Toward Machine Interpreting: Lessons from Human Interpreting Studies

Matthias Sperber **Maureen de Seyssel** **Jiajun Bao** **Matthias Paulik**
Apple
{sperber, mdeseyssel, jbao3, mpaulik}@apple.com

Abstract

Current speech translation systems, while having achieved impressive accuracies, are rather static in their behavior and do not adapt to real-world situations in ways human interpreters do. In order to improve their practical usefulness and enable interpreting-like experiences, a precise understanding of the nature of human interpreting is crucial. To this end, we discuss human interpreting literature from the perspective of the machine translation field, while considering both operational and qualitative aspects. We identify implications for the development of speech translation systems and argue that there is great potential to adopt many human interpreting principles using recent modeling techniques. We hope that our findings provide inspiration for closing the perceived usability gap, and can motivate progress toward true machine interpreting.

1 Introduction

Even though speech translation (ST) research has celebrated great successes, the user experience when employing ST technology in real-world tasks is often still perceived to be inferior to the experience of receiving assistance from a human interpreter (Federico et al., 2024). This subjective impression is in contrast to the impressive accuracies reported on standard benchmarks. For example, Wein et al. (2024) report superiority to human interpreters as measured by common machine translation (MT) metrics against reference translations, and Cheng et al. (2024) report parity when measured by their proposed valid information proportion metric. Such findings imply that the gap in user experience largely stems from factors not captured by such benchmarks (Savoldi et al., 2025). Such factors likely include, among others, interpreters' flexibility in modes of operation, their situational awareness, advanced translation strategies that include cultural adaptation, effective error prevention or recovery strategies (Jones, 2002). Many of these characteristics and well-studied features of human interpreting have obtained little attention from MT researchers, perhaps partly because technical solutions to emulate human interpretation used to be out of reach.

Recent advances on large language models (LLMs) and their application to translation opens up an avenue towards closing the usability gap between ST[1] and human interpretation. For example, LLMs with long context may allow accumulating "increased knowledge about a communicative event" (Fantinuoli, 2024) in its entirety, allowing systems to effectively mimic situational awareness. Multimodal LLMs may learn to leverage audiovisual context, thereby extending situational awareness beyond the spoken word (Yin et al., 2024). Prompts can be engineered to flexibly inform the translation model about speaker/listener relationship and their cultural/topical knowledge gap, allowing for more helpful translations and appropriate cultural adaptation (Yao et al., 2024). Reinforcement learning and instruction tuning may enable systems to imitate human error prevention or recovery strategies (Goldberg, 2023).

To effectively make progress towards translation systems that provide a more pleasant, interpreting-like experience, a precise understanding of the nature and goals of human interpreting is crucial. To this end, this paper discusses human interpreting literature and draws out implications and opportunities for MT research. We hope that our work will serve as inspiration on the quest of advancing current ST technology toward a more interpreting-like experience, i.e. toward what we might call "machine interpreting".[2]

[1] In this paper, we refer to MT as automatic translation between any modality (speech or text), and ST as automatic translation from speech to any modality.

[2] We use this term cautiously: even though *machine interpreting* has occasionally been used as a synonym of *speech-*

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23350–23365
November 4-9, 2025

Feature	Description	Example
Temporal immediacy	Produces interpretation in real-time.	Maintains 1–2 seconds ear-voice-span.
Spatial immediacy	Operates in proximity of speaker & audience.	Shares stage with speaker.
Multimodality	Uses visual or gesture cues when available.	Refers to chart while speaker points.
Free/diverse actions	Dynamically adapts to any situation.	Adapts translation approach to content type.
Interaction/influence	Acts as a independent agent when needed.	Requests clarification, improves acoustics.
Intent translation	Interprets what is meant, not what is said.	Interpretation conveys hidden accusations.
Interpreter uncertainty	Maintains trust by signaling own uncertainty.	"Speaker may have said 'revenue'."
Speaker errors	Indicates or corrects unintentional speaker errors.	Corrects "million" to "billion" in context.
Adaptation/explanation	Adapts or explains culture-specific expressions.	"Break a leg!" → "Good luck!".
Explicitation	Explicitates logic, intent, order, viewpoints.	"..., according to X's view."
Brevity	Keeps sentences short and clear.	"The results were strong. More tests needed."
Rhetoric quality	Delivers exceptionally high rhetoric quality.	Adapts style to particular audience.
Pleasant experience	Works reliably; pleasant voice; eye contact.	Avoids hectic speech when falling behind.
Cognitive ergonomics	Minimizes audience stress and fatigue.	Avoids complex language.

Table 1: Summary of operational (immediacy, embodiment, agency) and qualitative (faithfulness, clarity, ease of comfort) interpreting goals.

2 Goals and Scope

For our purposes, we understand *interpreting* to mean real-time oral (speech-to-speech) translation. We focus on the well studied fields of simultaneous[3] interpreting (SI) and consecutive[4] interpreting (CI), leaving less formalized paradigms (e.g. dialog interpreting) to the side, although many insights are more generally applicable.

We contend that the objective of machine interpreting should not be to uncritically replicate human interpreters, but rather to identify and emulate those aspects that are both desirable and feasible within the context of machine-based applications. For instance, interpreting research includes techniques meant to address purely human limitations affecting the interpreter, such as cognitive overload and exhaustion. The limitations machines face are different. We therefore will not focus on describing interpreting principles and techniques meant to address such human limitations.

We also note that while many of the discussed principles immediately raise questions regarding how these might be evaluated, a systematic treatment of evaluation is beyond this paper's scope.

2.1 A Note on Prior Interdisciplinary Work

Although MT and human interpreting research have progressed largely independently (Pöchhacker, 2024), the possibility of learning from human interpreters has already been discussed in ST literature (Paulik, 2010; Shimizu et al., 2013; Grissom II et al., 2014; Cheng et al., 2024; Wein et al., 2024, etc.). In contrast to these prior works, which primarily focus on specific interpreting strategies, we wish to put additional emphasis on the deeper underlying principles that drive such interpreting strategies. We argue that only focusing on specifics is too limiting, for several reasons: (1) some strategies (e.g., passivization) are highly language dependent and hard to generalize. (2) A strategy-only view tends to over-simplify the nature of human interpreting, also because (3) it is difficult to exhaustively list all strategies employed by interpreters, as evidenced by the significant number of non-overlapping strategies mentioned in above papers.

3 Features of Interpreting

To categorize the characteristics of interpreting, and discuss their implications for MT, we will employ two orthogonal descriptive systems: first, a set of operational features proposed by Pöchhacker (2024) that sheds light on the "how" of interpreting; second, a set of complementary features following Jones (2002) that characterize high quality of the rendered interpretation itself, the "what" of interpreting. Both descriptive systems (summarized in Table 1, illustrated in detail in Appendix A) characterize the differences between human interpreting goals and existing ST solutions, contributing to our

to-speech translation (S2ST), it has been convincingly argued that given the current major differences between S2ST and human interpreting, true *interpreting* is not something that machines currently achieve, hence the term *machine interpreting* should be reserved for a future in which machines "achieve credible performance in all aspects of embodied and situated cognitive processing" (Horváth, 2021; Pöchhacker, 2024).

[3]Interpreting concurrently through a microphone/earphone setup.

[4]Taking turns with the speaker.

goal of identifying aspects that true machine interpreting systems would be expected to address. The development of such systems will require both the integration of suitable existing methods, and addressing unsolved problems. Accompanying this section, Table 2 therefore summarizes prior work and suggests areas for future research.

3.1 Operational View

Pöchhacker (2024) propose that the defining features of interpreting (human or machine) are high degrees of *immediacy*, *embodiment*, and *agency*. This definition purposefully goes further than most prior work and highlights that besides the often discussed immediacy aspect, interpreting is marked by additional core characteristics, the absence of which in ST systems partially explains the usability gap that exists despite MT having already achieved high degrees of immediacy (e.g. low latency).

3.1.1 Immediacy

Interpreting is characterized by a high degree of immediacy in both a temporal and a spatial sense. *Temporal* immediacy requires that the pace of translation be determined by the source speaker and the interaction between speaker and recipient, not by the translator (Pöchhacker, 2024). Temporal immediacy is usually referred to as *real-time* translation in MT literature (Papi et al., 2024). *Spatial* immediacy indicates that the interpreter should be physically present at the location of the communicative event (Jones, 2002). Immediacy is both a desirable property and a limiting factor: real-time cross-lingual communication in a particular place at a particular time are of high utility, while also limiting the interpreter's scope for repair or revision (Kade, 1968). Note that immediacy is critical in both SI and CI. SI tends to emphasize the temporal aspect, requiring results within a few seconds, and CI tends to emphasize the spatial aspect, with the interpreter often standing right next to the speaker. In MT, spatial immediacy might easily be achieved through a portable device (Eck et al., 2010), while temporal immediacy requires efficient algorithms and hardware – even in the less demanding consecutive setting.

Temporal immediacy in CI. In this scenario, the speaker and interpreter typically stand side by side and take turns, with the interpreter rendering the contents of the speaker's previous turn into the target language. Turns can be individual sentences or longer speech fragments of arbitrary length (15-minute fragments or longer are not uncommon). Turn taking between speaker and interpreter increases the duration of a speech considerably, hence interpreters are expected to deliver the interpreted speech as efficiently as possible. Specifically, interpreters **speak immediately** when it is their turn, and aim at delivering a speech that is **shorter and more concise** than the original speech (Pöchhacker, 2012). A good rule of thumb is to aim for 75% of the source speech duration, although the ideal ratio depends on many factors, such as the speed and verbosity of the source speech (Jones, 2002). Consecutive machine interpreting is relevant in situations where simultaneous interpretation via parallel channels is not feasible.

Temporal immediacy in SI. Here, interpreters speak concurrently with the source speaker, typically equipped with a soundproofing booth for the interpreter, a microphone, and earphones for the audience. Interpreters navigate an **ideal voice-ear-span** between too low and too high (Janikowski and Chmiel, 2025). Interpreting at extremely low latency deteriorates quality by pushing interpreters toward unnaturally sounding translationese and toward making errors due to the inadequate context. But interpreting at overly high latency also deteriorates quality: it increases the risk for forgetting contents of the speech, and may result in accumulated latencies over time, often followed by compromised quality as interpreters rush to catch up.

To help interpreters balance this trade-off, Jones (2002) recommends latencies substantially lower than 5 seconds,[5] and outlines several principles: interpreters should (1) speak as soon as possible, (2) aim at grammatical speech with natural pauses not mid sentence but between completed sentences,[6] and (3) start speaking only when a semantic unit is completely available. These principles can be summarized pragmatically by stating that interpreters should **wait to speak until confident that the sentence can be finished without making unnatural breaks** caused by waiting for additional information, a principle that has been partly modeled in MT systems trained on segmented data from human interpreters (Nakabayashi and Kato, 2019) (this does not mean that all information must be available

[5] Empirically, average interpreter ear-voice-spans ranging between 2 and 5 seconds (Seeber, 2011) have been reported.

[6] Relatedly, Pradas Macías (2006) cautions that lengthy pauses (>2 seconds) may unintentionally be perceived as disfluent speech or omission errors by listeners.

when *starting* to speak the interpreted sentence). Speaking in short sentences is an effective way to reduce latency in this context. We note that the summarization principle may be especially applicable for machine interpreting purposes because it is straightforward to operationalize, such that it could aid the design of simultaneous speech-to-speech translation systems, or of streaming text-to-speech modules that are used in a cascade on top of simultaneously generated text output.

There are two situations in which the interpreter aims for **especially low latency: the beginning of the speech, and the end of the speech** (Jones, 2002). At the beginning of the speech, starting to interpret immediately is important because it signals listeners that the interpreter is ready to operate, and listeners do not need to worry about missing any contents. In this case, it is even permissible for interpreters to invent light phrases ("hello", "ladies and gentlemen", etc.) that the speaker did not actually say, just in order to say *something*. At the end of the speech, low latency is important because there may be actions to take immediately after the speech (applause, preparing replies or questions, moving around the room, etc.), such that waiting for several seconds until the interpreted speech ends may put listeners in an awkward position.

In order to achieve low latency while maintaining high quality, interpreters aim to choose **simple sentence structures** that provide flexibility and control in how one might finish the sentence (Jones, 2002). For instance, interpreters avoid starting sentences with relative or subordinate clauses, because this limits options for continuing the sentence. Other strategies exist to minimize latency which may come at the cost of compromising quality or control and must therefore be used with care. Examples are passivization, generalization (replace "spin dryer, cooker, and vacuum cleaner" by "electrical appliances"), omission of non-crucial contents, and anticipation of future content. Such compromises can be preferable over situations in which interpreters end up producing overly fast or hectic speech (Chmiel et al., 2024).

3.1.2 Embodiment

Pöchhacker (2024) introduce embodiment as referring to factors including **spatial/geographic situatedness**, and usage of **multimodal** communication channels. Human interpreters naturally base translations not just on the speaker's words, but utilize the full multimodal context. This includes reading the speaker's body language, visual cues, shared context regarding the nature of the communicative event and of the geographical location, or of any writing and diagrams that are available on slides or elsewhere in physical space, all of which might explain or disambiguate the communicative intent (Arbona et al., 2024). Moreover, interpreters will themselves actively use body language and facial expressions to clarify their interpretation, to signal readiness or technical problems, or to navigate speech discourse in the case of CI (Ahrens, 2004).

It is clear that human-level embodiment, including aspects such as spatial situatedness, input multimodality, and output multimodality, is tremendously difficult for machines to achieve. Although initial steps have been made toward supporting multimodal inputs (Caglayan et al., 2020), progress is hindered by lack of data and sparsity of visual and other sensory signal in practice. Moreover, we may also wish to render *outputs* in an embodied or multimodal fashion. Machine interpreting could be seen both at a disadvantage and at an advantage in this regard: on the one hand, unless one employs a humanoid robot or avatar (Xie et al., 2015), the MT system lacks a body and can therefore not employ gestures and facial expressions. On the other hand, it can display information on screen concurrently to generating speech, in order to convey additional (or redundant) information, signal readiness, or call out technical problems. It may employ non-speech sounds for the same purpose, open multiple speech channels in parallel through headphones, or signal which of several source speakers is currently being interpreted by cloning the speaker's voice, to name just a few design options.

3.1.3 Agency

Human interpreters possess a high degree of agency (Llewellyn-Jones and Lee, 2013): they may choose to work in consecutive or simultaneous fashion, and may spontaneously switch, e.g., in case of technical problems. They might switch from merely bridging the language barrier to addressing knowledge gaps between speaker and recipient, e.g., by adding short explanations. They may choose to ask questions of clarification back to the speaker, and can even refuse to interpret in case of inadequate acoustic conditions. They are also required to exercise good judgement in case the speaker makes errors: if the error is unintended (e.g., a number or term that's clearly wrong and unintended given the context), a good interpreter would silently correct

the error, but only insofar as this provides no embarrassment for the speaker. Interpreting cannot easily be reduced to a fixed set of behaviors and techniques, because the number of unique situations to which the interpreter would act spontaneously and fittingly is unbounded.

Pöchhacker (2024) elaborates that "agency strongly implies intentionality and would therefore be closely associated with humanness, but it can accommodate the view that a degree of agency can also be attributed to machines". He refers to Engen et al. (2016), who generalize the concept of agency in a way applicable to both humans and machines, defining it as the capacity to perform activities in a particular environment according to certain objectives. The degree of agency would then strongly be determined by (a) the actions that can be performed, (b) their type (to what degree are these **actions free and diverse**?), and (c) the **ability to interact with and influence** other actors.

Current ST systems are highly restricted in their available actions and ability to influence. Existing efforts include multilingual models (actions are language pairs), automatic voice activity detection (actions are deciding when to start/stop translating). Controllable MT also increases the number of available actions, but usually needs to be triggered by users manually. However, such manually devised modeling approaches are unlikely to scale to the degree of agency expected by interpreters, and more flexible and scalable approaches are needed.

3.2 Qualitative View

Quality in MT is traditionally understood as achieving similarity to a human-created reference translation, a perspective which might lead researchers to ignore some quality aspects that are highly relevant to interpreting. Per Jones (2002), the interpreter's overarching goal can be summarized as being threefold, namely interpreting "(1) with greatest faithfulness to the original but also (2) greatest clarity and (3) ease of comfort for the listener". Here we understand faithfulness as related to meaning and speaker intent (both broad and subtle), clarity as related to wording and voicing for optimal comprehensibility, and ease of comfort as related to form of presentation. While interpreters will strive simultaneously for optimal faithfulness, clarity, and ease of comfort, and while there is much overlap between the three goals, there may also at times be tension between these goals, requiring the interpreter to take reasonable trade-offs.[7] At the same time, it is important to realize that interpreted speech can be of *higher* quality than the source speech, e.g., by bringing additional clarity that was lacking in the source speech. In the following, we will discuss each of the three aspects, heavily borrowing from Jones (2002)'s view on the subject.

3.2.1 Faithfulness

Faithfulness is related to the familiar concept of accuracy (or adequacy) in MT literature (White and O'Connell, 1993), but takes this concept quite far: interpreters aim to faithfully convey the **speaker's intent**, i.e., what source speaker *meant* matters more than what they said.[8] This means that at the heart of faithful interpreting lies the apparent paradox that "in order to be faithful to the speaker, the interpreter must betray them" (Jones, 2002).

One common way in which a faithful interpreter is expected to deviate from what the speaker said (but not what they meant) regards cultural references. The audience may not be familiar with certain places, public figures, currencies, conversion units, commonly used metaphors, etc. There is then a choice between literal translation, resorting to a non-literal equivalent in the target language **(adaptation), and/or explanation**. In many cases, explanations are preferable (Jones, 2002), but the level of detail can be tricky to get right: the interpreter must include just enough detail to convey the speaker's intent understandably to the particular audience, without adding so much explanation as to distract from the speaker's point.

Another common example for when the interpreter is expected to deviate from what the speaker said is when the speaker **unintentionally misspeaks** (Besien and Meuleman, 2014). Jones (2002) recommends interpreters not to simply translate such mistakes as-is, because the audience may think that the error was made by the interpreter instead of the speaker, causing the audience to lose trust in the interpreter. Instead, the interpreter must choose to either silently correct the error, or to explicitly state uncertainty (but in a manner that does not embarrass the source speaker). Moreover, in some cases interpreters may tone down rude remarks which the speaker regrets as soon as

[7] This is reminiscent of the trade-off between accuracy and fluency in traditional MT (Lim et al., 2024).

[8] Faithfulness as understood here is quite different from the faithful translation approach of Newmark (1981), which falls closer to the literal side of the translation spectrum.

having uttered them, which are in that sense unintentional. On the other hand, any deliberate speech (including flawed logical arguments, impoliteness, dishonesty) must always be translated unchanged.

As a general rule, faithful interpretation means that the interpreter should say what the source speaker *would have said* in the target language, if (s)he were fluent in that language. This requires interpreters to overcome **specific linguistic challenges** that are too numerous to list here. Examples of such challenges include the need to compile a technical glossary ahead of time in preparation of technical content interpretation, paying special attention to nuances in the source speech, staying consistent in style and vocabulary in the context of long speeches, etc.

It must be noted that a critical requirement in faithful interpretation is that the interpreter is trusted by the listener (and speaker) to always be rendering a reliable interpretation and that the interpreter resists the temptation of coloring their interpreted speech with any agenda, opinions and world view of their own. The interpreter establishes this trust both by adhering to a high standard of accuracy, and by transparently admitting **uncertainty/error**. Whenever in doubt, instead of guessing, interpreters adhere to a sort of error handling cascade: first, aim for perfection; if uncertain about minor details, simplify or generalize; if uncertain about important content, ask clarification questions to the source speaker if possible (usually in consecutive mode), or else admit failure; if failure becomes too frequent due to poor interpreting conditions (e.g. acoustic), warn the audience about it or even refuse to interpret until conditions are improved (Jones, 2002).

Conceivably, many of the particular interpreting strategies discussed (establishing trust through error handling cascades, toning down rude comments, adding explanations, etc.) could be solved through available techniques such as prompting tuning or preference optimization (Yu et al., 2025). An open question is whether these rules are too numerous to solve through individual strategies, in which case one may resort either to data driven approaches or to designing simple overarching prompts that implement general principles (e.g. rendering what the source speaker would have said in the target language). However, the biggest challenge may lie in choosing *when* to apply certain solutions and when not to. This brings us back to the notion of agency discussed in the previous section: machine interpreting systems would be required to proactively select appropriate translation and error recovery strategies depending on circumstances, a desideratum that is currently only achieved by human interpreters (and tends to be the main factor that distinguishes an excellent interpreter from a good interpreter). Even assuming that a system could be designed that successfully mimics a skillful human interpreter in all of these aspects, the question of whether or not this is even desirable must be addressed: perhaps most users would want silent correction of unintentional minor mistakes only from a human interpreter but not a machine? Or perhaps only if indicated on screen and not in fact silent?

3.2.2 Clarity

Given a source sentence X and target sentence Y, it has been suggested to regard adequacy as corresponding to the conditional probability $p(X|Y)$ and fluency[9] as $p(Y)$ (Teich et al., 2020). The same could be said of faithfulness and clarity, respectively, but the relationship is less direct: faithfulness goes beyond accuracy (see previous section), and similarly clarity also goes beyond fluency and requires an active effort of producing exceptionally clear speech. We might appropriately call such clear speech "interpretese", in a positive sense.

Clarity is achieved through clear pronunciation and well-formed language, but importantly also by **explicitating intent and discourse** (Gumul, 2017). Clear, well-formed and explicit speech counteracts inevitable comprehension gaps caused by the linguistic and cultural differences (Meyer and Poláková, 2013; Meyer and Webber, 2013). For instance, there is a risk that even skillfully interpreted jokes or persuasive rhetoric fail to carry their full intended effect, and clarity can be achieved by explicitly stating the speaker's intention to persuade, or to convey humor ("said the speaker jokingly"). Similarly, implicitly expressed points of view tend to get missed by listeners despite faithful interpretation, and need repeated clarification ("according to…"). Implicit (chrono-)logical connections should also be made explicit ("first", "then"; "therefore", "but").

Clarity also demands that complex sentences be broken up into **short, easy-to-digest sentences**.

[9] Also referred to as intelligibility or well-formedness by White and O'Connell (1993).

Principle	*Existing MT research*	*Research in related fields*	*Potential research gaps*
		Temporal immediacy (consecutive case)	
Speak immediately	Efficient decoding (Junczys-Dowmunt et al., 2018), model compression (Rajkhowa et al., 2025).	Low delay endpointing (Li et al., 2002; Zink et al., 2024); recovery from false endpointing triggers (Ma et al., 2025).	General inference efficiency; Explore simultaneous MT techniques → speak before decoding finished.
Short, concise speech	Summarization (Bouamor et al., 2013; Karande et al., 2024); length control (Lakew et al., 2019).	Speech-worthy language generation (Cho et al., 2024).	Techniques exist but are not commonly applied in practice; user preferences?
		Temporal immediacy (simultaneous case)	
Ideal voice-ear-span	Simultaneous S2TT (Fügen, 2009), S2ST (Zheng et al., 2020; Ma et al., 2024); mimic interpreters (Grissom II et al., 2014; Nakabayashi and Kato, 2019).	Incremental TTS (Liu et al., 2022), concurrent speech/text generation (Yang et al., 2024; Fang et al., 2024).	Low latency S2ST methods less mature than S2TT, some room for improvement; lacking studies measuring user preference.
Zero initial/final lag	–	–	Likely unaddressed.
No unnatural breaks	Semantic segmentation (Huang et al., 2023b).	–	Related work exists, core issue likely unaddressed.
Simple sentence structure	Data driven approaches (Shimizu et al., 2013; Cheng et al., 2024); preference learning (Yu et al., 2025).	–	Initial work exists.
		Embodiment	
Multimodality	Multimodal MT (Caglayan et al., 2020), directional audio (Chen et al., 2025).	Speech-to-pictogram (Macaire et al., 2024); avatar animation (Xie et al., 2015).	Large open-ended research space.
		Agency	
Free/diverse actions, interaction	Voice activity detection (Graf et al., 2015); controllable MT (Agrawal and Carpuat, 2019).	Interactive speech LLMs (Li et al., 2025).	Promising techniques exist but need exploration and integration with MT.
		Faithfulness	
Intent translation	Non-linear translation with LLMs (Yao et al., 2024); prosodic intent (Anumanchipalli et al., 2012), lexical choice (Tsiamas et al., 2024); expressive S2ST (Gong and Veluri, 2024).	Intent discovery (Song et al., 2023).	Promising techniques exist, progress potentially hindered by established evaluation methods not rewarding shift toward non-linear translation.
Interpreter uncertainty	Quality estimation: segment level (Specia and Giménez, 2010), speech (Han et al., 2024); trust (Savoldi et al., 2025); uncertainty disentangling (Zerva et al., 2022).	Confidence estimation (Papadopoulos et al., 2000); generating "answer unknown" through reinforcement learning (Goldberg, 2023).	Plenty of related research, but ST angle underexplored.
Speaker errors	Robust MT (Anastasopoulos et al., 2019); error correction (Koneru et al., 2024).	Factual text correction (Shah et al., 2020).	Sensitive topic with initial work, needs careful investigation.
Adaptation/explanation	Cultural adaptation (Cao et al., 2024; Conia et al., 2024); named entity explanation (Han et al., 2023; Peskov et al., 2021).	Culturally aligned LLMs (Li et al., 2024).	More holistic methods and user studies needed.
		Clarity	
Discourse/intent explicitation	Labeled data (Meyer and Poláková, 2013); corpus analysis (Lapshinova-Koltunski et al., 2022).	(Chrono)logic argument (Hulpus et al., 2019; Mendoza et al., 2024), humor (Peyrard et al., 2021) detection.	Not yet addressed directly, limited related work exists.
Short, easy-to-digest sentences	Simplification for text (Oshika et al., 2024), speech (Wu et al., 2025) translation; disfluency removal (Cho et al., 2016; Salesky et al., 2019).	Text (Laban et al., 2021), spoken language (Cho et al., 2024) simplification.	Initial work exists, need more investigation, user studies, application in practice.
Rhetoric quality	Expressive S2ST (Huang et al., 2023a); on-the-fly adaptation. (Morishita et al., 2022).	Expressive TTS (Cohn et al., 2021).	Recent work exists, but needs improvement especially in simultaneous scenario.
		Ease of Comfort	
Pleasant listening experience	User-centric MT (Liebling et al., 2020; Briva-Iglesias et al., 2023); eval. through questionnaires (Müller et al., 2016).	HCI design principles (Norman, 1983); cross-lingual voice cloning (Zhang et al., 2019).	Underexplored.
Cognitive ergonomics	Translation reading assistance (Minas et al., 2025); evaluation through interviews (Müller et al., 2016), eye tracking (Castilho and Guerberof Arenas, 2018; Guerberof Arenas et al., 2021).	Reduce cognitive strain through multimodality (Malakul and Park, 2023); LLM cognitive ergonomics (Wasi and Islam, 2024).	Underexplored.

Table 2: Exemplary prior work in MT and adjacent fields addressing selected characteristics. New abbreviations: S2TT (speech-to-text translation), TTS (text-to-speech synthesis).

Short sentences help prevent interpreting errors, but crucially also improve clarity and the overall listening experience. Speech becomes more comprehensible as complex sentence structures are replaced by simple grammar that conveys the speech intent more directly and accessibly. Short sentences can be obtained through splitting of long and complex sentences into shorter ones, but also through removing redundancy (e.g. rhetorical repetitions) and disfluent speech (e.g. hesitations).

Generalizing this idea further, interpreters aim at delivering an overall **rhetorically good speech** (Jones, 2002): using neither a bored nor overly intense but a natural intonation, avoiding disfluent speech and unnatural breaks, using effective placement of prosodic emphasis, speaking with a clear pronunciation and at a natural speed. Finally, clarity is improved through usage of terminology and expressions that the particular audience can relate with (e.g. academic, political, casual).

On the MT side, many of the considerations discussed for faithfulness apply here as well, including

the question of following data-driven, fine-grained strategies driven (most prior work), or overarching principles driven approaches; machine agency to invoke methods at appropriate moments; the need for appropriate evaluation and user studies. Wording/language generation and speech synthesis both contribute to clarity and most work hand in hand.

3.2.3 Ease of Comfort

Good interpreters provide a **pleasant listening experience** for the audience (Kurz, 2001) and **prevent cognitive strain**. While interpreting faithfully and with clarity contributes to these aspects, *ease of comfort* includes additional nuances (Jones, 2002). Among others, it is achieved through speaking with a pleasant voice, through establishing eye contact and a personal connection with the audience, and through signaling an "I got you covered" attitude. It also requires reliable and intuitive technical equipment such as headphones, volume control, and connectivity for the audience. Glitches in any of these areas would pose a distraction and cause a target-language audience to feel less well integrated and cared for than a source-language audience.

Ease of comfort is not a commonly used term in the MT community, and is more difficult to formalize than faithfulness and clarity. It is closer to a human-computer-interaction (HCI) research mindset and recognizes that a well-designed user interface (UI) is just as important to users as the quality of translations. While HCI work on MT is unfortunately sparse, human interpreting best practices may provide hints for what users' needs are, such as the engineering of robust and reliable "I got you covered" systems, and going out of one's way to reduce unnecessary cognitive strain from users which typically already find themselves struggling to navigate complex multicultural situations when employing speech translation tools.

4 Discussion and Future Work

Interpreting is a complex and multi-faceted activity, and the road to developing true machine interpreting systems that perform the nuanced communicative functions expected of human interpreters is difficult. Among the many aspects discussed above, some are already successfully addressed (e.g. parts of immediacy); some are relatively low hanging fruit in the sense that while not usually integrated into MT systems, NLP methods exist to address them (e.g. summarization); some aspects are very challenging to address (e.g. agency, embodiment), but even for the challenging aspects, meaningful first steps are in sight thanks to recent progress with LLMs, multimodal representations, zero-shot learning, etc. As a starting point for identifying promising avenues for future work, Table 2 provides a (non-comprehensive) overview of research that partially addresses some of the identified characteristics, as well as potential research gaps.

It is important to note that not all discussed interpreting goals are equally relevant in all types of machine interpreting use cases: Some use cases may lend themselves to a more interactive design that benefits from sophisticated machine agency, but others may by nature be restricted in the degree of input/output multimodality, and some use cases may require taking stronger trade-offs such as prioritizing immediacy over other factors. User studies may facilitate such design choices.

A common challenge with most of the discussed principles is that they are inherently difficult to evaluate, and may in fact detrimentally impact the established evaluation metrics based on similarity to reference translations. Even human evaluation will face difficulties uncovering all aspects, e.g. identifying an ideal immediacy-faithfulness trade-off may require task-based evaluation for holistic assessment. In addition, user studies that assessing whether certain interpreting principles are actually beneficial to users would be of high value. Such user studies need careful planning because users may not always be aware of their needs, as exemplified in a study on human interpreting that found users expressing no particular preference for natural intonation in the interpreted speech, but comprehension tests revealing a noticeable positive impact of good intonation (Shlesinger, 1994).

5 Conclusion

This paper discussed human interpreting literature, with the aim of drawing implications for MT research and addressing what has been characterized as a "lack of interaction and exchange" between the two research communities (Pöchhacker, 2024). Unlike prior ST work, we have focused on higher level ideals and goals as identified by the interpreting research community, which allows sidestepping the otherwise difficult question of which specific interpreter strategies should or should not be imitated by MT. We categorized insights into operational and qualitative axes, finding that among the various aspects, only immediacy can meet the standards

of interpreting. At the same time, recent modeling advances such as general-purpose LLMs and multimodal learning seem to have paved the way toward making significant progress on many of the remaining fronts, placing the development of more user-centric ST systems within sight.

Limitations

This paper aimed to give a high level perspective of human and machine interpreting. While we hope to have covered all important aspects, by nature this vast topic cannot be comprehensively treated within the given space constraints. Among others, we have not discussed that some aspects are subject to debate among human interpreting researchers, such as the question on how eager interpreters should be to correct speaker errors. We have also centered discussions on *conference* interpreting literature, mainly owing to the abundance of literature on this setting. While the operational and qualitative aspects generalize to most interpreting settings, covering literature on additional paradigms such as dialog interpreting may yield additional insights with regards to how to trade off and prioritize the discussed dimensions. On the technical side, the provided references on prior work are only a selective sample. For a systematic and comprehensive treatment of the technical aspects, we refer to overview papers such as Seligman and Waibel (2019); Sperber and Paulik (2020); Sulubacak et al. (2020); Savoldi et al. (2025); Sperber (2019).

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China.

Barbara Ahrens. 2004. Non-verbal phenomena in simultaneous interpreting: Causes and functions. In *Claims, Changes and Challenges in Translation Studies*, pages 227–237. John Benjamins Publishing Company.

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, Minnesota.

Gopala Krishna Anumanchipalli, Luís C. Oliveira, and Alan W. Black. 2012. Intent transfer in speech-to-speech machine translation. In *IEEE Spoken Language Technology Workshop (SLT)*, Miami, FL, USA.

Eléonore Arbona, Kilian G. Seeber, and Marianne Gullberg and. 2024. The role of semantically related gestures in the language comprehension of simultaneous interpreters in noise. *Language, Cognition and Neuroscience*, 39(5):584–608.

Fred Besien and Chris Meuleman. 2014. Dealing with speakers' errors and speakers' repairs in simultaneous interpretation. *The Translator*, 10:59–81.

Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2013. SuMT: A framework of summarization and MT. In *International Joint Conference on Natural Language Processing (IJCNLP)*, Nagoya, Japan.

Vicent Briva-Iglesias, Benjamin Cowan, and Sharon Brien. 2023. The impact of traditional and interactive post-editing on machine translation user experience, quality, and productivity. *Translation, Cognition & Behavior*, 6:58–84.

Ozan Caglayan, Julia Ive, Veneta Haralampieva, Pranava Madhyastha, Loïc Barrault, and Lucia Specia. 2020. Simultaneous machine translation with visual context. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2350–2361, Online.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics (TACL)*, 12:80–99.

Sheila Castilho and Ana Guerberof Arenas. 2018. Reading comprehension of machine translation output: What makes for a better read? In *Conference of the European Association for Machine Translation (EACL)*, Alicante, Spain.

Tuochao Chen, Qirui Wang, Runlin He, and Shyamnath Gollakota. 2025. Spatial speech translation: Translating across space with binaural hearables. In *Conference on Human Factors in Computing Systems (CHI)*. Association for Computing Machinery.

Shanbo Cheng, Zhichao Huang, Tom Ko, Hang Li, Ningxin Peng, Lu Xu, and Qini Zhang. 2024. Towards achieving human parity on end-to-end simultaneous speech translation via LLM agent. *arXiv preprint arXiv:2407.21646*.

Agnieszka Chmiel, Marta Kajzer-Wietrzny, Danijel Koržinek, Dariusz Jakubowski, and Przemysław Janikowski. 2024. Syntax, stress and cognitive load, or on syntactic processing in simultaneous interpreting. *Translation, Cognition & Behavior*, 5(2):195–220.

Eunah Cho, Jan Niehues, Thanh-Le Ha, and Alex Waibel. 2016. Multilingual disfluency removal using NMT. In *International Conference on Spoken Language Translation (IWSLT)*, Seattle, Washington, USA. International Workshop on Spoken Language Translation.

Hyundong Justin Cho, Nicolaas Paul Jedema, Leonardo F. R. Ribeiro, Karishma Sharma, Pedro Szekely, Alessandro Moschitti, Ruben Janssen, and Jonathan May. 2024. Speechworthy instruction-tuned language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Miami, Florida, USA.

Michelle Cohn, Kristin Predeck, Melina Sarian, and Georgia Zellou. 2021. Prosodic alignment toward emotionally expressive speech: Comparing human and Alexa model talkers. *Speech Communication*, 135:66–75.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Miami, Florida, USA.

Matthias Eck, Ian Lane, Ying Zhang, and Alex Waibel. 2010. Jibbigo: Speech-to-speech translation on mobile devices. In *IEEE Spoken Language Technology Workshop (SLT)*.

Vegard Engen, J. Brian Pickering, and Paul Walland. 2016. Machine agency in human-machine networks; impacts and trust implications. In *International Conference on Human-Computer Interaction (HCI)*. Springer.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. LLaMA-Omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Claudio Fantinuoli. 2024. Situational awareness in machine interpreting. *Globalization and Localization Association*.

Marcello Federico, Satoshi Nakamura, Hermann Ney, Sara Papi, Elizabeth Salesky, Alex Waibel, and Shinji Watanabe. 2024. Panel discussion held on Aug 15, 2024. International Workshop on Spoken Language Translation (IWSLT).

Christian Fügen. 2009. *A System for Simultaneous Translation of Lectures and Speeches*. Ph.D. thesis, Universität Karlsruhe (TH).

Yoav Goldberg. 2023. Reinforcement learning for language models. Accessed: 2025-05-10.

Hongyu Gong and Bandhav Veluri. 2024. SeamlessExpressiveLM: Speech language model for expressive speech-to-speech translation with chain-of-thought. *arXiv preprint arXiv:2405.20410*.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. 2015. Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015.

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar.

Ana Guerberof Arenas, Joss Moorkens, and Sharon Brien. 2021. The impact of translation modality on user experience: an eye-tracking study of the Microsoft word user interface. *Machine Translation*, 35.

Ewa Gumul. 2017. Explicitation and directionality in simultaneous interpreting. *Linguistica Silesiana*, 38.

HyoJung Han, Jordan Boyd-Graber, and Marine Carpuat. 2023. Bridging background knowledge gaps in translation with automatic explicitation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore.

HyoJung Han, Kevin Duh, and Marine Carpuat. 2024. SpeechQE: Estimating the quality of direct speech translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Miami, Florida, USA.

Ildikó Horváth. 2021. Speech translation vs. interpreting. *Language Studies and Modern Humanities*, 3(2):174–187.

Wen-Chin Huang, Benjamin Peloquin, Justine Kao, Changhan Wang, Hongyu Gong, Elizabeth Salesky, Yossi Adi, Ann Lee, and Peng-Jen Chen. 2023a. A holistic cascade system, benchmark, and human evaluation protocol for expressive speech-to-speech translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. 2023b. Speech translation with large language models: An industrial practice. *Preprint*, arXiv:2312.13585.

Ioana Hulpus, Jonathan Kobbe, Christian Meilicke, Heiner Stuckenschmidt, Maria Becker, Juri Opitz, Vivi Nastase, and Anette Frank. 2019. Towards explaining natural language arguments with background knowledge. In *1st Workshop on Semantic Explainability (SEMEX) co-located with the 18th International Semantic Web Conference (ISWC)*, Auckland, New Zealand.

Przemysław Janikowski and Agnieszka Chmiel. 2025. Ear–voice span in simultaneous interpreting: Text-specific factors, interpreter-specific factors and individual variation. *Interpreting: International Journal of Research and Practice in Interpreting*, 27(1):28–51.

Roderick Jones. 2002. *Conference interpreting explained*. Routledge.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Otto Kade. 1968. *Zufall und Gesetzmäßigkeit in der Übersetzung*, volume 1. VEB Verlag Enzyklopädie, Leipzig.

Pranav Karande, Balaram Sarkar, and Chandresh Kumar Maurya. 2024. Cross-lingual summarization of speech-to-speech translation: A baseline. In *Speech and Computer: 26th International Conference, SPECOM 2024*, Berlin, Heidelberg.

Sai Koneru, Thai Binh Nguyen, Ngoc-Quan Pham, Danni Liu, Zhaolin Li, Alexander Waibel, and Jan Niehues. 2024. Blending LLMs into cascaded speech translation: KIT's offline speech translation system for IWSLT 2024. In *International Conference on Spoken Language Translation (IWSLT)*, Bangkok, Thailand.

Ingrid Kurz. 2001. Conference interpreting: Quality in the ears of the user. *Meta*, 46(2):394–409.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Online.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *International Conference on Spoken Language Translation (IWSLT)*, Hong Kong.

Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Heike Przybyl. 2022. Exploring explicitation and implicitation in parallel interpreting and translation corpora. *The Prague Bulletin of Mathematical Linguistics*, 119:5–22.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CultureLLM: Incorporating cultural differences into large language models. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Qi Li, Jinsong Zheng, Augustine Tsai, and Qiru Zhou. 2002. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Transactions on Speech and Audio Processing*, 10(3):146–157.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, and 1 others. 2025. Baichuan-Audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*.

Daniel J. Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. 2020. Unmet needs and opportunities for mobile translation AI. In *Conference on Human Factors in Computing Systems (CHI)*, New York, NY, USA.

Zheng Wei Lim, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. 2024. Simpson's paradox and the accuracy-fluency tradeoff in translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand.

Danni Liu, Changhan Wang, Hongyu Gong, Xutai Ma, Yun Tang, and Juan Pino. 2022. From start to finish: Latency reduction strategies for incremental speech synthesis in simultaneous speech-to-speech translation. In *Annual Conference of the International Speech Communication Association (Interspeech)*.

Peter Llewellyn-Jones and Robert G. Lee. 2013. Getting to the core of role: Defining interpreters' role-space. *International Journal of Interpreter Education*, 5(2):54–72.

Zhengrui Ma, Qingkai Fang, Shaolei Zhang, Shoutao Guo, Yang Feng, and Min Zhang. 2024. A non-autoregressive generation framework for end-to-end simultaneous speech-to-any translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand.

Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2025. Language model can listen while speaking. In *Conference on Artificial Intelligence (AAAI)*, Philadelphia, Pennsylvania, USA.

Cécile Macaire, Chloé Dion, Didier Schwab, Benjamin Lecouteux, and Emmanuelle Esperança-Rodier. 2024. Towards speech-to-pictograms translation. In *Annual Conference of the International Speech Communication Association (Interspeech)*, Kos Island, Greece.

Sivakorn Malakul and Innwoo Park. 2023. The effects of using an auto-subtitle system in educational videos to facilitate learning for secondary school students: learning comprehension, cognitive load, and satisfaction. *Smart Learning Environments*, 10(1):4.

Daniel Mendoza, Christopher Hahn, and Caroline Trippel. 2024. Translating natural language to temporal logics with large language models and model checkers. In *Conference on Formal Methods in Computer-Aided Design (FMCAD)*. TU Wien Academic Press.

Thomas Meyer and Lucie Poláková. 2013. Machine translation with many manually labeled discourse connectives. In *Workshop on Discourse in Machine Translation (DiscoMT)*.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Workshop on Discourse in Machine Translation (DiscoMT)*.

Dimosthenis Minas, Eleanna Theodosiou, Konstantinos Roumpas, and Michalis Xenos. 2025. Adaptive real-time translation assistance through eye-tracking. *AI*, 6(1).

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2022. Domain adaptation of machine translation with crowdworkers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Abu Dhabi, UAE.

Markus Müller, Sarah Fünfer, Sebastian Stüker, and Alex Waibel. 2016. Evaluation of the KIT lecture translation system. In *International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France.

Akiko Nakabayashi and Tsuneaki Kato. 2019. Simulating segmentation by simultaneous interpreters for simultaneous machine translation. In *Pacific Asia Conference on Language, Information and Computation (PACLIC 2023)*. Waseda University.

Peter Newmark. 1981. Approaches to translation. *Studies in Second Language Acquisition, 7 (1)*, 114:115.

Donald A. Norman. 1983. Design principles for human-computer interfaces. In *Conference on Human Factors in Computing Systems (SIGCHI)*.

Masashi Oshika, Makoto Morishita, Tsutomu Hirao, Ryohei Sasano, and Koichi Takeda. 2024. Simplifying translations for children: Iterative simplification considering age of acquisition with LLMs. In *Findings of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand.

G. Papadopoulos, P.J. Edwards, and A.F. Murray. 2000. Confidence estimation methods for neural networks: A practical comparison. In *European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium.

Sara Papi, Peter Polak, Ondřej Bojar, and Dominik Macháček. 2024. How "real" is your real-time simultaneous speech-to-text translation system? *Preprint*, arXiv:2412.18495.

Matthias Paulik. 2010. *Learning Speech Translation from Interpretation*. Ph.D. thesis, Karlsruher Institut für Technologie (KIT).

Denis Peskov, Viktor Hangya, Jordan Boyd-Graber, and Alexander Fraser. 2021. Adapting entities across languages and cultures. In *Findings of the Association for Computational Linguistics (EMNLP)*, Punta Cana, Dominican Republic.

Maxime Peyrard, Beatriz Borges, Kristina Gligorić, and Robert West. 2021. Laughing heads: Can transformers detect what makes a sentence funny? In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Franz Pöchhacker. 2024. Is machine interpreting interpreting? *Translation Spaces*.

Macarena Pradas Macías. 2006. Probing quality criteria in simultaneous interpreting: The role of silent pauses in fluency. *Interpreting*, 8(1):25–43.

Franz Pöchhacker. 2012. Consecutive interpreting. In Kirsten Malmkjær and Kevin Windle, editors, *The Oxford Handbook of Translation Studies*, pages 294–299. Oxford University Press.

Tonmoy Rajkhowa, Amartya Roy Chowdhury, Achyut Mani Tripathi, Sanjeev Sharma, and Om Jee Pandey. 2025. Semi-supervised knowledge distillation framework towards lightweight large language model for spoken language translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Elizabeth Salesky, Matthias Sperber, and Alexander Waibel. 2019. Fluent translations from disfluent speech in end-to-end speech translation. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, Minnesota.

Beatrice Savoldi, Alan Ramponi, Matteo Negri, and Luisa Bentivogli. 2025. Translation in the hands of many: Centering lay users in machine translation interactions. *Preprint*, arXiv:2502.13780.

Kilian G. Seeber. 2011. Cognitive load in simultaneous interpreting: Existing theories — new models. *Interpreting*, 13(2):176–204.

Mark Seligman and Alex Waibel. 2019. Advances in speech-to-speech translation technologies. *Advances in Empirical Translation Studies*, pages 217–251.

Darsh J. Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. In *AAAI Conference on Artificial Intelligence*.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Constructing a speech translation system using simultaneous interpretation data. In *International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany.

Miriam Shlesinger. 1994. Intonation in the production and perception of simultaneous interpretation. In *Bridging the Gap: Empirical Research in Simultaneous Interpretation (S. Lambert and B. Moser-Mercer, eds)*, pages 225–236. Amsterdam and Philadelphia, John Benjamins.

Xiaoshuai Song, Keqing He, Pei Wang, Guanting Dong, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2023. Large language models meet open-world intent discovery and recognition: An evaluation of ChatGPT. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore.

Lucia Specia and Jesús Giménez. 2010. Combining confidence estimation and reference-based metrics for segment-level MT evaluation. In *Conference of the*

Association for Machine Translation in the Americas (AMTA), Denver, Colorado, USA.

Matthias Sperber. 2019. End-to-end neural speech translation. chapter 2.4. Karlsruher Institut für Technologie (KIT).

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34(2):97–147.

Elke Teich, José Martínez Martínez, and Alina Karakanta. 2020. Translation, information theory and cognition. In *The Routledge handbook of translation and cognition*, pages 360–375. Routledge.

Ioannis Tsiamas, Matthias Sperber, Andrew Finch, and Sarthak Garg. 2024. Speech is more than words: Do speech-to-text translation systems leverage prosody? In *Conference on Machine Translation (WMT)*, Miami, Florida, USA.

Azmine Toushik Wasi and Mst Rafia Islam. 2024. CogErgLLM: Exploring large language model systems design perspective using cognitive ergonomics. In *Workshop on NLP for Science (NLP4Science)*, Miami, FL, USA. Association for Computational Linguistics.

Shira Wein, Te I, Colin Cherry, Juraj Juraska, Dirk Padfield, and Wolfgang Macherey. 2024. Barriers to effective evaluation of simultaneous interpretation. In *Findings of the Association for Computational Linguistics (EACL)*, St. Julian's, Malta.

John S. White and Theresa A. O'Connell. 1993. Evaluation of machine translation. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Jing Wu, Shushu Wang, Kai Fan, Wei Luo, Minpeng Liao, and Zhongqiang Huang. 2025. Improve speech translation through text rewrite. In *International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE.

Lei Xie, Jia Jia, Helen Meng, Zhigang Deng, and Lijuan Wang. 2015. Expressive talking avatar synthesis and animation. *Multimedia Tools and Applications*, 74:9845–9848.

Yifan Yang, Ziyang Ma, Shujie Liu, Jinyu Li, Hui Wang, Lingwei Meng, Haiyang Sun, Yuzhe Liang, Ruiyang Xu, Yuxuan Hu, and 1 others. 2024. Interleaved speech-text language models are simple streaming text to speech synthesizers. *arXiv preprint arXiv:2412.16102*.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics (EMNLP)*, Miami, Florida, USA.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Donglei Yu, Yang Zhao, Jie Zhu, Yangyifan Xu, Yu Zhou, and Chengqing Zong. 2025. SimulPL: Aligning human preferences in simultaneous machine translation. In *International Conference on Learning Representations (ICLR)*.

Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. 2022. Disentangling uncertainty in machine translation evaluation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Abu Dhabi, United Arab Emirates.

Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R. J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. In *Annual Conference of the International Speech Communication Association (Interspeech)*, Graz, Austria.

Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, Jiahong Yuan, Kenneth Church, and Liang Huang. 2020. Fluent and low-latency simultaneous speech-to-speech translation with self-adaptive training. In *Findings of the Association for Computational Linguistics (EMNLP)*, Online.

Oswald Zink, Yosuke Higuchi, Carlos Mullov, Alexander Waibel, and Tetsunori Kobayashi. 2024. Predictive speech recognition and end-of-utterance detection towards spoken dialog systems. *arXiv preprint arXiv:2409.19990*.

A Human interpreting examples

This appendix provides a non-exhaustive list of concrete examples (Table 3) illustrating the principles of human interpreting discussed in the main paper. The examples demonstrate how specific interpreting phenomena relate to both the operational dimensions (immediacy, embodiment, agency) proposed by Pöchhacker (2024) and the qualitative dimensions (faithfulness, clarity, ease of comfort) described by Jones (2002).

This appendix grounds the concepts presented in the main paper in practical scenarios that interpreters encounter. Each example shows how interpreters might have to make decisions that concurrently address multiple dimensions: for instance, how accounting for the type of audience in order to correctly adapt the text might encompass both embodiment and faithfulness.

The table is organized into functional categories (information management, handling uncertainty/errors, cultural & pragmatic adaptation, delivery & fluency, and simultaneous MT specifics) to help comparison across different types of interpreting challenges.

Table 3: Illustrative examples of linguistic and pragmatic phenomena observed in human interpreting, categorized according to both the operational and qualitative views. The *mode* column indicates whether the phenomenon is primarily relevant to simultaneous interpreting (SI), consecutive interpreting (CI), or both modes. Examples show source text/speech and corresponding target interpretations.

Phenomenon	Explanation	Op. view	Qual. view	Mode	Example (Source)	Example (Target)
Information Management						
Compression (Simultaneous)	Reducing the source content to its essential meaning while omitting detail.	Agency	Ease of comfort	SI	"The rapid increase in global temperatures, coupled with the alarming rise in sea levels, has created an unprecedented challenge for coastal communities."	"Rising global temps and sea levels challenge coasts."
Simplification / Restructuring	Reformulating complex syntax or discourse into simpler, more accessible structures.	Agency	Ease of comfort	Both	"The man, who was wearing a hat, which was red, and had a feather, which was very long, walked into the room."	"The man wearing a red hat with a long feather walked into the room."
Making Implicit Content Explicit	Making implied meaning explicit to ensure clarity for the listener.	Agency	Clarity	Both	**SPK 1**: "Are you coming to the party?" **SPK 2**: "I have to work."	**SPK 1**: "Are you coming to the party?" **SPK 2**: "No, I have to work."
Short, easy-to-digest phrases	Segmenting output into brief, manageable chunks for ease of processing.	Agency	Clarity	Both	"The new climate report, which has been reviewed by over 200 scientists worldwide and covers a 30-year span, warns of unprecedented environmental changes."	"The new climate report warns of major environmental change. It was reviewed by 200 scientists. It spans 30 years."
Handling Uncertainty / Errors						
Speaker Uncertainty (in Source)	Recognising and conveying the speaker's lack of certainty.	Agency	Faithfulness	Both	"The company reported... uh... I think it was a 10 or 12 percent... mumble... increase."	"The company reported a profit increase of approximately 10 to 12 percent."
Interpreter/MT Uncertainty	Indicating hesitation, uncertainty about the input in the produced output.	Agency	Faithfulness	Both	*[Poor audio/mumble]* "[...] reported *[...]* percent increase..."	"[...] reported some type of percent increase that I am unsure of..." *[Signals uncertainty]*
Factual Error (in Source)	Correcting factual inaccuracies in the source speech during interpretation.	Agency	Faithfulness	Both	"We're excited to be in Paris. Aachen is such a nice city."	"We're excited to be in Paris. **Paris** is such a nice city." *[Error corrected]*
Source Misspeaks / Self-Correction	Conveying the speaker's final, corrected intent while omitting false starts.	Agency	Faithfulness	Both	"The meeting is on Wednes- uh, Thursday"	"The meeting is on Thursday"
Resolving errors interactively	Revising or correcting prior output in response to listener feedback or clarification.	Agency	Faithfulness	CI	**USER**: "Turn left." **MT OUTPUT**: "Turn right." **USER**: "No, left!"	**MT OUTPUT**: *[After previous utterances]* "Sorry, turn left."
Cultural & Pragmatic Adaptation						
Cultural References / Idioms	Identifying culture-specific items/phrases and adapting them for target audience understanding.	Agency	Faithfulness	Both	"He really moved the goalposts."	**[OPTION A]** "He changed the requirements unfairly." **[OPTION B]** "He moved the goalposts. In other words, he changed the rules unexpectedly."
Play on Words / Humor	Reproducing the humour or rhetorical effect of a pun or wordplay using target-language strategies.	Agency	Faithfulness	Both	"I'm reading a book about anti-gravity. It's impossible to put down!"	**[OPTION A]** "I'm reading a book about anti-gravity. It's impossible to put down!" *[Verbatim if pun works in the target language]* **[OPTION B]** "I'm reading a book about time travel. It's about time!" *[Different pun suited to the target language]*
Understanding of Audience	Adjusting language, complexity, and detail based on the target audience.	Embodiment	Faithfulness	Both	"Gene expression was upregulated."	"The activity of the gene was increased."
Handling Offensive/Sensitive Language	Deciding whether/how to translate sensitive language based on context, audience, and ethics.	Agency	Faithfulness	Both	"That was a *** disaster."	**[OPTION A]** "That was a total mess." *[Mitigated]* **[OPTION B]** "That was a *** disaster." *[Verbatim]*
Meaning conveyed through tone	Identifying and conveying the source's rhetorical intent, tone, or style.	Agency	Faithfulness	Both	"Oh, that was a brilliant idea!" *[said with heavy sarcasm]*	**[OPTION A]** "Oh, that was a brilliant idea!" *[Maintains sarcastic tone]* **[OPTION B]** "Well, that was obviously a terrible idea." *[Explicates the sarcasm]*
Delivery & Fluency						

(Continued on next page)

(Continued from previous page)

Phenomenon	Description	Pöchhacker Concepts	Jones Concepts	Mode	Example (Source)	Example (Target)
Disfluent Speech / Handling Disfluencies	Smoothing, simplifying, or retaining speaker disfluencies depending on context.	Agency	Faithfulness	Both	“I, um, was very ha-ha-happy to at-attend.”	“I was very happy to attend.”
Incorrect Linguistic Capabilities (Source)	Choosing whether to correct or reproduce errors in the speaker’s grammar or usage.	Agency	Faithfulness	Both	"The children **is** playing."	**[Option A]** "The children **are** playing." *[Corrected]* **[Option B]** "The children **is** playing." *[Retained to reflect speaker’s level]*
Simultaneous MT Specifics						
Start speaking immediately	Beginning interpretation output as soon as speech begins, without waiting for complete context.	Immediacy	Ease of comfort	SI	*[The speaker begins talking without hesitation.]*	*[The interpreter or MT system begins output immediately, even before full sentence is heard.]*
Wait to speak until sentence can be finished	Delaying output until enough source context is available to ensure accurate output.	Immediacy	Clarity	SI	*[Speaker begins with a fragment that could be misleading without the full sentence.]*	*[Interpreter delays output until full sentence meaning is available.]*
Speak at low latency	Maintaining a continuous, minimal-delay output that closely tracks the source speech.	Immediacy	Ease of comfort	SI	*[Speaker talks continuously with minimal pauses.]*	*[Interpreter or MT output follows closely behind, maintaining fluid, real-time delivery.]*