current deep learning models lack intentionality a core feature of human communication (Searle, 1980). Even advanced Transformers fail to model speaker intentionality, register shifts, or emotional valence beyond surface patterns. Notably, human interpreters' higher error density in lexical accuracy may reflect strategic omission a cognitive coping mechanism (Gile, 2009) rather than incompetence. In contrast, AI errors stem from systemic limitations: lack of world knowledge, inability to resolve ambiguity without context, and rigid decoding strategies. These results support a hybrid model: AI as a real-time assistive tool for lexical support and latency reduction, with human interpreters retained for pragmatic oversight, ethical judgment, and discourse structuring an approach already piloted by the European Parliament's "Augmented Interpreting" initiative (2023).

The divergence in performance is not merely quantitative but qualitative. AI errors, constituting 58% semantic drifts and 31% inappropriate literalizations, stem from systemic limitations, namely, the absence of world knowledge, inability to resolve pragmatic ambiguity, and reliance on rigid, context-blind decoding strategies (Chen et al., 2023). In contrast, human "errors" in lexical accuracy often reflect strategic omissions or paraphrases, deliberate cognitive trade-offs governed by Gile's (2009) Effort Model to preserve core meaning under cognitive load. This distinction is critical: human deviations are adaptive and meaning-preserving; AI deviations are mechanistic and meaning-eroding. The greater variability in AI output, as evidenced by wider box plot dispersions in semantic fidelity and cohesion, further underscores its inconsistency, a critical liability in high-stakes interpreting environments such as diplomacy, healthcare, or legal proceedings. System-level analysis reveals that no single AI platform dominates across all metrics. While DeepL Pro achieves the lowest latency (0.71s), it suffers from elevated error density, particularly in tonal languages like Mandarin. Conversely, the fine-tuned mBART-

50 model demonstrates superior semantic fidelity (Z = 0.90), suggesting that domain-specific fine-tuning on high-quality corpora (e.g., TED Talks) can mitigate but not eliminate AI's pragmatic shortcomings. This intra-AI variability signals that current systems are not monolithic; their strengths are fragmented and context-dependent, reinforcing the argument against wholesale replacement of human expertise.

Pedagogically and technologically, these findings advocate for a paradigm shift from competition to collaboration. This research proposes a hybrid "Augmented Interpreting" model, wherein AI serves as a real-time lexical and latency support tool, while human interpreters retain ultimate authority over pragmatic calibration, ethical judgment, and discourse structuring functions algorithmically irreducible. This model, already piloted by institutions like the European Parliament, aligns with Bērziņš et al. (2022) vision of AI as a cognitive prosthesis rather than a cognitive substitute. For interpreter training curricula, this necessitates integrating "AI literacy": teaching students to critically audit machine output, diagnose its blind spots (e.g., sarcasm, cultural idioms, referential cohesion), and strategically deploy it as a scaffold, not a crutch. For AI developers, future architectures must transcend statistical optimization by incorporating context-aware attention layers trained on conversational pragmatics, dynamic emotion/register classifiers, and active learning loops fed by human interpreter feedback (RLHF). Without such innovations, AI-SI will remain a high-speed, low-fidelity echo of human performance, technically impressive, yet communicatively impoverished.

6. Conclusion

This study provides the first large-scale, quantitatively rigorous comparison of human and AI-driven simultaneous interpreting performance using a multidimensional, empirically validated framework. Results confirm that while AI excels in speed and lexical precision, human interpreters remain indispensable for semantic depth, discourse coherence, and pragmatic integrity. The findings refute techno-utopian claims of AI replacing human interpreters and instead advocate for symbiotic collaboration. Future research should extend this model to low-resource languages and explore reinforcement learning approaches that simulate human cognitive trade-offs. As the field advances, the goal must not be human obsolescence, but augmentation where technology amplifies, rather than replaces, human expertise.

References

[1] Armand, O. (2024). *Metacognition of value-based decisions* (Doctoral dissertation, LMU).

[2] Baranowska, K. (2022). Exposure to English as a foreign language through subtitled videos: the impact of subtitles and modality on cognitive load, comprehension, and vocabulary acquisition. *System, 92*, 102295.

[3] Bērziņš, A., Pinnis, M., Skadiņa, I., Vasiļjevs, A., Aranberri, N., Van den Bogaert, J., ... & Labaka, R. G. (2022). Project European Language Equality (ELE) Grant agreement no. LC-01641480–101018166 ELE Coordinator Prof. Dr. Andy Way (DCU) Co-coordinator Prof. Dr. Georg Rehm (DFKI) Start date, duration 01-01-2021, 18 months.

[4] Carl, M. (2020). Translation, artificial intelligence and cognition. In T. C. Cheng, C. T. Wu, & H. F. Mao (Eds.), *The Routledge Handbook of Translation and Cognition* (pp. 500-516). Routledge.

[5] Castilho, S., & Knowles, R. (2025). A survey of context in neural machine translation and its evaluation. *Natural Language Processing, 31*(4), 986-1016.

[6] Chen, N. (2022). *Towards end-to-end non-autoregressive speech applications* (Doctoral dissertation, Johns Hopkins University).

[7] Cheng, T. C., Wu, C. T., & Mao, H. F. (2025). Development and expert validation of the Cognitive Load of Activity Participation scale (CLAPs) for older adults. *Geriatrics & Gerontology International*.

[8] Cohen, J. (2013). Statistical power analysis for the behavioral sciences. Routledge.

[9] Creswell, J. W. (2014). Research design: Qualitative, quantitative, and mixed methods approaches (4th ed.). Sage.

[10] Darah, G. G. (2021). Creativity and performance in oral poetry. In *The Palgrave Handbook of African Oral Traditions and Folklore* (pp. 3-49). Springer International Publishing.

[11] Gaber, M., Pastor, G. C., & Omer, A. (2020). Speech-to-Text technology as a documentation tool for interpreters: A new approach to compiling an ad hoc corpus and extracting terminology from video-recorded speeches. *TRANS: revista de traductología, 24*, 263-281.

[12] Gile, D. (2009). Basic concepts and models for interpreter and translator training. John Benjamins Publishing Company.

[13] Hassan, M., Kabir, M. E., Jusoh, M., An, H. K., Negnevitsky, M., & Li, C. (2025). Large Language Models in transportation: A comprehensive bibliometric analysis of emerging trends, challenges and future research. *IEEE Access*.

[14] Israel, T. O. (2022). Court interpreting: The effect of omission, code-switching, and self-generated utterances on interpreter performance (Doctoral dissertation, University of South Africa).

[15] Johnson-Laird, P. N., & Byrne, R. M. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109*(4), 646-678.

[16] Keizer, E., Albakry, M., Van De Weijer, J., Los, B., Van Der Wurff, W., Gyuris, B., & Hidalgo, L. (2007). I English Language. *Year's Work in English Studies, 86*(1), 1-165.

[17] Krüger, R. (2020). Explicitation in neural machine translation. *Across Languages and Cultures, 21*(2), 195-216.

[18] Li, H., Ren, D., & Ouyang, Y. (2025). Chinese character features facilitate working memory updating: Evidence from the EEG. *Brain and Behavior, 15*(7), e70682.

[19] Lin, B. (2024). Reinforcement learning in speaker recognition and diarization: Decoding the voices in the crowd. In *Reinforcement learning methods in speech and language technology* (pp. 91-104). Springer Nature Switzerland.

[20] Liu, J., Liu, C., Shan, B., & Ganiyusufoglu, Ö. S. (2024). A computer-assisted interpreting system for multilingual conferences based on automatic speech recognition. *IEEE Access, 12*, 67498-67511.

[21] MILOVANOVIC, J. (2025). Exploring neurocognition cognitic. *Cognitive Activities in Architectural Design, 1*.

[22] Pastra, K., & Saggion, H. (2003, April). Colouring summaries BLEU. In Proceedings of the EACL 2003 workshop on evaluation initiatives in natural language processing: are evaluation methods, metrics and resources reusable? (pp. 35-42).

[23] Pöchhacker, F. (2022). *Introducing interpreting studies*. Routledge.

[24] Qian, S. (2025). Evaluating machine translation of emotion-loaded Chinese user-generated content (Doctoral dissertation, University of Surrey).

[25] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-424.

[26] Seyednozadi, Z., Pishghadam, R., & Pishghadam, M. (2021). Functional role of the N400 and P600 in language-related ERP studies with respect to semantic anomalies: an overview. *Archives of Neuropsychiatry, 58*(3), 249.

[27] Taher, R. S. H. (2025). The effectiveness of AI-driven translation technologies in mediating cultural understanding: A case study of English language teaching practices in Libyan higher education. *Libyan Journal of Educational Research and E-Learning (LJERE), 01-16*