Linguistics and Education 80 (2024) 101273

Contents lists available at ScienceDirect

Linguistics and Education

journal homepage: www.elsevier.com/locate/linged

# Multidimensional comparison of Chinese-English interpreting outputs from human and machine: Implications for interpreting education in the machine-translation age

Yiguang Liu, Junying Liang *

*Department of Linguistics and Translation, School of International Studies, Zhejiang University, Hangzhou 310058, China*

ARTICLE INFO

*Keywords:*
Interpreting education
Machine translation
Human interpreter
Corpus

ABSTRACT

Interpreting (i.e., oral translation) is facing challenges due to the prevalence of machine translation, and one urgent question for interpreting educators is what knowledge and skills should be taught in the machine-translation age. Focusing on this issue, the present corpus-based study systematically quantified the differences between interpreting outputs from expert interpreters and two machine translation systems. Using text analysis tools, significant differences in multidimensional linguistic features including lexical, syntactic, and cohesive ones were shown between humans and machines but not between two artificial systems. Finer-grained statistical analyses indicated that human-machine differences in certain indices deviated in varied interpreting modes. Our data collectively revealed the strengths of human interpreters in audience-oriented communicative mediation but limitations in cognitive resources. By relating the findings to interpreting competence, the current research provides important implications for empowering students in adaptively resorting to human strengths and/or embracing machine translation technologies.

## 1. Introduction

Interpreting is an oral translational activity to enable communications across barriers of language and culture (Pöchhacker, 2016), which plays an essential role in international conferences and a range of community settings, e.g., healthcare interpreting, legal interpreting, and signed interpreting for deaf people. Interpreting features cross-linguistic processing and immediacy (i.e., the target-language output is rendered on the basis of one-time presentation of source-language input; Kade, 1968), which make it a linguistically and cognitively demanding bilingual activity (Christoffels et al., 2006). Given its importance as well as processing difficulties, interpreting education developed since the mid-to-late twentieth century (Sawyer & Roy, 2015), with the aim of preparing trainee interpreters for high requirements. Recent years have witnessed the dramatic advances of machine translation (MT), which pose both challenges and opportunities to translation and interpreting professions. The concept of translation, as well as subtasks and skillsets involved, has accordingly changed (Pym & Torres-Simón, 2021). Placing interpreting education in the context of the rise of MT, one urgent question lies in what knowledge and skills are required and should be taught for interpreters to compete and/or collaborate with machines. Focusing on this issue, the present study conducts a product-oriented investigation comparing human interpreters and MT systems in terms of linguistic features. Our efforts could inspire interpreting education in helping trainees play to their strengths and avoid showing limitations in the machine-translation age.

## 2. Literature review

### 2.1. Multicomponent and dynamic construct of interpreting competence

The acquisition of interpreting competence, a set of abilities, knowledge, and skills required for accurate and smooth interpreting (Lee, 2020), is the stated objective of interpreting education (Kalina, 2000), and thus competence-related issues form a key concern of interpreting education research (Grbić & Pöchhacker, 2015).

Interpreting competence is generally regarded as a multicomponent construct, encompassing competences on various levels, including linguistic, cognitive, cultural, etc. (Albl-Mikasa, 2013; Kalina, 2000; Pöchhacker, 2000; Wang et al., 2020). Also, the construct of interpreting

Funding sources: This work was supported by the China Postdoctoral Science Foundation [grant number 2022M710134].

* Corresponding author.
*E-mail address:* jyleung@zju.edu.cn (J. Liang).

https://doi.org/10.1016/j.linged.2024.101273
Received 31 May 2023; Received in revised form 27 January 2024; Accepted 29 January 2024
Available online 15 February 2024

competence is dynamic, that is, it evolves with academic and social progress over time. Interpreting competence was, in earlier research, understood as coextensive with the natural abilities of bi-/multi-linguals to use more than one language (Harris & Sherwood, 1978), suggesting that all required for successful interpreting is linguistic knowledge and skills. This view echoed the fact that interpreting services have been performed by non-professionals for centuries, and interpreting teaching at the early stage of interpreting education was limited to training students "simply by trying and practising until it worked" (Kalina, 2000).

Nevertheless, the increasingly high demands of and varied requirements for interpreting have called for systematic trainings and a profession-oriented view of competence, which regarded numerous skills and abilities beyond bilingual proficiency as essentials (Lörscher, 1991; Toury, 1995). As an intense language processing activity where cognitive processing and language use interact, converging evidence suggests that interpreting is constrained by sociocultural and cognitive factors (Liang et al., 2018). On the one hand, to reach mutual understanding under given sociocultural contexts, qualified interpreters should have a good command of cultural knowledge and communication skills (Angelelli, 2004; Hatim & Mason, 1997). An audience-oriented awareness is also important to guarantee the audience's perception of intended meanings (Li & Dong, 2022). On the other hand, interpreting is a cognitively demanding activity whose difficulties mainly include the temporal overlap of multitasks (i.e., perception and analysis of inputs, production of outputs, mental storage of information, and coordination of various operations; Darò & Fabbro, 1994; Gile, 2009; Moser, 1978), and interferences between subtasks (Seeber, 2011), concurrently activated two languages (Gerver, 1976), as well as modalities of information (Seeber, 2017). Accordingly, interpreters face varied kinds of and levels of cognitive loads during different processing stages (Seeber, 2011; Seeber & Kerzel, 2012), and commonly work close to the limit of their processing capacity (Christoffels et al., 2006; Gile, 1999). These support the argument that relevant cognitive functions (e.g., working memory, monitoring, shifting) and appropriate use of interpreting strategies are important for interpreters to relieve processing loads and perform well (Dong & Li, 2020). Furthermore, varied cognitive demands are inherent to different interpreting modes, mainly consecutive interpreting (CI) and simultaneous interpreting (SI). SI is more demanding in time pressure as it requires the simultaneous execution of the above-mentioned subtasks. By contrast, comprehension and production happen serially during CI, and thus it taxes interpreters more in mnemonic resources since the total memory load is accumulating and accelerating prior to the integration of stored constituents (Liang et al., 2017; Lv & Liang, 2019). Consistent with the above-reviewed studies, interpreting competence models highlight the integration of sociocultural and physio-psychological components into the competence construct (e.g., Kaczmarek, 2010; Pöchhacker, 2000; Wang et al., 2020).

More recently, the advances of interpreting technologies have attracted interest from academic and industry circles (Fantinuoli, 2018). Efforts have been made to develop digital tools and technology-assisted working modes to empower interpreters in interpreting preparation (Fantinuoli, 2017), number rendition (Frittella, 2022; Pisani & Fantinuoli, 2021), input speech recognition (Chen & Kruger, 2022; Defrancq & Fantinuoli, 2020), etc. Consistent with the increased introduction of technologies into translator training (Schnell & Rodríguez, 2017), the ability to use emerging technological tools has been additionally deemed an important part of interpreters' skillset (Wang & Li, 2022).

### 2.2. Prevalence of machine translation

Machine translation can be dated back to the 1930s (Hutchins & Lovtskii, 2000), but it has not exerted significant impact on translational communication until recent years due to dramatic advances driven by artificial intelligence (AI) technologies (Forcada, 2017). It was even claimed that a neural network MT system has reached human parity performance (Hassan et al., 2018), though this claim has been questioned (e.g., Toral et al. 2018). Likely, recent years have witnessed much progress of machine interpreting (also known as automatic speech-to-speech translation), whose typical architecture combines three modules, namely, automatic speech recognition, text-to-text translation by MT system, and automatic speech synthesis (Jekat, 2015).

MT excels in automatic and extremely efficient processing of words. Nowadays, texts in many languages can be translated into target languages simply by feeding them into free online MT systems (e.g., Google Translate, DeepL Translator, Baidu Translate) and waiting for a few seconds. Considering the impressive improvements of MT quality since the arrival of neural machine translation (Forcada, 2017), translation and interpreting professions are under great threat (Downie, 2019; Pym & Torres-Simón, 2021). However, on the other side of the coin, MT shows potential to empower (human) translation. Professional translators adopt MT in workflow commonly and even on a daily basis since MT can lead to speed gains, be effective for certain text types and language pairs, and provide assistance with terminology-related translation challenges (Cadwell et al., 2018). Also, translation companies actively pursue in-house machine translation and other language automations to optimize workflow and improve productivity (ELIA, 2022).

Being both a challenge and opportunity, the concept of translation itself, as well as subtasks and skillsets involved, has changed with the prevalence of MT (Pym & Torres-Simón, 2021). For example, the growing supply of MT-performed routine tasks has made what MT cannot manage more valuable. For translation and interpreting educators, they should future-proof their programs and figure out how to align education with the changes and updated requirements in the face of MT prevalence (Loock, 2020; Bulut, 2019). One important but open question is what (sub)competences should be prioritized to empower interpreting students who are supposed to be qualified for what MT cannot do and smoothly work with MT systems. The present study aims to provide empirical evidence for this question.

### 2.3. Linguistic analysis of human-machine differences in translation outputs

Comparative analysis of human and machine interpreting outputs provides a pathway to answering the urgent question above. The rationale is that such human-machine comparison makes it possible to define the strengths of human interpreters and MT systems (Loock, 2020), with the former speaking to something that cannot be automated within interpreting and the latter pointing to the potentials of human-machine doubling. Thus, suggestions can be offered on what (sub)competences to focus on when empowering students to outperform or/and collaborate with MT.

Concerning methodology, corpus-based linguistic analysis of interpreting outputs was conducted in the present study out of two considerations: First, the corpus-based approach tends to effectively assess interpreting quality based on commonly-recognized quality criteria, and, more importantly, it shows advantages in assessment consistency and reliability by reducing human bias caused by subjectivity (Liu, 2021; Ouyang et al., 2021). For example, Liu (2021) selected a group of linguistic features, which are associated with three major quality criteria respectively (i.e., information accuracy, output fluency, and audience acceptability), to quantitatively analyze interpreting outputs and found that linguistic data can effectively distinguish interpretations of different qualities. Second, beyond the evaluation of "good" interpretation, corpus-based linguistic analysis excels in uncovering the characteristics inherent to translated renditions (i.e., translationese) from both translators and MT systems and thus match our research aim. Drawing on corpus linguistics to do comparisons, prior studies identified the differences between MT and human outputs in certain features (e.g., Bizzoni et al. 2020, De Clercq et al. 2021, Kuo 2019, Loock 2020, Vanmassenhove et al. 2019), and revealed compromised tendency of hypothesized (human) translation universals in the MT texts, including explicitation (Krüger, 2020), simplification (Luo & Li, 2022), and normalization (Lapshinova-Koltunski, 2015).

However, there is a paucity of endeavors to touch upon human-machine comparison in oral translation (i.e., interpreting). Given the extreme cognitive difficulties and time pressure specific to interpreting, distinct patterns may be revealed in interpreting compared with written translation. Secondly, the above-reviewed studies usually limited their analysis to sparse linguistic features, with the sentence- and discourse-level information mostly under-researched (but see Bizzoni et al. 2020, Jiang & Niu 2022). Thirdly, the connection of empirical data to translation/interpreting education issues is scant yet.

### 2.4. The present study

Here, we sought to quantitatively compare the interpreting outputs from expert human interpreters and state-of-the-art MT systems in terms of linguistic features. To have a systematic and comprehensive picture, large-grained and fine-grained measures from text analysis tools were chosen here, covering multiple dimensions including lexical, syntactic, cohesive measures. More importantly, our empirical investigation is situated in the context of interpreting education, with the aim of providing insights for empowering interpreting students in the machine-translation age. The following questions are to be addressed:

**Q1**: To what extend and in what dimensions do Chinese-English interpreting outputs from human interpreters and two MT systems significantly differ? Does interpreting mode (CI and SI) modulate the possible differences?

**Q2**: What strengths and limitations do human interpreters hold compared with MT?

**Q3**: What (sub)competences should be prioritized to prepare interpreting students for competing and/or collaborating with MT?

## 3. Method

### 3.1. Materials

To quantify the differences between human and machine interpreting outputs in linguistic features, we self-built a Chinese-English parallel interpreting corpus for four sub-corpora: 1) an interpreting input corpus consisting of to-be-interpreted speeches given by Chinese government leaders in international conferences. Specifically, 20 texts were transcripts of Chinese Premiers' speeches, which were interpreted in the consecutive manner, at the annual Premier Press Conferences (from 2002 to 2021); while another 20 texts were transcripts of speeches delivered on international forums including the Davos Forum and the Boao Forum for Asia, where the speeches were interpreted in the simultaneous manner. Similar materials have been used in corpus-based interpreting studies (e.g., Jia and Liang 2020, Liang et al. 2017, Lv and Liang 2019); 2) a human interpreting corpus for transcribed interpretations corresponding to each source text in the above-mentioned corpus. These interpretations were all rendered on the spot by professional staff interpreters from the Department of Translation and Interpretation of China's Ministry of Foreign Affairs; 3) a Google interpreting corpus consisting of translated outputs based on the interpreting inputs using Google Translate; 4) a Baidu interpreting corpus generated by translating the same interpreting inputs using Baidu Translate.

All the interpreting outputs were from Chinese (the interpreters' mother tongue) to English (their second language). On the premise of the significant differences between CI and SI in interpreting process and cognitive demands (e.g., Liang et al., 2017), both of them were involved to cover mainstream interpreting modes and to further investigate the possible modulation of interpreting mode on human-machine comparison. Google Translate and Baidu Translate were selected since they represent the state-of-the-art MT technologies, with the former as the most famous and frequently used system globally (Wu et al., 2016) and the latter particularly good at Chinese-English translation (Sun et al., 2019). To note, though a whole procedure of machine interpreting, consisting of speech recognition, text-to-text translation, and speech synthesis (see Section 2.2 for more information), is more analogous to human interpreting, text-to-text translation outputs were analyzed in this study out of the following consideration: Text-to-text translation is the core module of machine interpreting because it is where translational activity really happens (Downie, 2019). By contrast, the other two modules are more about voice processing technologies, but not the main concern of interpreting education research. The inclusion of them may blur the attribution of our data to human-machine differences concerning translational matters.

**Table 1**
Overview of materials for Coh-Metrix analyses.

| Sub-corpus | Language | Number of texts | Size (tokens) |
|---|---|---|---|
| Interpreting input | Chinese | 40 | 25,209 |
| Human interpreting | English | 40 | 18,596 |
| Google interpreting | English | 40 | 16,957 |
| Baidu interpreting | English | 40 | 16,922 |

In addition, due to the strict 15,000-character limit per text by the mainly used text analysis tool (i.e., Coh-Metrix 3.0), shorter texts were randomly extracted based on each long text for analysis. These short texts were selected without splitting a complete paragraph to avoid interfering text cohesion measures. An overview of the materials for Coh-Metrix analyses is shown in Table 1, and a total of 120 output texts formed our focus of analysis.

### 3.2. Instrument and measures

Linguistic features encompass multilevel measures (Green, 2019). To systematically compare human and machine interpreting outputs, Coh-Metrix 3.0 was mainly used in the present study. As a reliable computational tool, Coh-Metrix 3.0 excels in text analysis since it provides a one-stop solution to automated text analysis through 106 indices, covering extensive measures on the lexical, syntactic, and cohesive level (Elfenbein, 2011; Graesser & McNamara, 2011). Furthermore, the emphasis on cohesion is a pronounced feature of Coh-Metrix 3.0. *Cohesion* is defined as an explicit text feature related to the coherent representation in readers' mind and thus a high-level text cohesion contributes to readers' comprehension and memory for the text (McNamara et al., 2010). Coh-Metrix makes it possible to capture both referential (i.e., overlap of linguistic forms) and deep (i.e., causal and logical relationships) cohesion of text (McNamara et al., 2014), and thus has been widely used in corpus studies on the intersection of applied linguistics and language education (e.g., Crossley and McNamara 2012, Maamuujav et al. 2021).

To further validate the analyses through Coh-Metrix 3.0 and extend the human-machine comparison, the corpus in this study was also processed using other text analysis tools, including *Tool for the Automatic Analysis of Lexical Sophistication* (TAALES; Kyle et al., 2018) and *Tool for the Automatic Analysis of Syntactic Sophistication and Complexity* (TAASSC; Kyle, 2016). These two more recent systems focus on lexical and syntactic sophistication/complexity respectively with larger-grained and finer-grained indices, which can complement the analyses based on Coh-Metrix (see Supplementary Material for the analyses using TAALES and TAASSC).

### 3.3. Data analysis

Using Coh-Metrix 3.0 Web Tool,[1] we calculated 106 indices measuring linguistic features for the interpreting outputs. For each index, a one-way repeated measures ANOVA was performed to test the

[1] Coh-Metrix 3.0 Web Tool is freely available at http://cohmetrix.com/.

differences among interpreting outputs from human interpreters, Google Translate, and Baidu Translate. Given the involvement of multiple comparisons, the *p*-value was adjusted based on the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) to avoid false positive results. When the cross-corpus differences reached statistical significance in certain indices, pairwise comparisons were employed to test whether there were significant differences in each pair, and two-way mixed ANOVAs were employed to test whether interpreting mode (i.e., CI and SI) modulated the human-machine differences.

## 4. Results and interim discussion

The one-way ANOVA results identified 30 (out of 106) indices showing significant differences among outputs from human interpreters (HI), Google Translate (GT), and Baidu Translate (BT). These indices cover four dimensions, i.e., descriptive, lexical, syntactic, and cohesive measures. The more detailed results, as well as corresponding interim discussions, are presented below.

### *4.1. Descriptive measures*

Seven indices related to descriptive statistics showed significant human-machine differences but no significant differences between the two MT systems, including sentence count, word count, sentence length, word length, Flesch Reading Ease, and Flesch-Kincaid Grade Level, see Table 2.

Pairwise comparisons revealed that human interpreters rendered significantly more sentences (Human: 23.7 ± 6.95; Google: 19.0 ± 4.70; Baidu: 19.2 ± 4.65) and words (Human: 468 ± 107.0; Google: 430 ± 78.7; Baidu: 429 ± 73.0) per text than Google Translate and Baidu Translate, see Fig. 1 and Table 2. Based on the same source language (SL) inputs, human interpreters produced more informative contents in the target language (TL) outputs. This tends to derive from *explicitation*, which is defined as “a stylistic translation technique which consists of making explicit in the target language what remains implicit in the source language” (Vinay & Darbelnet, 1995). Explicitation shifts prevail during interpreting as a result of cross-linguistic differences and/or audience-oriented efforts to improve communication (Li & Dong, 2022; Li & Halverson, 2020). Although the results here cannot rule out the existence of explicitation in MT outputs, our data indeed showed that human interpreters outperformed MT systems in the use of explicitation. Case analyses are presented below to complement the quantitative results.

**Example 1**: **Source input** 与其他发展中国家分享减贫经验, 提供更多帮助.

**Human output** We are ready to share our poverty reduction experience with and offer more assistance to fellow developing countries.

**Google output** Share poverty reduction experience with other developing countries and provide more assistance.

**Baidu output** Share the experience of poverty reduction with other developing countries and provide more help.

**Example 2**: **Source input** 苟利国家生死以, 岂因祸福避趋之.

**Human output** Remaining committed to the conviction that **I shall dedicate myself to the interests of the country** in life and death irrespective of personal weal and woe.

**Google output** Going for the life and death of the country, how can it be avoided by misfortune or fortune.

**Baidu output** I will benefit the country in life and death, and I will not avoid it because of misfortunes and blessings.

Chinese allows zero subjects, and the omitted subjects are generally added in the English counterpart during Chinese-English interpreting (Li & Halverson, 2020). In Example 1, “we” was added as the subject in the human output to accommodate to the linguistic norm of the target language. However, such explicitation shift was not shown in the MT outputs. Furthermore, explicitation is also motivated by the audience-oriented tendency (Li & Dong, 2022). Example 2 provides a case in this regard. “苟利国家生死以, 岂因祸福避趋之” is an excerpt from a famous Chinese ancient poetry written by Zexu Lin, a patriotic general in the dynasty of Qing. Without relevant cultural information, it is difficult to infer the speaker’s intention (i.e., the Premier Jiabao Wen), who cited the verse for expressing his determination to serve the country. In this case, the interpreter explicitly clarified the contextually- and culturally-dependent connotations of the cited verse, particularly the speaker’s intention (i.e., “I shall dedicate myself to the interests of the country”). This explicitation operation benefited the nonnative audience by minimizing comprehension barriers, but was beyond the reach of the two MT systems since their word-to-word translations were neither pragmatically accurate nor linguistically readable.

In terms of sentence length and word length, HI used shorter sentences (Human: 20.6 ± 3.79; Google: 23.6 ± 4.87; Baidu: 23.3 ± 5.14) and words, with the latter measured by both the number of syllables (Human: 1.69 ± 0.12; Google: 1.74 ± 0.14; Baidu: 1.74 ± 0.15) and letters (Human: 5.03 ± 0.32; Google: 5.16 ± 0.38; Baidu: 5.13 ± 0.39), see Fig. 1 and Table 2. These results indicated that the interpreters appeared to use simpler sentences and words compared to MT systems.

Furthermore, two readability indices showed significant differences, namely, Flesch Reading Ease (FRE; Flesch, 1948) and Flesch-Kincaid Grade Level (FKGL; Kincaid et al., 1975). The HI outputs had a higher averaged FRE score (Human: 43.3 ± 12.4; Google: 35.8 ± 15.0; Baidu: 36.2 ± 16.0) than those from GT and BT. As for the FKGL score, the HI outputs had the lowest score, followed by the BT and GT outputs (Human: 12.3 ± 2.45; Google: 14.1 ± 3.15; Baidu: 14.0 ± 3.36). These two metrics are calculated based on sentence length and word length,[2] with a higher FRE score but lower FKGL score indicating a lower-level comprehension difficulty. Although readability measures are not tailored for listening comprehension, it is plausible to suppose that shorter words and shorter sentences are also easier to be understood even when they are orally presented since word and sentence length can affect the processing capacity to quickly recognize and recall words (Smith & Taffler, 1992). In addition, *listenability* measured by the Ease Listening Formula was found to be highly correlated with *readability* measured by FRE (Fang, 1966). Therefore, the results here likely revealed that the HI outputs were more understandable at the surface level.

In sum, although human interpreters (versus two MT systems) rendered more sentences and words per text, they tended to use shorter sentences and words in interpreting outputs.

### *4.2. Lexical measures*

Lexical indices showing significant differences mainly involved the measure of lexical diversity, word information related to psychological ratings, and word categories (Table 3).

Type-token ratio (TTR) and measure of textual lexical diversity (MTLD) are two indices to reflect how diverse the words used in a text are. In the present study, the HI outputs had a lower TTR score than the MT outputs (Human: 0.12 ± 0.02; Google: 0.13 ± 0.02; Baidu: 0.13 ± 0.02). Conversely, the MTLD results showed that human interpreters used a wider range of words (Human: 0.69 ± 0.06; Google: 0.67 ± 0.05; Baidu: 0.66 ± 0.05). Such an inconsistency may be due to the fact that TTR score is sensitive to text length and the more words used in the HI outputs (see Section 4.1) confounded the lexical diversity measured by TTR. MTLD, however, can overcome the potential confounding of text length by using sampling and estimation methods (McCarthy & Jarvis, 2010). Thus, the MTLD tends to be a more reliable index for lexical diversity here, which revealed that the HI excelled in lexical diversity. For example, “推进 (meaning *promote*)”, a frequently-used word in the

[2] Flesch Reading Ease is computed as [206.835 – (1.015 * sentence length) – (84.6 * word length)], and Flesch-Kincaid Grade Level is computed as [(0.39 * sentence length) + (11.8 * word length) – 15.59].

**Table 2**
ANOVA analyses and pairwise comparisons on descriptive measures.

| Index | ANOVA | | Pairwise comparisons | | | | | |
|---|---|---|---|---|---|---|---|---|
| | F | *p* | Human-Google | | Human-Baidu | | Google-Baidu | |
| | | | *p* | *d* | *p* | *d* | *p* | *d* |
| Sentence count, number of sentences | 34.70 | <0.001 | <0.001 | 0.94 | <0.001 | 1.03 | .650 | 0.07 |
| Word count, number of words | 12.17 | .002 | .001 | 0.62 | .002 | 0.55 | .840 | 0.03 |
| Sentence length, number of words | 13.75 | <0.001 | <0.001 | 0.63 | <0.001 | 0.59 | .444 | 0.12 |
| Word length, number of syllables, mean | 27.36 | <0.001 | <0.001 | 0.96 | <0.001 | 0.88 | .520 | 0.10 |
| Word length, number of letters, mean | 13.47 | <0.001 | <0.001 | 0.76 | .001 | 0.58 | .299 | 0.17 |
| Flesch Reading Ease | 41.24 | <0.001 | <0.001 | 1.17 | <0.001 | 1.04 | .520 | 0.10 |
| Flesch-Kincaid Grade Level | 28.83 | <0.001 | <0.001 | 0.92 | <0.001 | 0.86 | .450 | 0.12 |

Note: *d* refers to the value of Cohen'd as an effect size measure.

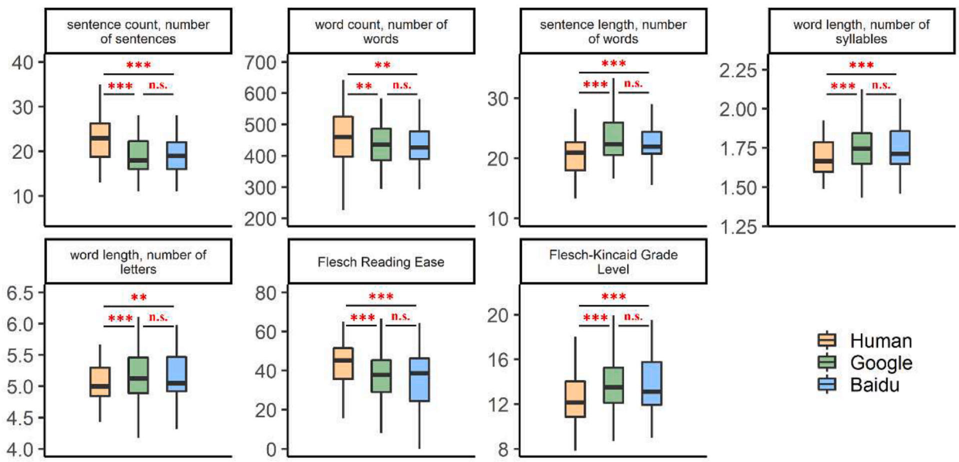


**Fig. 1.** Comparison of interpreting outputs from human interpreters, Google Translate, and Baidu Translate in terms of descriptive measures. $^{**}p < 0.01$, $^{***}p < 0.001$, n.s. = non-significant.

**Table 3**
ANOVA analyses and pairwise comparisons on lexical measures.

| Index | ANOVA | | Pairwise comparisons | | | | | |
|---|---|---|---|---|---|---|---|---|
| | F | *p* | Human-Google | | Human-Baidu | | Google-Baidu | |
| | | | *p* | *d* | *p* | *d* | *p* | *d* |
| Lexical diversity, type-token ratio | 8.20 | .002 | .007 | 0.47 | .004 | 0.55 | .158 | 0.23 |
| Lexical diversity, MTLD | 11.53 | .002 | .009 | 0.45 | .001 | 0.62 | .009 | 0.44 |
| Age of acquisition for content words | 24.66 | <0.001 | <0.001 | 0.84 | <0.001 | 0.86 | .540 | 0.10 |
| Familiarity for content words | 5.23 | .032 | .056 | 0.34 | .035 | 0.42 | .193 | 0.21 |
| Text Easability, word concreteness | 8.77 | .006 | .001 | 0.62 | .054 | 0.31 | .013 | 0.44 |
| Polysemy for content words | 25.80 | <0.001 | <0.001 | 0.86 | <0.001 | 0.84 | .890 | 0.02 |
| Noun incidence | 8.96 | .001 | .031 | 0.38 | .001 | 0.63 | .048 | 0.32 |
| Pronoun incidence | 42.81 | <0.001 | <0.001 | 1.14 | <0.001 | 1.16 | .720 | 0.06 |
| First person plural pronoun incidence | 14.71 | <0.001 | <0.001 | 0.67 | <0.001 | 0.62 | .346 | 0.15 |
| Third person plural pronoun incidence | 16.50 | <0.001 | <0.001 | 0.64 | <0.001 | 0.69 | .517 | 0.10 |

Note: *d* refers to the value of Cohen'd as an effect size measure.

inputs, was translated into "promote" or "advance" in most cases by the MT systems, while it was also translated into other various expressions by human interpreters, including "pursue", "progress", "deepen", "boost", "press ahead", and "push ahead", etc.

Human-machine differences were also showed in indices related to psycholinguistic word information. One of them is *age of acquisition* (AOA), which reflects how early a word is acquired by children. Words with higher AOA scores are learnt later by children and typically related to higher-level lexical difficulty (McNamara et al., 2014). The HI outputs had a lower AOA score (Human: 386 ± 16.9; Google: 396 ± 19.2; Baidu: 397 ± 19.8). Another two indices, *Familiarity for content words* and *word concreteness*, reflect how familiar and how concrete a word is respectively. If a word is more familiar and/or more concrete, it is generally easier to understand. The HI outputs (versus MT outputs) had a higher score in terms of both word familiarity (Human: 565.06 ± 7.07; Google: 563.60 ± 6.43; Baidu: 562.93 ± 6.67) and word concreteness (Human: 54.39 ± 21.79; Google: 45.72 ± 21.92; Baidu: 49.38 ± 23.35). Furthermore, the significant differences lay in *polysemy for content words*, which measures the number of senses a word has. For example, the word "bank" has at least two senses, with one referring to a monetary institution and the other the side of a river. Accordingly, polysemy is closely linked to text ambiguity, and higher polysemy scores reflect more use of simple words in a text since simple words often have more meanings (McNamara et al., 2014). In this study, the HI outputs had a highest polysemy score among the three corpora (Human: 4.16 ± 0.32; Google: 3.88 ± 0.32; Baidu: 3.88 ± 0.31). The above-mentioned results

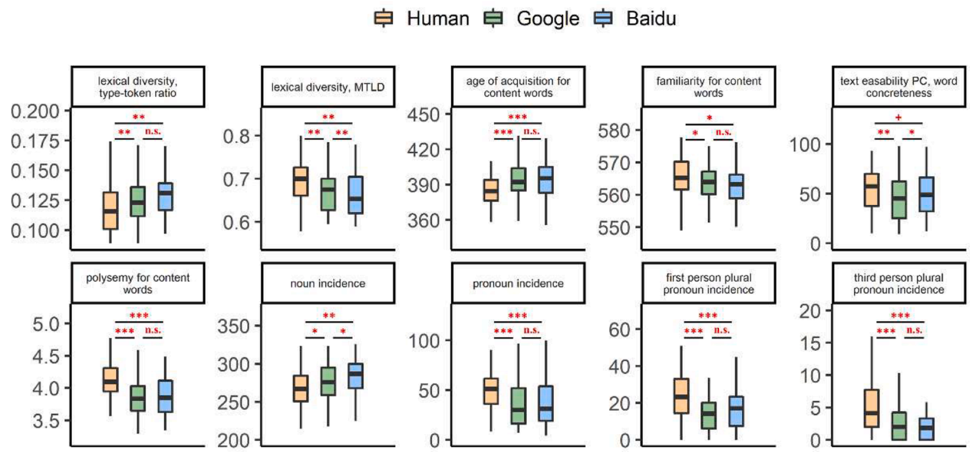


**Fig. 2.** Comparison of interpreting outputs from human interpreters, Google Translate, and Baidu Translate in terms of lexical measures. + $p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, n.s. = non-significant.

generally revealed that the interpreters tended to select words that are easier to process.

Coh-Metrix 3.0 also provides information on occurrence of each word category, and our data showed significant differences in the incidence[3] of noun, pronoun, first person plural pronoun, and third person plural pronoun. There was a lower proportion of nouns in the HI outputs than in the MT outputs (Human: 269.60 ± 28.09; Google: 275.70 ± 25.91; Baidu: 281.42 ± 25.76). However, human interpreters used more pronouns than two MT systems in terms of all pronouns (Human: 50.19 ± 23.28; Google: 36.06 ± 22.05; Baidu: 36.54 ± 21.23), first person plural pronouns (Human: 24.24 ± 13.68; Google: 15.26 ± 11.80; Baidu: 16.36 ± 10.79), as well as third person plural pronoun (Human: 5.22 ± 4.90; Google: 2.76 ± 3.46; Baidu: 2.60 ± 3.72), see Fig. 2 and Table 3.

The reversed results in terms of noun and pronoun incidence could result from the fact that the interpreters more frequently used pronouns to refer to corresponding nouns. The use of pronouns appears to improve text cohesion since binding a pronoun with its referent can guide the audience to connect ideas in the text (McNamara et al., 2014). Furthermore, pronominal usage, especially in political settings, is highly dependent on the specific context, and plays an important pragmatic role in conveying speakers' stances and intentions (De Fina, 1995). For instance, first person plural pronouns (e.g., "we", "us", "ourselves") can function as a linguistic device to build solidarity, and, through the overuse of these pronouns, politicians can fulfill persuasive goals (Zupnik, 1994).

It was found here that first person plural pronouns were more frequently utilized by the interpreters. In Example 3 below, the two MT systems literally translated "中国 (meaning *China*)" into "China", whereas the interpreter used "we" to refer to "China". By using "we", a united community including the government of China and Chinese people was built, which implies that all Chinese people collectively take efforts to promote the reform program of China. Similarly, "us" was used to refer to "世界各国 (meaning *all countries in the world*)" by the interpreter rather than the two MT systems in Example 4. The pronoun "us" here contributes to the construction of a community including China and other countries. Such an interpretation could shorten the distance between China and other countries in the world, consistent with the China's concept of building a global community with shared future for mankind.

**Example 3: Source input** 我可以肯定地说，中国会坚定不移的推进改革.

**Human output** One thing is certain, that is **we** are determined to push ahead with our reform agenda.

**Google output** I can say with certainty that **China** will unswervingly push forward reform.

**Baidu output** I can say for sure that **China** will unswervingly promote reform.

**Example 4: Source input** 在经济全球化条件下，**世界各国**的命运已紧紧联系在一起.

**Human output** Economic globalization has linked **us** closely together.

**Google output** Under the conditions of economic globalization, the destinies of **all countries in the world** are closely linked.

**Baidu output** Under the condition of economic globalization, the fate of **all countries** **in the world** has been closely linked.

### 4.3. *Syntactic measures*

Five syntactic indices showed significant differences, including number of modifiers per noun phrase, sentence syntax similarity, verb phrase incidence, infinitive density, and negation density, see Table 4.

Both *the average number of modifiers per noun phrase* and *sentence syntax similarity* are related to syntactic complexity. Pairwise comparisons revealed that the outputs of interpreters had fewer modifiers per noun phrase (Human: 1.05 ± 0.15; Google: 1.11 ± 0.19; Baidu: 1.10 ± 0.20) but higher syntax similarity (Human: 0.104 ± 0.024; Google: 0.088 ± 0.017; Baidu: 0.092 ± 0.016). Given that the syntax in the text is easier when there are fewer modifiers and more uniform syntactic constructions (McNamara et al., 2014), the current results indicated that human interpreters preferred sentences which were less complex and had more similar syntactic structures.

The density of particular phrase types and syntactic patterns is also indicative of syntactic complexity and certain text features. In this regard, pairwise comparisons showed that human interpreters yielded a higher density of verb phrase (Human: 207.24 ± 32.88; Google: 197.94 ± 32.37; Baidu: 194.98 ± 32.35) as well as infinitives (i.e., *to do* structures; Human: 23.20 ± 13.78; Google: 19.51 ± 10.41; Baidu: 17.20 ± 9.24), but a lower negation density (Human: 4.66 ± 4.88; Google: 7.23 ± 6.55; Baidu: 7.93 ± 7.50), see Table 4 and Fig. 3. As the crucial constituent to build the skeleton of sentences, verb phrases were more frequently yielded by the interpreters. This indicates that the interpreters prioritized skeleton building and conveying of main contents. In terms of infinitives, they are prevalent in discourses with a high density of intentional contents, where the goals of certain actions are more explicitly presented (McNamara et al., 2014). Thus, the more use of infinitives suggested that the interpreters attached importance to the demonstration of causal relations within text.

[3] An incidence score refers to occurrence per 1,000 words.

**Table 4**
ANOVA analyses and pairwise comparisons on syntactic measures.

| Index | ANOVA | | Pairwise comparisons | | | | | |
|---|---|---|---|---|---|---|---|---|
| | F | *p* | Human-Google | | Human-Baidu | | Google-Baidu | |
| | | | *p* | *d* | *p* | *d* | *p* | *d* |
| Number of modifiers per noun phrase | 5.52 | .020 | .015 | 0.47 | .048 | 0.35 | .429 | 0.13 |
| Sentence syntax similarity | 12.04 | .001 | <0.001 | 0.64 | .006 | 0.48 | .033 | 0.35 |
| Verb phrase density | 7.85 | .006 | .020 | 0.41 | .006 | 0.53 | .189 | 0.21 |
| Negation density | 11.52 | <0.001 | <0.001 | 0.61 | <0.001 | 0.67 | .327 | 0.16 |
| Infinitive density | 9.39 | .004 | .032 | 0.35 | .001 | 0.61 | .012 | 0.44 |

Note: *d* refers to the value of Cohen'd as an effect size measure.

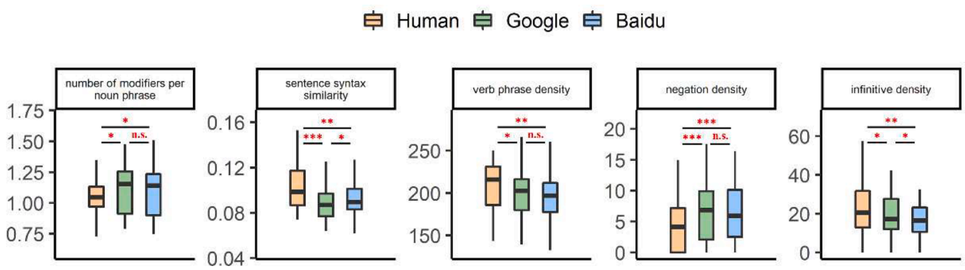


**Fig. 3.** Comparison of interpreting outputs from human interpreters, Google Translate and Baidu Translate in terms of syntactic measures. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, n.s. = non-significant.

### 4.4. Cohesive measures

Beyond the lexical and syntactic levels, Coh-Metrix 3.0 excels in quantifying text cohesion. In the present study, eight cohesion indices revealed significant differences, see Table 5.

Connectives are crucial for linking ideas and clauses, and provide clues about text organization (Cain & Nash, 2011; Millis & Just, 1994). It is shown in Fig. 4 that Human interpreters rendered more connectives in general (Human: 75.76 ± 17.62; Google: 61.30 ± 12.86; Baidu: 59.35 ± 12.63) as well as causal connectives (e.g., because, so; Human: 82.50 ± 17.66; Google: 71.39 ± 10.99; Baidu: 68.55 ± 12.39). Consistently, the HI outputs scored higher in the deep cohesion dimension of text easability (Human: 43.63 ± 24.59; Google: 32.87 ± 21.85; Baidu: 34.35 ± 23.84), which reflects the use of causal and intentional connectives within the text. The more frequent use of connectives revealed HI outperformed MT in clarifying causal relations. Example 5 provides a case where the cause-consequence relation is implicit in the SL input, but explicitly clarified through the addition of "so" in the HI output rather than the MT outputs. With the inter-clause causal relation clarified by the causal connective, it is less effortful for the audience to connect clauses and to understand that "economic relations and trade have been affected" is the result of "sluggish global economic recovery and anemic growth of global trade". What is worth noting, however, the HI outputs had the lowest incidence of negative connectives (e.g., however, but; Human: 62.21 ± 14.20; Google: 68.49 ± 12.95; Baidu: 68.10 ± 13.50), which aligned with the lower negation density in the HI outputs reported above (see Section 4.3).

**Example 5: Source input** 这些年世界经济复苏低迷，贸易增长乏力，确实中俄关系在经贸方面由于能源价格下跌等因素也受到了一些影响。

**Human output** Over the years, we have faced sluggish global economic recovery and anemic growth of global trade. **So** naturally our economic relations and trade have been affected.

**Google output** The world economic recovery has been sluggish in recent years and trade growth has been sluggish. Indeed, Sino-Russian relations have also been affected by factors such as falling energy prices.

**Baidu output** The world economic recovery has been sluggish in recent years, and trade growth has been sluggish. Indeed, the economic and trade relations between China and Russia have also been affected by factors such as the decline in energy prices.

Four additional cohesive indices reflect the situation model comprehension of text, and they pay more attention to semantic and deeper connections beyond linguistic forms (Graesser & McNamara, 2011; Zwaan & Radvansky, 1998). Among this line of indices, LSA (latent semantic analysis) overlap measures the conceptual similarity between any two text excerpts and can tap both textbase cohesion and situation model coherence (McNamara et al., 2007). The results of LSA overlap showed that the HI outputs scored higher than the outputs of two MT systems (Human: 0.31 ± 0.19; Google: 0.22 ± 0.17; Baidu: 0.25 ± 0.18). Furthermore, the HI outputs also had a higher score in causal verbal incidence (Human: 29.25 ± 8.48; Google: 23.68 ± 7.99; Baidu: 23.94 ± 7.45), causal verbs and causal particles incidence (Human: 39.17 ± 8.77; Google: 31.73 ± 8.97; Baidu: 32.30 ± 9.18), and intentional verbs incidence (Human: 18.29 ± 7.13; Google: 13.73 ± 5.70; Baidu: 15.27 ± 6.97). Causation and intentionality describe the cause-consequence relations at a broader sense, albeit the subtle distinction between the two dimensions.[4] The higher incidence of causal/intentional verbs and particles (e.g., connectives) in the HI outputs can form more cohesive contents.

### 4.5. Modulation of interpreting mode on human-machine comparison

To further examine whether interpreting mode modulated the human-machine differences in the above-mentioned 30 indices, two-way mixed ANIOVAs were employed with interpreting source (HI vs. GT vs. BT) as a within-group factor and interpreting mode (CI vs. SI) as a between-group factor. The analyses indicated significant two-way (interpreting source × interpreting mode) interactions in seven indices, indicating that most indices showed similar tendencies for CI and SI but this was not universal. Consequently, multiple pairwise comparisons were conducted to identify the human-machine differences at each level of interpreting mode (i.e., CI and SI). The results are shown in Fig. 5.

Human-machine differences reached significant levels in SI outputs exclusively when involving five indices, with the HI outputs of SI having

[4] *Intentionality* refers to the actions of animate agents in pursuit of goals, whereas *causation* refers to events, in the material world, that may or may not be driven by goals (McNamara et al., 2014).

**Table 5**
ANOVA analyses and pairwise comparisons on cohesive measures.

| Index | ANOVA | | Pairwise comparisons | | | | | |
|---|---|---|---|---|---|---|---|---|
| | F | *p* | Human-Google | | Human-Baidu | | Google-Baidu | |
| | | | *p* | *d* | *p* | *d* | *p* | *d* |
| All connectives incidence | 34.63 | <0.001 | <0.001 | 1.03 | <0.001 | 0.99 | .180 | 0.22 |
| Causal connectives incidence | 41.31 | <0.001 | <0.001 | 0.92 | <0.001 | 1.17 | <0.001 | 0.57 |
| Negative connectives incidence | 9.72 | .002 | .001 | 0.61 | .004 | 0.50 | .755 | 0.05 |
| Text Easability, deep cohesion | 4.88 | .043 | .031 | 0.43 | .067 | 0.33 | .544 | 0.10 |
| LSA overlap, adjacent sentences | 7.06 | .006 | .009 | 0.50 | .016 | 0.42 | .249 | 0.18 |
| Causal verb incidence | 13.78 | <0.001 | <0.001 | 0.60 | <0.001 | 0.65 | .714 | 0.06 |
| Causal verbs/causal particles incidence | 17.55 | <0.001 | <0.001 | 0.70 | <0.001 | 0.67 | .405 | 0.13 |
| Intentional verbs incidence | 7.69 | .006 | .004 | 0.55 | .046 | 0.35 | .062 | 0.30 |

Note: *d* refers to the value of Cohen'd as an effect size measure.

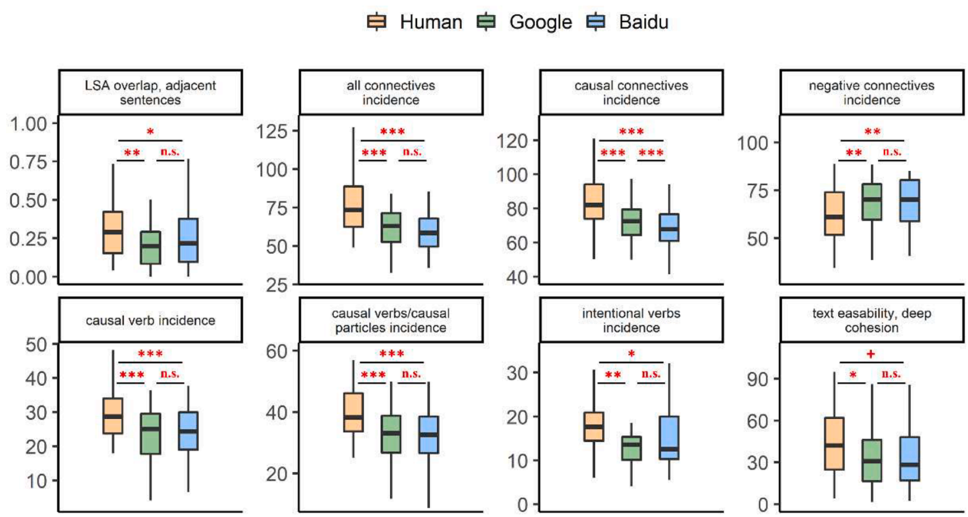


**Fig. 4.** Comparison of interpreting outputs from human interpreters, Google Translate, and Baidu Translate in terms of cohesive measures. + $p < 0.1$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, n.s. = non-significant.

shorter sentences (HI vs. GT: $p = 0.003$, $d = 0.87$; HI vs. BT: $p = 0.002$, $d = 0.92$), shorter words (HI vs. GT: $p = 0.043$, $d = 0.64$; HI vs. BT: $p = 0.043$, $d = 0.65$), a higher Flesch Reading Ease score (HI vs. GT: $p = 0.007$, $d = 0.82$; HI vs. BT: $p = 0.007$, $d = 0.84$), a lower Flesch-Kincaid Grade Level score (HI vs. GT: $p = 0.002$, $d = 0.89$; HI vs. BT: $p = 0.002$, $d = 0.93$), and a higher lexical diversity level (HI vs. GT: $p = 0.017$, $d = 1.07$; HI vs. BT: $p = 0.015$, $d = 1.14$). The first four indices suggested that human interpreters (versus MT systems) produced significantly easier outputs during SI, but such tendencies did not reach statistical significance for CI. The inconsistence between CI and SI could be rooted in their deviations in process and cognitive demands. In the face of higher difficulties in time constraint and perception-production overlap during SI (Seeber, 2011), interpreters tended to produce shorter words and sentences for buying time and resources to concurrently manage multitasks involved, but such the demand may not be so urgent during CI. Higher lexical diversity observed in the SI outputs from interpreters corroborates the previously reported pattern that SI texts outscored CI texts in lexical diversity (Lv & Liang, 2019), which could be attributed to the fact that the higher memory loads inherent to CI cause the preference for retrieving stereotypical and frequent words rather than diverse ones.

The other two indices showed significant human-machine differences only in CI rather than SI outputs. The CI outputs from interpreters had more words (HI vs. GT: $p = 0.015$, $d = 0.56$; HI vs. BT: $p = 0.016$, $d = 0.58$) and a lower negation density (HI vs. GT: $p = 0.019$, $d = 0.68$; HI vs. BT: $p = 0.004$, $d = 0.76$). The results could reflect interpreters' outperformance (versus MT) in using explicitation and lowering comprehension difficulty, but such human advantages were only prominent in CI rather than SI. Given the extreme temporal and multitasking constraints in SI, interpreters tend to activate and process smaller chunks of inputs at a time, and in this case, they keep the input sequences as much as possible. This processing style is referred to as "form-based interpreting" (Dam, 2001) or "literal translation" (Moser-Mercer, 2000), which resembles the word-to-word processing style of MT systems. On the contrary, CI formulates target speech in a relatively independent and self-paced way (Gile, 2005). Thus, CI outputs are less constrained by the forms of the source speech, and interpreters could have more room to construct outputs with both complete meaning transfer and easy reception of the audience considered.

## 5. General discussion

Results of the present study showed differences between human and machine interpreting outputs in multidimensional linguistic features, including descriptive, lexical, syntactic, and cohesive ones. Significant differences mostly lay between human interpreters and MT systems rather than between the two MT systems. The additional analyses using TAALES and TAASSC (see Supplementary Material) showed consistent patterns and complemented the findings above, which will be discussed in the following sections as well. Also, finer-grained statistical analyses indicated that the observed human-machine differences were not always the same for CI and SI (7 of 30). Since the comparison was conducted between well-trained expert interpreters and state-of-the-art MT systems, our data could point to relatively inherent strengths and limitations of human interpreters and MT systems, which will be specified below. Data-supported insights into interpreting competence currently required and interpreting education in the machine-translation age are

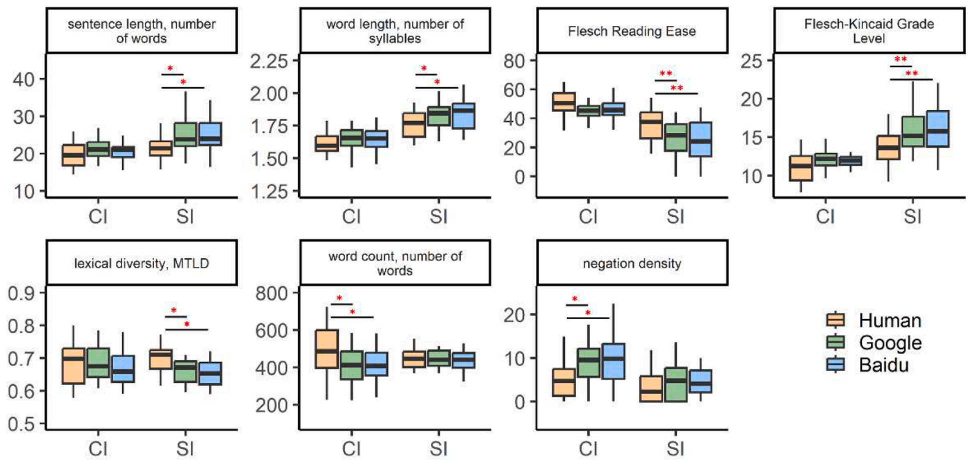


**Fig. 5.** Human-machine differences when it comes to CI and SI respectively. CI = consecutive interpreting, SI = simultaneous interpreting, $^{*}p < 0.05$, $^{**}p < 0.01$.

accordingly presented.

### 5.1. Human interpreters as better audience-oriented communication mediators

Our data indicated that human interpreters outperformed MT systems in three respects, but such outperformance was constrained by interpreting mode. First of all, human interpreters lowered the processing difficulty of comprehenders. Word and sentence length, altogether with readability measures, are indicative of text comprehension difficulty (McNamara et al., 2014). With shorter words, shorter sentences, as well as more understandable discourses (Section 4.1), the HI outputs were easier-to-process. But human-machine differences in these surface elements reached significance only in SI but not CI, possibly revealing that human interpreters' efforts in this regard were constrained by extreme temporal and coordination loads in SI (Section 4.5). As for word information, Coh-Metrix analyses (Section 4.2) and TAALES analyses (Section S2.1) showed similar results that human interpreters preferred the words that were relatively familiar and acquired at an earlier age. Word neighborhood and word recognition norms indices, included in TAALES, suggested that words in the HI outputs were generally recognized and named faster and/or more accurately (Section S2.1). Nevertheless, human interpreters used fewer negations (Section 4.3) or negative connectives (Section 4.4) which are more difficult to process during language comprehension (Darley et al., 2020; Just & Carpenter, 1971). Regarding syntactic measures, the HI outputs were also characterized by lower syntactic sophistication in terms of phrasal (Coh-Metrix and TAASSC analyses converged on the HI outputs' fewer dependents of noun phrases, Section 4.3 and Section S2.2) and syntactic complexity (indicated by the patterns concerning syntax similarity from Coh-Metrix, and those concerning the uses of parataxis and T-units from TAASSC, Section 4.3 and Section S2.2). These findings consistently suggested that the HI outputs were more audience-friendly in lowering comprehension difficulty. To note, the found simplifications in terms of lexical and syntactic measures may also derive from the interpreters' alleviation of cognitive efforts for themselves (see Section 5.2 below for detailed discussion).

Second, human interpreters outperformed MT systems in text cohesion. According to the Coh-Metrix analyses, the higher incidences of connectives, verbs and particles related to causality or intentionality revealed the interpreters' efforts in clarifying cause-consequence relations (Section 4.4). The higher occurrence of conjunction "or" (Section S2.2) showed by the TAASSC analyses provided further evidence in this regard. These corroborate the observations that interpreters tend to make implicit logic relations between lines explicit by adding causal devices, e.g., "because", "so", "so that", and "so as to" (Tang & Li, 2016; Wang, 2012). But the appropriate use of causal devices is likely beyond the capacity of MT systems due to the current AI systems' incapacities to do causal inference (Pearl & Mackenzie, 2018). Furthermore, the higher scores of LSA overlap and deep cohesion of the HI outputs indicated the human advantages in constructing semantic and deeper connections within the text (Section 4.4). The above-mentioned results point to the higher text cohesion of the HI outputs, which can benefit the audience in figuring out the semantic and logical relations between clauses/sentences (Cain & Nash, 2011; Millis & Just, 1994), and having a more coherent discourse-level representation of interpreting outputs (McNamara et al., 2007; Zwaan & Radvansky, 1998).

Third, human interpreters outperformed MT systems in contextual awareness and pragmatic competence. Our quantitative and qualitative analyses showed the interpreters' more frequent and appropriate use of explicitation (Example 1 and 2), although such a tendency was only significantly indicated in the CI outputs (Section 4.5). Krüger (2020) reported similar patterns, which were interpreted as the evidence for the compromised contextual awareness and pragmatic limitations of MT. Our results extend this study by showing that the HI outputs also excelled in the use of pronouns and words of lower frequency. On the one hand, lower score of word frequency (Section S2.1) suggested that the interpreters mastered distinguishing among synonyms and selecting semantically, culturally, and pragmatically appropriate expressions. By contrast, the MT's preference for frequent words can be attributed to the "algorithmic bias" that MT systems over-generate frequent patterns in their training set (Vanmassenhove et al., 2019). The higher word diversity in the HI outputs provided further evidence in this regard, but this pattern was only found in SI possibly because of the high memory load during CI (Section 4.5). On the other hand, through the addition of pronouns (Example 3 and 4), human interpreters, instead of MT systems, played their pragmatic role in highlighting the speakers' stances and intentions. The indices measuring contextual distinctiveness also revealed that the interpreters tended to put more contextual constraints on word meanings, benefiting disambiguation and efficient comprehension (Section S2.1). These can collectively aid the audience to accurately and adequately infer the intended and contextually-dependent meanings from the speaker.

Although different criteria have been used to assess interpreting quality, it appears that three key quality elements are commonly highlighted: (1) content (or fidelity); (2) delivery (or fluency); (3) language quality (or audience acceptability) (Han, 2018). The *content* criterion values the sense consistency between the source text and interpretations (Kurz, 1993). Our data covers some rubrics of this criterion, including accurate/complete meaning transfer and creation of cohesion and coherence, etc. (Liu, 2021). The index of word count can reflect information completeness (Ouyang et al., 2021) and our examples further revealed that the appropriate use of explicitation in the HI outputs benefited meaning and intention transfer (Example 1 and 2). Together with the better cohesion of the HI outputs (Section 4.4), it seems plausible to say the interpreters rendered higher-quality outputs in conveying the meanings and intentions of SL inputs. As for *language quality*, it refers to the extent TL expressions are correct and idiomatic, serving a better audience reception (Han, 2015; Liu, 2021). The HI outputs also outscored in this respect due to the interpreters' flexible operations (e.g., addition of subject in an English sentence, more use of pronouns and low-frequency words).

To sum up, by contrasting state-of-the-art MT systems with expert interpreters, our data systematically point to the strengths of human interpreters in the cross-cultural communicative mediation. But to note, such strengths are constrained by varied cognitive demands involved in different interpreting modes. Specifically, human interpreters can adapt outputs to the audience's easy, coherent, and contextually adequate perception of intended meanings from the source text. Whereas MT systems stick to word-to-word translation and their outputs are more dependent on the context-free aspects of inputs and statistical regularities inherent to their training data. The strengths of human have been regarded as the key for translators to go beyond machine translation and to future-proof their profession (Pym & Torres-Simón, 2021). The current findings here provide empirical evidence for the speculated added values of human (Massey & Ehrensberger-Dow, 2017), and extend some preliminary empirical efforts (Jiang & Niu, 2022; Loock, 2020; Vanmassenhove et al., 2019) with a more systematic and comprehensive investigation.

### 5.2. *Human interpreters constrained by cognitive resources*

In view of the intense nature of (human) interpreting, the lexical and syntactic simplifications revealed in the HI outputs[5] also likely reflect the interpreters' tendency to avoid possible cognitive saturation faced by themselves. Consistent with this assertion, empirical evidence, based on similar interpreting corpus, correlates high cognitive load with interpretation simplification in terms of lexical (Lv & Liang, 2019) and syntactic traits (Liang et al., 2017). Similarly, when facing high cognitive burdens, interpreters tend to prioritize the occurrence of verbs to scaffold sentence construction (Jia & Liang, 2020). The higher incidence of verb phrases in the HI outputs reported above (Section 4.3) and in the supplementary material (Section S2.2) could reveal the interpreters' adaptive allocation of processing resources in face of extreme cognitive pressures.

Furthermore, finer-grained analyses of several indices showed that the observed human-machine differences deviated in varied interpreting modes (Section 4.5). Specifically, HI outputs were shaped differently by varied cognitive demands of CI and SI, with interpreters mainly challenged by memory load during CI (Liang et al., 2017) but by extreme time pressure and multitasking load during SI (Gile, 2009; Seeber, 2011). On the one hand, the interpreters' preference for shorter words and sentences was found in the SI outputs, which may arise from the urgent need to keep up with the source speech and lower processing difficulty. But this preference was not salient when involving CI where comprehension and production are separated. Furthermore, efforts towards adequate and easy comprehension of the audience were significantly revealed when interpreters can render outputs in a self-paced way during CI rather than SI. On the other hand, the outscored lexical diversity in the HI outputs was not reported when interpreters faced increased memory load during CI. Thus, our data provide further evidence for that human interpreting is highly constrained by limited cognitive resources.

The findings echo the converging evidence that human cognition, as a limited-capacity system, is constrained and challenged in language processing in general and interpreting particularly (Liang et al., 2018). This is rooted in the biological boundaries (Liu, 2018). In comparison with MT systems as machines which work efficiently and effortlessly, the limited cognitive resources can be regarded as a comparative limitation of human during interpreting.

### 5.3. *Implications for interpreting education*

The discussions above point to the human advantages in communicative mediation but limitations in cognitive resources. This is in line with the view of (human) language as a human-driven system which adapts to both sociocultural diversity and human cognition (Liu, 2018), and emphasizes the call to highlight the added value of adaptive expertise in the translation community (Massey & Ehrensberger-Dow, 2017).

In the context of interpreting education, our findings provide important information on what competences are crucial for interpreters in the machine-translation age. On the one hand, the communicative competence in general and cross-cultural communication skills specifically, which AI-boosted machines do not master yet, deserve priority. Student interpreters should be equipped with communicative awareness, and then communication skills and cultural knowledge to achieve audience-oriented renditions in certain contexts, helping them remain superior to artificial systems. On the other hand, given that cognitive limitations are more likely to be the result of biological boundaries rather than limited skills, to complement human cognition through computer-assisted interpreting (CAI) technologies could be promising. Therefore, technological competence to work with MT is in the current demand (Wang & Li, 2022), which can extend human cognition (Pym, 2011) and relieve interpreters of the tedious, repetitive aspects of their work (O'Brien, 2012).

With the above-highlighted interpreting competences in mind, pedagogical implications can be proposed as follows. First, interpreting students need to be informed of MT basics as well as human strengths, so that they can make appropriate judgement on when and how to resort to human advantages, and when and how to embrace new technologies (Massey & Ehrensberger-Dow, 2017). Introducing academic findings (e.g., reported in this study) is useful in this regard. Also, guiding students to do self-analysis and self-induction can play a complementary role (see Bulut, 2019). Second, to ensure the interpreter trainees' abilities to make full use of human strengths, it is necessary to prioritize the cross-cultural communication competence. In more detail, it is important to develop the students' awareness of situated nature of text production and reception. Acquisition of communication strategies and cultural knowledge also needs attention in class so that students can match their language choice with contextual, pragmatic, sociocultural, and audience-related reality. Awareness-raising and skill-acquisition teaching can be employed by introducing situated interaction environments into classroom tasks (Hamilton & Woodward-Kron, 2010; Chan, 2022) and by conducting in-depth source and target text analysis (Károly, 2014). Third, abilities to utilize CAI technologies should be fostered to free up cognitive resources. To note, the utilization of technologies should match varied pain points across situations and interpreting modes. On the premise of keeping abreast of emerging tools and CAI-supported working modes, interpreting teachers should carefully select suitable ones for specific training task. For example, an AI-supported CI working mode was proposed by Chen and Kruger (2022), incorporating an automatic speech recognition (ASR) tool to replace note-taking in the conventional CI. The rationale is that the use of ASR can create a better storage of the original speech and help interpreters to handle the extreme memory loads in CI. Similarly, for SI tasks, the supports from technologies are supposed to target the time pressure and multitasking loads specific to SI. Furthermore, it is also important to guide students to reflect on the deployment of technologies

[5] These observations are also discussed in Section 5.1, which are interpreted as the evidence for the interpreters' tendency to ease comprehension difficulty for the sake of the audience.

and to use them critically, with machines' potentials and limits weighted (Massey & Ehrensberger-Dow, 2017). Finally, given that interpreting is by nature a situated and complex interaction, the authentic experiential learning and practice should be encouraged to empower students in adaptively managing their resources and competences (Kiraly et al., 2016), especially in light of the value of adaptability in the AI-dominant age.

## 6. Conclusion

The present corpus-based study quantified the multidimensional differences of interpreting outputs from human interpreters and MT systems. Our data suggested the strengths of human interpreters to achieve cross-linguistic and cross-cultural communication, but such strengths were conditioned by varied cognitive demands specific to different interpreting modes. The limitations of MT systems were revealed in narrowing down interpreting into a word-to-word information transfer. To future-proof interpreting profession in the machine-translation age, the awareness of adaptively using certain skills to ensure successful interpreter-mediated communication should be highlighted in interpreting education. Meanwhile, as the use of AI-boosted technological can complement the limited cognitive resources of human, the AI-supported human interpreting is promising, which is worth further research to develop effectively and ergonomically optimized working mode.

Despite the encouraging findings from the present study, two limitations shall be pointed out. First, interpreting is in essence a speech-to-speech activity, but our analyses focused on the transcribed texts and relied on linguistic features. Therefore, fluency as an important quality assessment criterion and auditory features which may affect interpretation reception were not covered here. Robust speech analysis of interpreting outputs is expected to complement our findings. Second, interpretations included in our corpus were all from Chinese to English and thus the results were confined to one language pair (i.e., Chinese-English) and one interpreting direction (i.e., from A language to B language). We call for further studies to cover more language pairs and both interpreting directions. It is hoped that our study could stimulate more researchers to pay attention to and conduct research on interpreting education in the rise of AI technologies.

## Funding information

The National Social Science Fund of China [grant number 17BYY068].

## CRediT authorship contribution statement

**Yiguang Liu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Junying Liang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Formal analysis, Conceptualization.

## Declaration of competing interest

None.

## Data availability

Data will be made available on request.

## Acknowledgements

We thank two anonymous reviewers for their insightful comments on earlier versions of this manuscript. Thanks also go to Dr. Qianxi Lv for polishing the manuscript.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.linged.2024.101273.

## References

Albl-Mikasa, M. (2013). Developing and cultivating expert interpreter competence. *The Interpreters' Newsletter, 2013*(18), 17–34. https://doi.org/10.21256/zhaw-4081

Angelelli, C. V. (2004). *Medical interpreting and cross-cultural communication*. Cambridge University Press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.

Bizzoni, Y., Juzek, T. S., España-Bonet, C., Dutta Chowdhury, K., van Genabith, J., & Teich, E. (2020). How human is machine translationese? Comparing human and machine translations of text and speech. In *Proceedings of the 17th international conference on spoken language translation* (pp. 280–290). https://doi.org/10.18653/v1/2020.iwslt-1.34

Cadwell, P., O'Brien, S., & Teixeira, C. S. C. (2018). Resistance and accommodation: Factors for the (non-) adoption of machine translation among professional translators. *Perspectives, 26*(3), 301–321. https://doi.org/10.1080/0907676X.2017.1337210

Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology, 103*, 429–441. https://doi.org/10.1037/a0022824

Chan, V. (2022). Using a virtual reality mobile application for interpreting learning: Listening to the students' voice. *Interactive Learning Environments.*. https://doi.org/10.1080/10494820.2022.2147958

Chen, S., & Kruger, J. L. (2022). The effectiveness of computer-assisted interpreting: A preliminary study based on English-Chinese consecutive interpreting. *Translation and Interpreting Studies*. https://doi.org/10.1075/tis.21036.che

Christoffels, I. K., de Groot, A. M. B., & Kroll, J. F. (2006). Memory and language skills in simultaneous interpreters: The role of expertise and language proficiency. *Journal of Memory and Language, 54*(3), 324–345. https://doi.org/10.1016/j.jml.2005.12.004

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Dam, H. V. (2001). On the option between form-based and meaning-based interpreting: The effect of source text difficulty on lexical target text form in simultaneous interpreting. *The Interpreters' Newsletter, 11*, 27–55.

Darley, E. J., Kent, C., & Kazanina, N. (2020). A 'no' with a trace of 'yes': A mouse-tracking study of negative sentence processing. *Cognition, 198*, Article 104084. https://doi.org/10.1016/j.cognition.2019.104084

Darò, V., & Fabbro, F. (1994). Verbal memory during simultaneous interpretation: Effects of phonological interference. *Applied Linguistics, 15*(4), 365–381. https://doi.org/10.1093/applin/15.4.365

De Clercq, O., Sutter, G., Loock, R., Cappelle, B., & Plevoets, K. (2021). Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated french. *Translation Quarterly, 101*, 21–45.

De Fina, A. (1995). Pronominal choice, identity, and solidarity in political discourse. *Text - Interdisciplinary Journal for the Study of Discourse, 15*(3). https://doi.org/10.1515/text.1.1995.15.3.379

Defrancq, B., & Fantinuoli, C. (2020). Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers. *Target, 33*(1), 73–102. https://doi.org/10.1075/target.19166.def

Dong, Y., & Li, P. (2020). Attentional control in interpreting: A model of language control and processing control. *Bilingualism: Language and Cognition, 23*(4), 716–728. https://doi.org/10.1017/S1366728919000786

Downie, J. (2019). *Interpreters vs machines: Can interpreters survive in an AI-dominated world?* Routledge.

Elfenbein, A. (2011). Research in text and the uses of Coh-Metrix. *Educational Researcher, 40*(5), 246–248. https://doi.org/10.3102/0013189x11414181

ELIA. (2022). *European language industry survey 2022–Trends, expectations and concerns of the European language industry*. European Language Industry Association. https://elis-survey.org/wp-content/uploads/2022/03/ELIS-2022-report.pdf?utm_source=elis-repository&utm_medium=website&utm_campaign=elis-report22&utm_id=elis-report-22 accessed 10 October 2023.

Fang, I. E. (1966). The "Easy listening formula". *Journal of Broadcasting, 11*(1), 63–68. https://doi.org/10.1080/08838156609363529

Fantinuoli, C. (2017). Computer-assisted preparation in conference interpreting. *Translation & Interpreting, 9*(2), 24–37.

Fantinuoli, C. (2018). Interpreting and technology: The upcoming technological turn. In C. Fantinuoli (Ed.), *Interpreting and technology* (pp. 1–12). Language Science Press.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. https://doi.org/10.1037/h0057532

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces, 6*(2), 291–309. https://doi.org/10.1075/ts.6.2.06for

Frittella, F. (2022). CAI tool-supported SI of numbers: A theoretical and methodological contribution. *International Journal of Interpreter Education, 14*, 32–56. https://doi.org/10.34068/ijie.14.01.05

Gerver, D. (1976). Empirical studies of simultaneous interpretation: A review and a model. In R. W. Brislin (Ed.), *Translation. applications and research* (pp. 165–207). Gardner Press.

Gile, D. (1999). Testing the effort models' tightrope hypothesis in simultaneous interpreting–A contribution. *HERMES - Journal of Language and Communication in Business, 12*(23), 153–172. https://doi.org/10.7146/hjlcb.v12i23.25553

Gile, D. (2005). Directionality in conference interpreting: A cognitive view. In R. Godijns, & M. Hindedael (Eds.), *Directionality in interpreting. The 'retour' or the native?* (pp. 9–26). Communication and Cognition.

Gile, D. (2009). *Basic concepts and models for interpreter and translator training.* John Benjamins Pub. Co. Rev. ed.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*(2), 371–398. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Grbić, N., & Pöchhacker, F. (2015). Competence. In F. Pöchhacker (Ed.), *Routledge encyclopedia of Interpreting studies* (pp. 69–70). Routledge.

Green, C. (2019). A multilevel description of textbook linguistic complexity across disciplines: Leveraging NLP to support disciplinary literacy. *Linguistics and Education, 53*, Article 100748, 10/gr2sbm.

Hamilton, J., & Woodward-Kron, R. (2010). Developing cultural awareness and intercultural communication through multimedia: A case study from medicine and the health sciences. *System, 38*(4), 560–568. https://doi.org/10.1016/j.system.2010.09.015

Han, C. (2015). Investigating rater severity/leniency in interpreter performance testing: A multifaceted Rasch measurement approach. *Interpreting. International Journal of Research and Practice in Interpreting, 17*(2), 255–283. https://doi.org/10.1075/intp.17.2.05han

Han, C. (2018). Using rating scales to assess interpretation: Practices, problems and prospects. *Interpreting. International Journal of Research and Practice in Interpreting, 20* (1), 63–101. https://doi.org/10.1075/intp.00003.han

Harris, B., & Sherwood, B. (1978). Translating as an innate skill. In D. Gerver, & H. W. Sinaiko (Eds.), *Language interpretation and communication* (pp. 155–170). Springer US. https://doi.org/10.1007/978-1-4615-9077-4_15.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Zhou, M. (2018). *Achieving Human Parity on Automatic Chinese to English News Translation.* 10.48550/arXiv.1803.05567.

Hatim, B., & Mason, I. (1997). *The translator as communicator.* Routledge.

Hutchins, J., & Lovtskii, E. (2000). Petr Petrovich Troyanskii (1894–1950): A forgotten pioneer of mechanical translation. *Machine Translation, 15*(3), 187–221. https://doi.org/10.1023/A:1011653602669

Jekat, S. (2015). Machine interpreting. In F. Pöchhacker (Ed.), *Routledge encyclopedia of iInterpreting studies* (pp. 69–70). Routledge.

Jia, H., & Liang, J. (2020). Lexical category bias across interpreting types: Implications for synergy between cognitive constraints and language representations. *Lingua, 239*, Article 102809. https://doi.org/10.1016/j.lingua.2020.102809

Jiang, Y., & Niu, J. (2022). A corpus-based search for machine translationese in terms of discourse coherence. *Across Languages and Cultures, 23*(2), 148–166. https://doi.org/10.1556/084.2022.00182

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior, 10*(3), 244–253. https://doi.org/10.1016/S0022-5371(71)80051-8

Kaczmarek, L. (2010). Modelling competence in community interpreting: Expectancies, impressions and implications for accreditation PhD Thesis, University of Manchester.

Kade, O. (1968). *Zufall und gesetzmäßigkeit in der übersetzung.* Verlag Enzyklopädie.

Kalina, S. (2000). Interpreting competence as a basis and a goal for teaching. *The Interpreters' Newsletter, 10*, 3–32.

Károly, A. (2014). Translation in foreign language teaching: A case study from a functional perspective. *Linguistics and Education, 25*, 90–107, 10/gr2sc7.

Kincaid, J., Fishburne, R., Rogers, R., & Chissom, B. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) for navy enlisted personnel. *Institute for Simulation and Training.* https://stars.library.ucf.edu/istlibrary/56.

Kiraly, D., et al. (2016). *Towards authentic experiential learning in translator education.* V&R Unipress, Mainz University Press.

Krüger, R. (2020). Explicitation in neural machine translation. *Across Languages and Cultures, 21*(2), 195–216. https://doi.org/10.1556/084.2020.00012

Kuo, C. (2019). Function words in statistical machine-translated Chinese and original Chinese: A study into the translationese of machine translation systems. *Digital Scholarship in the Humanities, 34*(4), 752–771. https://doi.org/10.1093/llc/fqy050

Kurz, I. (1993). Conference interpretation: Expectations of different user groups. *The Interpreters' Newsletter, 5*, 13–21.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral Dissertation]. Georgia State University.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods, 50*(3), 1030–1046, 10/gdqrvq.

Lapshinova-Koltunski, E. (2015). Variation in translation: Evidence from corpora. In C. Fantinuoli, & F. Zanettin (Eds.), *New directions in corpus-based translation studies* (pp. 93–114). Langugae Science Press.

Lee, J. (2020). Competence, interpreting. In M. Baker, & G. Saldanha (Eds.), *Routledge encyclopedia of translation studies* (pp. 84–89). Routledge.

Li, Y., & Dong, Y. (2022). Use of explicitation by interpreting students and its contribution to consecutive interpreting performance: A developmental perspective. *Perspectives, 30*(1), 103–119. https://doi.org/10.1080/0907676X.2021.1892158

Li, Y., & Halverson, S. L. (2020). A Corpus-based exploration into lexical bundles in interpreting. *Across Languages and Cultures, 21*(1), 1–22. https://doi.org/10.1556/084.2020.00001

Liang, J., Fang, Y., Lv, Q., & Liu, H. (2017). Dependency distance differences across interpreting types: Implications for cognitive demand. *Frontiers in Psychology, 8*, 2132. https://doi.org/10.3389/fpsyg.2017.02132

Liang, J., Lv, Q., & Liu, Y. (2018). Interpreting as a mirror for language foundations. *Physics of Life Reviews, 26-27*, 139–141. https://doi.org/10.1016/j.plrev.2018.06.002

Liu, H. (2018). Language as a human-driven complex adaptive system: Comment on "Rethinking foundations of language from a multidisciplinary perspective. by T. Gong et al. *Physics of Life Reviews, 26-27*, 149–151. https://doi.org/10.1016/j.plrev.2018.06.006

Liu, Y. (2021). Exploring a corpus-based approach to assessing interpreting quality. In J. Chen, & C. Han (Eds.), *Testing and assessment of interpreting* (pp. 159–178). Springer Singapore. https://doi.org/10.1007/978-981-15-8554-8_8.

Loock, R. (2020). No more rage against the machine: How the corpus-based identification of machine-translationese can lead to student empowerment. *The Journal of Specialised Translation, 34*, 150–170.

Lörscher, W. (1991). *Translation performance, translation process, and translation strategies: A psycholinguistic investigation.* Gunter Narr.

Luo, J., & Li, D. (2022). Universals in machine translation?: A corpus-based study of Chinese-English translations by WeChat Translate. *International Journal of Corpus Linguistics, 27*(1), 31–58. https://doi.org/10.1075/ijcl.19127.luo

Lv, Q., & Liang, J. (2019). Is consecutive interpreting easier than simultaneous interpreting?–A corpus-based study of lexical simplification in interpretation. *Perspectives, 27*(1), 91–106. https://doi.org/10.1080/0907676X.2018.1498531

Maamuujav, U., Olson, C. B., & Chung, H. (2021). Syntactic and lexical features of adolescent L2 students' academic writing. *Journal of Second Language Writing, 53*, Article 100822, 10/gnhng8.

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

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McNamara, D. S., Cai, Z., & Louwerse, M. (2007). Optimizing LSA measures of cohesion. In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.). *Handbook of latent semantic analysis* (pp. 379–400). Erlbaum.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292–330. https://doi.org/10.1080/01638530902959943

Millis, K. K., & Just, M. A. (1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language, 33*, 128–147. https://doi.org/10.1006/jmla.1994.1007

Moser, B. (1978). Simultaneous interpretation: A hypothetical model and its practical application. In D. Gerver, & H. W. Sinaiko (Eds.), *Language interpretation and communication* (pp. 353–368). Springer US. https://doi.org/10.1007/978-1-4615-9077-4_31.

Moser-Mercer, B. (2000). Simultaneous interpreting: Cognitive potential and limitations. *Interpreting, 5*(2), 83–94. https://doi.org/10.1075/intp.5.2.03mos

O'Brien, S. (2012). Translation as human–computer interaction. *Translation Spaces, 1*(1), 101–122. https://doi.org/10.1075/ts.1.05obr

Öner Bulut, S. (2019). Integrating machine translation into translator training: Towards 'human translator competence'? *TransLogos Translation Studies Journal, 2*(2), 1–26. https://doi.org/10.29228/transLogos.11

Ouyang, L., Lv, Q., & Liang, J. (2021). Coh-Metrix model-based automatic assessment of interpreting quality. In J. Chen, & C. Han (Eds.), *Testing and assessment of interpreting* (pp. 179–200). Springer Singapore. https://doi.org/10.1007/978-981-15-8554-8_9.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (1st ed.). Basic Books.

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.

Pöchhacker, F. (2000). *Dolmetschen: Konzeptuelle grundlagen und deskriptive untersuchungen (in German).* Stauffenburg-Verlag.

Pöchhacker, F. (2016). *Introducing interpreting studies (Second edition).* Routledge.

Pym, A. (2011). What technology does to translating. *Translation & Interpreting, 3*(1). Article 1.

Pym, A., & Torres-Simón, 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

Sawyer, D. B., & Roy, C. B. (2015). Education. In F. Pöchhacker (Ed.), *Routledge encyclopedia of interpreting studies.* Routledge.

Schnell, B., & Rodríguez, N. (2017). Ivory tower vs. workplace reality. *The Interpreter and Translator Trainer, 11*(2–3), 160–186. https://doi.org/10.1080/1750399X.2017.1344920

Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories–New models. *Interpreting, 13*(2), 176–204, 10/fgxpcc.

Seeber, K. G. (2017). Multimodal processing in simultaneous interpreting. In J. W. Schwieter, & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 461–475). Wiley. https://doi.org/10.1002/9781119241485.ch25.

Seeber, K. G., & Kerzel, D. (2012). Cognitive load in simultaneous interpreting: Model meets data. *International Journal of Bilingualism, 16*(2), 228–242. https://doi.org/10.1177/1367006911402982

Smith, M., & Taffler, R. (1992). Readability and understandability: Different measures of the textual complexity of accounting narrative. *Accounting, Auditing & Accountability Journal, 5*(4). https://doi.org/10.1108/09513579210019549

Sun, M., Jiang, B., Xiong, H., He, Z., Wu, H., & Wang, H. (2019). Baidu neural machine translation systems for WMT19. In *Proceedings of the fourth conference on machine translation (Volume 2: Shared Task Papers, Day 1)* (pp. 374–381). https://doi.org/10.18653/v1/W19-5341

Tang, F., & Li, D. (2016). Explicitation patterns in English-Chinese consecutive interpreting: Differences between professional and trainee interpreters. *Perspectives, 24*(2), 235–255. https://doi.org/10.1080/0907676X.2015.1040033

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the third conference on machine translation: Research papers* (pp. 113–123). https://doi.org/10.18653/v1/W18-6312

Toury, G. (1995). *Descriptive translation studies–and beyond*. John Benjamins.

Vanmassenhove, E., Shterionov, D., & Way, A. (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of machine translation summit XVII: Research track* (pp. 222–232). https://aclanthology.org/W19-6622.

Vinay, J. P., & Darbelnet, J. (1995). *Comparative stylistics of French and English*. John Benjamins.

Wang, B. (2012). A descriptive study of norms in interpreting: Based on the Chinese-English consecutive interpreting corpus of Chinese premier press conferences. *Meta : Journal Des Traducteurs /Meta: Translators' Journal, 57*(1), 198–212. https://doi.org/10.7202/1012749ar

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. https://doi.org/10.1080/1750399X.2022.2101850

Wang, W., Xu, Y., Wang, B., & Mu, L. (2020). Developing interpreting competence scales in China. *Frontiers in Psychology, 11*, 481. https://doi.org/10.3389/fpsyg.2020.00481

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation (arXiv: 1609.08144). arXiv. 10.48550/arXiv.1609.08144.

Zupnik, Y. J. (1994). A pragmatic analysis of the use of person deixis in political discourse. *Journal of Pragmatics, 21*(4), 339–383. https://doi.org/10.1016/0378-2166(94)90010-8

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185. https://doi.org/10.1037/0033-2909.123.2.162