AI vs Human Simultaneous Interpreting: Quality Assessment

Kristina Vesselskaya

Submitted in partial fulfillment of the requirements for the degree of

Master of Arts

in

Translation Studies

MAQSUT NARIKBAYEV UNIVERSITY

School of Liberal Arts

May, 2025

Word Count: 20,748

DECLARATION

I, the undersigned Kristina Vesselskaya grant to MAQSUT NARIKBAYEV UNIVERSITY the right to store and distribute my submission in print and electronic format.

I confirm that I am the sole author of this thesis, and that it does not infringe any copyright. This thesis is the result of my own original work, except where due acknowledgement has been made.

MAQSUT NARIKBAYEV UNIVERSITY will clearly identify my name(s) as the author(s) of the submission, and will not make any alteration, other than as allowed by this agreement, to your submission.

I hereby accept the terms of the above Author Agreement.

Author's signature:

Date:

07.05.2025

Abstract

Despite recent advancements in neural machine interpreting, limited research has assessed its performance as compared to human simultaneous interpreting, particularly in the English-Russian language pair. Addressing this gap, the present study investigates the quality of machine interpreting by Yandex as compared to professional human simultaneous interpreting. The research aims to identify the strengths and weaknesses of machine interpreting outputs and reveal typical errors through a convergent mixed-methods design. Utilizing a quality assessment methodology, the study integrates quantitative scoring to measure the performance and qualitative feedback to reveal particular mistakes. Thus, a quality assessment scale consisting of score-based and feedback components was adapted to evaluate six audio fragments interpreted by both professional human interpreters and Yandex neural networks. Five expert assessors were sampled to ensure objective evaluation of the interpretations. Quantitative results indicated that human interpreters outscored machine interpreting in terms of logical cohesion, terminology, and style. In contrast, Yandex scored higher than humans in completeness and fluency of delivery, successfully handling strong regional accents and high speed of delivery. Qualitative analysis identified that while machine interpreting demonstrated no cognitive limitations typical for human interpreters, it resulted in lexical and grammatical redundance, producing overloaded sentences difficult for comprehension. Yandex also misinterpreted numbers, leading to significant meaning distortions. Unnatural delivery, marked by a robotic, monotonous voice and lack of prosody further diminished the output quality. Additionally, machine interpreting struggled with context recognition, resulting in inaccurate word choices and terminological inconsistencies. This study concludes that while machine interpreting cannot yet fully replicate human expertise, it holds potential as supportive technologies – particularly in assisting human interpreters or offering cost-

effective solutions for low-stakes communicative events. Large-scale empirical studies should be conducted in the future to evaluate professional-grade machine interpreting tools in real-time conditions and to consider user perceptions.

Key words: artificial intelligence, machine interpreting, simultaneous interpreting, quality assessment.

Аннотация

Несмотря на недавние достижения в области нейронного машинного устного перевода, мало исследований посвящено оценке его эффективности по сравнению с синхронным переводом человека, особенно в паре английский-русский. Устраняя данный пробел, в настоящем исследовании рассматривается качество машинного устного перевода Яндекса по сравнению с профессиональным синхронным переводом человека. Цель исследования – определить достоинства и недостатки машинного устного перевода и выявить типичные ошибки, используя конвергентный смешанный метод исследований. Применяя методологию оценки качества, исследование объединяет количественную оценку для измерения эффективности и качественную обратную связь для выявления конкретных ошибок. Для оценки шести аудиофрагментов, переведенных как профессиональными переводчиками, так и нейросетями Яндекса, была адаптирована шкала оценки качества, состоящая из балльной системы оценки и отдельного компонента для обратной связи. Для объективной оценки переводов были отобраны пять экспертов. Количественные результаты показали, что переводчики превзошли машинный перевод по логической связности, терминологии и стилю. Однако Яндекс превзошел по полноте и беглости перевода, успешно справляясь со сложными региональными акцентами и высокой скоростью. Качественный анализ показал, что, хотя машинный перевод не имеет когнитивных ограничений, характерных для переводчиков, это приводит к лексической и грамматической нагроможденности, усложняя понимание сложных предложений. Яндекс также неправильно перевел числа, искажая смысл. Неестественная подача в следствии роботизированного, монотонного голоса и отсутствия смысловых ударений ухудшала качество перевода. Кроме того, машинный перевод затруднялся с распознаванием контекста, что приводило к

неточному выбору слов и терминологическим несоответствиям. Данное исследование позволяет сделать вывод, что, хотя машинный перевод пока не может полностью повторить опыт человека, его можно потенциально использовать как вспомогательная технология в качестве помощника переводчика или использоваться как бюджетное решение для менее серьезных коммуникативных задач. Крупные эмпирические исследования необходимо провести в будущем для оценки профессиональных инструментов машинного устного перевода в условиях реального времени и с учетом восприятия пользователей.

Ключевые слова: искусственный интеллект, машинный устный перевод, синхронный перевод, оценка качества.

Аңдатпа

Жасанды нейрондық жүйелерге негізделген машиналық ілеспе аударма саласындағы соңғы жетістіктерге қарамастан, әсіресе ағылшын-орыс тіл жұбы бойынша оның кәсіби ілеспе аудармамен салыстырғандағы сапасы әлі де жеткіліксіз зерттелген. Осы ғылыми олқылықты толтыру мақсатында бұл зерттеу Яндекстің машиналық ілеспе аудармасының сапасын кәсіби аудармашылардың орындауындағы ілеспе аудармамен салыстыра отырып талдайды. Зерттеудің мақсаты – машиналық аударманың артықшылықтары мен әлсіз тұстарын анықтау және жиі кездесетін қателерді ашып көрсету. Осы мақсатта конвергентті аралас әдіснамалық тәсіл қолданылды. Сапаны бағалау әдістемесі аясында зерттеу сандық бағалау арқылы аударманың сапасын өлшеп, сапалық кері байланыс арқылы нақты қателерді анықтауға ұмтылды. Кәсіби аудармашылар мен Яндекстің нейрондық жүйесі орындаған алты аудиофрагментке баға беру үшін баллдық жүйе мен жазбаша кері байланысты қамтитын арнайы сапа шкаласы бейімделіп қолданылды. Аудармаларды объективті бағалау үшін бес сарапшы іріктелді. Сандық нәтижелер кәсіби аудармашылардың логикалық тұтастық, терминология және стиль бойынша машиналық аудармадан жоғары нәтиже көрсеткенін дәлелдеді. Ал Яндекс, керісінше, мазмұнның толықтығы мен жеткізудің жатықтылығы жағынан жақсырақ нәтиже көрсетті – әсіресе күшті аймақтық акценттер мен жоғары сөйлеу қарқынын тиімді өңдеді. Сапалық талдау көрсеткендей, машиналық ілеспе аударма адам аудармашыларына тән когнитивтік шектеулерден бос болғанымен, лексикалық және грамматикалық артықтықтарға жол беріп, қабылдауға қиын әрі ауыр сөйлемдер туғызды. Сондай-ақ, Яндекс сандарды қате аударып, мағына бұрмалануларына себеп болды. Роботтық, монотонды дауыс және просодияның болмауы аударманың табиғилығына теріс әсер етті. Бұған қоса, контексті танудағы қиындықтар сөз

таңдаудағы дәлсіздіктер мен терминологиялық сәйкессіздіктерге әкелді. Бұл зерттеу машиналық ілеспе аударманың әлі де кәсіби мамандардың шеберлігін толық алмастыра алмайтынын көрсетсе де, оны көмекші құрал ретінде немесе маңыздылығы төмен коммуникациялық жағдайларда үнемді шешім ретінде пайдалануға болатынын айғақтайды. Алдағы уақытта нақты уақыт режимінде жұмыс істейтін кәсіби деңгейдегі машиналық ілеспе аударма құралдарын кең ауқымды эмпирикалық зерттеулер негізінде, сондай-ақ пайдаланушылардың пікірлерін ескере отырып бағалау қажет.

Түйін сөздер: жасанды интеллект, машиналық ауызша аударма, ілеспе аударма, сапаны бағалау.

Table of Contents

List of Figures

List of Tables

Introduction

The influence of artificial intelligence (AI) on humanity is comparable to the invention of electricity hundreds of years ago or even the discovery of fire, as stated by a number of prominent figures in the field (Sheikh, Prins & Schrijvers, 2023). AI marks a new milestone of technological development where a wide range of professional fields are on the verge of a fundamental shift. Diverse language-related services such as translation and interpretation are not an exception. Nowadays there are over 700 technological language solutions focused on machine translation and interpreting, automatic speech recognition, transcription, and other services, according to Nimdzi Language Technology Atlas (2023). Even simultaneous interpretation which is believed to be one of the most comprehensive cognitive tasks the human brain is capable of (Garbovsky, 2021; Wu, 2010; Zhang, 2017) is subject to automation by the latest breakthrough AI tools playing the life-changing role for the language specialists. It results in a skyrocketing number of books, research papers, training programs, and conferences aimed at exploring the role of technologies in the translation and interpretation market (Guo et al., 2023). These endeavors are actively backed by language service departments of some government bodies. For example, The Knowledge Centre on Interpretation at the European Commission has founded a research and technology space to promote the introduction of technologies in interpreting (Guo et al., 2023). Contributing to this emergent and hotly debated field, the present study "AI vs Human Simultaneous Interpreting: Quality Assessment" states that the recent technologies are promising, but still quite understudied in the context of machine interpreting, particularly in the English-Russian language pair. Thus, machine interpreting requires careful attention in the form of quality control and empirical output analysis in close relation to traditional human interpreting. This section substantiates this thesis statement by providing the necessary research background,

formulating a problem statement, research objectives, and questions while highlighting the high significance and novelty of the topic.

Research Background

Simultaneous interpreting has traditionally been conceptualized as a cognitively demanding task fundamentally reliant on human agency and expertise (Fantinuoli & Dastyar, 2022). Being one of the most comprehensive brain exercises, it has always required immense creative and social human intelligence (Garbovsky, 2021) and as a result "has been historically reluctant to technological change" (Fantinuoli, 2019, p. 5). Even after the creation of the machine cross-lingual algorithms which marked the beginning of the machine translation (MT) era and boomed the translation market decades ago (Garbovsky, 2021), these technologies did not impact the interpreting profession. Until recently they were analyzed in isolation without its direct impact on the interpreting field (Fantinuoli, 2023). Nowadays it has changed since much progress has been achieved in technologization of interpreting (Liu & Liang, 2024) attracting research interest of academia and industry circles (Fantinuoli, 2018). However, it triggers major concerns about AI fully substituting human translators and interpreters in the near future, as stated by Elon Musk (2021), a famous businessman and investor. This statement seems to be reasonable, especially considering the technologies thanks to which cross-language communication barriers might be erased without the support of translators and interpreters. For example, a Russian technology company Yandex (2021) has started to provide in their browser free real-life captioning, machine interpreting, and full-fledge simultaneous interpretation of live broadcasts in beta version from several foreign languages into Russian. Thus, users are able to consume foreign language video content anytime for free and without waiting for translators or interpreters to provide the services. One more example is Apple Inc., one of the biggest technological corporations in the world. It has

announced a new AirPods feature with a function of live translation of conversations (Gurman, 2025). Samsung (2024) has successfully introduced a similar function – AI-powered live translation of phone calls. Real-time multilingual communication is streamlined through this feature, which transforms the device into a personal translator by enabling immediate translation of phone calls. Therefore, people are already able to communicate irrespectively of language barriers and without human interpreters just by using their smartphone devices.

Interestingly, machine interpreting tools approach not only daily foreign language content or simple inter-language communication but also professional conference interpreting services at bilateral or multilateral meetings and large-scale events. Such companies as KUDO, Meta, Stenomatic, Wordly, Interprefy actively develop AI-driven tools for full automation of interpreting services. KUDO, for example, created a special AI for automated simultaneous interpreting which is available on a commercial basis for both small and large-scale multilingual events. According to the company's website, there is a growing number of success stories when KUDO AI was introduced instead of human interpreters in different kinds of events (KUDO, n.d.). More technologies that might potentially replace the work of consecutive or simultaneous interpreters are on the way out as technology continues to advance rapidly. Still, some experts firmly believe that machines will not be capable of high-quality interpreting and literary translations, according to the Rector of Moscow State Linguistic University, also known as MGLU (2020). High-stakes confidential or political negotiations are also out of the equation and will continue to be interpreted without technology integration, as argued by Fantinuoli (2019), a prominent expert in technologization of interpreting and founder of KUDO AI. Nevertheless, human substitution at least to some degree seems to be irreversible, since the

recent breakthrough development of neural AI networks resulted in fully automated and quite competitive simultaneous interpretation tools, as exemplified above.

This emerging AI-integration trend raises the question of the quality of the provided interpretation in comparison to the traditional human-performed services, especially considering the continuous technological upgrades. Therefore, it is extremely important to keep track of the output improvements and empirically examine its benefits and limitations. However, unlike machine translation which counts plenty of research studies focused on diverse aspects of human versus machine output in multiple language pairs, research activity in the field of machine interpreting is still in its infant stage (Carl & Braun, 2017; Guo et al., 2023; Liu, 2023). Certainly, there is a growing number of studies circling around computer-assisted interpreting (CAI) tools aimed at simplifying the interpreters' duty by integrating the latest technologies such as speech recognition, terminology management, and others, as illustrated in the works of Defrancq and Fantinuoli (2021), Desmet et al. (2018), Fantinuoli et al. (2022), Prandi (2023), Wang and Wang (2019), to name just a few. Yet, as far as the literature demonstrates, fully automated MI in relation to human performance, especially in simultaneous mode, is the focus of a relatively smaller number of studies, such as Müller et al. (2016), Fantinuoli and Prandi (2021), Liu and Liang (2024) and other. All the abovementioned papers demonstrate notable improvements in machine outputs across different languages including English, Chinese, German, and Italian while also indicating major linguistic and non-linguistic elements machines struggle with, as specified in the literature review chapter. However, to the best of the researcher's knowledge and at the moment of writing, there are very few open and available studies touching upon the application of technological disruptions specifically in English-Russian interpreting. In light of research endeavors in other

language pairs, it constitutes a gap and is directly related to the problem statement of the present paper as described in the following paragraph.

Problem Statement and Significance of the Study

As mentioned above, research activity in the field of technologies in interpreting is still in its early stage of development. Thus, it remains unclear, particularly in the English-Russian language pair, where very few studies approached this issue, to what extent machine interpreting tools are really competitive compared to conventional human interpreting and what are their typical benefits and limitations. At the moment of writing, there are only several publicly available studies and articles that analyzed the future of English-Russian interpreting in the view of technologization, for example, Avedova and Miteleva (2016), Garbovsky and Kostikova (2019), Sarmanova (2022). Yet, the existing limited number of studies did not empirically examine the output of the newest AIs developed specifically for interpreting automation which underpins the research problem of this paper. In the era of AI disruption where no field, including interpreting, will be untouched, knowledge of AI operation, strengths, and weaknesses is extremely important. It presents both theoretical and practical significance. In terms of theoretical significance, the present study developed quality assessment framework in order to estimate quality of different versions of machine and human outputs. This assessment framework might be potentially used in the further studies in this field aimed at quality assessment and comparison of simultaneous interpreting. This Master's thesis also has practical significance for all stakeholders of the interpreting market. For instance, practitioners and academia need to be aware of the practical strengths and weaknesses of machine interpreting in order to make an appropriate judgment in which subject domains humans will probably be replaced due to equivalent machine output or vice versa, where human interpreters still outperform AI. It also serves the interests of employers and end-users of

interpreting services who need to be informed when to resort to traditional human interpreters, and when to embrace AI-driven machine interpreting solutions that are more financially accessible (Massey & Ehrensberger-Dow, 2017). Thus, it proves the novelty, high relevance, and practical significance of the present study which intends to address the existing research problem related to the lack of information on how machine interpreting tools operate in English-Russian language combination. The next paragraph detailly describes the main goal of this Master's thesis underpinning the research questions as well as the methodology utilized to answer them.

Research Purpose and Questions

Taking into consideration the existing research gap, this study aims at a critical analysis of AI-driven machine interpreting tools in the English-Russian language pair. It intends to reveal the strengths and weaknesses of machine interpreting output quality in comparison to traditional human interpreting. Thus, as a result of the study, it will be clear, to what extent machine interpreting is competitive, what are the comparative advantages and common error patterns. To meet this purpose, the study needs to answer two research questions:

1. How does the performance of AI-driven machine interpreting in the English-Russian language pair measure up against human interpreters?
2. What are the typical mistakes and limitations in AI-driven machine interpreting in the English-Russian language pair?

First research question requires quantitative approach in order to measure the outputs while the second question is about more in-depth qualitative considerations to find examples of mistakes. Thus, mixed-method approach is used in this Master's thesis. In order to answer these questions, the study needs to empirically test a machine interpreting tool and critically analyze, and assess performance in relation to the corresponding human

interpreting. Quality assessment is a research method that falls under the category of translation studies (Saldanha & O'Brien, 2014) and provided a pathway to answering the research questions. This method fully corresponds to the nature of this study aimed at revealing the quality of the different versions of interpretation. This method can be both quantitative and qualitative in nature (Saldanha & O'Brien, 2014) and implies an assessment framework with specific criteria for consideration created or adapted from the existing assessment scales enabling the human assessors to carefully analyze and evaluate the performance based on the scale indicators (Han, 2021; Saldanha & O'Brien, 2014). Thus, an analytical assessment scale with detailed criteria descriptors for each score was adapted specifically for this study in order to answer the research questions in depth. The assessment framework utilized in this study is described in the methodology chapter.

Summary

This introductory chapter provided a solid foundation for further exploration of the topic. The overview of the research background highlighted the current trends in the field and the existing research gap to be addressed in this paper. The clear delineation of the problem statement underscores the lack of knowledge on the quality of AI-generated simultaneous interpreting in the English-Russian language pair, motivating the need for empirical research articles examining the strengths and weaknesses of AI language solutions. The research purpose and questions have been articulated in relation to the gap of knowledge to guide the investigation towards meaningful insights and contributions to the field. Throughout the section, the practical significance of the topic for the wide range of stakeholders has been strongly emphasized. The next paragraph concludes the introduction chapter by providing the structure of this Master's thesis chapter by chapter.

This paper consists of six chapters. The first is the Introduction chapter as provided in the previous sections with a detailed outline of the research background, problem

statement, and its significance as well as research purpose and questions. The Literature Review chapter reviews the relevant literature on machine interpreting in different language pairs, including the development path of machine interpreting, empirical studies and real-life implementation cases as well as existing quality assessment methods. This extensive chapter is summarized in the conceptual framework explaining how all the major findings are interrelated highlighting the research gap and pathways to fill it. The next chapter is Methodology, it details and justifies the selected mixed-method approach, including the criteria used for quality assessment method, participants selection, quantitative and qualitative data collection procedures, as well as data analysis approaches. The Results chapter presents the findings, beginning with the quantitative results from the t-test, followed by an analysis of the qualitative feedback to identify common error patterns. The Discussion chapter discusses the major findings in relation to the research questions, relates them to the previous studies in the same area, and considers further implications. Finally, the last chapter is the Conclusion, it summarizes the thesis, outlining the key insights, acknowledging limitations, and suggesting future research directions. In accordance with this structure, the next chapter reviews existing literature on the topic of this Master's thesis.

Literature Review

Recognized as a highly intensive cognitive task (Garbovsky, 2021), simultaneous interpreting has traditionally been framed as a human-dependent process grounded in cognitive operations (Fantinuoli & Dastyar, 2022). Aa a result, interpreting, unlike other professions, was not a subject of automation for quite a long time (Fantinuoli, 2019). However, due to skyrocketing development of artificial intelligence and neural machine translation, some researchers believe that human dominance in the translation and interpreting fields is questioned these days (Downie, 2019; Pym & Torres-Sim´on, 2021). Recent developments in interpreting technologies have captured the attention of both academic researchers and industry professionals (Fantinuoli, 2018), leading to the emergence of various AI-driven solutions for automated interpretation. The purpose of this chapter is to review the growing body of research surrounding technological turn in the interpreting field and to provide the solid literature background for the present empirical research “AI vs Human Simultaneous Interpreting: Quality Assessment”. This literature background helps to predict and estimate possible outcomes of the practical part of this paper aimed at answering the following research questions: How does the performance of AI-driven machine interpreting in English-Russian language pair measure up against human interpreters? What are the typical mistakes and limitations in AI-driven machine interpreting in English-Russian language pair? This chapter concludes that there is a significant gap in knowledge on how machine interpreting operates in English-Russian language pair. Yet, based on the existing literature in other language pairs and the latest real-life implementation cases, it is plausible to assume quite accurate and adequate interpreting which nevertheless inferior to human performance in many aspects. This chapter also provides critical review of the existing quality evaluation criteria in simultaneous interpreting field and presents original quality assessment framework based

on the examined literature. This quality assessment framework was used in the empirical part of this study to estimate the performance of machine automatic output in relation to human interpreting. The paragraphs of this chapter are organized according to the major literature review findings based on the studies of the prominent and credible figures of the field over the last 10-15 years. The relevant literature was searched primarily in Google Scholar, Research Gate, Taylor & Francis Online, Wiley Online Library and other credible databases. The articles were selected based on their credibility, relatedness with the topic and expertise of the authors. The first section of this chapter introduces explanations of the major concepts widely used throughout the paper. The second section provides an overview of the development path of automated interpreting. The third section reviews and analyses previous empirical studies on machine interpreting output in different language pairs followed by the fourth section devoted to some real-life cases of machine interpreting implementation. The final paragraph reviews the existing quality assessment principles in order to adopt a list of assessment criteria necessary for the methodology section. This chapter is summarized in the concluding paragraph with the main literature take-aways presenting the visual conceptual framework in relation to this Master's thesis purpose and questions.

Concepts

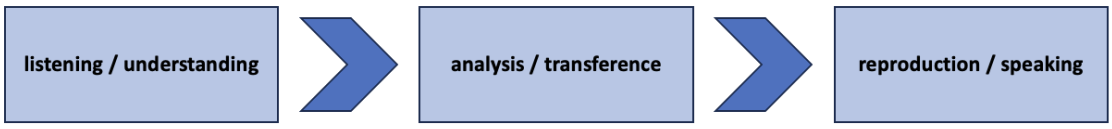
The topic of this study is directly related to the interpreting technologies. Therefore, it entails special terms which could be unclear for the unfamiliar audience. To be able to comprehend the intricate subject matter of the field, it is necessary, first and foremost, to define the major concepts which will be used throughout the paper. They are: simultaneous interpretation (SI), artificial intelligence (AI) and machine interpreting (MI).

Simultaneous Interpretation (SI)

The concept of SI has a paramount significance to this thesis. Sometimes confused with other types of oral translations by unfamiliar audience, SI can be defined as "a mode of interpreting in which the speaker makes a speech and the interpreter reformulates the speech into a language his audience understands at the same time" (Pearson, 2022, para. 2). Such practice requires special interpreting equipment, headphones and microphone for all the speakers, listeners and interpreters (Tirosh, 2023) which perform their job in a sound-proof cabin, also known as booth (Wu, 2010). As explained by Selescovitch, a prominent conference interpreter and theoreticians in the field of SI, human interpreting (HI) is a triangular process involving active listening and understanding in the source language, analysis and structuring of the message, and reproduction of the meaning in the target language altogether accomplished in a matter of few seconds (Zhang, 2017). Similarly, Herbert (as cited in Pöchhacker, 2024) also envisioned HI as a sequence of three steps: understanding, transference and speaking. Therefore, the act of simultaneous interpreting can be visually represented as follows in the Figure 1.

Figure 1

Simultaneous Interpreting Process



This is one of the most difficult linguistic tasks the human brain is capable of since it requires deep concentration and the ability to listen, process the information and reformulate it another language at the same time (Wu, 2010). That is why simultaneous interpreters work in a group of two interpreters and change every 15–30 minutes since brain is incapable to perform this task for a long time (Moser-Mercer et al., 1998). Undoubtedly, SI requires extensive preparation, aptitude and can be successfully

performed only upon years of training (Zhang, 2017). Yet, with the advancement of the language technologies and artificial intelligence, the first automated interpreting tools gradually emerged in the field to simplify such a complex human activity (Corpas Pastor, 2018). In this regard, artificial intelligence plays the central role in the development of interpreting technologies.

Artificial Intelligence (AI)

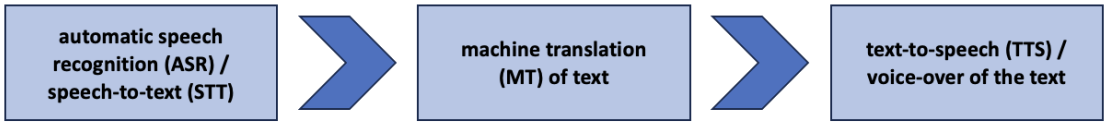
The concept of AI has quite a complicated nature. The term was officially coined and accepted as a field of study during the Dartmouth University Workshop in 1956 and was initially defined as "the science and engineering of making intelligent machines" (Manning, 2020, para. 1). Certainly, nowadays literature evidences much broader definition. Due to continuous technological development, huge investments, and countless improvements, AI has gained more advanced meaning and become closely interrelated to other complex disruptive technologies (Sheikh et al., 2023). It resulted in the absence of one universally accepted definition of AI these days, as noted by a number of authors (Horváth, 2022; Zhang, 2017). AI is rather an umbrella term that covers a huge variety of technologies such as machine learning, algorithms, natural language processing (NLP), and other (Coursera, 2023; Zhang, 2017). Focusing on its key features, one of the possible AI definitions is "the ability of a machine to display human-like capabilities such as reasoning, learning, planning, and creativity" (European Parliament, 2023, para. 3). The advanced AI tools also include automatic speech recognition, conversion of speech into a text, neural translation of the texts into a target language and its voice-over (Corpas Pastor, 2018; Prandi, 2023). Therefore, it can automate the work of a human interpreter, as argued by Wang and Wang (2019). This AI-powered process is also known as machine interpreting, as explained in the next paragraph.

Machine Interpreting (MI)

As well as in the case of AI, there is no single established definition of the MI concept either, since machine interpreting has much shorter history compared to widespread machine translation (Fedorova, 2023). As noted by Pöchhacker (2024), there is a "terminological mess" (p. 10) in the field where the same concept of automated simultaneous interpreting is referred to in various studies differently using inconsistent terminology. Machine interpreting sometimes also referred to as automated interpreting, speech translation, spoken language translation or speech-to-speech translation. Despite the terminological inconsistencies, all of the listed terms are practically synonyms and denote the same concept of automated interpreting using cutting-edge technologies. This study mainly utilizes the "machine interpreting" term, but some of the other existing terms denoting the same idea might be used throughout the paper synonymously. One of the possible definitions of MI is "the transmission of a spoken message in one language into a spoken message in a different language using AI, without the input of a human interpreter" (Hynes, 2022, para. 2). MI involves a combination of technologies, also called cascading system (Corpas Pastor, 2021). Cascade is a sequence of consecutively performed steps to produce simultaneous interpreting, where the first step is the speech-to-text conversion (STT), also known as automatic speech recognition (ASR) performed in order to transfer spoken language into a written text (Prandi, 2023; Wang et al., 2022). The second step of the cascade is machine translation (MT) of written text from a source language intro a target language (Corpas Pastor, 2021; Prandi, 2023; Wang et al., 2022). The final step is text-to-speech synthesis (TTS), in other words the voice-over of the translation (Corpas Pastor, 2021; Prandi, 2023; Wang et al., 2022). Therefore, the process of machine interpreting can be visually represented as follows in the Figure 2 below.

Figure 2

Machine Interpreting Process



Interestingly, all three cascade steps of MI have certain flaws hampering the interpreting output. For example, the speech recognition step is significantly impeded in case of a regional dialects, speaker's errors, speech disfluency, external sounds and noises, and other factors (Krenz, 2008). As of the second step of the cascade, automatic machine translation systems are used, which are not always able to translate a highly difficult text in a given context and might be still inferior to more careful human translation (Drozdova, 2015; Li & Chen, 2019). At the third stage of speech synthesis the greatest flaw is harmonization of the acoustic design of the translated on the second stage text with the source speech, which might also be performed inaccurately given the differences in intonation and pitch manners in a given language pair (Radwinski, 2017). The majority of the automated interpreting tools use the cascading system detailly explained above, yet, there are attempts to develop an end-to-end model which does not require ASR step where most of the mistakes come from (Prandi, 2023; Wang et al., 2022). It could significantly improve the output quality and reduce error propagation, i.e., repetition of the same mistake throughout interpreting process (Wang et al., 2022). End-to-end systems are explored by Google and are extremely challenging to setup since a huge amount of data needs to be trained for the model (Wang et al., 2022). The next section seeks to explore the development path and reasons for rather slow integration of such technologies in the interpreting market.

History and Prospects of MI

Despite being introduced in the 1990s, primitive types of interpreting technologies did not receive such popularity as machine translation (MT) which appeared by the 1950s (Ahmed, 2022; Horváth, 2022). Claudio Fantinuoli, Ph.D., founder of InterpretBank and KUDO AI, in his interview for Dastyar (Fantinuoli & Dastyar, 2022) explained the reason for rather low practical usage of interpreting technologies. He stated, "interpreting has, on one hand, a long tradition of being a very human-centric and cognitive-oriented activity, and on the other, it is rather a small profession if you think for example of the relatively limited numbers of practitioners" (Fantinuoli & Dastyar, 2022, p. 186). A similar point has been made by Garbovsky (2021), a professor at the Moscow State University, who stated that SI takes complex cognitive abilities, and creative and social intelligence requiring an exclusively human approach. As a consequence, demand and trust in interpreting technologies remained low and has led to the fact that interpreters compared to translators, have not benefited from the major technological breakthrough, relying on the long-established manual methods, according to Costa et al. (2014). One more possible reason for slow integration lies in economic considerations since such tools might decrease interpreters' remuneration, as marked by Ahmed (2022). According to statistics, payments for written translation have been reduced by approximately 50% since 2008 because of MT introduction and economic optimization (DePalma et al., 2013). Sarmanova (2022), Ph.D. of Translation Studies and prominent Kazakhstani freelance interpreter stated that the interpreting market is not immune to the similar developments. Despite the relatively small impact of technology on the interpreting sector, the pressure to adopt new technologies increases year by year and can eventually lead to the gradual substitution of humans in some contexts (Sarmanova, 2022). There are also concerns about low quality of interpreting and dehumanisation of the profession, as marked by Jourdenais and Mikkelson

(2015). These concerns were supported by Wang and Li (2022) who conducted a large survey and interviews among Chinese interpreters and found out that half of the respondents had doubts and skepticism toward interpreting technology. Similarly, researchers in the European context such as Pym (2011) and Fantinuoli (2018) also argue that interpreters mostly distrust the promising language technologies. Therefore, due to the reluctance and limited practical usage of interpreting technologies, the research body has started its formation considerably later and resulted in a comparatively small number of empirical studies in this field, as marked by Corpas Pastor (2018). The next section provides the review of the existing limited number of research focused on evaluation of MI performance. It gives some insights on the actual machine output and helps to predict the possible outcomes of the present study.

MI Output Quality

The limited number of empirical studies on MI and computer-assisted interpreting (CAI) evidence both strengths and weaknesses of the provided output. Fantinuoli (2017) tested the InterpretBank that transfers speech into a terminological glossary that can be subsequently used by an interpreter in the preparation process. The experiment resulted in successfully recognized 11 out of 11 numerals and 113 out of 119 terms in the English-Italian language pair (Fantinuoli, 2017). Similar evidence is found in the Doctoral thesis of Sarmanova (2022) who conducted an experiment to test Russian-English speech recognition and automatic translation integrated into Google Doc, YouTube, and Google Assistant. The speeches contained terms, abbreviations, numbers, names, different accents, and speeds. Sarmanova (2022) as well as Fantinuoli (2017), marked very accurate recognition of all the numbers and major terms. Yet, the tested applications failed to identify some abbreviations, names and collocation which were misarticulated by the speaker. For example, the name “Calea Power” was interpreted in Russian as “*Коля*

Power". The equivalent mistakes were found in the study of by Belenkova (2019), associate professor at People's Friendship University of Russia (RUDN), when testing machine interpreting provided by Google, Microsoft and Yandex. Using the similar methodology, Belenkova (2019) as well as Sarmanova (2022) comparatively assessed the performance of several automatic speech recognition and translation tools. The study brought resembling results in terms of ASR. To illustrate, the proper name "Foni Joyce" was recognized as "funny J's" and subsequently interpreted as "*Веселье Джи*" by Google Translator (Belenkova, 2019). Microsoft Translator recognized this proper name as "20 days" while Yandex did not interpret this part of the speech at all (Belenkova, 2019). Moreover, in the study of Sarmanova (2022), "CIPE" organization was wrongly recognized as "cite" and interpreted as "*сайт*". "Kazakhstan" was initially recognized by the tool as "calloused and" and later as "Kurdistan" (Sarmanova, 2022). Even relatively simple words were occasionally misrecognized: "prices" turned into "crisis", "corporation" into "cooperation", etc. (Sarmanova, 2022). Unfortunately, Belenkova (2019) did not provide the specific examples in her study as Sarmanova (2022), except of the mentioned above, but also marked low lexical and grammatical validity of the tested tools, especially Yandex which failed in almost every semantic segment. Thus, both studies in the English-Russian language pair (Belenkova, 2019; Sarmanova 2022) highlight the issue of incorrect speech recognition and as a result, incorrect translation. This type of mistakes is related to the limitations of the ASR step of the MI cascade system as was mentioned in the previous paragraphs. Sarmanova (2022) summarized that it is caused by the regional accents, spontaneous, rapid and misarticulated speech which is not always recognized by automated tools correctly. Yet, the clearly articulated and slow speeches were successfully recognized and interpreted with up to 99% accuracy (Sarmanova, 2022). Belenkova (2019), oppositely, did not explore the reasons for poor MI output, but expressed the need for

further studies in this domain to fill the evident research gap. Generally, in the opinion of Fedorova (2023), human interpreters still grasp the meaning better and are able to identify the essential nuances at the level of subtext or intonation, and to correct the speaker's mistakes and reservations, if needed. Unfortunately, to the best of the researcher's knowledge, the studies of Belenkova (2019) and Sarmanova (2022) are the only published and publicly accessible papers on MI in the Russian-English language pair at the moment of writing, making the research gap evident. Therefore, reviewing studies in other language pairs is important to better understand typical MI output.

Considering the literature in other language pairs, similar outcomes are evident. Some MI weaknesses seem to be present across different contexts. For example, Liu (2023) reviewed several studies in this area in the Chinese context and came to a conclusion consistent with Sarmanova's (2022). Liu (2023) made comparative analysis of MI and human interpreting (HI) outputs based on the previous studies in China. As well as Sarmanova (2022), Liu (2023) highlighted that such factors as accent, tone, sound volume, speed and background noises negatively affect MI output since the existing technologies still struggle in overcoming these challenges. The author noted that AI-powered tools outperform human interpreters (HI) in terms of memory volume and precision of numbers, figures and unknown concepts. Liu (2023) explained it by the fact that cognitive abilities of interpreters are limited and information completeness might be worse than the AI-powered tools demonstrate due to fatigue, stress, occasionally forgotten terms and other human factors. Very similar results with slightly different interpretation were presented by Fantinuoli and Prandi (2021) who examined these issues in the Italian-English language pair. They utilized quality assessment methodology and unlike Sarmanova (2022) and Belenkova (2019) not one, but six raters were asked to evaluate both human and machine performances from the communicative perspective based on the provided assessment

framework. As a result, Fantinuoli and Prandi (2021) found out that interpreters compared to machines tend to generalize the source information in order to keep up with the speed and cope with a cognitive load, therefore, skipping some units of information (Fantinuoli & Prandi, 2021). Humans stive to interpret the key message explicitly, however, it might be less precise than AI-powered tools which interpret every single unit, if articulated clearly, as the study demonstrates (Fantinuoli & Prandi, 2021). Nevertheless, some speech recognition issues present in this research as well which seems to be a problem across multiple language pairs, as exhibited in the previous studies of Belenkova (2019), Sarmanova (2022), and Liu (2023). Generally, Fantinuoli and Prandi (2021) note that humans are superior in terms of fluency, clarity and adequacy of the renditions while AI-powered interpreting is slightly better in informativeness, in other words "content completeness".

It seems logical that the MI limitations reflected above will be gradually eliminated with the advancements of AI technologies. Nonetheless, even three years later some of these patterns are still in place, as displayed in the recent study conducted by Liu and Liang (2024) who tested the output of human interpreters, Google Translate and Baidu Translate in the English-Chinese context. Unlike the methodology of Belenkova (2019), Sarmanova (2022) and Fantinuoli and Prandi (2021), these authors did not use human participants approach, but utilized automatic metrics to comparatively assess the human and machine performances. Trying to avoid human assessors' subjectivity, Liu and Liang (2024) used Coh-Metrics 3.0 and TAASSC (Tool for the Automatic Analysis of Syntactic Sophistication and Complexity). The authors transcribed the simultaneously interpreted speeches and let the automatic metrics evaluate the human and machine outputs on the lexical, syntactic and cohesive levels based on the 106 indices. The results demonstrated that despite continuous technological updates, the tested services "stick to word-to-word

translation" and unable to recognize the contexts of speeches (Liu & Liang, 2024, p. 10). In contrast, humans tend to simplify the speech, use more understandable, cohesive and shorter expressions while also demonstrating lexical diversity by adding causal prepositions (Liu & Liang, 2024). These results are consistent with the study of Fantinuoli and Prandi (2021) where the similar patterns were revealed. MI was incapable to creatively generate some language constructions unless they were explicitly stated in the original speech (Liu & Liang, 2024). Yet, there is a slight disagreement in terms of interpretation of the results among different authors. For example, the informativeness strength of MI output reflected in the study of Fantinuoli and Prandi (2021) is interpreted as rather negative factor by Liu and Liang (2024) since adherence to the input language for the sake of precision sometimes results in too literal and confusing outputs. Generally, this is the matter of the results' interpretation and authors might comprehend and represent them differently depending on their individual backgrounds.

Finally, there is also an issue of latency – "the time delay from when an utterance is pronounced in the source language to when it gets translated in the target language" (Fantinuoli & Prandi, 2021, p. 245). In technical terms, it refers to the time span, measured in seconds, from when a word is spoken in the source language to when its translation is delivered (KUDO, n.d.). It can be quantified by the number of words or seconds the listener must wait through before receiving the translation (KUDO, n.d.). HI latency is about 2-4 seconds (Pochhacker, 2015), while MI latency is usually longer. It presents a notable drawback since the interpretation recipients need to wait longer to understand the message. Longer machine latency is explained by extensive processing time and speech output synthesis which also takes several seconds. Reduction of machine latency requires algorithms for context prediction and reliability check of the obtained prediction (Fedorova, 2023). This issue is highly relevant for the language pairs with immense syntax

differences (Grissom, 2017). The issue of latency was not explored in the studies of Sarmanova (2022), Belenkova (2019), Fantinuoli and Prandi (2021), Liu and Liang (2024) and all the aforementioned articles. As a result, current MI latency in comparison to the conventional human interpreting is unexplored and requires more attention.

To summarize, MI output quality varies across different languages, but there are common failure patterns. Despite quite adequate and generally relevant performance in the reviewed studies, MI might still be unacceptable in high stakes situations (Carl & Braun, 2017). As marked in the referenced literature, irrespectively of the utilized methodology and the selected machine interpreting tools, AI struggles with high speed, strong accents, speech disfluencies and contexts recognition. The common themes present in most of the reviewed studies are: occasionally incorrect automatic speech recognition, literalism or word-to-word interpretation of some renditions and poor context recognition. Nevertheless, MI is generally better in terms of message completeness and informativeness compared to humans which due to limited cognitive abilities tend to simplify and shorten the complex messages. Yet, all reviewed studies were conducted in the experimental setting without the consideration of real audience perception. The next section aims to analyze a number of documented situations where people embraced MI in order to explore a real-life machine performance.

Real-Life Implications of MI

Apart from the research papers, there are some documented real-life cases of machine interpreting. For example, the AI-powered tool called InterACT was introduced at the lectures for international students at a German university to interpret the speeches simultaneously (Cho et al., 2013). Unfortunately, these MI recordings are not publicly available, but professors commended the general outcomes, marking a few imperfections (Landgraf, 2012). However, this experiment took place more than 12 years ago when MI

quality expectations were considerably lower than today. To illustrate, few years later the large-scale events resulted in negative feedback. Boao Forum 2018 and Translation Automation User Society Asia Conference 2018 introduced MI instead of human interpreters in Chinese-English language pair. As reported by Wang and Wang (2019), the tools utilized in these events demonstrated serious inaccuracies. For example, the name "Road and Belt Initiative" was interpreted as "road for transportation and belt for the waist" respectively (Wang & Wang, 2019). MI also made considerable ASR-related mistakes as well as did not recognize the context. Such failures were harshly criticized by the participants of the conference and widely discussed in the Internet. In contrast, five years later, the most recent cases indicated considerable improvements. In June 2023 the WIRED YouTube channel counting over 10 million subscribers invited two professional Spanish-English conference interpreters to assess several speeches interpreted by KUDO AI. Surprisingly, the performance was very good. The invited experts commended the completeness, terminology usage, high accuracy while marking only few shortcomings such as pauses, longer latency, occasional struggles in word choice, and minor grammar inconsistencies (WIRED, 2023). After the video went viral, Claudio Fantinuoli – founder KUDO AI, commented that the model assessed in the video has already been improved by 25% as of the end of 2023 (Santos, 2023). Moreover, Tzachi Levy – the product manager of KUDO, stated that their product is already used by over 20 corporate clients and continues to gain its popularity (Santos, 2023). In this regard, more tests and assessments are needed to keep track the recent breakthrough improvements, especially in other language pairs where the research gap is evident.

To summarize, MI output quality of moderate difficulty speeches is good, but in more complicated cases there is still room for improvement, as noted in both empirical studies and real-life implication cases. Thus, it is viable to suggest that in the present study

the tested MI will also demonstrate quite satisfying results in case of clear and straightforward speeches while struggling with more complicated ones where accents, misarticulations and intricate expressions are accompanied with high speed. Yet, considering the continuous technological updates, it is difficult to build any exact predictions, especially in the English-Russian language pairs where the research gap is vivid. The final section is aimed at reviewing the existing interpreting quality evaluation approaches. It is necessary since this research paper will utilize quality assessment methodology. Thus, clear-established quality criteria are needed.

Interpreting Quality Assessment

Quality assessment of interpreting varies widely due to subjective perspectives (Pöchhacker, 2001). Despite decades of debate, no consensus exists on defining "good" interpretation (Pagura, 2019). Gile, a famous author and emeritus professor in translation, interpreting and linguistics notes that a unified, valid, and reliable assessment metric remains elusive in interpreting research (Niska, 1999). There is no universally accepted assessment rubric or recognized components within translation and interpreting studies (Angelelli & Jakobson, 2009). Instead, debates continue over the most effective evaluation criteria. Kahane (2015), an associate member of the International Association of Conference Interpreters (AIIC), one of the first and most influential communities founded in 1953, describes quality as an “elusive concept” (p. 1), while Wadensjö (1998) also emphasizes the absence of clear, universal standards. Addressing this issue, in 1996 AIIC explored a possibility to obtain ISO (International Standardization Organization) quality certification for conference interpreting (Pagura, 2019). However, specialists concluded that quality in this field is challenging to standardize due to its dependence on numerous external factors beyond the interpreter’s control (Luccarelli & Gree, 2007, cited in MacDonald, 2013). International organizations recruiting interpreters and educational

institutions often develop their own criteria to address these complexities, particularly for certification purposes (Pagura, 2019). Many research was conducted to reveal the most essential parameters for these assessment frameworks. A common misconception is that interpreting assessment should mainly focus on linguistic factors. For instance, prominent professor of interpreting studies, publishing author and conference interpreter Franz Pöchhacker (2001) notes that the concept of quality extends far beyond linguistic criteria and must also consider communicative effectiveness, specific situational and institutional contexts. Therefore, the criteria for quality assessment of interpreting should include both linguistic and extra-linguistic aspects. Extra-linguistic characteristics include delivery, intonation, pausing and other factors which often are not taken into account in assessment grids (Pöchhacker, 2001).

Essential Assessment Criteria

Expanding on existing studies, He Huiling surveyed organizers, speakers, interpreters, and audiences from nine international academic conferences to find the top priority of interpreting quality (Ma, 2021). Accurate information delivery was the most significant criteria, followed by correct terminology and prior document preparation for interpreters (Ma, 2021). Similarly, Ru Mingli questioned participants of five conferences in commerce, engineering, and religion highlighting a shared focus on content accuracy over linguistic form, with business conference users demanding the highest faithfulness to speakers and precise terminology (Ma, 2021). As summarized by Ma (2021), professional users of interpreting services consider that correct terminology, completeness of information, consistency, logical cohesion and synchronicity are among the most important criteria while intonation and correct grammar are the least valuable factors. Likewise, Gile (1990), Moser (1995), Kurz (1989 and 2001) also tried to reveal the most significant criteria for evaluation of interpreters' performance through questionnaires for both

recipients and interpreters. The terminology used to give the idea of certain criteria varies immensely from one study to another. Yet, such criteria as fidelity or in other words accuracy, language appropriateness, coherence, voice and intonation, terminology, significance of mistakes and several others are present in most of the beforementioned research (Pagura, 2019). Trying to indicate the most valuable criteria out of this outlined set of priorities, Kurz (2001) surveyed the end-users and members of AIIC, where the significance of nine quality criteria was rated by the respondents. Answers demonstrated the following priority order: "fluency, logical cohesion, sense consistency, completeness, grammar, terminology, accent, and voice" (Kurz, 2001, p. 406).

Interestingly, Zwischenberger (2010) tried to replicate the study of Kurz (2001) and also surveyed the members of AIIC and the German Association of Conference Interpreters (VKD) to identify the order of priority in quality assessment criteria. His framework consisted of 11 criteria summarized into three categories: "Content-related criteria: sense of consistency with original, logical cohesion, completeness. Form-related criteria: correct terminology, correct grammar, appropriate style. Delivery-related criteria: fluency of delivery, lively intonation, pleasant voice, synchronicity, native accent" (Zwischenberger, 2010, p. 9). According to the results which are somewhat consistent with the outcomes of Kurz (2001), content and form-related criteria are the most significant, while native accent and synchronicity are the least important. Due to the complex nature of this field, where a criterion considered valuable in one study might be deemed less significant in another, clear and consistent categorizations as the one of Zwischenberger (2010) are quite challenging to find. Thus, this Master's thesis aimed at the quality assessment and comparison of AI and human interpreting will adopt the classification of Zwischenberger (2010) since it logically reflects all the main quality aspects of SI and contains many similar criteria outlined as important by the other beforementioned credible

authors (Pöchhacker, as cited in Kahane, 2015; Gile, as cited in Wu, 2010). However, criteria like pleasant voice and synchronicity are excluded due to potential subjectivity and their relatively lower significance according to Zwischenberger (2010).

Assessment Scale Format

In terms of the format of the assessment grid, there are different possible options. Quality is typically assessed using automated systems such as BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit Ordering), which compare system outputs against source referenced texts (Fantinuoli & Prandi, 2021). However, these metrics are not suitable for this study because the quality they measure automatically is confined to the linguistic similarity between the provided option of a translation and the reference exemplary translation. They also overlook the extra-linguistic aspects and fail to account for the translation process within a communicative event (e.g., Angelelli, 2002). While such automatic evaluations are useful for comparing systems, they do not consider the communicative context or how users perceive their effectiveness (Fantinuoli & Prandi, 2021). Considering the nature of this research it is better to select and apply a manual assessment framework based on certain criteria frequently employed for assessing human interpretation despite some inevitable degree of subjectivity related to the grading process.

Apart from adopting the list of criteria, it is also important to decide on the grading system and description of the assessment rubrics. While various assessment grids have been created by organizations and training programs globally, they often do not provide detailed criteria descriptions and lack validity and reliability in their testing methods (Angelelli & Jakobson, 2009). Undoubtedly, scoring should be based on fixed and objective criteria, ideally with the detailed descriptions for each possible score range (Angelelli & Jakobson, 2009). Score-based assessment grids with detailed descriptions for

each possible score are considered more reliable than the point-adding or point-deducting systems commonly used in some organizations and professional associations (Angelelli & Jakobson, 2009). Following these recommendations, performance needs to be marked strictly based on the provided criteria descriptors. Therefore, this Master's thesis utilized the score-based assessment format where each of the selected criteria is described in relation to the possible score to ensure transparency of the assigned scores and avoid subjectivity of the assessors.

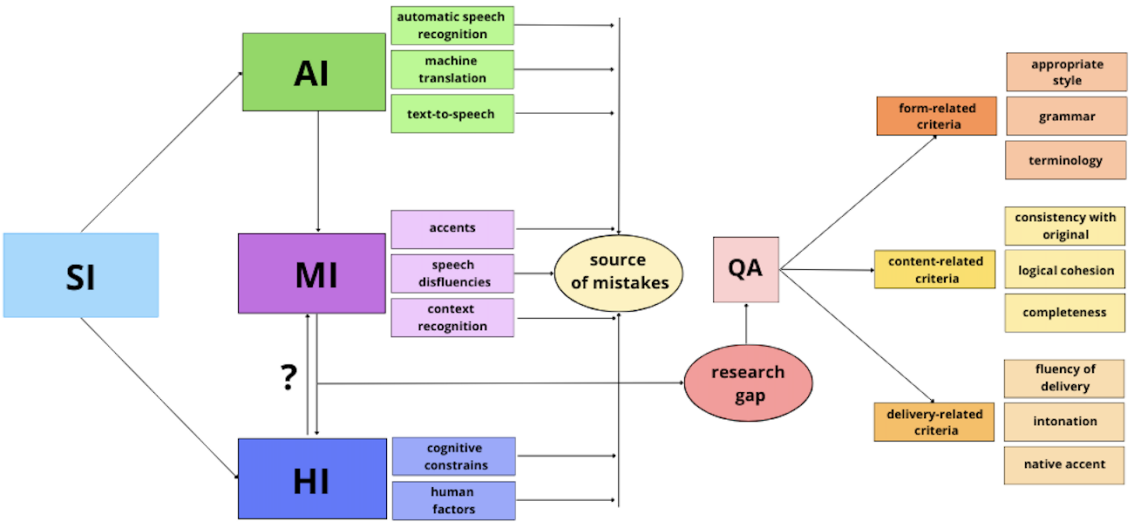
In terms of the number of possible scores, according to Angelelli and Jakobson (2009), a specific number of performance levels makes grading more manageable. Too many levels could confuse both graders and candidates, while too few would fail to clearly differentiate competencies needed for certification decisions (Angelelli & Jakobson, 2009). Therefore, this study will utilize four-fold grading system where one is the worst result and four is the best. The created scale attached in Appendix A is based on the nine most important quality criteria of simultaneous interpreting, adapted from Zwischenberger (2010). The criteria descriptions are summarized from the related studies of Han (2021), Lee (2008), Lee (2015), Liu (2013), Nadir (2017), Shang (2021), Wang et al. (2015), but rewritten in a succinct and concise manner. Each criterion can be assessed from one to four points and the maximum total score is 36. The scale also contains a specific column for notes, as adapted from Lee (2015). In this part, assessors will write possible errors, inaccuracies, etc., when listening to the interpretations. Further step-by-step details in relation to the assessors, data collection and analysis are presented in the methodology chapter. The next and final paragraph presents the conceptual framework as a logical summary of the chapter.

Conceptual Framework

Given the identified research gap, this study seeks to critically evaluate the performance of machine interpreting in the English-Russian language pair. The objective is to identify the strengths and weaknesses of AI-generated interpretation compared to those produced by human interpreters, to determine the extent of AI's competitiveness, its comparative advantages, and its recurring error patterns. To achieve this goal, the study focuses on answering two questions: How does the performance of AI-driven machine interpreting in English-Russian language pair compare to that of human interpreters? What are the typical errors and limitations of AI-driven machine interpreting in the English-Russian language pair? Therefore, the central concepts of this paper are artificial intelligence, simultaneous interpreting, and machine interpreting which were defined in details in a separate paragraph. Review of the existing studies in this field revealed notable gaps in knowledge and several recurring themes. Current literature highlights such challenges as AI's inability to correctly recognize rapid or poorly articulated speech and strong regional accents, challenges in interpreting intricate expressions, and other limitations discussed earlier. Human interpreters, in turn, mainly struggle with limited cognitive capabilities and human factors trying to keep up with the immense cognitive load. This study uncovers these and other common interpreting issues using an assessment scale based on selected key criteria. There is a strong interrelatedness of the central concepts of this paper, typical drawbacks and limitations of MI, and quality criteria adopted for the assessment framework of this paper. It can be visually represented in the following scheme as presented in Figure 3 below. This scheme maps out the relationship between the concepts overall reflecting the conceptual framework of this Master's thesis. The next section finalizes this chapter outlining the key insights from the reviewed literature.

Figure 3

Conceptual Framework



Summary

In conclusion, this literature review analyzed the theoretical and empirical studies on the performance of machine simultaneous interpreting technologies. Such cognitive-intensive activity as human interpreting (HI) nowadays can be performed by AI-powered tools capable of replicating human intelligence patterns including speech recognition, transcribing, translation, and voice-over (Corpas Pastor, 2018; European Parliament, 2023; Prandi, 2023). Literature evidences rather slow practical implementation of interpreting technologies due to long-established beliefs of an exclusively human approach and skepticism among the practitioners (Ahmed, 2022; Fantinuoli & Dastyar, 2022; Wang & Li, 2022). It resulted in a relatively small number of papers devoted to practical implementation and analysis of MI performance (Corpas Pastor, 2018) making the research gap self-evident. The existing empirical studies and real-life cases of MI integration demonstrate notable mistakes and inaccuracies, starting from speech recognition failures and lack of context understanding, to excessive word-to-word precision undermining the overall sense (Belenkova, 2019; Fantinuoli & Prandi, 2021; Liu, 2023; Liu & Liang, 2024; Sarmanova, 2022; Wang & Wang, 2019). However, the latest data demonstrated quite

adequate and commendable overall performance (Liu & Liang, 2024; WIRED, 2023) stipulated by the constant technological updates. It remains clear that the research field is still at an early development stage and more studies are needed to objectively assess the quality of continuously improved MI tools. Contributing to this emergent and hotly debated field, the present study states that the recent technologies are promising, but still quite understudied in the context of simultaneous interpreting, especially in the English-Russian language pair which counts very few studies and still no publicly available documented real-life cases. To address this research gap, the study adapted the criteria of Zwischenberger (2010) including content, form, and delivery-related categories. There is a strong connection between the core concepts of this paper, common interpreting challenges, and quality assessment criteria which altogether underpin the conceptual framework of this paper as visually summarized in Figure 3. The next chapter presents and justifies the methodology with the step-by-step plan for data collection and analysis.

Methodology

In accordance with the topic of the Master's thesis, the purpose of the study is to evaluate and analyze the quality of machine interpreting in comparison to traditional human simultaneous interpreting. The literature review preceding the methodology chapter demonstrated high relevance of this topic and the lack of similar studies in the English-Russian language pair. Addressing this research gap, the study examines the machine interpreting output, compares it to human performance, analyzes errors, and reveals common patterns. Thus, the research questions are focused on how AI-driven machine interpreting performs compared to human interpreters in the English-Russian language pair. It also explores the typical mistakes and limitations found in machine interpreting within this language combination. This chapter presents the methodology selected to achieve the research purpose and answer the research questions. The chapter includes information about the research design, outcomes of piloting, sampling procedures, data collection and analysis as well as ethical considerations and limitations of the selected approach.

Research Design

Considering the research purpose and questions, the quality assessment method was the most appropriate for this study. The quality assessment method falls under the category of product-based translation studies research methods (Saldanha & O'Brien, 2014). Product-based methods mean focusing on the translated text or interpreted speech as the main object of analysis, rather than the translation process or the agents (Saldanha & O'Brien, 2014). Quality assessment within product-based research methods implies adapting or creating an assessment framework in accordance with the research context and can be both quantitative and qualitative in nature depending on the scale construct (Saldanha & O'Brien, 2014). In order to answer the research questions and ensure a deeper

analysis, the study needed both types of data – an assessment scale consisting of quantitative and qualitative components. Therefore, this study applied mixed methodology as defined by Creswell and Creswell (2017). Selection of mixed methodology is justified by the fact that quantitative data was needed in order to answer the first research question focused on measurement of machine interpreting against human interpreters. At the same time, qualitative data was also necessary to answer the second research question aimed at revelation of typical mistakes in machine interpreting. Additionally, pilot study showed that solely qualitative data might not be enough for complex analysis, as addressed in the following section. Thus, the quality assessment scale included both quantitative and qualitative measurements. The quantitative part of the scale enabled comparison of AI and human scores via t-test while the qualitative part in the form of a separate column for identified mistakes and inaccuracies allowed to reveal the typical limitation patterns. This mixed methodology was identified as convergent in nature since quantitative and qualitative data was collected approximately at the same time to ensure a comprehensive analysis of the research problem, as described by Creswell and Creswell (2018). The next paragraph demonstrates outcomes of the pilot study conducted to test the selected methodological approach.

Piloting

A small-scale pilot study was conducted before the main Master's thesis writing. There were several purposes for the piloting: to test the assessment scale construct, to define the optimal number of participants, and to test different machine interpreting tools. In the initial stages, it was decided not to include the quantitative component in the assessment scale in order to find whether it is possible to simplify the methodology and obtain necessary results based on solely qualitative data. Three sampled assessors working as conference interpreters and lecturers of translation courses were invited to listen to both

versions of AI-generated and human interpretations of six fragments sampled from free UN Web TV website and left their feedback. It was found that three is the adequate sample size of participants to obtain general impressions and remarks. Yet, it was decided to expand the number in the final Master's thesis to retrieve more comments and as a result to deepen the analysis. Six selected fragments were enough to collect the necessary data. It was also found that the simplified assessment scale without numerical evaluations provided limited insights insufficient to reveal how machine interpreting really measures up against human performance. Solely qualitative data was not enough to answer all the research questions and to ensure more comprehensive analysis. Thus, the final assessment scale used in this Master's paper included both quantitative and qualitative components as described in the next sections.

An additional purpose of piloting was to test different machine interpreting tools and identify the one to be used in the final Master's thesis. Initially, KUDO AI was selected. However, since this AI is available on a commercial basis and no free access was provided for research purposes, other tools were analyzed. Stenomatic and Yandex neural networks were tested, revealing notable differences. Piloting showed that Stenomatic provided slightly worse machine interpretation in the English-Russian language pair with the latency up to one minute. Additionally, only 10 minutes were available in the free trial version which was not enough for the effective analysis. It produces full-fledged machine simultaneous interpreting in live format where AI starts to interpret when the speaker has not yet finished. However, serious mistakes were found in the given language pair. Conversely, Yandex demonstrated remarkably better overall performance, though it lacked a fully live format. Its machine interpretation required several minutes to process video fragments, prepared in advance. Live Yandex interpreting is also available in beta version but it is restricted to specific platforms. Given Stenomatic's limited trial period and

unsatisfactory output, Yandex was chosen for the final study. The next sections address the final study's methodology taking into account the piloting outcomes.

Materials for the Assessment

This Master's thesis required the selection of several video or audio materials interpreted by humans and AI to be assessed and compared against each other. First, to measure machine interpreting against humans, existing recordings of interpretation performed by people were sampled. Since the recordings were selected from an open-access website, no special permission to use them for analysis was needed. Nevertheless, to avoid any ethical issues, the study did not use recordings if the names of interpreters were directly indicated. Additionally, this study did not criticize specific human errors analyzing predominantly machine output to prevent ethical concerns. Sampling existing and open-access human simultaneous interpretations was the best option for this study considering the fact that simultaneous interpreting requires special equipment and the physical presence of interpreters, as was described in the literature review.

The study utilized Yandex neural networks that provide free machine interpreting of videos and streams when opened in the Yandex browser. Yandex was chosen as a result of the piloting where several platforms were tested. The pilot study showed overall adequate and effective work of Yandex neural networks. Expanding the sample size to several AIs might be financially and time-consuming. Thus, comparison among several AIs should be the focus of further studies in this area in the future. This study selected six video fragments which were given to Yandex for machine interpreting. In order to assess and compare Yandex and human performances, it was necessary to sample already interpreted and publicly accessible videos, as justified above. The UN Web TV is a website that stores recordings of various UN agencies meetings available in five languages, including Russian and English. The website was used in the pilot study and demonstrated

its relevancy. Interpretation is typically performed by the UN staff interpreters or freelancers after complex eligibility examinations pass rate of which is less than 15% (United Nations Department for General Assembly and Conference Management, 2017). Thus, MI performance was measured against professionals. The sample size of six video fragments up to three or four minutes was enough for the analysis as was found at the piloting stage. Despite the fact that there are no time limitations in Yandex machine interpreting, longer fragments could make the assessment process very time-consuming for the assessors and as a result, might affect the judgments.

Importantly, to analyze how Yandex navigates in various subject domains, it was necessary to select the videos from different areas, ensuring terminological, lexical, and contextual variety. To ensure a comprehensive evaluation of machine interpreting, six video fragments were selected from different UN sessions. Each fragment was chosen to present distinct interpreting challenges, including high speed of delivery – words per minute rate (WPM), specialized terminology, proper names, numerical data, and diverse speaker accents. This variety was intended to test the system's performance under realistic and demanding conditions. Full transcripts of the fragments are provided in Appendix B, while the table below presents a concise overview to contextualize the nature of each fragment.

Table 1

Video Fragments for Interpreting

No.	Duration	WPM	Topic	Challenges
1	1 min 6 sec	163	Terrorism	No significant challenges except high speed of delivery
2	3 min 34 sec	142	International Relations	High speed of delivery, proper names and figures

3	3 min 28 sec	119	Migration	A lot of numbers, abbreviations, proper names with prompt delivery
4	3 min 56 sec	122	Middle East	Numbers, terms, and regional accent
5	1 min 29 sec	137	Administration and Budgets	Strong regional accent, high speed of delivery, names and terms
6	3 min 22 sec	149	Drugs and Crime	Strong regional accent, speech articulation, terms, dates and abbreviations delivered at high speed

Sample

The study purposively sampled five assessors working as freelance interpreters and teaching simultaneous interpreting courses. Following recommendations of Saldanha and O'Brien (2014), in the quality assessment method, researchers should not participate in the evaluation process, if possible, since it might affect the reliability of judgments. Therefore, a purposive sampling strategy was utilized to ensure that the participants met the necessary eligibility criteria as described by Creswell and Creswell (2017). The first criterion was solid interpreting experience to ensure their familiarity with the quality requirements of the market. The second criterion required a participant to teach Translation Studies courses to ensure their ability to objectively assess the interpretations using evaluation grids. To the best of the researcher's knowledge, the minimal sample size of assessors is not specified in the methodological literature. According to Saldanha and O'Brien (2014), the sample size

of assessors depends on time and financial constraints, but a greater number ensures a higher reliability of judgments. As was noted in the literature review, previous studies typically employed between three and six raters. Based on the results of the pilot study where three participants were sampled, it was decided to increase the number to ensure a sufficent number of evaluations. Thus, five assessors were sampled for the final Master's thesis. This number was selected to balance methodological rigor with practical feasibility. While a larger sample might increase reliability, it would also complicate data collection and processing, especially given the time-consuming nature of both quantitative scoring and qualitative feedback. A sample of five assessors ensured a reasonable level of reliability while keeping the evaluation process manageable. However, it is acknowledged that with a small sample, inferential statistics must be interpreted cautiously. Thus, effect sizes (Cohen's d) were reported alongside *p*-values in the results chapter to provide greater insight into the magnitude of observed differences. Inter-rater reliability was calculated in the Jamovi statistical program (Jamovi Project, 2024) to ensure the consistency and reliability of participants' judgments. Permission to distribute recordings and assessment scales among sampled assessors was received from the Research Committee of the School of Liberal Arts of Maqsut Narikbayev University. The informed consent forms explaining the research purpose, voluntariness, and confidentiality was provided and signed by the participants, as attached in Appendix C.

Data Collection Tool

The main data collection instrument was the quality assessment scale. Thus, the primary step was to adapt or create the grid in accordance with the research context and literature review. There are many different forms of quality assessment scales, but the research used analytical score-based scale which is quite reliable and objective since it implies specific criteria description to be followed when grading (Han, 2021; Lee, 2015),

reducing possible subjectivity, as was described in the literature review. The created scale attached in Appendix A was based on the nine most important quality criteria of simultaneous interpreting, adapted from Zwischenberger (2010). As mentioned in the literature review chapter, the criteria descriptions were summarized from the related studies of Han (2021), Lee (2008), Lee (2015), Liu (2013), Nadir (2017), Shang (2021), Wang et. al (2015), but rewritten in a succinct and concise manner. Each criterion could be assessed from one to four points and the maximum total score is 36. The scale also contained a specific column for notes, as adapted from Lee (2015). In this part, assessors wrote possible errors, inaccuracies, etc., when listening to the interpretations.

Data Collection Procedures

Data collection consisted of three steps. The first step was to develop the assessment scale, as was described in the previous paragraph. The next step was to collect machine interpreting. The machine interpreting session required at least two devices. The first device was a laptop which played the original video fragment in the Yandex browser with activated machine interpreting function. The second device was an audio recorder to tape machine interpreting output. This setup minimized the risk of audio overlap, system interference, or data loss that could occur if both playback and recording were performed on the same device. By separating input and output functions, the method prevented possible technical issues. The third step was assessment collection. The sampled assessors received six original video fragments, six machine interpretations, and six human interpretations as well as printed or digital assessment scales. The assessors needed to carefully listen the both versions of interpretation for each original fragment, read the criteria descriptions, and following them, evaluate the performance. However, as mentioned in the methodological literature, there was a chance that assessors ignored the score descriptors and grade based on personal subjective beliefs (Han, 2021; Saldanha &

O'Brien, 2014). Moreover, there was a possibility that not all assessors used a column for notes, identified mistakes, and inaccuracies. To address these issues, as recommended by Han (2021) and Saldanha and O'Brien (2014), before collecting evaluation, debriefing sessions were conducted with each assessor. They were explained how it is important to strictly follow the score descriptions and make necessary notes in the special column. As a result, the assessors complied with the requirements, and no violations of the assessment procedures were found among sampled participants.

Data Analysis

The data analysis procedure consisted of three steps. First, the total scores of all seven assessors were manually tabulated in the Jamovi statistical program version 2.4.14 (Jamovi Project, 2024) as continuous variables. Then, inter-rater reliability also known as Interclass Correlation Coefficient (ICC) was calculated in Jamovi (Jamovi Project, 2024) to analyze the consistency of grades. ICC served as a validity-checking tool by measuring the degree of agreement among assessors, thus ensuring scoring reliability. As recommended by Koo and Li (2016), this research conducted ICC test selecting two-way mixed, average measure, absolute agreement which are relevant for such type of studies. In accordance with the methodological literature, if the coefficient is greater than 0.5, the raters reliability can be considered valid (Koo & Li, 2016). Typically, the closer the ICC to 1, the stronger reliability is (Koo & Li, 2016). The results of ICC indicated excellent degree of reliability, as reported in the results chapter.

Secondly, the interpreters were tabulated in Jamovi (Jamovi Project, 2024) as nominal variable with two categories (AI and Human) and the sub-scores for each criterion were tabulated as continuous variables to conduct the independent sampled t-test. It demonstrated the means of AI and human sub-scores to be compared. While the ICC provided only general validity of the scores, inferential analysis, specifically the

independent-sample t-test, was applied to determine the specific differences between human and machine interpreting scores. The results were analyzed and reported descriptively with the focus on what specific criterion was assessed significantly higher between AI and human, as reported in the next chapter. The analysis of the t-test was supported by descriptive plots and tables from Jamovi (Jamovi Project, 2024).

The third step was analysis of the notes section. Thematic coding of qualitative feedback was utilized to find the common themes in the assessors' feedback. It complemented the quantitative findings by providing contextual insights into recurring patterns of error and the nature of specific mistakes. After careful reading, open codes were highlighted. Since note-taking was involved where not all comments were clear, member checking was used as a qualitative validity strategy to ensure clarity, as proposed by Patten and Newhart (2017). Member checking validity strategy means that the researcher contacts participants to confirm whether their responses were correctly understood and interpreted (Patten & Newhart, 2017). Then, the open codes were summarized based on similarity into axial codes. Axial codes were manually placed on the digital stickers in Canva blackboard (Canva Pty Ltd, 2025) in order to identify and categorize them based on the common patterns, i.e., what the most typical limitations are related to. The results indicated five thematic codes. The analysis was supported by the example of the filled assessment scale with highlighted open codes (Appendix D) and a screenshot with digital stickers where axial codes were summarized into thematic codes (Appendix E).

Ethical Considerations

The ethical risks associated to this Master's thesis were reduced as much as possible. The study used publicly available human interpretations from the website where no interpreters' names were indicated. However, complete anonymity was impossible due to the fact that their voices might still be recognized. Yet, the ethical implications are

minimal since the recordings are already in the public domain, and no evaluative judgments were made about the interpreters themselves. Voices were not linked to identifiable individuals beyond what is publicly accessible. Moreover, the study was focused on limitations of machine interpreting and specific human interpreter output was not openly criticized and speculated on in this paper. Before conducting the study, permission from the Research Committee of the School of Liberal Arts of Maqsut Narikbayev University was obtained as was mentioned above. The sampled assessors received the informed consent form explaining the purpose, ensuring confidentiality, and the right to withdraw at any time (Appendix C). Neither real names nor the information about the participnts' particular place of work was indicated to ensure confidentiality. Debriefing sessions were conducted to meet any possible concerns and to explain all the procedures. The collected assessment scales are stored in the passworded folder accessible to the researcher only.

Summary

This chapter outlined the methodology used to assess the quality of machine-generated interpreting compared to human simultaneous interpreting. A product-based translation studies method for quality assessment was selected to achieve the research goal. A mixed convergent methodology was applied to collect both quantitative scores and qualitative feedback. The study involved five expert assessors who evaluated both machine and human interpretations using a quality assessment scale. The materials selected for evaluation included a range of complex audio fragments sampled from an open-access website that stores UN agency meeting records. Assessors rated the interpretations based on established criteria, providing both quantitative scores and qualitative feedback on specific errors. Intraclass correlation (ICC) was calculated to assess rater reliability, while an independent samples t-test was conducted to determine if there were significant

differences in the quality scores between machine and human interpretations. Additionally, thematic analysis of the feedback sections complemented the quantitative findings by identifying common error patterns and specific examples of mistakes. Ethical considerations were adhered to throughout the research process, ensuring the highest possible level of confidentiality. Ethical approval was obtained from the relevant body, and informed consent forms were collected in accordance with established ethical procedures. The next chapter presents the results of the data collection and analysis based on the applied methodology described in this chapter.

Results

With the rapid development of technological advancements, AI-driven interpreting is gaining attention as a potential alternative to human simultaneous interpretation. This study evaluates Yandex machine interpreting compared to human performance in the English-Russian language pair, identifying key strengths and weaknesses based on expert assessments. The purpose of this section is to present the quantitative and qualitative findings based on the collected data from the sampled expert assessors. The findings indicate that AI outperforms human interpreters in completeness and fluency of delivery criteria but it lacks the nuance required for accurate context recognition, numerical precision, and natural delivery quality as well as some other subtle factors. The chapter consists of two main parts. The first is the quantitative findings including inter-rater reliability and independent samples t-test results. The second is the qualitative part which presents the results of thematic analysis of participants' feedback. Altogether, these findings provide valuable insights into the current state of machine interpreting and its feasibility for professional use, as follows in the discussion chapter.

Inter-Rater Reliability

After all the participants filled in the assessment scales, total grades of each assessor were tabulated as continuous variables in order to calculate interclass correlation coefficient (ICC) which demonstrated inter-rater reliability and degree of agreement among sampled five raters across six subjects. ICC estimates and their 95% confident intervals were calculated using Jamovi statistical program version 2.4.14 (Jamovi Project, 2024) in Seolmatrics module (Seol, 2024) based on a mean-rating ($k = 5$), absolute-agreement, two-way mixed-effects model, as recommended in the methodological literature (Koo & Li, 2016). Table 2 and 3 demonstrate that ICC indicated excellent reliability and agreement among the assessors for both AI and human outputs. First, ICC

for the AI scores was calculated. The results indicated a high level of reliability as shown in Table 1, *ICC*(3, k) = 0.938**,** 95% CI [0.403, 1.40] (Gamer et al., 2019). Thus, the assessment results deemed to be highly reliable. Similarly, ICC analysis for human interpreting scores was conducted to assess inter-rater reliability using the same parameters. The results also indicated excellent reliability as demonstrated in Table 2, *ICC(*3, k) = 0.900, 95% CI [0.359, 1.12] (Gamer et al., 2019). Therefore, the sampled expert assessors were consistent and reliable in their judgements as evidenced by high ICC value in both cases. Yet, the confidence intervals (CI) were quite wide suggesting some variability in ratings, potentially due to the limited number of subjects and differences in individual rater judgments (Gamer et al., 2019).

Table 2

Interclass Correlation Coefficient for AI Scores

Bootstrap confidence intervals of ICC agreement

	95% CI	
	Lower	Upper
Value	0.403	1.40

Intraclass Correlation Coefficient(ICC)

Model	Type	Unit	Subjects	Raters	ICC
twoway	agreement	average	6	5	0.938

Note. The analysis was performed by 'irr::icc' function.

Table 3

Interclass Correlation Coefficient for Human Interpreters Scores

Bootstrap confidence intervals of ICC agreement

	95% CI	
	Lower	Upper
Value	0.359	1.12

Intraclass Correlation Coefficient(ICC)

Model	Type	Unit	Subjects	Raters	ICC
twoway	agreement	average	6	5	0.900

Note. The analysis was performed by 'irr::icc' function.

Independent Samples T-test

As a next step, independent samples t-test was conducted in order to compare the evaluation scores of AI and human interpretations across multiple criteria in six fragments (N = 30). As shown in Table 4 and 5 and descriptive plots below, the t-test results demonstrated that AI was slightly better than humans in completeness, grammar and intonation criteria and significantly better in fluency of delivery criteria. Human interpreters, in turn, were a bit better than AI in consistency with original, logical cohesion, terminology, style and native accent. Yet, there were very few cases of statistical significance, as reported in the next paragraphs. Before analyzing the t-test results, Levene's test for equality of variances was examined. It was necessary to test the assumption of homogeneity of variance using Levene's test which is typically indicated in the tables as Student's test in Jamovi (Jamovi Project, 2024). For several variables, Levene's test was significant ($p < .05$), indicating a violation of the homogeneity assumption. In these cases, in accordance with the common statistical procedures, Welch's t-test was reported instead of Student's t-test, as reported in the next sections.

Criteria where AI Outperformed Humans

As indicated in the descriptive plot (Figure 4), AI scored significantly higher in delivery fluency ($M = 3.67$, $SD = 0.48$) compared to human interpretations ($M = 3.17$, $SD = 0.59$), with a mean difference of 0.50, $t(55.6) = 3.595$, $p < .001$, $d = 0.928$ (Table 4 and 5). AI also performed slightly better in completeness criteria ($M = 3.20$, $SD = 0.66$) than human interpretations ($M = 3.00$, $SD = 0.53$), though this difference was not statistically significant, $t(58) = 1.293$, $p = .201$, $d = 0.334$, with a mean difference of 0.20 (Table 4 and 5). Since Levene's test was not significant in both criteria, Student's t-test was reported.

Figure 4

Descriptive Plot for Fluency of Delivery and Completeness Criteria

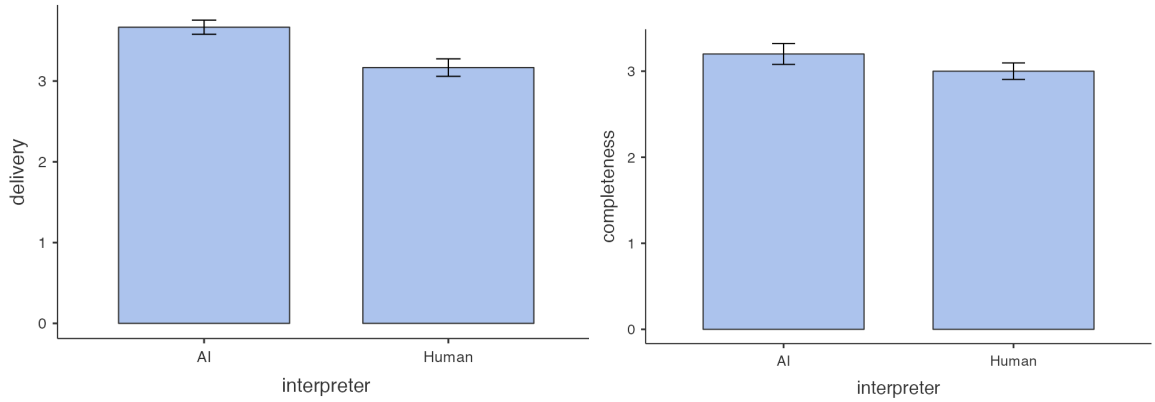


Table 4

Independent Samples T-Test

Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference		Effect Size
consistency	Student's t	−0.453	58.0	0.652	−0.0667	0.1473	Cohen's d	−0.117
	Welch's t	−0.453	55.3	0.653	−0.0667	0.1473	Cohen's d	−0.117
logical cohesion	Student's t	−1.939 [a]	58.0	0.057	−0.3000	0.1547	Cohen's d	−0.501
	Welch's t	−1.939	53.4	0.058	−0.3000	0.1547	Cohen's d	−0.501
completeness	Student's t	1.293 [a]	58.0	0.201	0.2000	0.1546	Cohen's d	0.334
	Welch's t	1.293	55.1	0.201	0.2000	0.1546	Cohen's d	0.334
terminology	Student's t	−2.470 [a]	58.0	0.016	−0.3000	0.1215	Cohen's d	−0.638
	Welch's t	−2.470	56.5	0.017	−0.3000	0.1215	Cohen's d	−0.638
grammar	Student's t	1.201 [a]	58.0	0.235	0.1000	0.0833	Cohen's d	0.310
	Welch's t	1.201	50.6	0.235	0.1000	0.0833	Cohen's d	0.310
style	Student's t	−2.408 [a]	58.0	0.019	−0.1667	0.0692	Cohen's d	−0.622
	Welch's t	−2.408	29.0	0.023	−0.1667	0.0692	Cohen's d	−0.622
delivery	Student's t	3.595	58.0	<.001	0.5000	0.1391	Cohen's d	0.928
	Welch's t	3.595	55.6	<.001	0.5000	0.1391	Cohen's d	0.928
intonation	Student's t	1.460	58.0	0.150	0.2000	0.1370	Cohen's d	0.377
	Welch's t	1.460	57.4	0.150	0.2000	0.1370	Cohen's d	0.377
native accent	Student's t	−1.000 [a]	58.0	0.321	−0.0333	0.0333	Cohen's d	−0.258
	Welch's t	−1.000	29.0	0.326	−0.0333	0.0333	Cohen's d	−0.258

Note. $H_a\ \mu_{AI} \neq \mu_{Human}$

[a] Levene's test is significant ($p < .05$), suggesting a violation of the assumption of equal variances

Table 5

Group Descriptives of T-Test

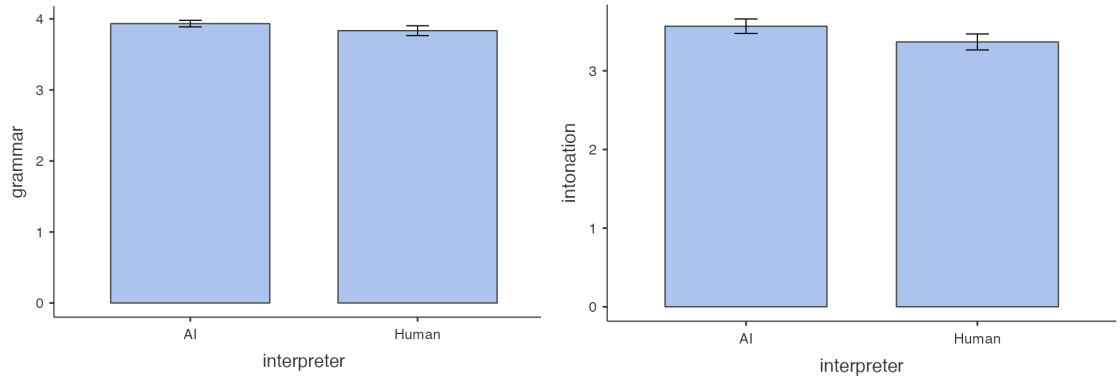
Group Descriptives

	Group	N	Mean	Median	SD	SE
consistency	AI	30	3.50	4.00	0.630	0.1150
	Human	30	3.57	4.00	0.504	0.0920
logical cohesion	AI	30	3.47	4.00	0.681	0.1244
	Human	30	3.77	4.00	0.504	0.0920
completeness	AI	30	3.20	3.00	0.664	0.1213
	Human	30	3.00	3.00	0.525	0.0959
terminology	AI	30	3.47	3.00	0.507	0.0926
	Human	30	3.77	4.00	0.430	0.0785
grammar	AI	30	3.93	4.00	0.254	0.0463
	Human	30	3.83	4.00	0.379	0.0692
style	AI	30	3.83	4.00	0.379	0.0692
	Human	30	4.00	4.00	0.000	0.0000
delivery	AI	30	3.67	4.00	0.479	0.0875
	Human	30	3.17	3.00	0.592	0.1081
intonation	AI	30	3.57	4.00	0.504	0.0920
	Human	30	3.37	3.00	0.556	0.1015
native accent	AI	30	3.97	4.00	0.183	0.0333
	Human	30	4.00	4.00	0.000	0.0000

Additionally, AI performed slightly better in grammar and intonation as shown in the descriptive plot below (Figure 5). AI interpretations in grammar ($M = 3.93$, $SD = 0.25$) scored slightly higher than human interpretations ($M = 3.83$, $SD = 0.38$), but the difference was not statistically significant, $t(50.6) = 1.201$, $p = .235$, $d = 0.310$, with a mean difference of 0.10 (Table 4 and 5). As well as in the previous criteria, Welch's t-test was used due to unequal variances. Similarly, in intonation criterion AI interpretations ($M = 3.57$, $SD = 0.50$) outperformed human ($M = 3.37$, $SD = 0.56$), though this difference was also not statistically significant, $t(57.4) = 1.460$, $p = .150$, $d = 0.377$, with a mean difference of 0.20 (Table 4 and 5). Levene's test demonstrated equal variance in this criterion, therefore, Student's t-test value was used.

Figure 5

Descriptive Plot for Grammar and Intonation Criteria

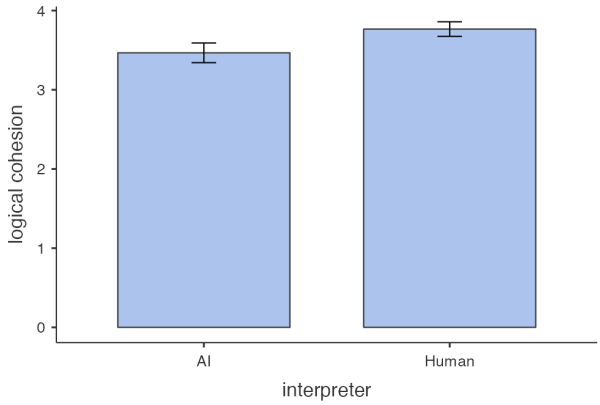


Criteria Where Human Interpreters Outperformed AI

T-test results demonstrated that human interpretations scored higher than machine interpreting in logical cohesion, terminology, style, consistency with original and native accent. The descriptive plot in Figure 6 and results in Table 5 indicated that human interpretations (M = 3.77, SD = 0.50) scored higher in logical cohesion compared to AI interpretations (M = 3.47, SD = 0.68), with a mean difference of -0.30 (Table 4 and 5). However, this difference was not statistically significant, $t(53.4) = -1.939$, $p = .057$, $d = -0.501$ (Table 4). Since Levene's test was significant, Welch's t-test was used.

Figure 6

Descriptive Plot for Logical Cohesion Criterion

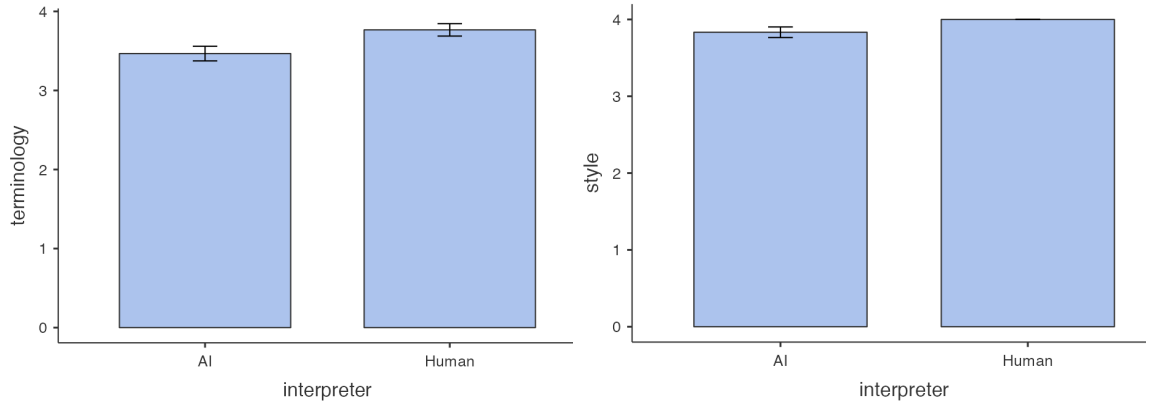


Terminology scores, as indicated in the plot below (Figure 7), were significantly higher for human interpreters (M = 3.77, SD = 0.43) than for AI (M = 3.47, SD = 0.51),

$t(56.5) = -2.470$, $p = .017$, $d = -0.638$, with a mean difference of -0.30 (Table 4 and 5). Similarly, the plot below (Figure 7) indicates that human interpretations ($M = 4.00$, $SD = 0.00$) slightly outperformed AI ($M = 3.83$, $SD = 0.38$) in style (Table 5), $t(29.0) = -2.408$, $p = .023$, $d = -0.622$, with a very small mean difference of -0.17 (Table 4). Again, Welch's test was used due to unequal variances in both cases. Remarkably, for both terminology and style criteria, the difference was statistically significant, as indicated by $p < .05$.

Figure 7

Descriptive Plot for Terminology and Style Criteria



The remaining criteria, including consistency with the original and native accent, showed very minor differences between AI and human scores, with mean differences ranging only from -0.10 to 0.20, none of which were statistically significant ($p > .05$). Thus, due to very small difference between AI and humans, these two criteria were not reported in this section.

The effect sizes (Cohen's d) for the statistically significant comparisons reported above ranged from -0.622 to 0.928, suggesting moderate to strong effects. Yet, despite few cases of statistical significance, the small size of the sampled fragments and assessors does not allow for relevant generalizations but rather provides indications for the general performance of two types of interpreters (AI and human) as well as indicates effectiveness of the chosen evaluation framework which demonstrated its reliability based on the ICC

value. Additionally, generalization of results is practically irrelevant in the context of this research paper even in the case of a large sample size and statistically significant *p-value* due to the fact that machine neural translation networks used in Yandex are continuously updated. Therefore, after each update the statistically significant results will not be longer relevant.

To sum up the quantitative findings, Table 6 was prepared to show the winners in the criteria where the highest mean score differences were found. For better understanding in which particular aspects AI or humans outperformed each other, the table contains short descriptions from the original assessment scale as attached in Appendix A given to participants to guide their judgments more objectively. In accordance with quantitative findings, human interpreters outperformed AI in terms of logical cohesion (coherence and cohesion of interpretations), terminology (equivalence of terms and subject matter nuances), and style criteria (idiomatic and relevant register and context). AI, in turn, outperformed humans in completeness (amount of conveyed information) and fluency of delivery (diction, articulation, unfilled pauses). Thus, the table below reflects the key quantitative findings with the focus on the criteria where the notable mean score differences were found.

Table 6

Key Difference in AI and Human Interpretations

Criteria	Descriptions from the assessment scale	AI wins	Human wins
Logical Cohesion	Excellent coherence and cohesion of the interpretation. No misleading or redundant elements that might affect the overall logic.	-	✓

Terminology	Highly accurate interpretation of the specialized terminology, subject matter nuances and subtleties while ensuring consistent equivalence between source and target language terms	-	✓
Completeness	The original content is conveyed in full volume. Accurate rendition of numbers, names, titles, as well as minor details.	✓	-
Fluency of Delivery	Fluent pace of delivery with seamless diction and articulation. No distinct time lags or (un)filled pauses.	✓	-
Style	The register is fully appropriate to the context. All the expressions sound completely natural, idiomatic and stylistically relevant.	-	✓

Thematic Analysis of the Feedback Section

After quantitative analysis in Jamovi (Jamovi Project, 2024), qualitative data was analyzed. Qualitative results demonstrated more detailed information on human and AI strengths and weaknesses, mistakes, and inaccuracies. As was mentioned in the methodology chapter, the quality assessment scale consisted of two parts, numerical grades and feedback, where raters left their comments. This column was carefully analyzed qualitatively using thematic coding approach. The most meaningful pieces of information

were highlighted as open codes. After initial analysis 273 open codes were identified. Afterward, open codes were connected into broader categories in 78 axial codes. All the axial codes were copied, and inserted into digital stickers on the interactive blackboard in Canva website (Canva Pty Ltd, 2025). Next, the stickers were united and categorized based on their similarity and broader recurring themes. Appendix D demonstrates an example of the filled assessment scale with highlighted open codes and Appendix E shows the screenshot from the Canva board (Canva Pty Ltd, 2025) with the axial codes placed on the stickers used for thematic analysis. These axial codes were categorized into five thematic codes – major themes based on the most common trends identified from AI and human outputs as presented below.

No Cognitive Limitations

The first theme was not directly related to a specific type of mistake but approaches cognitive limitations among human interpreters which affected completeness and delivery criteria. Most of the raters noted, that due to the natural cognitive constraints human interpreters used different optimization strategies such as generalization, compression, omission, etc. in order to keep up with the speed of delivery and regional accents. Participants marked that machine interpreting did not experience any struggles related to high speed of delivery irrespectively of strong regional accents and managed to convey information fully with very few omissions, unlike human interpreters who tend to skip details more often to reduce cognitive load. For example, most of the assessors consistently noted that humans interpreted the term “unmanned aircraft systems” briefly as “*беспилотник*” instead of full the version “*беспилотные авиационные системы*” to optimize time in the Fragment 1 (Appendix B). The word “executive director” was interpreted shortly as “*испол. директор*” instead of the full term “*исполнительный директор*” in the Fragment 3. Generally, humans often applied generalization, reduction

or omission strategies instead of full terms or phrases and skipped less important details to win more time and withstand cognitive burden. For example, in Fragment 3 the original phrase "to provide some numbers that provide a sobering picture of the dimensions of the problem" was shortly interpreted by human interpreter as "*осветить различные аспекты этой проблемы*". In comparison, Yandex interpreted the fragments fully without using any translation techniques or optimization, therefore resulting in slightly better completeness of the original message.

Additionally, in very difficult fragments containing strong regional accents with high WPM for example, in Fragment 6, assessors paid great attention to heavy sighs, misarticulations, pauses and fillers among human interpreters. Thus, due to fatigue resulting from significant cognitive load, humans performed worse in terms of completeness and delivery. Unlike human interpreters, speech intensity and accents did not affect the pace of delivery and completeness in Yandex output as was highlighted by the participants. This topic identified from the thematic analysis resonates with the quantitative results presented above where AI scored higher in completeness and fluency of delivery criteria compared to human interpreters who conveyed less amount of information, misarticulated some words, and used filled and unfilled pauses. Interestingly, it also aligns with the previous studies in other language pairs as presented in the discussion section.

Lexical and Grammatical Redundancy

The second theme identified from the thematic analysis was the logical consequence of the first theme. Since machine interpreting did not experience any human factors such as cognitive constraints, fatigue, etc. and conveyed the full message, its interpretations occasionally sounded lexically and grammatically redundant. Most of the assessors highlighted that machine interpreting was too comprehended and overloaded,

unlike human interpretations which were more succinct, concise and optimized. For example, Yandex provided the following interpretation in the Fragment 6:

> *В прошлом году моя страна приняла новый закон, вносящий поправки в национальное законодательство о борьбе с незаконным оборотом наркотических средств и психотропных веществ, введя национальную классификацию наркотических средств и психотропных веществ, которая будет дополнять международную классификацию запрещенных веществ, и для борьбы с распространением наркотиков, таких как прегабалин и трамадол, которые используются в медицинских целях, но могут вызывать зависимость у подростков и молодежи.*

Such endless sentence was grammatically and lexically redundant and was difficult to comprehend for the assessors. Yandex interpreted this fragment straightforwardly without reduction or omission of repetitive elements such as "*наркотических средств и психотропных веществ"* which should better be omitted when repeated several times. In comparison, human interpreters divided this sentence in several parts and made it sound easier for comprehension. Most of the assessors marked these significant limitations of machine interpreting in their comment section and opted for human interpreting options over Yandex'.

Failed Interpretation of Numbers

The third theme is the most critical one since it presents the very serious type of mistake. Despite the fact that Yandex interpreted more completely, it misinterpreted almost all the complex numbers higher than 1,000 as was marked by the participants. For example, Yandex interpreted the utterance "in 7000 miles away" as "*в 7 милях отсюда*" instead of the correct "*в 7000 милях отсюда*" in the Fragment 2. Additionally, the number "44,000 Palestinians" in Fragment 4 was interpreted as "*44 Палестинца*" which was also

completely irrelevant in the opinion of the assessors. One more example was when the original utterance in Fragment 3 "in addition to more than 45,000 people from Libya" was misinterpreted by Yandex as "*это люди старше 45 лет из Ливана*" which was incorrect and entirely changed the meaning of the sentence as was noted by the participants. Similarly, the fragment "with the vast majority, more than 130,000 people" was interpreted as "*подавляющее число людей старше 130 лет*" in Fragment 3 which was not just incorrect but also completely illogical, therefore disrupting the overall consistency with the original message, participants reported. Interestingly, this finding contradicts the previous studies in the same or different language pairs where most of the numbers were successfully recognized and correctly interpreted by the machine interpreting tools. This significant type of mistakes might be explained by the fact that machine neural networks use cascading system, as was mentioned in the literature review. Possible explanations for such serious disadvantage and discussions on discrepancies with the previous studies are presented in the next chapter.

Unnatural Delivery

Further analysis of participants' feedback showed some delivery issues. A common view amongst assessors was that machine interpreting often sounded very unnatural and even unpleasant. In spite of the fact that machine interpreting scored higher than human interpreters in fluency of delivery – primarily due to clearer diction, articulation, and the absence of filled pauses – this advantage was limited to technical aspects. As previously reported, human interpreters, under extreme cognitive load, occasionally misarticulated words and used fillers, leading to lower scores. For example, human interpreters misarticulated the following words: "*тож(е) самое", "касающих(ся)", "у всех нуждающих(ся)"* in the Fragment 5 and 6. Trying to keep up with the flow and pronounce as many words as possible, they lost the endings "*ся*". Unlike Yandex's output, human

interpreters also filled the pauses with such fillers as "*emm*" or "*ahh*" in Fragments 2 and 5. As a result, the scores for fluency of delivery were lower among human interpreters as was explained in their feedback section.

Nevertheless, all expert assessors expressed the opinion that AI-generated speech was much worse than human outputs in terms of general perception since it was too monotonous and robotic. Moreover, Yandex's robotic voice lacked necessary word stresses and failed to emphasize key phrases, affecting the overall tone. For instance, when interpreting enumerations in Fragment 6, Yandex did not use appropriate intonation, making the list sound like separate, disjointed sentences, which disrupted logical flow. One more example of unnatural delivery and disjointed sentences was in Fragment 4 where the original excerpt "we also echo the call for the resolution issued by the recent extraordinary Arab Islamic Summit to mobilize international support to suspend Israel's participation in the UN General Assembly" was interpreted by Yandex as "*Мы также присоединяемся к призыву резолюции принятой на недавнем чрезвычайном арабо-исламском саммите. Мобилизовать международную поддержку для приостановления участия Израиля в мероприятии. Собрание".* Irrelevant segmentation of one sentence into several unrelated ones as well as lack of intonation and robotic voice altogether resulted in extremely unnatural delivery precepted unsatisfactory by the assessors.

Another instance of unnatural delivery was the case when machine interpreting failed to reflect grief and sadness in its interpretation, as was marked by Assessor 4 in Fragment 3. When the original speaker reported the number of victims and deaths in the Middle East, in the opinion of the expert participant, it was necessary to reflect a certain degree of grief in voice but machine interpreting sounded too positive for such a tragic message.

To sum up, this finding partially contradicts the quantitative results where AI outperformed humans in delivery criterion. However, more detailed analysis of the feedback section demonstrated that the assessors scored AI higher only in terms of better articulation and absence of pauses. Detailed qualitative analysis of participants' feedback demonstrated generally worse delivery due to unnatural robotic and unpleasant voice in machine interpreting. It also lacked word stresses to highlight keywords or phrases and struggled with correct segmentation of sentences leading to disrupted logical flow.

Irrelevant Context Recognition

Finally, a recurrent theme in the assessment sheets was a sense amongst participants that machine interpreting occasionally struggled in context recognition and as a result, sometimes misinterpreted certain terms. Raters indicated, that Yandex misrecognized the UN context in Fragments 2, 4 and 6. As a result such term as "Madame President" was interpreted as "*Госпожа Президент*" which was inaccurate in the given context where "president" did not refer to "*президент*" but "*председатель*". One more example was the context of Israeli-Palestine relationships which was also not fully recognized by Yandex in Fragment 4. Consequently, Yandex interpreted the utterance "implementation of two-state solution" as "*реализация решения о двух государств*" instead of "*двухгосударственное решение*" which was a significant inaccuracy in the given context. Formally, these terms were interpreted correctly, however, in the given context such interpretation sounded less appropriate and inequivalent, as expressed by the participants. This type of inaccuracy aligns with the quantitative results where AI mean scores in terminology criterion were lower compared to human interpreters. Thus, due to worse context recognition Yandex interpreted some of the terms inaccurately.

To sum up the qualitative findings, deeper analysis demonstrated that the most common mistakes and inaccuracies were related to inability of machine interpreting to

optimize and reduce redundant elements, wrong segmentation of sentences and irrelevant interpretation of numbers, ineffective context recognition leading to less appropriate word choices, and unnatural delivery. Yandex produced more complete interpretations and, unlike human interpreters, did not struggle with high speed or strong accents. While human interpreters occasionally misarticulated words or omitted less critical information using various interpreting strategies, Yandex maintained a more fluent pace. However, this led to grammatical and lexical redundancy, making its output overloaded and difficult to comprehend. The most critical issue was misinterpretation of numerical data, distorting the original message. Additionally, Yandex struggled with context recognition, leading to less accurate word choices and terminological inconsistencies. Participants also perceived its robotic, monotonous voice negatively, noting a lack of word stress and improper sentence segmentation.

Summary

To sum up the chapter, the findings revealed complementary strengths and weaknesses in human and machine interpretations. Inter-rater reliability analysis showed excellent agreement among assessors, confirming the reliability of the evaluations. However, the wide confidence intervals pointed to some variability in rater judgments, likely due to the limited sample size. Quantitative findings indicated that human interpreters significantly outperformed AI in logical cohesion, terminology accuracy, and stylistic appropriateness, whereas AI demonstrated superior performance in completeness and fluency of delivery. Qualitative analysis supported these findings, uncovering strengths such as more complete and clear machine interpretation due to absence of typical human cognitive constraints and persistent issues in machine output such as misinterpretation of numbers, lexical and grammatical redundancy, lack of contextual sensitivity, and unnatural delivery patterns. The next chapter interprets and explains the

findings reported in this chapter, provides answers to the research questions as well as discusses how the results align or contradict the existing literature in this field.

Discussion

The present study was designed to determine the quality of machine interpreting in comparison to the traditional human interpreting in the English-Russian language combination. It was aimed to reveal the strengths and weaknesses of machine interpreting output quality in comparison to traditional human interpreting as well as to find to what extent machine interpreting is competitive, what are the comparative advantages and common error patterns. A mixed method approach applying quality assessment method was selected to meet the research goal and answer the research questions. Five expert assessors with strong background in conference interpreting and teaching translation studies courses were invited to listen and assess six original audio fragments in English with corresponding interpretation in Russian produced by Yandex neural networks and human interpreters. Analytical score-based assessment scale with criteria descriptors and four-fold evaluation system was distributed among the participants to assess the outputs and provide constructive feedback. Before data analysis procedures, inter-rater reliability test was conducted to evaluate the consistency of grades and reliability of judgments. Quantitative t-test and qualitative analysis provided meaningful results sufficient to answer the research questions. Thus, the purpose of this chapter is to interpret the results and answer the research questions as well as to connect the results with the relevant literature in this field. The first section summarizes and interprets the results and explains possible reasons for the identified mistakes and inaccuracies. The second section provides the answers the research questions. The third section discusses how the results agree or contradict the previous studies in this field. The chapter is summarized in a concluding paragraph with the main discussion insights.

Interpretation of Findings

The findings highlight distinct strengths and weaknesses in both human and AI-generated interpretations. Inter-rater reliability demonstrated excellent agreement among assessors proving relevancy of the provided assessments. Therefore, the sampled expert assessors were consistent and reliable in their judgments as evidenced by high ICC value. Yet, the confidence intervals (CI) were quite wide suggesting slight variability in ratings and indicating some uncertainty in case if the same study will be repeated multiple times. It might be explained by the relatively small number of subjects and differences in individual rater judgments altogether resulted in wider intervals (Gamer et al., 2019). A bigger sample of the speech fragments might potentially solve this limitation in future studies. Independent-sample t-test was conducted to compare the difference of mean scores between human and AI performance. Quantitative analysis showed that human interpreters outperformed AI in logical cohesion, terminology, and style, while AI excelled in completeness and fluency of delivery and slightly in grammar and intonation. Qualitative analysis further reinforced these trends, highlighting AI's struggles with sentence segmentation, numerical accuracy, and context recognition. Although machine interpretations were more complete and free from different human factors, they often included redundant elements and unnatural phrasing, making them harder to follow. In contrast, human interpreters occasionally omitted less critical information but employed strategies that ensured coherence and appropriateness. Additionally, Yandex's robotic delivery and lack of prosody negatively impacted overall perception. These differences indicate that machine interpreting produces more fluid and uninterrupted speech but lacks the nuanced cohesion and contextual adaptation achieved by human interpreters. Thus, there are both positive and negative sides of machine interpreting output.

Interestingly, quantitative and qualitative results are interrelated and reflect each other in terms of the possible explanations of the results. Explanations of some of the quantitative findings were found during the qualitative analysis where participants reflected on the interpreting outputs. For example, much better performance of humans in terms of logical cohesion might be explained by the fact that Yandex struggled with segmentation of long sentences and context recognition, as was found in the thematic analysis. Unlike humans who have critical thinking and context awareness, Yandex machine interpreting sometimes divided long excerpts into shorter sentences improperly while the word choices were occasionally irrelevant in the given contexts. These inaccuracies affected overall logical flow which was sometimes interrupted in Yandex machine interpreting. Likewise, worse performance in terms of terminology and style where statistically significant difference in the mean scores was found, might also be explained by the challenged context recognition. Since the context of some of the speeches was recognized improperly, the terminological choices and stylistic devices were less precise and not as idiomatic as in humans' outputs. One more explanation for the disrupted logical flow in the MI's outputs is wrong interpretation of numbers in almost every fragment. Participants significantly reduced the marks for Yandex in the logical cohesion criterion in the fragments where thousands were inappropriately interpreted as decimals. This trend was observed in many fragments and cannot be accepted as an occasional inaccuracy. False interpretation of numerical data is a significant mistake which undermined the original message of the sentences and interrupted the logical flow causing misleading conclusions. Such a mistake can be explained by the typical limitations of the cascading system as was explained in the literature review. Cascade means that there are several steps of machine interpreting when the original speech is converted into text, then translated and finally voiced-over. Perhaps, at some stage of the cascade the neural

networks removed “extra” zeros in the numbers, for example 4,700 people in Fragment 3. By removing these zeros, the result was 4,7 people which was completely wrong. Yet, it does not explain the fact that the original speakers clearly mentioned the word “thousands” but still was misinterpreted by Yandex. More empirical tests are needed to determine whether this issue is an inherent flaw in Yandex or merely a temporary system bug.

Additionally, it was found that machine interpreting excelled in such criteria as completeness and fluency of delivery. It means that the original content was conveyed fuller and more complete by Yandex rather than by humans. Assessors indicated that machine interpreting had almost seamless diction and articulation, without distinct time lags or pauses, unlike humans’ performance. This finding might be explained by the fact that humans, in contrast to machine neural networks, have inherent cognitive limitations and such human factors as fatigue, stress, etc., due to which interpreters use different techniques and strategies to reduce the cognitive burden. For example, the thematic analysis demonstrated that humans tend to use generalization, reduction, omission and other optimizations. In contrast, Yandex did not apply any optimizations and often adhered to literal word-by-word interpretation. As a result, it provided more complete and fuller machine interpretations. Similarly, in case of very rapid speeches and strong regional accents, humans struggled to withstand the speed and keep up with the flow. It led to misarticulations, filled pauses and exhales of fatigue. Yandex, on the other hand, successfully handled even the most complicated fragments, recognizing strong accents irrespectively of high speed of delivery. Therefore, Yandex outperformed humans in terms of completeness and fluency, and withstood speed and accents better since it does not have inherent human factors. This finding marks a significant advantage of machine interpreting over humans.

However, it was also found that machine interpreting was often lexically and grammatically redundant. Yandex sometimes provided complex sentences difficult for comprehension often with unnatural phrasing. This drawback might be explained by literal and word-by-word precision of machine outputs aligning with the previous finding. Since Yandex provided more complete interpretation, it sometimes resulted in cumbersome, overly explicit sentences that hindered comprehension due to its redundancy. Unlike human interpreters, who purposely omit or rephrase information to maintain clarity and to reduce the cognitive load, Yandex retained every element of the source speech, occasionally overloading the original message. MI's difficulty in distinguishing between essential and supplementary information contributed to unnatural phrasing. This rigid adherence to exhaustive interpretation, while beneficial for completeness, ultimately diminished comprehensibility and made the interpretation hard to follow. It means that in simultaneous interpreting it is better to avoid redundancy and optimize comprehensive information instead of maintaining exact word-by-word accuracy. Human interpreters, unlike AI understand better how it is crucial to ensure a balance between completeness and clarity, allowing the audience to process the information more effectively.

Lastly, it was found that although Yandex performed better in terms of fluency of delivery, overall perception of machine output by the participants was negative. Robotic, monotonous, emotionless voice and absence of necessary word stresses were identified as significant drawbacks of machine interpretation. It means that human audience preferred to listen other humans with lively voice, natural rhythm and prosody rather than synthetic and monotonous voice of AI even though it was much better in terms of completeness and fluency. Moreover, occasionally improper sentence segmentation means that Yandex struggles to navigate through long sentences with interrupted intonation, resulting in irrelevant division of excerpts into smaller phrases. Furthermore, the lack of emphasis on

keywords and phrases in Yandex's output contributed to difficulties in extracting meaning, as stress patterns served as important cues for highlighting essential information. Without these elements, speech can feel lifeless and disconnected from the communicative intent, leading to lower engagement. These results suggest that future improvements in machine interpreting should not only focus on linguistic accuracy but also on enhancing the naturalness and expressiveness of synthesized speech to make AI-generated interpretations more pleasant and engaging for listeners.

To sum up, both human and machine interpretations exhibited unique advantages and limitations. Table 7 below provides a visual summary of the results' interpretation discussed in this section, illustrating how the qualitative findings reflect and support the quantitative results. Based on the interpretation of findings provided in this section, the next section answers the original research questions.

Table 7

Visual Summary of the Results' Interpretation

Criterion	Quantitative Results	Qualitative Results
Logical Cohesion	AI < Human	Wrong interpretation of numbers and illogical segmentation of sentences by AI.
Completeness	AI > Human	Information was conveyed fuller and more complete by AI. Humans occasionally omitted less critical information while AI preserved all the details.

Terminology	AI < Human	Less equivalent vocabulary of AI in comparison to humans due to worse context recognition.
Grammar	AI > Human	General grammar in AI output is slightly better. Yet, there are grammatical, lexical redundancy and overloaded sentences in AI outputs while humans optimized complex sentences using different strategies.
Style	AI < Human	Stylistic devices of AI are less precise and not as idiomatic as in humans' outputs in the given context.
Fluency of Delivery	AI > Human	AI has almost seamless diction and articulation, without distinct time lags or pauses, unlike humans who struggled to withstand the speed and accents. Yet, AI' has unnatural robotic delivery and lacks prosody.
Intonation	AI > Human	AI has slightly more persuasive tone as compared to humans, but there is monotonous, emotionless voice and absence of necessary word stresses in AI output.

Answers to the Research Questions

Summarizing the key findings of this study it is possible to answer the research questions. The first question was: How does the performance of AI-driven machine interpreting in the English-Russian language pair measure up against human interpreters? With respect to the first research question, statistical analysis revealed significant differences. The independent-sample t-test showed that human interpreters achieved higher mean scores in logical cohesion, terminology, and style, whereas machine interpreting performed significantly better in completeness and fluency of delivery. These results indicate that while AI-generated interpretations provide sufficiently accurate and relevant output for communication, they remain inferior in areas requiring more nuanced linguistic choices. Thus, it was found that AI-driven machine interpreting in English-Russian language pair competitively measures up against human interpreters providing adequate and relevant output sufficient for successful communicative event. On one hand, it is increasingly competitive with human interpreters, excelling in completeness and fluency of delivery but on the other hand, it lags behind in logical cohesion, terminology accuracy, and stylistic appropriateness.

The second research question was: What are the typical mistakes and limitations in AI-driven machine interpreting in the English-Russian language pair? The most prominent mistakes and inaccuracies include misinterpretation of numerical data, robotic unnatural delivery and occasionally irrelevant context recognition resulted in imprecise terminology usage in context-sensitive scenarios. Additionally, AI-generated outputs tend to be lexically and grammatically redundant, struggle with sentence segmentation and correct prosody leading to hindering comprehension. Its inability to distinguish essential from non-essential content and its failure to apply prosodic features diminish the overall listeners perception. While machine interpreting handles high speed of delivery and regional accents

successfully, in some cases even better than human interpreters, it still lacks clarity and cohesion as compared to humans. These findings suggest that while machine interpreting has clear advantages, further improvements are needed to enhance context recognition, speech optimization and segmentation, numerical accuracy, and natural delivery. Therefore, this thesis states that machine interpreting is potentially feasible for professional use since it produces adequate output sufficient for effective communicative event. However, it remains overall inferior to human interpreters in some aspects crucial for high-level events. The study underscores the evolving nature of AI-powered interpreting tools, indicating that while they may not yet be suitable for high-stakes settings, they offer valuable support for less demanding tasks. Compared to the previous studies in this field, there are both significant improvements and some contradictions as presented in the next section.

Alignment and Contradiction to the Literature

The results chapter demonstrated interesting findings to some extent aligning with the reviewed literature, yet contradicting in some areas. The literature review chapter highlighted AI's challenges with high speed, strong accents, speech disfluencies, and context recognition. Studies across various language pairs revealed occasional errors in automatic speech recognition, a tendency toward literal translation, and poor context recognition. Based on this, the literature review suggested that Yandex would perform well with clear, straightforward speech but struggle with more complex cases involving accents, misarticulations, and high speed. However, as discussed in this chapter, the findings of this study both align with and contradict prior research, reflecting improvements in machine interpreting output. Yandex successfully handled even very complex speeches with high speed of delivery and strong regional accents. MI produced generally satisfactory and overall adequate output comparable to human interpreters with several exceptions where

serious disfluencies and mistakes were found. Thus, this section presents how these results align or contradict to the existing literature.

Some of the findings align with existing literature on similar topics. For example, Liu (2023) and Fantinuoli and Prandi (2021) concluded that interpreters' cognitive limitations can negatively impact information completeness, especially due to fatigue, stress, and other human factors – an area where AI-powered tools tend to perform better. The results of this thesis support these conclusions, demonstrating that AI notably outperforms human interpreters in terms of information completeness and precision. Compared to AI, human interpreters often generalize source information to keep up with speed and manage cognitive load, which can result in the omission of certain details, as observed in this study and in the research of Fantinuoli and Prandi (2021). While humans strive to convey the key message explicitly, their output may lack the precision of AI-powered tools, which translate every unit of information without optimization. However, both this study and Liu and Liang (2024) highlight that MI's strict adherence to the input language for the sake of precision frequently results in overly literal and confusing outputs. In this paper, this tendency led to grammatical and lexical redundancy. Additionally, ineffective context recognition – another major limitation of MI was emphasized by Liu and Liang (2024) and is consistent with the findings of this study as well. Fedorova (2023) marked that human interpreters comprehend meaning better, accurately detecting subtle nuances in subtext and intonation, and, when necessary, correct the speaker's errors or hesitations – something which MI is still unable to do as was also found in this study. As was concluded by Sarmanova (2022), on the technical side the main issues of machine interpreting are related to speech recognition while on the communicative side, systems struggle with context and fail to interpret implicit information, such as speaker attitudes and emotions. Generally, this thesis supports this statement since the similar results were

found in terms of ineffective context recognition and nonverbal cues in delivery. On the other hand, technical side has notably improved since this study did not find any serious mistakes related to improper speech recognition.

Therefore, the findings of the present study partially contradict some of the previous research. It signals the remarkable improvements in machine interpreting over time and the impact of technological updates on the overall quality of machine interpreting. For example, Belenkova (2019) also tested quality of Yandex and found very low lexical and grammatical validity of the tested tool which failed in almost every semantic segment. It was found that some parts were not interpreted by Yandex at all (Belenkova, 2019) which is completely opposite to the present results where Yandex neural networks demonstrated high completeness and general adequacy in most of the criteria. Additionally, Sarmanova (2022) summarized in her study that some of the key mistakes were caused by the regional accents, spontaneous, rapid and misarticulated speech which are not always recognized by AI correctly. However, the present study also contained quite difficult speeches with regional accents and high speed of delivery. Nevertheless, Yandex managed to recognize the accents and successfully interpreted the original messages keeping up the rapid delivery, sometimes even better than humans. Thus, it might be suggested that thanks to the continuous technological updates, machine interpreting has significantly improved its performance in terms of ASR, high speed and accents recognition. This improvement may be attributed to continual algorithmic enhancements that are not reflected in earlier studies. These discrepancies highlight a time lag in the literature, as much of the existing research relies on already outdated versions of machine interpreting systems that no longer reflect current performance capabilities. Yet, despite significant machine performance improvements as compared to the previous studies, the existing literature does not explain the fact of misinterpreted numbers in the given sample

of speeches. The previous studies of Sarmanova (2022) and Fantinuoli (2017) demonstrated very accurate recognition of all the numbers with up to 99% accuracy. It contradicts the present findings where almost all the complex numbers were misinterpreted, most likely due to mistakes in the cascading system Yandex applies, as was explained in the previous sections. Thus, it requires more attention in the future studies to monitor the further improvements.

Interestingly, some literature predicts that machines will be able to interpret on the same level as human interpreters just in few years. For example, Ray Kurzweil (1999) in his book "The Age of Spiritual Machines" forecasted that machine interpreting will become daily routine by 2029. However, he also noted that while such technologies already exist, they remain limited in adoption, as they are unsuitable for situations requiring meticulous attention to nuances. The same point was made by Carl and Braun (2017) seven years ago. Both Kurzweil (1999) and Carl and Braun (2017) believed that machine interpreting might be still unacceptable in high-stakes situations. This Master's thesis generally supports these claims, as the results showed adequate and relevant machine interpretation which nevertheless is inferior to humans in some aspects making it still unsuitable for high-level events where accuracy is critical and errors carry significant consequences.

Summary

To conclude the chapter, the findings demonstrate that while machine interpreting has made significant advancements, it remains a tool with distinct strengths and limitations. Machine interpreting excels in completeness and fluency, successfully handling rapid speech and strong accents without cognitive fatigue. However, its challenges with logical cohesion, context recognition, numerical accuracy, and prosody hinder its ability to fully replicate the nuanced of human interpretation. Despite noticeable

improvements in machine interpreting technology, human interpreters continue to outperform in key areas that require critical thinking, adaptability, and contextual awareness. The study underscores the continuous development of AI-powered interpreting tools, suggesting that while they are not yet ideal for high-stakes professional settings, they can serve as useful aids for less complex tasks. These findings align with some existing research while also revealing contradictions, particularly in areas where AI has demonstrated unexpected strengths, such as handling complex accents and rapid delivery. Ultimately, while AI-powered interpreting tools are increasingly competitive, they are not yet a complete substitute for human expertise. The next and final chapter concludes this thesis, discussing broader implications, future research directions, study limitations, and practical significance.

Conclusion

This Master's thesis was underpinned by the existing problem that research on interpreting technologies is still in its early stages, particularly for the English-Russian language pair, where only a few studies have explored this area. As a result, there is a lack of understanding on how competitive machine interpreting tools are compared to human interpreters, as well as their specific advantages and limitations. While some studies have examined the future of English-Russian interpreting in the context of technological advancements, they have not empirically assessed the latest AI systems developed for interpreting automation. Therefore, this timely study meaningfully contributes to an underexplored area underscoring the need for a deeper understanding of the typical output of machine interpreting tools. The purpose of this Master's thesis was to evaluate the quality of machine interpreting in comparison to human interpreting in the English-Russian language pair. The research problem stemmed from the limited research on machine interpreting technologies, particularly regarding their competitiveness, advantages, and typical limitations in this language combination. Addressing this gap, the study focused on assessing the overall performance of machine interpreting, identifying the strengths and weaknesses, and determining how it is relevant as compared to human interpreters. Quality assessment research method combining quantitative and qualitative analyses was used to provide a comprehensive evaluation. The study employed an analytical score-based quality assessment framework, inviting expert assessors to listen six original English audio fragments and evaluate the corresponding Russian interpretations produced by Yandex's neural networks and human professional interpreters as sampled from the UN Web TV website. Five expert assessors with extensive experience in conference interpreting and teaching were invited to analyze the output based on predefined quality criteria, ensuring a structured and objective evaluation. An inter-rater reliability test was conducted to confirm

high consistency among assessors, and a t-test was applied to find the statistical differences in the mean scores between human and machine interpretations. The adopted methodological approach generated the findings outlined in the next section.

Findings Outline

Quantitative findings revealed that human interpreters outperformed machine interpretation in such aspects as logical cohesion, terminology accuracy, and stylistic appropriateness. Machine interpreting, however, excelled in completeness and fluency of delivery. Statistical significance was observed in criteria such as terminology, style, and fluency of delivery. Yet, the generalizability of the quantitative findings is limited by the relatively small sample size and the evolving nature of machine interpreting tools, which are regularly updated, making the results relevant only for a short timeframe. Further qualitative analysis uncovered specific mistakes and critical weaknesses in machine interpretations. Thematic analysis of the participants' feedback revealed five major topics. The most common issues included ineffective context recognition, misinterpretation of numerical data, unnatural delivery, and excessive redundancy in phrasing. Yandex often failed to optimize sentence structures, leading to an overloaded output that reduced comprehensibility. While MI managed to maintain a steady pace and showed resilience to strong accents or high-speed speech, it struggled with sentence segmentations, nuanced lexical choices and terminological inconsistencies. Assessors also highlighted its robotic voice and lack of natural prosody as a major drawback, making it less effective for professional communication settings.

Overall, the research successfully achieved the purpose, provided a detailed quality assessment analysis of machine and human interpreting and answered the research questions comprehensively, as provided in the previous chapter. The findings align with existing literature while also revealing contradictions, particularly in areas where machine

interpreting has shown unexpected strengths, such as handling complex accents and fast-paced speech more effectively than human interpreters and much better than it was anticipated. While previous studies suggested AI would struggle with misarticulations and rapid delivery, this study demonstrated improvements in machine interpreting technology. Nevertheless, the persistent weaknesses, such as contextual accuracy, numerical precision, etc. reinforce the view that machine interpreting remains unsuitable for high-stakes events where accuracy is paramount. The study contributes to the growing body of research on machine interpreting, emphasizing the need for further technological advancements and ongoing human oversight in professional settings. The next paragraph outlines the specific contributions this study makes to the existing body of knowledge.

Contribution to Knowledge

This study offers several novel contributions to the field of interpreting studies. Most notably, at the moment of writing and to the best of the researcher's knowledge, it presents the first empirical quality assessment of Yandex's machine interpreting in the English–Russian language pair, thereby addressing a significant gap in current research body. Furthermore, this study contributes to the field by developing a tailored assessment scale for simultaneous interpreting quality, addressing the long-standing absence of a universally accepted evaluation framework. By offering a replicable quality assessment framework adapted based on the existing prominent studies (Han, 2021; Lee, 2008, 2015; Liu, 2013; Nadir, 2017; Shang, 2021; Wang et al., 2015), it advances efforts toward more consistent and transparent quality assessment. Future research could utilize this framework to assess interpreting quality in diverse settings, contributing to a more standardized approach in evaluating simultaneous interpretation. Additionally, the findings challenge prevailing assumptions by identifying unexpected strengths in MI performance, particularly in the accurate rendering of accented speech with high speed of delivery, thus

evidencing significantly improved output of machine output as compared to the previous studies (Belenkova, 2019; Fantinuoli & Prandi, 2021; Liu, 2023; Liu & Liang, 2024; Sarmanova, 2022). Finally, the study extends and nuances the previous works on this topic by Belenkova (2019), Fantinuoli and Prandi, (2021), Sarmanova (2022), Liu and Liang (2024), and others, offering updated empirical evidence that corroborates, refines, and questions earlier claims about the low efficacy and reliability of MI systems. Given this contribution, the next section explores the real practical implication of the study, considering the potential application of machine interpreting and limitations in real-life professional settings.

Practical Implications

The findings of this study have significant practical implications valuable for all stakeholders in the interpreting market, including AI developers, professional interpreters, clients relying on interpreting services, and the academic community. The identification of recurring errors in AI-generated machine interpretations, particularly in terminology accuracy, style consistency, and fluency, highlights areas requiring targeted improvements in machine learning models. Thus, developers should focus on refining terminology databases and enhancing contextual adaptation mechanisms to improve AI performance, especially in specialized domains. The findings regarding less pleasant and unnatural delivery suggest recommendations that future improvements in machine interpreting should not only focus on linguistic accuracy but also on enhancing the naturalness and expressiveness of synthesized speech to make AI-generated interpretations more pleasant and engaging for listeners.

For professional conference interpreters, these findings offer valuable practical insights into the extent to which AI-generated interpretations can be considered comparable or competitive to human performance. As AI continues to evolve, interpreters

must stay informed about its capabilities and limitations, particularly in areas where human expertise remains superior, such as nuanced contextual adaptation and complex discourse management. Understanding AI's strengths and weaknesses can help interpreters position themselves in a rapidly changing market and leverage technology where appropriate.

For clients and end users of interpreting services, the study provides a clearer picture of AI's practical applicability. Given the observed limitations in fluency, terminology precision, stylistic coherence, and others, users must carefully assess whether machine interpreting meets their specific needs or whether human interpretation remains the preferred option in high-stakes or linguistically complex settings.

This study also carries theoretical significance and is beneficial for the academic community. As was mentioned in the previous section, the refined assessment scale has the potential for broader practical application in future research, serving as a structured tool for evaluating simultaneous interpretations. Its applicability extends beyond the current study, allowing for further quality assessment analysis in different contexts and languages. Despite the valuable practical implications, this study entails several limitations as presented in the next section.

Limitations

While this study provides valuable insights into the quality of machine interpreting, several limitations must be acknowledged. The key limitation is the relatively small sample size, which restricts the generalizability of the findings. A larger dataset, encompassing a wider range of interpreting contexts and more fragments for consideration, would enhance the reliability of future studies. Additionally, the study's timeframe was constrained by the frequent updates of machine interpreting tools, meaning that the results reflect a specific technological stage rather than long-term trends.

Another limitation concerns the nature of the utilized machine interpreting tool. The issue of latency was not considered in the present research since Yandex provided interpretation based on the pre-recorded videos which did not allow to measure the degree of delay in seconds. As a result, the findings do not fully reflect the challenges AI might face in live interpreting scenarios. Furthermore, this study was primarily focused on the quality of the produced output without taking into account the users' satisfaction which should be a focus of future research.

Finally, the selected methodology presents certain constraints. Although the assessment framework was carefully adapted from existing research, expert assessors' evaluations involved a certain degree of subjectivity. This is an inherent limitation of the selected methodology. While efforts were made to ensure consistency, minor variations in judgment may have influenced the results. In light of these limitations, several recommendations for further research can be identified to address the existing gaps in knowledge.

Recommendations for Future Research

This thesis underscores the importance of continued empirical research in AI interpreting. As machine interpreting tools evolve, their capabilities and limitations will shift, necessitating ongoing assessment. Building upon the identified limitations, future research can provide valuable insights by addressing several areas for improvement. First, larger datasets should be employed, encompassing a broader range of interpreting contexts and more audio fragments to enhance the reliability and generalizability of the findings. Additionally, conducting longitudinal studies would help track machine interpreting performance over time and assess the effects of technological advancements.

Future studies should also focus on AI tools designed for professional, real-time commercial use, particularly addressing latency in live interpreting scenarios. This would

provide a more accurate understanding of their applicability and competitiveness in real-world environments. It is also recommended to examine users' satisfaction alongside output quality to present a more holistic view of machine interpreting effectiveness.

In terms of refining the evaluation framework, future research could simplify some criteria utilized in this study to streamline the assessment process without sacrificing analytical depth, thereby improving efficiency and reducing the time-consuming process of assessment. Building on the limitations and recommendations discussed, the next and final section offers original reflections and suggestions aimed at advancing the discourse on machine interpreting and its future development.

Suggestions

As AI continues to evolve, its role in simultaneous interpreting will likely expand, raising important questions about its reliability and practical applicability. While this study highlights both the strengths and limitations of machine interpreting, it also underscores the need for continuous research and industry awareness. Based on the findings, this thesis argues that machine interpreting tools are not yet capable of fully replicating human expertise, however, with ongoing advancements, AI could become a valuable tool in specific contexts, such as assisting human interpreters or providing low-cost solutions for less critical communication needs. To ensure responsible development and implementation of machine interpreting, collaboration between researchers, developers, and industry professionals is essential. Future research should prioritize large-scale empirical studies that will test professional machine interpreting tools, evaluate real-time interpreting conditions, and take into account users' perception. Additionally, ethical considerations surrounding AI's role in interpreting, including issues of accountability and quality assurance, should not be overlooked.

Ultimately, the question is not whether AI will replace human interpreters but rather how it can be integrated effectively while maintaining high-quality communication. By fostering critical discussions and empirical research, the interpreting field can better prepare for the evolving landscape of human-AI interaction, ensuring that technological advancements serve rather than disrupt the profession.

References

Ahmed, S. A. A. (2022). Technology and artificial intelligence in simultaneous interpreting: A multidisciplinary approach. *CDELT Occasional Papers in the Development of English Education*, *78*(1), 325-353. https://www.doi.org/10.21608/opde.2022.249945

Angelelli, C. (2002). Interpretation as a communicative event: A look through Hymes' lenses. *Meta*, *45*(4), 580-592. https://doi.org/10.7202/001891ar

Angelelli, C. V., & Jacobson, H. E. (Eds.). (2009). *Testing and assessment in translation and interpreting studies*. John Benjamins Publishing Company. https://doi.org/10.1075/ata.xiv

Avedova, R. & Miteleva, V. (2016). Sovremennye tekhnologii v sfere sinkhronnogo perevoda. [Modern Technologies in Simultaneous Interpreting]. *Sovremennye tendetsyy razvitiya nauki i tekhnologii*, *10*(2), 5-9.

Belenkova, N. (2019, November). Assessing the capacity of machine interpreting technologies: The Russian experience. In *Conference Proceedings. Innovation in Language Learning 2019*.

Braun, S. (2019). Technology and interpreting. In M. O. Hagan (Ed.), *Routledge handbook of translation and technology* (pp. 271-288). Routledge. https://doi.org/10.4324/9781315311258-19

Canva Pty Ltd. (2025). Canva [Graphic design software]. Retrieved from https://www.canva.com

Carl, M., & Braun, S. (2017). Translation, interpreting and new technologies. In *Routledge handbook of translation studies and linguistics* (pp. 374-390). Routledge.

Cho, E., Fugen, C., Herrmann, T., Kilgour, K., Mediani, M., Mohr, C., Niehues, J., Rottmann, K., Saam, C., Stuker, S., & Waibel, A. (2013). A real-world system for

simultaneous translation of German lectures. *INTERSPEECH, 13*, 3473 – 3477. https://doi.org/10.21437/Interspeech.2013-612

Corpas Pastor, G. (2018). Tools for interpreters: The challenges that lie ahead. *Current Trends in Translation Teaching and Learning E*, 5, 157 – 182. https://www.doi.org/10.5281/zenodo.5940648

Corpas Pastor, G. (2021). Interpreting and technology: Is the sky really the limit? *Proceedings of Translation and Interpreting Technology Online*, pp. 15–24. http://dx.doi.org/10.26615/978-954-452-071-7_003

Coursera. (2023, July 29). What is Artificial Intelligence? Definition, uses, and types. https://www.coursera.org/articles/what-is-artificial-intelligence

Costa, H., Corpas Pastor, G., & Durán Muñoz, I. (2014). Technology-assisted interpreting. *MultiLingual*, *143*(25), 3.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications.

Defrancq, B., & Fantinuoli, C. (2021). Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers. *Target*, *33*(1), 73-102. http://dx.doi.org/10.1075/target.19166.def

Desmet, B., Vandierendonck, M., & Defrancq, B. (2018). Simultaneous interpretation of numbers and the impact of technological support. In *Interpreting and Technology* (pp. 13-27). Language Science Press. http://dx.doi.org/10.5281/zenodo.1493281

DePalma, D. A., Pielmeier, H., Stewart, R. G., & Henderson, S. (2013). The language services market. *Common Sense Advisory*.

Doherty, S. (2016). The impact of translation technologies on the process and product of translation. *International Journal of Communication, 10, 947–969.*

Downie, J. (2019). *Interpreters vs machines: Can interpreters survive in an AI-dominated world?* Routledge. https://doi.org/10.4324/9781003001805

Drozdova, K. A. (2015). Mashinnyi perevod: Istoriya, klassifikatsiya, metody [Machine translation: History, classification, methods]. *Filologicheskie Nauki v Rossii i Za Rubezhom*, *3*(7), 139–141.

European Parliament. (2023, June 20). *What is artificial intelligence and how is it used?* News: European Parliament. https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used

Fantinuoli, C. (2017). Speech recognition in the interpreter workstation. *Proceedings of the Translating and the Computer*, 39, 25-34.

Fantinuoli, C. (2018). Interpreting and technology: The upcoming technological turn. In C. Fantinuoli (Ed.), *Interpreting and technology* (pp. 1–12). Language Science Press. http://dx.doi.org/10.5281/zenodo.1493289

Fantinuoli, C. (2019, November). The technological turn in interpreting: The challenges that lie ahead. In *Proceedings of the Conference Übersetzen und Dolmetschen* (Vol. pp. 334-354).

Fantinuoli, C., & Prandi, B. (2021). Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. In *Proceedings of the 18th International Conference on Spoken Language Translation* (pp. 245–254). https://doi.org/10.48550/arXiv.2103.08364

Fantinuoli, C., & Dastyar, V. (2022). Interpreting and the emerging augmented paradigm. *Interpreting and Society, 2*(2), 185-194. https://doi.org/10.1177/27523810221111631

Fantinuoli, C., Marchesini, G., Landan, D., & Horak, L. (2022). KUDO Interpreter Assist: Automated real-time support for remote interpretation. In J. Esteves-Ferreira, R. Mitkov, M. Recort Ruiz, O.-M. Stefanov, D. Chambers, J. M. Macan, & V. Sosoni (Eds.). In *Proceedings of the 43rd Conference Translating and the Computer* (pp. 68–77). https://doi.org/10.5281/zenodo.7143056

Fantinuoli, C. (2023). Towards AI-enhanced computer-assisted interpreting. In *Interpreting Technologies–Current and Future Trends* (pp. 46-71). John Benjamins. http://dx.doi.org/10.1075/ivitra.37.03fan

Fedorova, E. A. (2023). Perspektivy ustnogo mashinnogo perevoda kak al'ternativy professii "perevodchik" [Prospects of machine interpreting as an alternative to the profession of "translator"]. In *Perevod i inostrannye yazyki v global'nom dialoge kul'tur: Sbornik statei.* (pp. 119–122).

Gamer, M., Lemon, J., & Singh, I. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. [R package]. https://CRAN.R-project.org/package=irr.

Garbovsky, N. & Kostikova, O. (2019). Intelligence in translation: Artful or artificial? *Vestnik of Moscow University. Series 10: Translation Theory*, (4), 3–25.

Garbovsky, N. K. (2021). Chetvertaya promyshlennaya revolutsiya, obrazovanie i kultura. [Fourth industrial revolution, education and culture]. *Pedagogica*, 11, 83-92.

Gile, D. (1990). L'évaluation de la qualité de l'interprétation par les délégués: Une étude de cas [The evaluation of interpretation quality by delegates: A case study]. *The Interpreters' Newsletter, 3*, 66–71.

Grissom, A. C. (2017). *Incremental prediction and decision-making for simultaneous machine translation*. Boulder.

Guo, M., Han, L., & Anacleto, M. T. (2023). Computer-assisted interpreting tools: Status quo and future trends. *Theory and Practice in Language Studies*, *13*(1), 89-99. http://dx.doi.org/10.17507/tpls.1301.11

Gurman, M. (2025, March 13). *Apple plans AirPods feature that can live-translate conversations.* Bloomberg. https://www.bloomberg.com/news/articles/2025-03-13/apple-plans-ios-19-feature-that-lets-airpods-live-translate-conversations

Han, C. (2021). Analytic rubric scoring versus comparative judgment: A comparison of two approaches to assessing spoken-language interpreting. *Meta*, *66*(2), 337-361. https://doi.org/10.7202/1083182ar

Horváth, I. (2022). AI in interpreting: Ethical considerations. *Across Languages and Cultures, 23*(1), 1–13. https://doi.org/10.1556/084.2022.00108

Hynes, R. (2022, December 19). *The state of machine interpreting.* Nimdzi. https://www.nimdzi.com/machine-interpreting-mt/

"Elon Musk predskazal ischeznoveniye professii perevodchika v budushchem iz-za razvitiya II." (2021, May 21). [Elon Musk predicted the disappearance of the translator profession in the future due to the development of AI]. TASS. https://tass.ru/ekonomika/11435349

Jamovi Project. (2024). Jamovi (Version 2.4) [Computer software]. Retrieved from https://www.jamovi.org

Jourdenais, R. & Mikkelson, H. (2015). *The Routledge handbook of interpreting*. Routledge. http://dx.doi.org/10.1075/intp.18.1.07gil

Kahane, E. (2015, December 2). Thoughts on the quality of interpretation. *Communicate*, 4.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Krenz, M., & Ramlow, M. (2008). *Maschinelle Übersetzung und XML im Übersetzungsprozess. Prozess der Translation und Lokalisierung im Wandel* [Machine translation and XML in the translation process. Process of translation and localization in transition]. Berlin.

KUDO. (n.d.). *AI speech translation: A-Z guide to quality*. https://kudo.ai/blog/ai-speech-translation-a-z-guide-to-quality/

KUDO. (n.d.). *24% increase in translation quality and new languages on KUDO AI v.3.0* https://kudo.ai/newsroom/press-release/24-increase-in-translation-quality-and-new-languages-on-kudo-ai-v-3-0/

Kurz, I. (1989, October). Conference interpreting: User expectations. In *Coming of Age: Proceedings of the 30th Annual Conference of the American Translators Association* (pp. 143-148). Learned Information.

Kurz, I. (2001). Conference interpreting: Quality in the ears of the user. *Meta*, *46*(2), 394-409. https://doi.org/10.7202/003364ar

Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence*. Viking. https://doi.org/10.1016/S0308-5961(99)00064-6

Landgraf, M. (2012). *Simultaneous translation: University without language barriers.* Karlsruhe Institute of Technology. https://www.kit.edu/kit/english/pi_2012_10978.php

Lee, J. (2008). Rating scales for interpreting performance assessment. *The interpreter and translator trainer*, *2*(2), 165-184. http://dx.doi.org/10.1080/1750399X.2008.10798772

Lee, S. B. (2015). Developing an analytic scale for assessing undergraduate students' consecutive interpreting performances. *Interpreting*, *17*(2), 226-254. http://dx.doi.org/10.1075/intp.17.2.04lee

Li, H., & Chen, H. (2019). Human vs. AI: An assessment of the translation quality between translators and machine translation. *International Journal of Translation, Interpretation, and Applied Linguistics (IJTIAL)*, *1*, 43–54. http://dx.doi.org/10.4018/IJTIAL.2019010104

Liu, M. (2013). Design and analysis of Taiwan's interpretation certification examination. In C. V. Angelelli (Ed.), *Assessment issues in language translation and interpreting* (pp. 163–178). John Benjamins.

Liu, W. (2023). Reflection on the history of connecting simultaneous interpreting with artificial intelligence in China: 2017–2021. In *Proceedings of Voronezh State University. Series: Linguistics and Intercultural Communication* (pp. 142–150). https://doi.org/10.17308/lic/1680- 5755/2023/2/142-150

Liu, Y., & Liang, J. (2024). Multidimensional comparison of Chinese-English interpreting outputs from human and machine: Implications for interpreting education in the machine-translation age. *Linguistics and Education*, *80*, 101273. https://doi.org/10.1016/j.linged.2024.101273

Ma, D. (2021, October). Quality Assessment Criteria in Business Conference Interpreting from the Perspective of Loyalty Principle. In *2nd International Conference on Language, Communication and Culture Studies (ICLCCS 2021)* (pp. 55-64). Atlantis Press. https://doi.org/10.2991/assehr.k.211025.009

Macdonald, P. (2013). It don't mean a thing... Simultaneous interpretation quality and user satisfaction. *Interpreters Newsletter, 18*, 35–59.

Manning C. (2020). *Artificial Intelligence Definitions.* Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf

Massey, G., & Ehrensberger-Dow, M. (2017). Machine learning: Implications for translator education. *Lebende Sprachen*, 62(2), 300–312. https://doi.org/10.1515/les-2017-0021

Moser, P. (1996). Expectations of users of conference interpretation. *Interpreting*, *1*(2), 145-178. https://doi.org/10.1075/intp.1.2.01mos

Moser-Mercer, B., Künzli, A., & Korac, M. (1998). Prolonged turns in interpreting: Effects on quality, physiological and psychological stress (Pilot study). *Interpreting*, *3*(1), 47-64. https://doi.org/10.1075/intp.3.1.03mos

Müller, M., Nguyen, T. S., Niehues, J., Cho, E., Krüger, B., Ha, T. L., ... & Waibel, A. (2016, June). Lecture translator-speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 82-86). https://doi.org/10.18653/v1/N16-3017

Nadir, A. (2017). *Simultaneous interpreting quality assessment at the Higher Arab Institute of Translation from teachers' and students' perspectives* [Doctoral dissertation, University of Algiers 2 - Abou EL Kacem Saâdallah]. https://doi.org/10.13140/RG.2.2.19729.43362

Nimdzi (2023). *Language Technology Atlas*. https://www.nimdzi.com/language-technology-atlas/

Niska, H. (1999). Quality issues in remote interpreting. In *Anovar-Anosar: Estudios de traducción e interpretación* (pp. 109–122). Servizo de Publicacións.

Pagura, R. J. (2019). An overview of assessment in interpreting: A conversation with our colleagues in language testing. *Avaliação de Traduções: Sala de Aula e Crítica*, *139*, 139. https://doi.org/10.14393/LL63-v35n2-2019-8

Patten, M. L., & Newhart, M. (2017). *Understanding research methods: An overview of the essentials* (10th ed.). Routledge. http://dx.doi.org/10.4324/9781315213033

Pearson, C. (2022, November 25). *Simultaneous Interpreting*. Knowledge Centre on Interpretation. European Commission. https://knowledge-centre-interpretation.education.ec.europa.eu/en/conference-interpreting/simultaneous-interpreting

Pisani, E., & Fantinuoli, C. (2021). Measuring the impact of automatic speech recognition on number rendition in simultaneous interpreting. In C. Wang & B. Zheng (Eds.), *Empirical studies of translation and interpreting* (pp. 181–197). Routledge. http://dx.doi.org/10.4324/9781003017400-14

Pöchhacker, F. (2001). Quality assessment in conference and community interpreting. *Meta*, *46*(2), 410-425. https://doi.org/10.7202/003847ar

Pöchhacker, F. (2015). *Routledge encyclopedia of interpreting studies*. Routledge.

Pöchhacker, F. (2024). Is machine interpreting interpreting? *Translation Spaces*. https://doi.org/10.1075/ts.23028.poc

Prandi, B. (2015). Use of CAI tools in interpreters' training: A pilot study. In J. Esteves-Ferreira, J. Macan, R. Mitkov, & O. M. Stefanov (Eds.), *Proceedings of the 37th Conference Translating and the Computer* (pp. 48–57). Tradulex.

Prandi, B. (2018). An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation. In C. Fantinuoli (Ed.), *Interpreting and technology* (pp. 29–59). Language Science Press. https://doi.org/10.5281/zenodo.1493293

Prandi, B. (2023). *Computer-assisted simultaneous interpreting: A cognitive-experimental study on terminology*. Language Science Press. https://doi.org/10.5281/zenodo.7143056

Pym, A. (2011). What technology does to translating. *Translation & Interpreting: The International Journal of Translation and Interpreting Research*, *3*(1), 1-9.

Pym, A., & Torres-Sim´on, E. (2021). Is automation changing the translation profession? *International Journal of the Sociology of Language, 2021*(270), 39–57. https://doi.org/ 10.1515/ijsl-2020-0015

Radwinski, M. C. (2017). *Maschinelles Dolmetschen — Zukünftige Entwicklung und deren Bedeutung für den Berufsstand „Dolmetschende"* [Machine interpreting — Future developments and their significance for the profession of interpreters].

"Rektor MGLU zayavila, chto iskusstvennyy intellekt ne smozhet polnost'yu zamenit' perevodchika." (2020, July 10). [The rector of Moscow State Linguistic University stated that artificial intelligence will not be able to completely replace a translator]. TASS. https://tass.ru/obschestvo/8931011

Saldanha, G., & O'Brien, S. (2014). *Research methodologies in translation studies*. Routledge. https://doi.org/10.1556/084.2016.17.1.8

Samsung. (2024). *How to use Live translate for phone calls with Galaxy AI*. Samsung Caribbean. https://www.samsung.com/latin_en/support/mobile-devices/how-to-use-live-translate-for-phone-calls-on-the-galaxy-s24/

Sarmanova, Zh. (2022). *Digitalization of interpreting: Synergy of the interpreter's professional activity and automation*. [Doctoral dissertation, L.N. Gumilyov Eurasian National University]. National Digital Library of the Republic of Kazakhstan.

Seol, H. (2024). Seolmatrix: Correlations suite for jamovi. (Version 3.8.6) [Jamovi module]. https://github.com/hyunsooseol/seolmatrix.

Shang, X. (2021). Developing a weighting scheme for assessing Chinese-to-English interpreting: Evidence from native English-speaking raters. In *Testing and assessment of interpreting: Recent developments in China* (pp. 45–65). https://doi.org/10.1007/978-981-15-8554-8_3

Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: the new system technology* (p. 410). Springer Nature. https://doi.org/10.1007/978-3-031-21448-6

Tirosh, O. (2023, March 30). *Simultaneous interpretation – how it works*. Tomedes. https://www.tomedes.com/translator-hub/simultaneous-interpretation

Tongpoon-Patanasorn, A., & Griffith, K. (2020). Google Translate and translation quality: A case of translating academic abstracts from Thai to English. *PASAA: Journal of Language Teaching and Learning in Thailand, 60*, 134–163. https://doi.org/10.58837/CHULA.PASAA.60.1.5

United Nations DGACM. (2017, April 28). *Language Competitive Examinations for Interpreters.* [Video]. YouTube. https://www.youtube.com/watch?v=_Rkj7mn0Azg&t=1s

Wadensjö, C. (1998). *Interpreting as interaction*. Longman

Wang, J., Napier, J., Goswell, D., & Carmichael, A. (2015). The design and application of rubrics to assess signed language interpreting performance. *The Interpreter and Translator Trainer*, *9*(1), 83-103. https://doi.org/10.1080/1750399X.2015.1009261

Wang, X., & Wang, C. (2019). Can computer-assisted interpreting tools assist interpreting? *Transletters. International Journal of Translation and Interpreting*, (3), 109-139.

Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in machine translation. *Engineering*, *18*, 143-153. https://doi.org/10.1016/j.eng.2021.03.023

Wang, H., & Li, Z. (2022). Constructing a competence framework for interpreting technologies, and related educational insights: an empirical study. *The Interpreter and Translator Trainer*, *16*(3), 367-390. http://dx.doi.org/10.1080/1750399X.2022.2101850

WIRED. (2023, June 19). *Pro Interpreters vs. AI Challenge: Who Translates Faster and Better?* [Video]. Youtube. https://www.youtube.com/watch?v=pwOxlpGYJAY&list=WL&index=90&t=2s

Wu, S. C. (2010). *Assessing simultaneous interpreting. A study on test reliability and examiners' assessment behaviour* [Doctoral dissertation, Newcastle University].

Yandex. (2021, July 7). *Smotrite po-russki: Yandex zapustil zakadrovyi perevod video* [Watch in Russian: Yandex launched voice-over translation for videos]. Yandex Blog. https://yandex.ru/blog/company/smotrite-po-russki-yandeks-zapustil-zakadrovyy-perevod-video

Zhang, A. (2017). Simultaneous interpreting (SI): the holy grail of artificial intelligence–an SI practitioner's perspective. *Lebende Sprachen*, *62*(2), 253. https://doi.org/10.1515/les-2017-0023

Zhang, J., & Qian, H. (2023, September). The impact of machine translation on the translation quality of undergraduate translation students. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track* (pp. 99-108).

Zwischenberger, C. (2010). Quality criteria in simultaneous interpreting: an international vs. a national view. *The Interpreters' Newsletter*, 15, 127-142.

Appendix A

Assessment Scale

Criteria	Band score with description *(Please, read carefully and mark the most appropriate score based on the criteria descriptions)*				Comments / Feedback *(Please, indicate the identified mistakes, inaccuracies, general impression both positive and negative or any other notes emerged during the assessment)*
Content-related criteria					
Consistency with the original	4 The message in the interpretation is the same as in the original speech. Accurate rendition of all the main ideas and details. No unjustified additions, omissions or substitutions.	3 The message in the interpretation is similar to the original speech. The main ideas are conveyed, though minor details are missed. There might be slight unnecessary additions, omissions or substitutions.	2 The message in the interpretation is slightly different from the original speech. Not all the main ideas are conveyed and significant number of details are missed. There are notable unnecessary additions, omissions or substitutions.	1 The message in the interpretation is inconsistent with the original speech. Few main ideas are preserved and the most of the details are missed. There are numerous unjustified and unnecessary additions, omissions or substitutions.	
Logical cohesion	4 Excellent coherence and cohesion of the interpretation. No misleading or redundant elements that might affect the overall logic.	3 The interpretation is coherent and cohesive, though insignificant deviations that might affect the overall logic are present.	2 Adequate coherence and cohesion of the interpretation, but the overall logic is affected due to some misleading or redundant elements.	1 Low coherence and cohesion of the interpretation due to significant deviations that compromise the overall logic.	
Completeness	4 The original content is conveyed in full volume. Accurate rendition of all the numbers, names, titles, as well as minor details.	3 Substantial amount of original content is conveyed, including the most important numbers, names, titles, etc., though minor details are omitted.	2 About a half of the original content is effectively conveyed. Some numbers, names, titles, etc. are omitted. Most minor details are missed.	1 Small portion of the original content is conveyed. Frequent omissions of both major and minor information.	
Form-related criteria					
Terminology	4 Highly accurate interpretation of the specialized terminology, subject matter nuances and subtleties while ensuring consistent equivalence	3 Quite precise interpretation of the specialized terminology. Equivalence between source and target language terms is overall preserved, yet some	2 Contextually appropriate interpretation of the specialized terminology, but the equivalence between the source and target language terms is low.	1 Specialized terminology is interpreted incorrectly. There is no equivalence between source and target language terms and subject matter nuances and	

	between source and target language terms.	subject matter nuances and subtleties might be unaddressed.	Subject matter nuances and subtleties are largely overlooked.	subtleties are disregarded.	
Grammar	4 Interpretation is grammatically correct. All the sentences are complete and clear. Target language grammar structure sounds natural.	3 Interpretation is overall grammatically correct and all the sentences are complete. Minor inaccuracies or slightly unnatural target language grammar structure are possible.	2 Interpretation is acceptable, but contains markable grammatical inaccuracies or incomplete sentences. Some source language grammar structures are preserved in the target language and sound unnatural.	1 Interpretation contains critical grammatical errors and incomplete sentences. Interpretation contains a lot of source language grammar patterns which sound unnatural in target language.	
Style	4 The register is fully appropriate to the context. All the expressions sound completely natural, idiomatic and stylistically relevant.	3 The register is appropriate and overall corresponds to the setting, but few expressions might sound slightly less idiomatic, natural or stylistically consistent.	2 The register does not fully correspond to the context. Significant number of expressions sound unnatural, unidiomatic, resulting in stylistic errors.	1 Inappropriate register and a great number of unnatural and unidiomatic expressions result in a serious mismatch between the source and target language styles.	
Delivery-related criteria					
Fluency of delivery	4 Fluent pace of delivery with seamless diction and articulation. No distinct time lags or (un)filled pauses.	3 Overall fluent pace of delivery with a few cases of slightly unclear diction or articulation. Minimal or no (un)filled pauses. Acceptable time lag.	2 Acceptable pace of delivery, but diction and articulation leave some room for improvement. Markable (un)filled pauses and time lag disrupt the flow.	1 Slow pace of delivery. Unclear diction and articulation throughout the interpretation. Notable time lags, prolonged (un)filled pauses.	
Intonation	4 Very confident and persuasive voice. Highly appropriate use of stress to emphasize important points. Steady rhythm throughout the interpretation.	3 Overall confident and persuasive voice. There are slight inaccuracies in appropriate stress of the key points or minor rhythm intermittency.	2 Little voice confidence with notable signs of hesitation. Key points are stressed inconsistently or inappropriately. Rhythm is somewhat steady, but with occasional disruptions.	1 Unconfident and monotonous voice. Limited or inappropriate use of stress to emphasize key points. Rhythm is irregular, with frequent disruptions in flow.	
Native accent	4 Flawless native-like pronunciation. No signs of a foreign accent.	3 Overall native-like pronunciation with a very subtle sign of a foreign accent.	2 Generally understandable pronunciation, but foreign accent is obvious.	1 Non-native accent significantly deviating from the target language pronunciation norms.	

Appendix B

Transcripts of Video Fragments

Fragment 1:

English:
Thank you, mister president. I have a right speech in front of me now and apologies for its length. As Chair of the Security Council Committee established pursuant to resolution 1373 (2001) concerning counter-terrorism, I have the honour to brief the Council on key aspects of the work of the Committee, supported by the Counter-Terrorism Committee Executive Directorate (CTED). The past year has seen a continued evolution of threats and challenges posed by terrorist activities across the globe. The terrorist threat from Da'esh and Al-Qaida, respectively, remains concentrated in conflict zones, where fragilities are more easily exploited, even though counter-terrorism measures have effectively been mitigating their activities elsewhere. Their respective terrorist activities in the Middle East, Asia and Africa have become more decentralized and are often framed by local dynamics. As technologies have become cheaper and easier to access, terrorist groups have become increasingly adept at exploiting them, including by planning and conducting attacks with unmanned aircraft systems.

Russian:
В качестве Председателя Комитета Совета Безопасности, учрежденного резолюцией 1373 (2001) о борьбе с терроризмом, я имею честь кратко проинформировать Совет о важнейших аспектах работы Комитета, осуществляемой при поддержке Исполнительного директората Контртеррористического комитета (ИДКТК). Прошедший год ознаменовался непрерывным развитием угроз и проблем, исходящих от террористической деятельности по всему миру. Основными источниками террористической угрозы, исходящей от ДАИШ и «Аль-Каиды», по-прежнему являются зоны конфликтов, где легче использовать нестабильность в своих целях, при том, что контртеррористические меры эффективно снижают деятельность этих организаций в других регионах. Их террористическая деятельность на Ближнем Востоке, в Азии и Африке соответственно приобрела более децентрализованный характер и часто определяется развитием событий на местах. По мере удешевления технологий и упрощения к ним доступа террористические группировки стали все более умело их использовать, в том числе планировать и проводить теракты с применением беспилотных авиационных систем.

Fragment 2:

English:
Jared Cohen 142 WPM
It is a privilege to appear before the Security Council today. This meeting follows a crisis that happened 760 days ago, 7,000 miles away. It has largely faded from the headlines, but today is an important reminder of the ongoing plight of the Afghan people and continuing crises around the world. We all remember when Kabul fell to the Taliban. Hundreds of thousands of Afghans were put at risk. Most of them had nowhere to go, and a significant portion of the world's countries closed their doors. But a few countries opened them up, including some represented around this table today. I remember at the time talking to journalists, philanthropists, and a number of world leaders, asking what could be done. Overnight, that group formed what could only be described as a multisectoral, multinational network of goodwill. It was Sheikh Mohammed bin Abdulrahman bin Jassim Al-Thani, then Foreign Minister of Qatar and now Prime Minister, who mobilized his country. Qatar evacuated, resettled, and transported tens of thousands of Afghan women, girls, and civil-society leaders to safety. Sheikh Mohammed and his team did so with extraordinary attention to detail, and his staff worked through

some extreme diplomatic complexities, as well as physical risks on the ground, to achieve those goals. After the evacuations, the then President of Iraq, Burhan Salih, brought hundreds of Afghan university students to the American University of Iraq in Sulaymaniyah. And in bringing those hundreds of students to Iraq, he managed to ensure that while the Taliban occupied their university campus, those students missed not more than two weeks of school. Prime Minister Edi Rama of Albania personally greeted Afghan refugees on the tarmac in Tirana. He transformed some of the country's crown jewels, its precious resort towns, into new homes so that Afghans could rebuild their lives. I want to thank Albania for championing humanitarianism through its presidency of the Council and thereafter. Those are just a few examples of the people and countries that stepped up two years ago. Around the world, countless chief executive officers (CEOs), as well as individuals with no business interests and no connections, covered planes, resettlement, living expenses, and so much more. I think the fact that I last addressed the Security Council more than a year ago as an executive at Google and that today, a year later, I am addressing it as an executive at Goldman Sachs, is just one small point that shows the breadth of private-sector commitment to humanitarian assistance and crisis response. What happened in Afghanistan was nothing short of a tragedy, and in many respects represented humankind at its worst, particularly considering that young girls are now no longer allowed to go to school. But responses such as those I have mentioned were a reminder to me, and I hope to everyone in the Chamber, that such moments can also bring out the best of humankind and what we can all achieve together. That is why I am here today. The task of crisis response is growing much more urgent. The world is facing the greatest moment of geopolitical uncertainty in more than two decades, perhaps since the Cold War. But we cannot let humanitarian crises become even more geopoliticized than they already are. The mission of this body is the maintenance of international peace and security. Many of today's crises would look fairly familiar to the leaders who founded the Council 78 years ago — pandemics, food shortages, floods, and of course the tragic 6.8 magnitude earthquake that occurred in Morocco just last Friday. Other challenges are much newer — cyberattacks, climate change, disinformation and misinformation, and even lethal drugs such as fentanyl. And it is no secret that Russia's war on Ukraine and the competition between great Powers are making the Council's objective that much more challenging.

Russian:
Для меня большая честь выступать сегодня в Совете Безопасности. Это заседание посвящено последствиям кризиса, произошедшего 760 дней назад за 7000 миль отсюда. Средства массовой информации уже практически забыли об этой истории, но сегодняшнее заседание — это важное напоминание о тяжелом положении афганского народа и о продолжающихся кризисах по всему миру. Мы все помним, как талибы взяли Кабул. Под угрозой оказалась жизнь сотен тысяч афганцев. Большинству из них было некуда бежать, и значительная часть стран мира их не принимала. Однако несколько стран открыли им свои двери — в том числе страны, представленные сегодня за этим столом. Помню, как тогда общался с журналистами, благотворителями и рядом мировых лидеров и спрашивал, что можно сделать. В одночасье эта группа сформировала то, что можно назвать многоотраслевой и многонациональной сетью единомышленников, действующих в духе доброй воли. Именно шейх Мухаммед бен Абдель Рахман бен Джасем Аль Тани, который занимал в то время должность министра иностранных дел Катара, а ныне находится на посту премьер-министра, побудил свою страну действовать. Усилиями Катара были эвакуированы, переселены и перевезены в безопасное место десятки тысяч афганских женщин, девушек и лидеров гражданского общества. Шейх Мухаммед и его сотрудники проделали эту работу с исключительным вниманием к деталям, и для достижения этих целей им пришлось преодолеть чрезвычайные дипломатические сложности, а также физические риски на местах. После эвакуации тогдашний президент Ирака Бархам Салех помог сотням афганских студентов попасть в Ирак и возобновить учебу в Американском университете Ирака в Сулеймании. Благодаря его усилиям на протяжении того времени, пока талибы оккупировали комплекс университета, где учились эти студенты, они пропустили лишь две

недели занятий. Премьер-министр Албании Эди Рама лично приветствовал афганских беженцев в аэропорту в Тиране, встретив их прямо возле приземлившегося самолета. Он превратил некоторые из главных достояний страны — драгоценные курортные города Албании — в новые дома для афганцев, чтобы они могли заново построить свою жизнь. Я хочу поблагодарить Албанию за то, что она отстаивает идеи гуманизма во время своего председательства в Совете и в последующий период. Это лишь несколько примеров людей и стран, проявивших солидарность два года назад. По всему миру бесчисленное множество руководителей высшего звена, а также лиц, не имеющих никаких деловых интересов и связей, покрывали расходы афганцев на перелеты, переезд, проживание и многое другое. По моему мнению, тот факт, что в последний раз я выступал в Совете Безопасности более года назад в качестве одного из членов руководства компании 'Гугл', а сегодня, год спустя, я выступаю в качестве одного из членов руководства компании 'Голдман Сакс', является лишь одним из небольших свидетельств того, насколько масштабной является приверженность частного сектора делу оказания гуманитарной помощи и реагирования на кризисы. То, что произошло в Афганистане, — это не что иное, как трагедия, которая во многих отношениях продемонстрировала худшую сторону человечества, особенно с учетом того, что теперь девочкам запрещено посещать школу. Однако ответные меры, подобные тем, о которых я упомянул, послужили для меня и, надеюсь, для всех присутствующих в этом зале напоминанием о том, что такие моменты могут также позволить выявить лучшие стороны человечества и раскрыть наш общий потенциал. Именно поэтому я выступаю здесь сегодня. Вопрос реагирования на кризисы становится все более насущным. Мир переживает период самой серьезной геополитической неопределенности за последние два десятилетия — возможно, со времен холодной войны. Однако мы не можем допустить еще большей геополитизации гуманитарных кризисов. Миссией данного органа является поддержание международного мира и безопасности. Многие из современных кризисов показались бы лидерам, основавшим Совет 78 лет назад, довольно знакомыми: пандемии, нехватка продовольствия, наводнения и, конечно же, трагическое землетрясение силой в 6,8 балла, произошедшее в прошлую пятницу в Марокко. Другие проблемы возникли гораздо позже: кибератаки, изменение климата, распространение дезинформации и ложной информации, а также даже продажа таких смертельно опасных препаратов, как фентанил. Не секрет, что война России на Украине и соперничество великих держав делают достижение целей, стоящих перед Советом, еще более сложной задачей.

Fragment 3:

English:
I would like to express the appreciation of the Office of the United Nations High Commissioner for Refugees (UNHCR) for this opportunity to brief the members of the Security Council and other invited participants on critical issues relating to the protection and human rights of refugees and migrants involved in irregular sea movements from North Africa to Europe.
As a front-line humanitarian agency — and despite our advocacy, assistance and other efforts with States to alleviate human suffering — we continue to bear witness to the tragedies of lives lost at sea and on land routes with no end in sight. Please bear with me, Mr. President, while I provide some numbers that provide a sobering picture of the dimensions of the problem. Between January and August of this year, it is estimated that more than 102,000 refugees and migrants attempted to cross the central Mediterranean Sea to Europe from Tunisia alone, a 260 per cent increase compared to last year, in addition to more than 45,000 people from Libya. Some 31,000 people were rescued at sea or intercepted and disembarked in Tunisia, in addition to 10,600 in Libya. Departures from Algeria were more limited, with almost 4,700 arrivals in Spain by August, an increase of 18 per cent compared to last year. In addition, a total of 3,700 people were rescued or intercepted by the Algerian authorities during the same period, a 68 per cent increase from last year. In total, between January and 24

September some 186,000 people reached Southern Europe by sea — in Italy, Greece, Spain, Cyprus and Malta with the vast majority, more than 130,000 people, arriving in Italy, an increase of 83 per cent compared to the same period in 2022.
By 24 September, more than 2,500 people were accounted dead or missing in 2023 alone. That number represents a two-thirds increase over the total of 1,680 people for the same period in 2022. Lives are also being lost on land, away from public attention. The journey from West Africa, or from the eastern Horn of Africa, to Libya and onward to points of departure on the coast, remains one of the most dangerous in the world. Refugees and migrants travelling along the land routes from sub-Saharan Africa risk death and gross human rights violations at every step. The high departure rates in Tunisia result from perceptions of insecurity among refugee communities, following incidents of racially motivated attacks and hate speech, as well as collective expulsions from Libya and Algeria. That is taking place in a broader context of the deterioration in the security situation of several countries neighbouring North Africa, triggering more secondary movements, with land arrivals and asylum-seeker registrations in Tunisia seeing a marked increase this year.

Russian:
Я хотела бы выразить признательность Управлению Верховного комиссара Организации Объединенных Наций по делам беженцев (УВКБ) за предоставленную мне возможность проинформировать членов Совета Безопасности и других приглашенных участников о крайне важных аспектах, касающихся защиты и прав человека беженцев и мигрантов, вовлеченных в нелегальную переправку по морю из Северной Африки в Европу.
Как работающая на местах гуманитарная структура мы, несмотря на проводимую нами информационную работу и оказываемую помощь и несмотря на другие усилия, предпринимаемые совместно с государствами в целях облегчения человеческих страданий, продолжаем становиться свидетелями все новых случаев трагической гибели людей на море и суше, которые будут продолжать происходить в обозримом будущем.
Г-н Председатель, прошу Вас запастись терпением, пока я приведу некоторые цифры, необходимые для получения объективного представления о масштабах проблемы. Согласно оценкам, в период с января по август этого года из одного лишь Туниса попытку попасть в Европу по маршруту через центральную часть Средиземного моря предприняли более 102 000 беженцев и мигрантов, что на 260 процентов больше, чем в прошлом году, а из Ливии — более 45 000 человек. Спасены в море или же перехвачены и высажены в Тунисе были примерно 31 000 человек, а в Ливии — 10 600 человек. Из Алжира в путь отправилось меньше людей — к августу в Испанию прибыли почти 4700 человек, что на 18 процентов больше, чем в прошлом году. Кроме того, за тот же период алжирскими властями были спасены или перехвачены в общей сложности 3700 человек, что на 68 процентов больше, чем в прошлом году.
Всего за период с января по 24 сентября морским путем в страны Южной Европы — в Италию, Грецию и Испанию, а также на Кипр и Мальту — прибыли около 186 000 человек, причем подавляющее большинство из них — более 130 000 человек — в Италию, что на 83 процента больше, чем за аналогичный период 2022 года.
За 2023 год по состоянию на 24 сентября сочтены погибшими или пропавшими без вести более 2500 человек. Это на две трети больше, чем за тот же период 2022 года, когда соответствующий показатель составил 1680 человек. Люди гибнут и на суше, однако мир этого не замечает. Маршрут из стран Западной Африки или восточной части Африканского Рога в направлении Ливии, а затем расположенных на побережье пунктов отправления остается одним из самых опасных в мире. Беженцы и мигранты, выбирающие сухопутные маршруты, начинающиеся в странах Африки к югу от Сахары, на каждом этапе своего перемещения подвергаются угрозе гибели и грубых нарушений прав человека.
Бросаться в путь из Туниса людей массово заставляет чувство незащищенности, царящее среди беженцев после инцидентов, связанных с нападениями и ненавистническими высказываниями на расовой почве, а также случаями коллективной высылки из Ливии и Алжира. Все это

происходит в условиях ухудшения ситуации в плане безопасности в целом в ряде стран, соседствующих со странами Северной Африки, что провоцирует рост случаев вторичной миграции, в рамках которой в Тунисе в нынешнем году заметно увеличилось число лиц, прибывающих сухопутным путем, а также число случаев регистрации лиц, ищущих убежища

Fragment 4:

English:
Madame President, we thank you for the resumption of the plenary meeting of the 10th emergency special session of the General Assembly today. Ruthless devastation and atrocities in Gaza continue, disregarding all calls and concerns of the International Community, including the United Nations system. So many deaths and damages are unacceptable to any person of conscience. Bangladesh aligns itself with the statements delivered on behalf of the OIC group and on behalf of the Non-Aligned Movement. Madame President, despite adopting a number of resolutions in the 10th emergency special session, we have been witnessing the continuation of one of the horrific massacres in the history of the world. We need to immediately stop the aggression, killing, and all other illegal practices in the occupied Palestinian territory by bringing an end to Israeli occupation and apartheid. Deplorably, after 7 October 2023, more than 44,000 Palestinians have been killed, and Gaza has been destroyed. Unfortunately, Israeli aggression and genocide have been intensified in other parts of the occupied (арт) Palestinian territory, the West Bank, and East Jerusalem, as well as in other countries in the Middle East. The arrangement for a cessation of hostilities to bring an end to the Israeli aggression against Lebanon is an encouraging development in this regard. We join others in calling for the full implementation of resolution 1701 of 2006. Madame President, at this moment, the International Community needs to take urgent, immediate action to end all hostilities and illegal occupation by Israel. We commend the continuous efforts to take forward the discussion towards a peaceful solution to the question of Palestine. In this regard, we welcome the launch of the Global Alliance for the implementation of the two-state solution. We demand full membership of the State of Palestine in the United Nations and call upon the UN Security Council to recommend the same without delay. We also echo the call for the resolution issued by the recent extraordinary Arab Islamic Summit to mobilize international support to suspend Israel's participation in the UN General Assembly. Madame President, without ensuring accountability and without bringing the perpetrators to justice, peace cannot be fully established. We recall the orders of provisional measures and advisory opinion of the International Court of Justice and emphasize ensuring full respect to those orders and their full implementation by immediately stopping the Israeli attack and by full withdrawal of Israeli forces from the occupied Palestinian territory and other occupied Arab territories in the Middle East. We also refer to the indictments by the International Criminal Court and call upon all member states to assist in its implementation. Accountability must be ensured for the perpetrators of mass atrocity crimes. We also welcome the decision to establish under the auspices of the United Nations the international mechanism to assist the investigation and prosecution of persons responsible for the most serious crimes under international law committed in the occupied Palestinian territory. Madame President, Gaza has been demolished, and it could take 350 years for Gaza to rebuild if it remains under blockade. In this destroyed territory, when many dead bodies are under the rubble and the remaining alive people are in dire need of humanitarian assistance and desperately seeking any kind of help to save lives, Israel has been using force, starvation, and forced displacement to continue their aggression against innocent civilians. Regrettably, Israel has been killing and attacking even humanitarian personnel and destroying their facilities. Furthermore, it has passed legislation to stop UNRWA's activities.

Russian:
Г-жа Председатель, мы благодарим Вас за возобновление сегодняшнего пленарного заседания 10-й чрезвычайной специальной сессии Генеральной Ассамблеи. Беспощадные разрушения и бесчинства в Газе продолжаются, игнорируя все призывы и опасения международного

сообщества, включая систему Организации Объединенных Наций. Такое количество погибших и разрушений неприемлемо для любого здравомыслящего человека. Бангладеш присоединяется к заявлениям, сделанным от имени группы ОИК и от имени Движения неприсоединения. Госпожа Председатель, несмотря на принятие ряда резолюций на 10-й чрезвычайной специальной сессии, мы являемся свидетелями продолжения одной из самых ужасных массовых убийств в истории мира. Мы должны немедленно прекратить агрессию, убийства и все другие незаконные действия на оккупированной палестинской территории, положив конец израильской оккупации и апартеиду. К глубокому сожалению, после 7 октября 2023 года более 44 000 палестинцев были убиты, а Газа разрушена. К несчастью, израильская агрессия и геноцид усилились и в других частях оккупированной палестинской территории, на Западном берегу и в Восточном Иерусалиме, а также в других странах Ближнего Востока. Договоренность о прекращении боевых действий, призванная положить конец израильской агрессии против Ливана, вселяет надежду в этом отношении. Мы присоединяемся к другим и призываем к полному выполнению резолюции 1701 от 2006 года. Г-жа Председатель, в данный момент международному сообществу необходимо принять срочные, незамедлительные меры, чтобы положить конец всем военным действиям и незаконной оккупации со стороны Израиля. Мы высоко оцениваем постоянные усилия по продвижению дискуссии в направлении мирного решения вопроса о Палестине. В этой связи мы приветствуем создание Глобального альянса за реализацию двухгосударственного решения. Мы требуем полноправного членства Государства Палестина в Организации Объединенных Наций и призываем Совет Безопасности ООН безотлагательно рекомендовать это. Мы также разделяем призыв к резолюции, принятой на недавнем чрезвычайном арабском исламском саммите, мобилизовать международную поддержку для приостановки участия Израиля в Генеральной Ассамблее ООН. Госпожа Председатель, без обеспечения подотчетности и привлечения виновных к ответственности мир не может быть полностью установлен. Мы напоминаем о постановлениях о временных мерах и консультативном заключении Международного суда и подчеркиваем необходимость обеспечения полного соблюдения этих постановлений и их полного выполнения путем немедленного прекращения израильского нападения и полного вывода израильских сил с оккупированной палестинской территории и других оккупированных арабских территорий на Ближнем Востоке. Мы также ссылаемся на обвинительные заключения Международного уголовного суда и призываем все государства-члены содействовать их выполнению. Необходимо обеспечить привлечение к ответственности лиц, виновных в совершении массовых злодеяний. Мы также приветствуем решение о создании под эгидой Организации Объединенных Наций международного механизма для содействия расследованию и преследованию лиц, ответственных за самые серьезные преступления по международному праву, совершенные на оккупированной палестинской территории. Мадам Председатель, Газа разрушена, и на ее восстановление может уйти 350 лет, если она останется в блокаде. На этой разрушенной территории, когда под обломками лежит множество мертвых тел, а оставшиеся в живых люди остро нуждаются в гуманитарной помощи и отчаянно ищут любую помощь для спасения жизни, Израиль использует силу, голод и насильственное перемещение, чтобы продолжать свою агрессию против невинных гражданских лиц. К сожалению, Израиль убивает и нападает даже на сотрудников гуманитарных организаций и разрушает их объекты. Кроме того, он принял закон о прекращении деятельности БАПОР.

Fragment 5:

English:
Muchas gracias, señora presidenta. Allow me to express our sincere thanks to you and your team for your outstanding leadership in guiding this committee to a successful and early completion. We also thank the Bureau and the Fifth Committee Secretariat for their support in our work. Additionally, we

deeply appreciate the hard work of all colleagues, especially Uganda and our brothers and sisters from the Group of Seventy-Seven and China.
Thanks to your efforts, this session has delivered many positive outcomes. Vietnam welcomes the outcomes on the RC system, the resolution for the implementation of the Pact for the Future, the policy guidance on strengthening the effectiveness of RPTC, and the Development Account. This session holds special significance for Vietnam, as the scale assessment adopted by the committee marks the highest increase in our membership history and represents one of the most significant percentage increases for any country in the upcoming cycle. As an active and responsible member of the United Nations, Vietnam reaffirms its commitment to shouldering the heightened responsibilities that accompany our new assessment rate.
During this session, I had the honor of facilitating the adoption of the Pension Fund resolution, which was successfully passed by consensus. I would like to take this opportunity to express my heartfelt thanks to my colleagues, brothers, and sisters for your dedication and support.
Once again, thank you, Madame Chair, and I wish all colleagues a Merry Christmas and a Happy New Year. Thank you.

Russian:
Большое спасибо, госпожа президент. Позвольте мне выразить вам и вашей команде нашу искреннюю благодарность за выдающееся руководство, которое привело этот комитет к успешному и скорейшему завершению. Мы также благодарим Бюро и Секретариат Пятого комитета за их поддержку в нашей работе. Кроме того, мы глубоко ценим напряженную работу всех коллег, особенно Уганды и наших братьев и сестер из Группы семидесяти семи и Китая. Благодаря вашим усилиям эта сессия принесла много положительных результатов. Вьетнам приветствует итоговые документы по системе КР, резолюцию по реализации Пакта на будущее, политическое руководство по повышению эффективности РПТС (Регулярной программы технического сотрудничества) и Счет развития. Эта сессия имеет особое значение для Вьетнама, поскольку оценка по шкале, принятая комитетом, является самым высоким увеличением за всю историю нашего членства и представляет собой одно из самых значительных процентных увеличений для любой страны в предстоящем цикле. Будучи активным и ответственным членом Организации Объединенных Наций, Вьетнам подтверждает свою готовность нести повышенную ответственность, которая сопровождает нашу новую ставку взноса. В ходе этой сессии мне выпала честь содействовать принятию резолюции по Пенсионному фонду, которая была успешно принята консенсусом. Пользуясь возможностью, я хотел бы выразить искреннюю благодарность своим коллегам, братьям и сестрам за вашу преданность и поддержку. Еще раз благодарю вас, госпожа Председатель, и желаю всем коллегам счастливого Рождества и Нового года. Спасибо.

Fragment 6:

English:
I thank you, Mr. Chair, for giving me the floor. Ladies and gentlemen, distinguished delegates, as we all know, the world drug problem is complicated and multifaceted, which needs a holistic and balanced approach to address it. The Single Convention on Narcotic Drugs of 1961, as amended by the Protocol of 1972, the Convention on Psychotropic Substances of 1971, the Convention of 1988 against illicit trafficking in narcotic drugs and psychotropic substances, the Political Declaration of 2009, the 2016 UNGA outcome on the world drug problem, and the Ministerial Declaration of 2019 constitute the cornerstone of our comprehensive work to counter the world drug problem. I take this opportunity to reiterate Algeria's full commitment to these international drug control instruments as well as its full support for the CND and INCB efforts to tackle the world drug problem. To efficiently face this serious problem, Algeria has taken concrete measures. Last year, my country enacted a new law

amending its domestic legislation against illicit trafficking in narcotic drugs and psychotropic substances by introducing a national classification of narcotic drugs and psychotropic substances, which will be in addition to the international schedule of controlled substances. This law aims to counter the proliferation of drug trafficking in new forms, like medications diverted from their medical use for addiction purposes among teenagers and young people, especially pregabalin and tramadol, without prejudice to the right of availability and access to these controlled substances so that they can relieve the pain of suffering people, in full accordance with international conventions that give state parties the possibility to take measures to protect public health, national security, and welfare. On the other side, the Algerian health authorities are working on the treatment and rehabilitation level by encouraging addicted people to start detoxification treatment and exempting them from any criminal prosecution under the condition of completing that detoxification treatment. We now have 48 medium treatment centers and five big ones. In addition to these existing centers, within the President's initiative of the pledge for action in this regard, Algeria has pledged to build four new modern, highly equipped treatment centers. We have finished the first step of this center establishment by choosing four cities to host them, with a view to adding a fifth one. Mr. Chair, Algeria reiterates its deep concern about the legalization of recreational cannabis. Even though very few countries currently allow this nonmedical and nonscientific use, the number of these countries may increase over the upcoming years. Cannabis has proved to be addictive and dangerous, so we think that allowing the recreational use of cannabis would be a dangerous step that would undoubtedly affect public health, welfare, and even national public security. Everyone has the right to enjoy the highest attainable standard of physical and mental health. To conclude, Algeria calls on UN member states and stakeholders to work closely together to address and counter the world drug problem. Thank you, Mr. Chair.

Russian:
Я благодарю вас, г-н Председатель, за предоставленное мне слово. Дамы и господа, уважаемые делегаты, как всем нам известно, мировая проблема наркотиков сложна и многогранна, что требует целостного и сбалансированного подхода к ее решению. Единая конвенция о наркотических средствах 1961 года с поправками, внесенными Протоколом 1972 года, Конвенция о психотропных веществах 1971 года, Конвенция 1988 года о борьбе против незаконного оборота наркотических средств и психотропных веществ, Политическая декларация 2009 года, итоговый документ ГА ООН 2016 года по мировой проблеме наркотиков и Министерская декларация 2019 года являются краеугольным камнем нашей комплексной работы по противодействию мировой проблеме наркотиков. Пользуясь этой возможностью, я хотел бы подтвердить полную приверженность Алжира этим международным документам по контролю над наркотиками, а также его полную поддержку усилий НКД и МККН по решению мировой проблемы наркотиков. Чтобы эффективно противостоять этой серьезной проблеме, Алжир принимает конкретные меры. В прошлом году в нашей стране был принят новый закон, вносящий изменения в национальное законодательство по борьбе с незаконным оборотом наркотических средств и психотропных веществ путем введения национальной классификации наркотических средств и психотропных веществ, которая будет дополнять международный список контролируемых веществ. Этот закон направлен на противодействие распространению наркоторговли в новых формах, таких как лекарственные препараты, отвлекаемые от медицинского использования в целях наркомании среди подростков и молодежи, особенно прегабалин и трамадол, без ущерба для права на наличие и доступ к этим контролируемым веществам, чтобы они могли облегчить боль страдающих людей, в полном соответствии с международными конвенциями, которые предоставляют государствам-участникам возможность принимать меры по защите общественного здоровья, национальной безопасности и благосостояния. С другой стороны, алжирские органы здравоохранения работают над лечением и реабилитацией, поощряя наркозависимых людей начать детоксикационное лечение и освобождая их от уголовного преследования при условии завершения детоксикационного лечения. В настоящее время у нас есть 48 средних и пять крупных лечебных центров. В

дополнение к этим существующим центрам, в рамках президентской инициативы по принятию обязательств в этой области, Алжир обязался построить четыре новых современных, хорошо оборудованных лечебных центра. Мы завершили первый этап создания этих центров, выбрав четыре города для их размещения и планируя добавить пятый. Г-н Председатель, Алжир вновь выражает глубокую озабоченность по поводу легализации каннабиса в рекреационных целях. Несмотря на то, что в настоящее время очень немногие страны разрешают это немедицинское и ненаучное употребление, в ближайшие годы число таких стран может увеличиться. Каннабис доказал, что вызывает привыкание и опасен, поэтому мы считаем, что разрешение рекреационного использования каннабиса было бы опасным шагом, который, несомненно, повлияет на здоровье, благосостояние и даже национальную общественную безопасность. Каждый человек имеет право на наивысший достижимый уровень физического и психического здоровья. В заключение Алжир призывает государства-члены ООН и заинтересованные стороны к тесному сотрудничеству для решения и противодействия мировой проблеме наркотиков. Спасибо, господин Председатель.

Appendix C

Informed Consent Form

Participant's nickname: _____

Title of study
AI vs Human Simultaneous Interpreting: Quality Assessment

Principal investigator
Kristina Vesselskaya
87719754686
k_vesselskaya@kazguu.kz

Purpose of study
The purpose of the research is s to evaluate and analyze the quality of simultaneous interpreting produced by artificial intelligence in comparison to the traditional human interpreting. It is aimed at examining the outputs and revealing inaccuracies, limitations and common error patterns of machine simultaneous interpreting.

Data collection procedure
After signing this consent form you will be provided with the link to the Google Drive folder. The folder contains six original audio fragments of different speeches. Each original audio fragment has two versions of simultaneous interpreting. The first is performed by a professional conference interpreter while the second version is provided by artificial intelligence. You will be asked to carefully listen to both versions of the provided simultaneous interpretations of each original fragment and to assess the results. To evaluate quality of the interpretations you will need to use the assessment scale uploaded in the same folder. For your convivence, you will receive the printed assessment scales and fill it out by hand. Alternatively, you might download the assessment scale from the folder and fill it out in digital format. Following the provided criteria descriptions in the assessment scale, you will need to grade the interpretations. When listening to and evaluating the audio fragments, you are also asked to take necessary notes in the separate column of the provided scale. These notes may include identified strengths, mistakes, or inaccuracies, and your general feedback. The whole assessment process might take up to 40-60 minutes of your time. At least one week will be provided to you to finalize the assessment process. You might contact the principal investigator at any time and ask for additional clarifications. The filled assessment scales will either be collected in person or sent via e-mail.

Confidentiality

You can be assured that your name and other personal information will remain confidential. There are no known risks if you decide to participate in this research study, nor are there any costs for participating in the study. Your collected assessments will not be additionally evaluated or judged by the principal investigator or any other party. All the responses you give will be kept strictly confidential. The information you provide will be used anonymously for internal publication for Kristina Vesselskaya's MA thesis and might be submitted for publishing in academic journals and conferences. If you have any comments or concerns about the ethics or procedures involved in this study, you can contact research supervisor, Olga Bainova, at the e-mail address o_bainova@kazguu.kz or Research and Ethics Committee of the School of Liberal Arts at the e-mail address rec_sla@kazguu.kz

Voluntary participation

Your participation in the assessment procedure is voluntary. It is entirely up to you whether or not to participate in this study. If you choose to participate, you will need to sign this informed consent form. After signing, you may withdraw at any moment and without explanation. Withdrawing from this study will have no effect on your relationship with the principal investigator, if you have one. If you withdraw before the data collection is finished, your assessment will either be returned to you or disposed.

Consent

I have read and understand the provided information and have had the opportunity to ask questions. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason. I understand that I will be given a copy of this consent form. I voluntarily agree to take part in this study.

Participant's signature _____ **Date** _____

Researcher's signature _____ **Date** _____

Appendix D

Example of Filled Assessment Scale

Human Interpreting

Criteria	Band score with description *(Please, read carefully and mark the most appropriate score based on the criteria descriptions)*				Comments / Feedback *(Please, indicate the identified mistakes, inaccuracies, general impression both positive and negative or any other notes emerged during the assessment)*
Content-related criteria					
Consistency with the original	4 The message in the interpretation is the same as in the original speech. Accurate rendition of all the main ideas and details. No unjustified additions, omissions or substitutions.	3 The message in the interpretation is similar to the original speech. The main ideas are conveyed, though minor details are missed. There might be slight unnecessary additions, omissions or substitutions.	2 The message in the interpretation is slightly different from the original speech. Not all the main ideas are conveyed and significant number of details are missed. There are notable unnecessary additions, omissions or substitutions.	1 The message in the interpretation is inconsistent with the original speech. Few main ideas are preserved and the most of the details are missed. There are numerous unjustified and unnecessary additions, omissions or substitutions.	Основной посыл оригинала полностью сохранен в переводе.
Logical cohesion	4 Excellent coherence and cohesion of the interpretation. No misleading or redundant elements that might affect the overall logic.	3 The interpretation is coherent and cohesive, though insignificant deviations that might affect the overall logic are present.	2 Adequate coherence and cohesion of the interpretation, but the overall logic is affected due to some misleading or redundant elements.	1 Low coherence and cohesion of the interpretation due to significant deviations that compromise the overall logic.	несмотря на высокую скорость и слышимые затруднения для переводчика, SPEED CHALLENGES ему удалось сохранить логически верно построенную структуру, благодаря используемым стратегиям перевода (генерализация, сокращения, компрессия) OPTIMIZATION TECHNIQUES
Completeness	4 The original content is conveyed in full volume. Accurate rendition of all the numbers, names, titles, as well as minor details.	3 Substantial amount of original content is conveyed, including the most important numbers, names, titles, etc., though minor details are omitted.	2 About a half of the original content is effectively conveyed. Some numbers, names, titles, etc. are omitted. Most minor details are missed.	1 Small portion of the original content is conveyed. Frequent omissions of both major and minor information.	пропущены пару деталей в начале, когда спикер извинялась за длину своего доклада. WORSE COMPLETENESS IN HUMAN OUTPUTS
Form-related criteria					

Terminology	4 Highly accurate interpretation of the specialized terminology, subject matter nuances and subtleties while ensuring consistent equivalence between source and target language terms.	3 Quite precise interpretation of the specialized terminology. Equivalence between source and target language terms is overall preserved, yet some subject matter nuances and subtleties might be unaddressed.	2 Contextually appropriate interpretation of the specialized terminology, but the equivalence between the source and target language terms is low. Subject matter nuances and subtleties are largely overlooked.	1 Specialized terminology is interpreted incorrectly. There is no equivalence between source and target language terms and subject matter nuances and subtleties are disregarded.	Использовались аббревиатуры и акронимы вместо полных расшифровок терминов для экономии времени – что очень хорошо. OPTIMIZATION TECHNIQUES Правильные переводы терминологии (ДАИШ=ИГИЛ, ДКТК – CTED). Кратко перевел «беспилотник» вместо «беспилотные авиационные системы» для оптимизации времени. OPTIMIZATION TECHNIQUES
Grammar	4 Interpretation is grammatically correct. All the sentences are complete and clear. Target language grammar structure sounds natural.	3 Interpretation is overall grammatically correct and all the sentences are complete. Minor inaccuracies or slightly unnatural target language grammar structure are possible.	2 Interpretation is acceptable, but contains markable grammatical inaccuracies or incomplete sentences. Some source language grammar structures are preserved in the target language and sound unnatural.	1 Interpretation contains critical grammatical errors and incomplete sentences. Interpretation contains a lot of source language grammar patterns which sound unnatural in target language.	С грамматической точки зрения все предложения построены верно, звучат натурально, хоть и были очень длинные предложения в переводе. Возможно стоило их разбить на несколько более коротких, но учитывая скорость речи, перевод все равно выполнен грамматически правильно. NATURAL DELIVERY LONG SENTENCES, NEED FOR SALAMI TECHNIQUE
Style	4 The register is fully appropriate to the context. All the expressions sound completely natural, idiomatic and stylistically relevant.	3 The register is appropriate and overall corresponds to the setting, but few expressions might sound slightly less idiomatic, natural or stylistically consistent.	2 The register does not fully correspond to the context. Significant number of expressions sound unnatural, unidiomatic, resulting in stylistic errors.	1 Inappropriate register and a great number of unnatural and unidiomatic expressions result in a serious mismatch between the source and target language styles.	
Delivery-related criteria					
Fluency of delivery	4 Fluent pace of delivery with seamless diction and articulation. No distinct time lags or (un)filled pauses.	3 Overall fluent pace of delivery with a few cases of slightly unclear diction or articulation. Minimal or no (un)filled pauses.	2 Acceptable pace of delivery, but diction and articulation leave some room for improvement. Markable (un)filled pauses	1 Slow pace of delivery. Unclear diction and articulation throughout the interpretation. Notable time lags,	Была заполненная пауза. Проглотил слово «децентрализованный». Иногда слишком торопится, чтобы успеть за спикером, а иногда наоборот замедляется и заметен этот переход. FILLED PAUSES

		Acceptable time lag.	and time lag disrupt the flow.	prolonged (un)filled pauses.	ARTICULATION ISSUES UNSMOOTH PACE
Intonation	4 Very confident and persuasive voice. Highly appropriate use of stress to emphasize important points. Steady rhythm throughout the interpretation.	3 Overall confident and persuasive voice. There are slight inaccuracies in appropriate stress of the key points or minor rhythm intermittency.	2 Little voice confidence with notable signs of hesitation. Key points are stressed inconsistently or inappropriately. Rhythm is somewhat steady, but with occasional disruptions.	1 Unconfident and monotonous voice. Limited or inappropriate use of stress to emphasize key points. Rhythm is irregular, with frequent disruptions in flow.	
Native accent	4 Flawless native-like pronunciation. No signs of a foreign accent.	3 Overall native-like pronunciation with a very subtle sign of a foreign accent.	2 Generally understandable pronunciation, but foreign accent is obvious.	1 Non-native accent significantly deviating from the target language pronunciation norms.	

Machine Interpreting

Criteria	**Band score with description** *(Please, read carefully and mark the most appropriate score based on the criteria descriptions)*				**Comments / Feedback** *(Please, indicate the identified mistakes, inaccuracies, general impression both positive and negative or any other notes emerged during the assessment)*
Content-related criteria					
Consistency with the original	4 The message in the interpretation is the same as in the original speech. Accurate rendition of all the main ideas and details. No unjustified additions, omissions or substitutions.	3 The message in the interpretation is similar to the original speech. The main ideas are conveyed, though minor details are missed. There might be slight unnecessary additions, omissions or substitutions.	2 The message in the interpretation is slightly different from the original speech. Not all the main ideas are conveyed and significant number of details are missed. There are notable unnecessary additions, omissions or substitutions.	1 The message in the interpretation is inconsistent with the original speech. Few main ideas are preserved and the most of the details are missed. There are numerous unjustified and unnecessary additions, omissions or substitutions.	All details preserved despite high speed of delivery unlike in human interpretation AI HANDLED SPEED BETTER
Logical cohesion	4 Excellent coherence and cohesion of the interpretation. No misleading or redundant	3 The interpretation is coherent and cohesive, though insignificant deviations that might affect the	2 Adequate coherence and cohesion of the interpretation, but the overall logic is affected due to	1 Low coherence and cohesion of the interpretation due to significant deviations that	

	elements that might affect the overall logic.	overall logic are present.	some misleading or redundant elements.	compromise the overall logic.	
Completeness	4 The original content is conveyed in full volume. Accurate rendition of all the numbers, names, titles, as well as minor details.	3 Substantial amount of original content is conveyed, including the most important numbers, names, titles, etc., though minor details are omitted.	2 About a half of the original content is effectively conveyed. Some numbers, names, titles, etc. are omitted. Most minor details are missed.	1 Small portion of the original content is conveyed. Frequent omissions of both major and minor information.	ИИ понял, что CTED (ДКТК) был расшифрован спикером и не стал переводить аббревиатуру, что тоже в целом правильно и нельзя назвать значительным пропуском. MINOR INACCURACY
Form-related criteria					
Terminology	4 Highly accurate interpretation of the specialized terminology, subject matter nuances and subtleties while ensuring consistent equivalence between source and target language terms.	3 Quite precise interpretation of the specialized terminology. Equivalence between source and target language terms is overall preserved, yet some subject matter nuances and subtleties might be unaddressed.	2 Contextually appropriate interpretation of the specialized terminology, but the equivalence between the source and target language terms is low. Subject matter nuances and subtleties are largely overlooked.	1 Specialized terminology is interpreted incorrectly. There is no equivalence between source and target language terms and subject matter nuances and subtleties are disregarded.	Вместо ИГИЛ оставил транслитерацию ДАИШ, что тоже правильно. Переводил название комитетов правильно, не используя сокращения для оптимизации времени, как человек. NO COGNITIVE CONSTRAINTS HUMAN FACTOR
Grammar	4 Interpretation is grammatically correct. All the sentences are complete and clear. Target language grammar structure sounds natural.	3 Interpretation is overall grammatically correct and all the sentences are complete. Minor inaccuracies or slightly unnatural target language grammar structure are possible.	2 Interpretation is acceptable, but contains markable grammatical inaccuracies or incomplete sentences. Some source language grammar structures are preserved in the target language and sound unnatural.	1 Interpretation contains critical grammatical errors and incomplete sentences. Interpretation contains a lot of source language grammar patterns which sound unnatural in target language.	Все предложения и грамматические структуры построены максимально правильно и звучат естественно. NATURAL GRAMMATICAL CONSTRUCTIONS
Style	4 The register is fully appropriate to the context. All the expressions sound completely natural, idiomatic and stylistically relevant.	3 The register is appropriate and overall corresponds to the setting, but few expressions might sound slightly less idiomatic, natural or stylistically consistent.	2 The register does not fully correspond to the context. Significant number of expressions sound unnatural, unidiomatic, resulting in stylistic errors.	1 Inappropriate register and a great number of unnatural and unidiomatic expressions result in a serious mismatch between the source and target language styles.	
Delivery-related criteria					
Fluency of delivery	4 Fluent pace of delivery with seamless diction	3 Overall fluent pace of delivery with a few cases	2 Acceptable pace of delivery, but diction and	1 Slow pace of delivery. Unclear diction and	Robotic and unnatural voice of delivery. NATURALNESS ISSUES

	and articulation. No distinct time lags or (un)filled pauses.	of slightly unclear diction or articulation. Minimal or no (un)filled pauses. Acceptable time lag.	articulation leave some room for improvement. Markable (un)filled pauses and time lag disrupt the flow.	articulation throughout the interpretation. Notable time lags, prolonged (un)filled pauses.	
Intonation	4 Very confident and persuasive voice. Highly appropriate use of stress to emphasize important points. Steady rhythm throughout the interpretation.	3 Overall confident and persuasive voice. There are slight inaccuracies in appropriate stress of the key points or minor rhythm intermittency.	2 Little voice confidence with notable signs of hesitation. Key points are stressed inconsistently or inappropriately. Rhythm is somewhat steady, but with occasional disruptions.	1 Unconfident and monotonous voice. Limited or inappropriate use of stress to emphasize key points. Rhythm is irregular, with frequent disruptions in flow.	Очень хорошая интонация и ритм перевода, в отличии от человека, где были короткие паузы и переменчивая скорость подачи, не «проглатывал» так как местами не успевал. ИИ в отличии от человека не испытывал проблем со скоростью и не оптимизировал структуры, чтобы успеть все проговорить. BETTER ARTICULATION NO COGNITIVE CONSTRAINTS NO OPTIMIZATION NEEDED HANDLED SPEED BETTER
Native accent	4 Flawless native-like pronunciation. No signs of a foreign accent.	3 Overall native-like pronunciation with a very subtle sign of a foreign accent.	2 Generally understandable pronunciation, but foreign accent is obvious.	1 Non-native accent significantly deviating from the target language pronunciation norms.	

Appendix E

Screenshot from Canva Board with Thematic Analysis