Comparative Analysis of Machine Interpreting and Human Interpreting: Insights into Consecutive Interpreting Teaching

Ying Cao

Wuhan Polytechnic University China, Huanhu Donglu 36, Dongxihu District, Wuhan, Hubei, China

E-mail: yingcao@whpu.edu.cn

Abstract. The fast development of artificial intelligence (AI) in the areas of automatic speech recognition and machine translation, in recent years, drives the application of Computer-aided interpretation (CAI) tools in various interpreting settings, such as consecutive, simultaneous and remote interpreting.

This paper is devoted to a comparative analysis between machine interpreting and human interpreting in the mode of English/Chinese consecutive interpreting. It is focused on a competence-based approach by adopting an empirical design and using mainly quantitative methods based on 300 minutes' audio record of interpreting practices. This study adopts the assessment criteria of interpreter [1] to compare the performances of Machine Interpreting (MI) and five MTI students in mock consecutive interpreting training (two passages in each language direction, and 10,521 words in total) by analysing their accuracy, the use of interpreting skills, and overall quality (i.e. fluency, logical cohesion). The author concludes that machine interpreter excels human interpreter in the areas of lexical chiice, accuracy and overall quality, especially in the direction of L1-L2. However, human interpreters show a stronger competence in using interpreting skills and dealing with the extralinguistic information, such as abbreviation and idiomatic expressions.

Keywords: Machine interpreting, Human interpreting, Deep-learning, Consecutive interpreting teaching

1 Introduction

Technologies have been widely applied to the professional practice of interpreting since 1920s [2]. More recently, however, the use of technology in interpreting has evolved and diversified at a much faster pace than in past decades, and technologies have been adopted to support or enhance interpreters' preparation, performance and workflow, leading to technology-supported interpreting [3]. Although doubted views that technologies are designed to replace human interpreters, leading to technology-generated or machine interpreting, keep resurfacing, some scholars have verified that machine inter

M. F. b. Sedon et al. (eds.), *Proceedings of the 2024 3rd International Conference on Social Sciences and Humanities and Arts (SSHA 2024)*, Advances in Social Science, Education and Humanities Research 851,
https://doi.org/10.2991/978-2-38476-259-0_76

preting's application to situations in which highly accurate professional language mediation is required remains a daunting challenge, and more efforts need to be devoted to improve machine interpreting [4].

Despites the divergent views on the (ir)replaceability of machine interpreting towards human interpreting, it is admitted that the proper use of AI in university translation and interpreting teaching can greatly motivate students, help them to think creatively, and inspire them to actively participate in mock practices as translators and interpreters [5]. Therefore, university teachers show an increasing willingness in applying technologies in translating and interpreting teaching. However, there remains some problems, especially in interpreting teaching. The lack of machine translation courses, the inadequate and improper use of technologies in training, the outdated training materials and training mode, as well as the lack of teachers with real interpreting experiences, impedes the improvement of students' interpreting performance [6]. In this vein, this study aims to conduct a comparative analysis of machine interpreting and human interpreting by drawing the merits of both machine interpreter and human interpreter, so as to identify a more effective and feasible training mode to improve the performance of student interpreters.

2 Artificial intelligence and Machine interpreting

Artificial intelligence (AI) is a field of computer science concentrating on building smart machines capable of performing tasks that are typically related with some form of intelligence when performed by humans [3]. Recent development in AI has been made possible by improvements in machine learning and deep learning, a field of AI that aims to teach machines how to actively learn and effectively carry out orders without being explicitly programmed to do so [7]. Deep learning has been widely applied in machine translation and interpreting.

Machine interpreting (MI), also known as speech-to-text or speech-to-speech translation, is the process of transferring a spoken text from one language into written or spoken text to another language in real-time, translating the original while it is still unfolding [7]. MI systems can be effectively applied to various live communicative settings, such as institutional events, lectures, conferences, negotiations, and so on. In addition, MI can make multilingual content accessible for receiver in real-time, thus increasing receives' feeling of inclusion when human services for language transference are not available, such as live interlingual subtitling or conference interpreting [8]. MI, especially the speech-to text transference, can also be integrated into CAI tools and be used to assist professional interpreters, for example to offer suggestions in real-time, and to train student interpreting as a mean to improve their interpreting performance [7].

Although MI is not a new phenomenon, it only receives scholarly attention in recent years as a branch of MT, with conferences, research institutions and major technology company, such as Google, Meta, Apple, Baidu, Tencent and so on working on it [5,9]. Many challenges related to the high complexity that is typical of the spoken language, both in the professional and in the everyday context, remain to be tackled. Language use and spoken language translation are, however, complex phenomena. While machines seem not to be able to perform interpretation from a communicative perspective, limited types of assessments are possible. For example, it is conceivable to verify the adherence to a specific terminology, the accuracy of number rendition, or the omissions of essential information of the source speech in a completely automatic way, generating a report for interpreters to reflect their own performance [10].

3 Methodology

3.1 The characteristics of machine interpreting and human interpreting

MI stems from Neural Machine Translation (NMT), and it relies on deep neural network to achieve the end-to-end overall translation mode. Through MI, translation language could become more natural and readable, and therefore, largely improves the accuracy and fluency of translation [11]. NMT uses machine deep-learning technology, which teaches computers to process data in a way that is similar with human brains, and it can recognize complex patterns in pictures, text, sounds, and other multimodal information to produce accurate predictions [12]. Basically, NMT is an encoder-decoder system, in which encoder encodes the source language sequence and extracts information, and then converts information to the target language, through decoder, so as to complete the process of translation. Based on NMT technology, MI excels human interpreter in ways of extracting vocabularies, constructing sentences, predicting the next sequence, and so on. However, it is undoubtable that the machine cannot process information in the same way as humans, lacking intentionality, and therefore, presents a weak handling capacity in integrating the contextual and dicoursal information when interpreting texts.

Although MI gains an increasing interest from scholars, industries and higher education institutions, human interpreting (HI) is still irreplaceable at the current stage. HI mainly take advantage of subjective initiative to convert one language to another language, so the distinct feature of HI is that human can analyze the deep-layered meaning of sentences, and constantly improve the grammar application, sentence fluency and syntactic structure with regard to the situated context. In this vein, HI presents a higher level of readability and a smoother transition between sentences.

In addition, HI shows a stronger capacity in processing social and cultural factors between two languages than MI, since the latter is restricted by computing algorithm and deep learning procedure. Admittedly, the shortcomings of HI are also obvious. To name a few, human interpreters can hardly maintain their optimal pace, speed, tone of voice and the volume of information they need to process due to the limit of their energy [13]. In this case, it is infeasible for HI to keep their performance and interpreting quality at a same high level throughout the whole meeting or interpreting task, especially when the meeting lasts for hours.

Considering the potential of both HI and MI, this study intends to conduct a comparative analysis of the performance of HI and MI in consecutive interpreting mode so as to shed lights on the improvements of MI and also the training mode of student interpreters.

3.2 Corpus and comparative interpreting study

Participants.

This study is a pilot study that look into new interpreting training modes assisted by technology under a large research project. Noticeably, all Ethical clarifications have been made before the research. Considering the explorative nature of the current study, only five participants were recruited at this stage. The participants were Master of Translation and Interpreting (MTI) students, majoring in interpreting in a key university under 211 progamme in China, who had finished at least five semesters of interpreting courses. All were females aged around 22, with a language combination of Chinese (L1) and English (L2). All of them have achieved TEM-8 certificates, and two received CATTI English-Chinese Translation level two certificates, and two got CATTI English-Chinese Interpreting level three certificates. All the participants are required to sign a consent form before they join the project.

Apparatus.

The MI softwares used in this study were Tencent Translation (i.e. Fanyi) and Dear Translate. The former one is developed by Tencent and the latter one is from NetEase, and both of them are the pioneers in the AI translation industry. Tencent Translation was firstly introduced in 2016, and released its simultaneous interpreting function in 2018, and provided AI interpreting service for 2018 Boao Forum for Asia. Dear Translate was released by NetEase on 2019, and was one of the first educational applications recognised by the Ministry of Education of China. Both these two MI systems adopted the most advanced YNMT technology, which boast a higher level of accuracy than traditional MI technologies, and provided interpreting services in more than 100 languages.

The criteria for selecting the MI system are as follows. First, the MI systems have the capability to process both Chinese and English to avoid emerging extra variables by using different systems for each language. Second, the systems' internet connection and performance should be smooth and stable during the training and experiments. Third, it should be achievable and accessible for most student users. Some MI service providers, such as google, and apple translation, are need to be accessed through a VPN for participants. However, *Tencent Translation* and *Dear Translate* can be easily accessed achieved for all the participants.

Procedure.

After eight weeks of training (1,440 minutes in total) on conference consecutive interpreting course, the participants officially joined the pilot experiment during which they were required to finish four interpreting tasks (two English speeches and two Chinese speeches). Table 1 presents the key features of the speeches, namely, English speeches last for about 5 minutes (approximately 700 words), while Chinese speeches last for a slightly longer, 5.5 minutes (approximately 850 words). Detailed instruction of the experiment will be given to the participants before they starting the recording. In what follows, participants performed the interpreting tasks, which is set as a consecutive mode, and the order of the recording is L2-L1 (general), L1-L2 (general), L1-L2 (technical), and L2-L1 (technical). It is noted that all the participants have been given the theme, the glossary, and background information of the speeches one day before the official recoding. When it comes to the recording session, five minutes are given for them to prepare before they interpreting. After all the tasks were completed, the participant were interviewed for their self-evaluations of interpreting performance. An average experiment session lasted about 1 hours.

Table 1. Summary of the speeches

No.	Language Direction	Level of technicality	Duration (m)	Word count	Speed (wpm)
1	English L2–L1	General	2.23	324	145.29
2	Chinese L1-L2	General	3.17	461	146.14
3	English L2–L1	Technical	2.78	392	141.01
4	Chinese L1-L2	Technical	2.39	351	146.86

Considering the wide application of MI in relevant fields, such as simultaneous interpreting in conference under a general theme, this study attempts to conduct a comparative analysis of the output of MI and HI under the guidance of Effort Model [14].

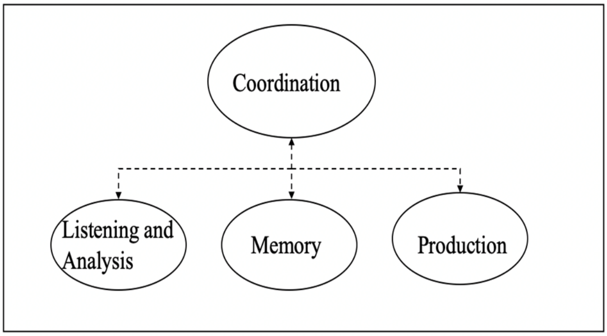

Fig. 1. : Effort Model of Conventional CI

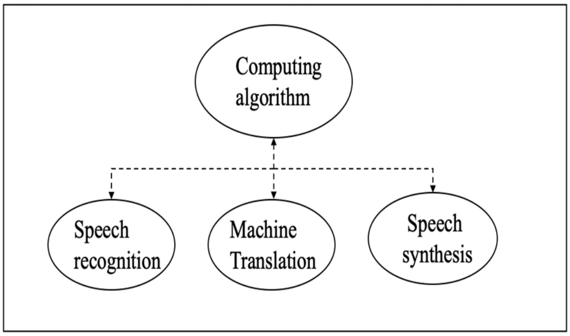

Fig. 2. : Workflow of MI

According to the Effort Model, CI consists four steps, namely listening and analysis, note-taking, short-term memory operations, and coordination (as depicted in Figure 1), while a typical workflow of MI includes the sub-processes of speech recognition, machine translation, speech synthesis (as depicted in Figure 2). In the process of machine translation, various sub-procedures may involve, depending on the use of translation technologies.

4 Data and analysis

Both HI and MI's interpreting process were recorded for comparative analysis. There were a total of 80 renditions, including HI (five participants) & MI × two directions × two levels of technicality). The audio recordings were rated by three advanced interpreters in line with the specific research needs. A scale rating mode developed by Han (2015) was used. A group of three advanced interpreters (lecturers teaching interpreting, who are also causal interpreters with at least five years working experience) participated in a half-day training session before they starting the rating (two interpreters rated HI and MI respectively, and the last interpreter rated 50% HI and MI randomly to ensure the error range is reasonable), following the procedures specified in previous interpreting comparative analysis studies [1,15].

Given that one of the main strengths of MI is supposed to be fluency (in the aspects of pauses, tone of voice, pace and so on) than HI because the output speech is automatically generated by AI, the current study focuses the accuracy and the overall quality of speeches, including the use of interpreting skills, and the logical coherence, completeness and naturalness of the speeches. In this vein, the translation of HI were transcribed with a clean version (deleting all disfluencies) used for paired sample t-tests (the focus of which was on accuracy alone).

Current study adopts SPSS statistics to analyse the statistic data and to make a comparisom between HI and MI with regard to both the process and the product of interpreting. The level of significance was set at 0.05, but due to the explorative nature of this study, results at the 0.1 level were deemed as potentially meaningful and need to be probed further.

This study adopts the following model: 1) incorporates multiple assessment facets; 2) conducts analysis for different measurement requirements; 3) generates diagnostic statistics on independent raters; 4) examines biased interaction between raters; and finally provides individual feedback to raters for further training [1].

4.1 Results

A comparison of HI and MI's accuracy via scale rating.

HI's performance on accuracy was rated based on the clean version, in which pauses and disfluencies are deleted, with a focus on the completeness of information and accuracy. The performance of both HI and MI when interpreting the eight speeches are rated and raters also calculated the average of the scores. An acceptable degree of reliability was found among three raters (Mean=13, Min=10.7, Max=14.8, SD=2.096). Therefore, the mean of the five scores given by the three raters was used to indicate the accuracy of the interpreting performance. The following Table 2 presents the results of Paired-samples t-tests. It showed that in both directions, $p>$ .05, it therefore, can conclude that MI excels HI in terms of accuracy. Interestingly, when interpreting from L1 to L2, the sample average of HI and MI is higher than those in the L2-L1 direction with regard to accuracy. A Cohen's d of 1.52 in the L1-L2 direction indicated that the magnitude of the effect was large, while in the L2-L1 direction, the effect size d is 0.41, which entailed that the magnitude of the effect was medium. The medium effect size revealed that MI enjoys a higher level of accuracy in the L1-L2 direction than HI.

Table 2. Differences of scores in terms of accuracy between HI and MI in each direction

Diretion	Level of Diffculty	HI		MI		95% confidence interval of the difference					
		M	SD	M	SD			t	df	p	Cohcn's d
L2-L1	General	12.7	2.95	11.25	5.19	± 1.826	± 5.084	0.3718	1.2699	.7625	0.41
L1-L2	Technical	11.7	2.63	15.5	19.1	± 1.628	± 1.877	-2.122	2.7033	.1338	1.52

A comparison of HI and MI's overall quality via scale rating.

HI was rated by considering their presentation (e.g. voice, pace), fluency, accuracy, overall coherence, and interpreting skills. The performance of both HI and MI when interpreting the eight speeches are rated, and raters also calculated the average of the scores. An acceptable degree of reliability was found among three raters (Mean=11.83, Min=9.5, Max=15.7, SD=3.37). Therefore, the scores given by the three raters was used to analyse the overall quality of HI and MI's interpreting performance. In Table 3, T-tests displayed that in both directions, $p>$ .05, it therefore, can conclude that MI excels HI in terms of overall quality. Interestingly, accuracy in the L2-L1 direction between the sample average of HI and MI is not big enough to be statistically significant. In other words, the difference between HI and MI was only significant in the L1–L2 direction. It clearly shows that MI's interpreting quality is higher than HI in the direction of L1-L2 only. In the L1-L2 direction, Cohen's *d* is calculated as 2.73, suggesting that the size effect was large, while in the L2-L1 direction, the effect size *d* is 0.42, which entails that the magnitude of the effect was medium. Different with the analysis of accuracy, the overall quality of MI in the L2-L1 direction is slightly higher than accuracy. It can be posited that MI performs significantly well than HI in the L2-L1 direction with a higher level of accuracy and obviously, fluency.

Table 3. Differences of rating scores in overall quality between HI and MI in each direction

Dire tion	Level of Diffcult y	HI		MI		95% confidence interval of the difference					
		M	SD	M	SD			*t*	*df*	*p*	Cohcn's *d*
L2-L1	General	11.5	2.11	10.25	5.19	±1.307	±5.084	0.3299	1.1354	.792	0.42
L1-L2	Tech-nical	9.9	1.84	14.75	17.1	±1.139	±1.674	-3.384	2.0318	.0756	2.73

5 Discussion

The results of the current study demonstrate interesting differences found between MI and HI in terms of both accuracy and the overall quality of interpreting. In terms of accuracy, this study finds that, MI increased the score of accuracy significantly higher than HI in both directions. Generally, MI performs much better than HI in the direction of L1-L2, and this does not necessarily hold true in the direction of L2-L1. Concerning the overall quality, MI's advantages in delivering smooth, fluent and high-quality interpretation in the L1-L2 direction become even more obvious.

This is partly due to the reason that when dealing with technical texts, the interpreter will have a heavier cognitive load in memorising technical terms, retrieving proper lexical choices, and then synthesising them into a sentence, however, this process can be easily achieved by MI through computing algorithm. However, MI does not necessarily perform better than HI in the direction of L2-L1 can be roughly summarised as the failure of audio recognition, and the failure to chunk the source text into correct translation units.

Although HI need to put more effort in listening and analysing L2 input, MI also have difficulty in decoding abbreviations and technical terms due to the limitation of its corpus, which will impact its accuracy. For example, when interpreting the sentence "China's youth employment rate now ranks with several G7 countries that have notorious problems with getting younger workers into the labor market, like Spain and Italy". Student interpreters can easily decode G7 countries as "七国集团", but one of the MI used in the study recognise "G7" as "g(ram) seven", so it translated it as "几克七". Besides, the main reason for the failure of accuracy and overall quality of MI is the failure to reproduce the information in a way adapting to the norms of target language. In a word, the main problem of MI is that the output is basically verbatim and therefore it is not as natural and hearer-friendly as those produced by HI, especially when dealing with complex sentence. Also, in the above mentioned example, one of the student interpreter interpret it as "中国的年轻人失业率与G7国不相上下，都问题严峻，其中最为严重的就是让年轻人进入劳动市场，这些情况西班牙和意大利也随处可见 (Lit. China's youth unemployment rate is almost the same with those in G7 countries. Both face with severe problems, especially the issue of letting youth entering into labor market, which are also pertinent in Spain and Italy". While MI interprets it as "中国的青年就业率与西班牙和意大利等七国集团国家一样，这些国家在让年轻工人进入劳动力市场方面存在着臭名昭著的问题 (Lit. China's youth employment rate is the same with G7 countries, such as Spain and Italy. These countries have bad-reputation problems in letting youth entering into labor market)". Although MI's translation boasts a higher quality than HI's in terms of accuracy and simpleness HI's interpretation is closer to the norms of target language (e.g. using more short sentences instead of long sentences) than MI's. Besides, in this example, the phrase "notorious problems" is used to emphasize the degree of severity of the problem, so HI's translation "问题严峻 (lit. severe problems)" is more natural and accurate than MI's translation "臭名昭著的问题 (lit. bad reputation problems)".

A possible explanation for the increase in quality and accuracy again lies in MI's capacity to create a better storage of the original speech than conventional notes and to relieve some of the burden on on-site interpreting (especially interpreting technical terms). The results reveled that computer-assisted human interpreting has great potential not only in the education field but also in the practical application field. This study proposes a new training mode of computer-assisted interpreting training by adopting MI technologies. As illustrated in Figure 3, Computer-assisted interpreting training can be used in two modes:1) using MI generating a written text of the source speech. Since MI generated text normally has a higher level of completeness in terms of information than notes in consecutive interpreting. In other words, the speed of speaking is faster than that of writing. Therefore, this mode is supposed to decrease interpreters' memory cognitive load and to outperform HI in terms of accuracy. In addition, MI will mute the speaker's voice, so interpreters will not be impacted by the surrounding environment. 2) using MI to perform the processes of listening and interpreting, and HI finish the last step of post-editing. If the MI is mature enough, HI only need to perform the step of post-editing to improve the minor syntactic and logical cohesion mistakes. However, this mode may not work perfectly as expected, since interpreter may pay much more attention on the correctness of MI's product, and which may incur a heavier work load than the first mode.

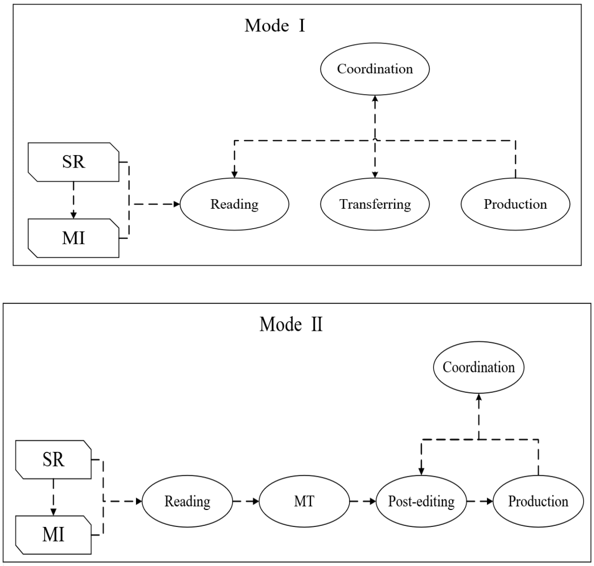

Fig. 3. Computer-assisted training mode for consecutive interpreting

6 Conclusion

The findings of this study indicate a demand of university education to adjust translation and interpreting courses to keep pace with the booming development of AI. Scholarly attention should not be restricted to how technology will replace human beings, but also to how wittingly exploit technology as powerful weapon to improve human's performance and ability. Interpreting education should place emphasis on the proper use of technology, and enlighten students to be a good user of technology rather than a competitor of technology [7]. University teachers need to rethink the use of CAI tools by fulling disclosing the bright and dark side of AI. It is their responsibility to educate students to apply technology to interpreting effectively and reasonably, and to better prepare them for the age of AI.

Undeniably, this study has several limitations. First and foremost, a sample size of five human interpreters and two machine interpreters is very small, rendering this an exploratory study with limited suggestive value. Second, this study only involves interpreting in the consecutive mode. More scholarly attention is required to find out how some of the recent technologies could help student interpreters deal with different and yet complex interpreting tasks and to find out new computer-assisted training modes for university interpreting teaching and training. Similarly, the research of liaison interpreting and sight translation are equally important. Third, this study only investigated the transference between two languages, and one group of participants (graduate students), and it remains to be further verified whether these findings are applicable to other languages. Also, the performance of professional interpreters might be different with what have been found in the current study. Forth, the comparative analysis resides on HI and MI's processing ability on political and economic texts. The performance of HI and MI in technology discourse, medical discourse and law discourse need to be further verified.

Despite various constraints, the findings nicely verify that technologies can assist interpreters in improving their performance before and even during their interpreting tasks. There remains numerous important and worthy topics in the field of interpreting that awaits further study –for example, the use of new interpreting modes in various interpreting settings, such as court interpreting and medical interpreting, and so on. It it hoped that the current study could shed light on future research on these topics/

Acknowledgment

The current study are funded by Social Science Fund of Hubei Province under Grant 22Q099 and Wuhan Polytechnic University under Grant XQ2023016.

Reference

1. Han, Chao. 2015. "Investigating rater severity/leniency in interpreter performance testing: A multifaceted Rasch measurement approach." *Interpreting* 17(2): 255–283. https://doi.org/10.1075/intp.17.2.05han.
2. Corpas Pastor, Gloria. 2022. "Technology Solutions for Interpreters: The VIP System." *Hermēneus. Revista de Traducción e Interpretación* 23: 91–123.
3. Fantinuoli, Claudio. 2017. "Computer-Assisted Preparation in Conference Interpreting." *Translation & Interpreting* 9 (2): 24–37.
4. Wang, Xinyu, and Caiwen Wang. 2019. "Can computer-assisted interpreting tools assist interpreting?" *Transletters. International Journal of Translation and Interpreting* 3: 109–139.
5. Deng, Juntao, Mianjun Xu, Tianyuan Zhao. 2021. "Frontiers of Interpreting Technologies and Informatization of Interpreting Education in the AI Era. *TEFLE* 3: 67-79.
6. Deng Juntao, Mianjun Xu, Tianyuan Zhao. 2023. "The review and enlightment of the roadmap of interpreting skills on British websites". *Foreign Language Studies* 1:75-80.
7. Fantinuoli, Claudio. 2018a. "Computer-assisted interpreting: Challenges and future perspectives." In *Trends in E-tools and Resources for Translators and Interpreters*, edited by Isabel Durán-Muñoz and Gloria Corpas Pastor, 153–174. Leiden: Brill.
8. Pöchhacker, Franz. 2016. *Introducing Interpreting Studies*. Manchester: Routledge.
9. Jia, Hao, Shuqin Gu, Yangbin Shi, Xiangyu Duan, Zhongkai Hu, Yuqi Zhang, and Weihua Luo. "Bilingual Terminology Extraction from Non-Parallel E-Commerce Corpora." ArXiv:2104.07398 [Cs], Jan 15, 2024. [Online] Available at http://arxiv.org/abs/2104.07398.
10. Fantinuoli, Claudio. 2018b. "Interpreting and technology: The upcoming technological turn." In Claudio Fantinuoli (ed.), *Interpreting and technology*, 1–12. Berlin: Language Science Press.
11. Braun, S. 2019. "Technology and Interpreting". In M. O'Hagan (Ed.), *Routledge Handbook of* Translation *and Technology*, 271-288. London: Routledge.

12. Fan, Jiamin, Yiting Ru, Yida Chen, Miao Su, Xiaoxiao Feng, and Mei Yang. 2022. "Comparative Analysis of NMT and Human Translation—Poverty Alleviation Perspective". *CCIS* 18, 188–198.
13. Waibel, A., M. Seligman, A. Joscelyne. 2017. TAUS Speech-to-speech translation technology. TAUS. Available at http://isl.anthropomatik.kit.edu/cmukit/
14. Gile, Daniel. 2009. Basic Concepts and Models for Interpreter and Translator Training: Revised Edition. 2nd ed. Amsterdam: John Benjamins.
15. Chen Sijia, Jan-Louis Kruger. 2023. " The effectiveness of computer-assisted interpreting: A preliminary study based on English-Chinese consecutive interpreting". *Interpreting* 18(3): 399–420.