Applied Linguistics, 2025, XX, 1–15
https://doi.org/10.1093/applin/amaf026
Advance access publication 3 May 2025
Article

Can sentiment analysis help to assess accuracy in interpreting? A corpus-assisted computational linguistic approach

Yujie Huang, Andrew K. F. Cheung, Kanglong Liu, and Han Xu*

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region, China

*Corresponding author. Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China. E-mail: hanxu@polyu.edu.hk

This study explores how sentiment analysis, a natural language processing technique, can help to assess the accuracy of interpreting learners' renditions. The data was obtained from a corpus consisting of 22 interpreting learners' performance over a training period of 11 weeks and comparable professional interpreters' performance used as a reference. The sentiment scores of learners' output were calculated using two lexicon-based sentiment tools and compared to the reference. The results revealed the learners' limited ability to convey the speaker's sentiment, which mainly resulted from their omission and distortion of key sentiment words and their intensity. Additionally, statistically significant correlations were found between the learner-reference sentiment gap of a given rendition and its accuracy level as perceived by the human raters, yet the extent of correlation is moderate. This suggests that the predictive power of sentiment analysis as a standalone indicator of accuracy is limited. Overall, the findings of this study have practical implications for the design of automated interpreting quality assessment tools and interpreting training.

Introduction

Accurately transferring messages from the source text to the target audience has always been a fundamental principle of the interpreting profession (Ozolins 2015). Since the emergence of interpreting studies as a distinct research field, there has been considerable interest in investigating the notion of accuracy (Cagigos 1990; Gile 1992; Pöchhacker 2004; Liu 2020). This interest stems from the recognition that accuracy plays a crucial role in facilitating effective communication across languages (Hale 2007; Xu 2021, 2024). In the early days, accuracy was perceived as the "faithful and full conveying of speakers' ideas" (Herbert 1952: 4). This definition implies the importance for interpreters to grasp the communicative intentions of the speakers in order to successfully convey their ideas to the other party. Thus, accuracy in interpreting encompasses the transmission of the intentional content of the source text.

© The Author(s) (2025). Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

Similarly, some researchers argued that accurate rendition requires interpreters to convey both the "sense" and the "style" of the message (Cagigos 1990; Gile 1992), emphasizing that interpreters should not only convey what the message is about but also how it is expressed by the speaker. In the same vein, Hale (2004, 2007) proposed that interpreters should adopt a pragmatic approach to interpreting, that is, in addition to preserving the propositional content of the source text, interpreters should maintain the pragmatic force of the message so that the same communicative effect can be created in the target language. This pragmatic approach has been recognized by professional organizations and laid the foundation for the formulation of accuracy norms in the industry (Tebble 2012). For instance, the Australian Institute of Interpreters and Translators (AUSIT)'s Code of Ethics stipulates that accurate rendition should "both preserve the content and the intent of the source message or text without omission or distortion" (AUSIT 2012: 5).

Given the critical role that accuracy plays in interpreting, adequate assessment of accuracy becomes vital for ensuring effective cross-lingual communication. The results of accuracy assessment also carry important practical implications for the interpreting profession, informing activities such as training, certification, and recruitment process (Han 2022). In light of the importance of conducting adequate assessment of accuracy, which is a major component of interpreting quality, extensive research efforts have been devoted to this area over the years. Researchers have examined accuracy and its assessment in different interpreting settings and scenarios (e.g., Lee 2005; Wang and Fang 2019; Jiménez Ivars 2020; Liu 2020). A wide array of methods, such as error analysis (Setton and Motta 2007), rubric-based scoring (Han and Shang 2022; Setton and Dawrant 2016), and comparative judgement (Han and Lu 2023), have been proposed to assess interpreting accuracy. These investigations highlighted some nuanced aspects of accuracy that need to be considered in the assessment process, such as the conveyance of illocutionary force and strategic omissions and additions to achieve accuracy (Liu and Hale 2018; Wang and Fang 2019). Notably, existing assessment approaches predominantly relied on human raters' evaluations, a method that has long been considered time-consuming and labour-intensive (Han 2022). Moreover, due to the inherently subjective nature of human-mediated assessment, its outcomes can be susceptible to the influence of numerous rater-related factors, such as rater expertise and experience, inter-rater variability, cognitive bias, fatigue and attention limitations (Mead 2005; Liu 2013; Han 2022). Given the numerous shortcomings of rater-based assessment, there is a pressing need for an objective approach to accuracy assessment. Although this approach may not entirely replace human raters, it holds the potential to complement existing methods, enhancing reliability and cost-effectiveness.

In consideration of this research backdrop, the present study proposes a novel approach to assessing accuracy in interpreting using sentiment analysis, a method that utilizes natural language processing techniques to analyse the sentiment polarities in a given text. It explores the possibility of developing an automated accuracy assessment approach from an interdisciplinary perspective. The data was obtained from a corpus consisting of interpreting learners' renditions in class and comparable professional interpreters' performance used as a reference. Adequate quality assessment is particularly important in interpreting training (Lee 2005; Han 2018; Han and Fan 2020). If less labour-intensive and more objective assessment tools can be used, they will allow trainers to provide immediate, comprehensive, and consistent feedback to help learners pinpoint their issues and seek solutions. This makes the present study relevant as the findings will contribute to the practice of interpreting quality assessment in interpreting training by exploring the potential of integrating sentiment analysis into the quality assessment process. Following this Introduction, the second section presents a brief review of the existing approach to accuracy assessment in interpreting. The third section introduces the concept of sentiment analysis. The fourth section describes the corpus used in this study and the data analysis method. The fifth and the sixth sections present the results and discussion, respectively. The last section concludes the study by summarising the key findings and pointing out limitations.

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

Assessing accuracy in interpreting

Accuracy has long been considered a major indicator of interpreting quality (Han 2022; Pöchhacker 2001). A popular approach to examine accuracy in previous research is to conduct error-based analysis of interpreting output by spotting and categorising interpreting errors (Gile 2009; Lee 2008; Su 2019; Turner, Lai and Huang 2010). This approach is effective in detecting error-related accuracy issues, such as omissions, additions, and distortions, providing useful feedback on an interpreter's performance. However, at the same time, it may also run the risk of neglecting the conveyance of the speaker's communicative intention, in other words, the pragmatic force of the utterance. A rendition that appears to be accurate at the semantic level may not contain the same illocutionary force of the original utterance (Hale 2007; Liu 2020). Therefore, some researchers attempt to add a pragmatic dimension to the measurement of accuracy to reflect its theoretical conceptions (Wang and Fang 2019; Liu 2020; Hale et al., 2022a, 2022b). Wang and Fang (2019), in a study that compared one professional interpreter's performance in onsite and remote interpreting, developed a more refined meaning unit-based accuracy assessment framework. A meaning unit is defined as a 'clause' in Halliday's (Halliday and Matthiessen 2013) term, representing a grammatical structure that not only serves as an independent syntactic entity but also has semantic meaning to form a message. Unlike the error-based assessment approach, Wang and Fang added strategic addition and strategic omission in the categories to show interpreter's coordination efforts, such as offering cultural explanations and explicitly stating the implied meanings in the source text, to facilitate the successful transfer of the speaker's communicative goal. Likewise, in an investigation of interpreter's performance in simulated police interviews, apart from examining the accuracy of propositional content, Hale and colleagues (Hale et al., 2022b) also considered whether the interpreter had accurately conveyed the speaker's manner and style by maintaining the original utterance's discourse markers, such as tone, intonation, hesitations, and repetitions.

Taking a different perspective, questioning the very subjective nature of accuracy assessment and its shortcomings, some researchers propose to use comparatively more objective assessment methods and automate the quality assessment process (Yu and van Heuven 2017; Ouyang, Lv and Liang 2021; Han and Lu 2023; Lu and Han 2023). Yet, in spite of its potential to assess quality in an efficient, objective, and affordable way, automated interpreting quality assessment as an area of research is "in its infancy, with many of its much-touted benefits being slow to materialise" (Han 2022: 40). Empirical investigations of automated solution appeared only a few years ago with a very limited number of studies. These studies demonstrate a strong interdisciplinary orientation, borrowing methodological tools and analysis methods from second language learning, computer science or statistics (Yu and van Heuven 2017; Steward et al., 2018; Ouyang, Lv and Liang 2021; Han and Lu 2023; Lu and Han 2023). While perceiving automated quality assessment as a substitute for human assessment remains a subject of controversy (Lu and Han 2023), numerous proposed methods have shown predictive capabilities in capturing at least certain aspects of interpreting quality. For instance, departing from a computational linguistic perspective, a group of researchers perceive that some aspects of interpreting quality can be predicted by the linguistic or paralinguistic characteristics of interpreted speech, such as acoustic features (Yu and van Heuven 2017) or textual-linguistic features (Ouyang, Lv and Liang 2021). For example, using Coh-Metrix, a text analysis system, Ouyang, Lv and Liang (2021) showed that a linear progression model with four entries variables (word count, lexical diversity, hypernymy of verbs and frequency of first person singular) can predict 60% of the variance in human scoring. In a divergent direction, some researchers explored the potential application of automatic assessment metrics used for machine translation in interpreting assessment (Chung 2020; Han and Lu 2023; Lu and Han 2023). For instance, Lu and Han's (2022) attempted to correlate the scores of five representative metrics—BLEU, NIST, METEOR, TER and BERT, with the human-assigned ones to see the validity of this method. Their findings suggest a moderate-to-strong correlation between three automated machine translation quality assessment metrics and human assessment in different assessment scenarios.

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

Sentiment analysis

The present study will explore the potential of sentiment analysis to assess accuracy in interpreting. Sentiment analysis, also known as opinion mining, is a field of natural language processing that involves determining the subjective information, such as opinions, feelings, evaluations, and attitudes, contained in a piece of text (Liu and Lei 2018; Poria et al., 2020; Lei and Liu 2021; Liu 2022). The primary goal of sentiment analysis is to identify the emotional tone of a given text by categorising its emotional disposition as positive, negative, or neutral.

There are two primary approaches to conducting sentiment analysis: lexicon-based and machine learning-based methods. The lexicon-based method depends on automatically or semi-automatically built or hand-ranked sentiment dictionaries that contain sentiment words with a given score. These dictionaries are often referred to as lexicons. The scores are calculated using specific rules or algorithms to determine the sentiment and its intensity associated with each word (Hu and Liu 2004; Esuli and Sebastiani 2006). Popular sentiment lexicons include *Liu & Hu*, *VADER*, *MPQA* (Hu, and Liu, 2004; Hutto, and Gilbert, 2014; Khoo and Johnkhan 2018). For this method, the way a lexicon is created and the rules according to which a sentiment score is assigned determine the sentiment analysis outcome (Lei and Liu 2021). For example, in an attempt to build a lexicon that considers the pragmatic dimension of language in use, Taboada et al., (2011) not only included words with semantic orientation annotations (polarity and strength) but also took into account the impact of sentiment intensification, which is mainly manifested by adjectives, adverbs (e.g. very, slightly) and some allocations like "a great deal of", as well as negation (e.g. not, none, nobody, never). The completed lexicon is more capable of processing sophisticated contextual information and delivering consistently accurate analysis results across domains. As for the machine learning-based approach, its polarity classification relies on training the classifier to learn from existing sentiment-labelled datasets. After the training process, the classifier is used to give sentiment scores to unlabelled data to test its performance (Jain and Dandannavar 2016). Compared to the lexicon-based method, the machine learning approach tends to have better performance within a specific domain. This is because the classifier is trained on manually labelled datasets specific to that domain. However, the machine learning classifier is susceptible to domain variation and the absence of contextual information. While it performs well within the domain it was trained on, its accuracy may significantly decline when applied to a different domain (Gamon 2005). Meanwhile, when having no access to sufficient contextual information, the accuracy of categorizing the polarity of sentiment would be undermined (Khoo and Johnkhan 2017).

The past decade has witnessed extensive application of sentiment analysis in various domains, such as finance, politics, and education, to address a wide range of practical problems (see Liu 2022). For example, sentiment analysis is frequently used to analyse social media data, such as Twitter posts and comments, to help companies gather customer feedback comments, understand market trends or monitor brand perceptions (Martínez-Cámara et al., 2014). In contrast, the application of sentiment analysis to tackle issues in language studies and related fields is only recent (Taboada 2016; Liu and Lei 2018; Jacobs et al., 2020; Wen and Lei 2022). For example, Jacobs et al., (2020) employed sentiment analysis to verify the Pollyanna hypothesis, a concept in psychology that describes a universal tendency for people to use positive words more frequently in their perception of events. Based on two corpora comprising children and youth literature, their findings supported the universality hypothesis and confirmed the validity of introducing sentiment analysis to the scientific study of literature. In another study, Wen and Lei (2022) adopted lexicon-based sentiment analysis to investigate the presence of linguistic positivity bias in academic writing. They conducted a comprehensive analysis of a substantial corpus comprising abstracts published in 123 scientific journals over a span of 50 years. The findings affirmed the existence of an overall linguistic positivity bias throughout the examined timeframe. Notably, the researchers observed a growing tendency among scholars to adopt a more positive tone in their abstracts, driven by various objectives, such as enhancing publication prospects, promoting research outcomes, and adhering to the principle of political correctness.

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

Considering the effectiveness of sentiment analysis in catching the semantic polarity of a given text, it carries the potential to become a useful tool for analysing interpreted speech. Conveying the speaker's communicative intention, which may include their attitudes, emotions, opinions, and sentiments, is an essential aspect of accuracy rendition Hale (2007). Therefore, accuracy in interpreting should include a complete transfer of the speaker's sentiment. Seen from this perspective, variations in the interpreter's conveyance of the speaker's sentiment may have implications for the achievement of accuracy. Yet very little is written about how sentiment analysis may be used to examine translational language. To the best of the authors' knowledge, only an unpublished doctoral dissertation (Liu 2023) adopted sentiment analysis to analyse three translated versions of Fairy Tales of Oscar Wilde. The study showed that the sentiment conveyed in the three translated works varied due to some social and historical background, providing an additional dimension to account for translators' different approaches and strategies.

To address this research gap, this study aims to explore how sentiment analysis can be utilized to assess accuracy in interpreting. The data was obtained from a corpus that includes interpreting learners' performance and that of professional interpreters, which serves as a reference. Specifically, this study will first investigate the extent to which learners convey the speaker's sentiment compared to the reference. Given learners may be less professionally competent in achieving accuracy, including the conveyance of the sentiment polarity of the message (Hale 2007; Lee 2005; Liu and Hale 2018), it is anticipated that a gap may exist between the sentiment scores of the learners' renditions and those of the reference. If this gap is confirmed, the study will explore what factors may lead to this gap. After that, this study explores how the learner-reference gap of a given rendition is associated with its accuracy level. This will involve examining the correlation between the gap of a given rendition and its accuracy level as perceived by human raters. A narrower gap indicates that the learner was able to convey a similar amount of sentiment as the reference, which may suggest a higher ability to achieve accuracy. Following this line of enquiry, this study will address the following research questions.

RQ1: How do the sentiment scores of learners' renditions vary from those of the reference?
RQ2: If there is a gap between the sentiment score of a learner's rendition and that of the reference (learner-reference sentiment gap), what factors cause this gap?
RQ3: How does the learner-reference sentiment gap of a given rendition correlate with its accuracy scores as perceived by human raters?

Methodology

Data description

The data used in this study was obtained from the Chinese-English Simultaneous Interpreting Learners Corpus (CESIL). CESIL consists of 22 interpreting learners' in-class performance in an 11-week advanced interpreting course. The 22 learners, involving 18 females and four males, were enrolled in a one-year Master's programme at the Hong Kong Polytechnic University. All of them were native Chinese speakers with high English proficiency. Prior to taking the advanced course, they had received one semester of interpreting training. The data used were the learners' renditions of 11 speeches delivered at United Nations Security Council (UNSC) meetings. These speeches were given by Chinese representatives to express opinions and attitudes on varying global peace and security issues, such as the crisis in Ukraine, the use of chemical weapons in Syria and women's rights in Afghanistan. Interpreting learners were required to simultaneously interpret the speeches into English. In addition, the CESIL also included professional interpreters' renditions of the same 11 speeches. The 11 speeches and the professional interpreters' renditions were obtained from the United Nations Digital Library. Due to the high-stakes nature of these meetings, interpreters were able to access the speech script prior to their interpreting assignment and make preparations to maintain optimum interpreting quality (Cheung 2019; Xu and

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

Liu 2024). The second and the fourth authors, both professionally certified interpreters and experienced trainers, have also checked the reference to ensure its accuracy.

The recordings of the learner's performance were transcribed using iFLYTEK, an automatic transcription tool with an accuracy rate exceeding 98 percent. The transcription results were manually checked to remove distracting features, such as mispronunciations, false starts, and fillers, that could potentially impact the sentiment analysis. To provide a more fine-grained analysis of the emotional disposition expressed in the interpreted speeches, sentiment analysis was conducted at the sentence level (Bonta, Kumaresh and Naulegari 2019; Eng et al. 2021; Khan et al., 2016). This helps to pinpoint how individual sentences contribute to the overall sentiment conveyed and their accuracy level. The 11 speeches were segmented into 83 sentences. As some sentences were omitted by the learners, the CESIL has a total of 1686 interpreted sentences from the 22 learners and 83 interpreted sentences from the professional interpreters. The two groups' renditions were aligned at the sentence level, with one professional interpreter's version corresponding to 22 learners' renditions of the same sentence. This helps to facilitate comparisons of their sentiment score and level of accuracy.

Lexicon-based sentiment analysis

The lexicon-based approach was chosen over machine learning methods mainly for two reasons. Firstly, its calculation of sentiment does not depend on contextual information (Khoo and Johnkhan 2017). Secondly, this approach is versatile across various fields, including diplomatic and international affairs settings (Lei and Liu 2021; Liu 2022). In sentiment analysis-based studies, it is common to use more than one sentiment tool to enhance the reliability and validity of the results, as different tools can capture diverse aspects of sentiment and mitigate the potential bias inherent in any single method (Lei and Liu 2021; Wen and Lei 2022). It also allows researchers to test the suitability and effectiveness of each tool. To follow their practice, the study used two established lexicon-based sentiment tools, *Liu & Hu* (Hu and Liu 2004) and *VADER* (Hutto and Gilbert 2014), to analyse the sentiment of the interpreted sentences. Both lexicons have demonstrated robustness across domains and provide reliable sentiment categorization outcomes, as evidenced by previous studies (Khoo and Johnkhan 2018; Bonta, Kumaresh and Naulegari 2019). *Liu & Hu* was composed of a list of positive words and a list of negative words. It calculates a final sentiment score that reflects the percentage of sentiment variation for a given sentence. *VADER* is built on three preset lexicons, including LIWC, ANEW, and General Inquirer (Hutto and Gilbert 2014). It is capable of extracting positive, negative, and neutral polarity and calculating a score that ranges from -1 (most negative) to + 1 (most positive) by adding the equivalent values of each detected word in the lexicon. The accuracy of these two tools in the Stanford dataset is 65% and 72%, respectively, outperforming other popular lexicons such as SentiWordNet (Al-Shabi 2020). These two sentiment analysis tools are readily accessible within Orange, a user-friendly data mining toolkit.

Rater-mediated accuracy assessment

To testify the validity of employing sentiment analysis in interpreting accuracy assessment, this study investigates how the learner-reference sentiment gap of a given rendition is associated with its level of accuracy as perceived by humans. Two independent raters, who are systematically trained professional interpreters and hold certifications from China Accreditation Test for Translators and Interpreters (CATTI), were recruited to assess the accuracy of the 22 learners' renditions. In the assessment tasks, the raters were required to assign two holistic scores to each interpreted sentence: one for the conveyance of propositional content and another for the conveyance of pragmatic force, using a customised accuracy rubric. The rubric was adapted with reference to the rating scales used by Hale et al. (2022a) and Wang and Fang (2019) to include both the semantic and pragmatic aspects of accuracy (Hale 2007). Specific criterion descriptors and subcategories are summarized in Table 1[1].

Prior to performing the formal rating tasks, there was a training session in which the researchers introduced the notion of accuracy, its two fundamental dimensions, criterion descriptors of

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

the rubric, and examples of inaccurate renditions to the two raters. Each rater was assigned benchmarked sample sentences to practice the use of the rubric. When there were sentences that received very different scores, the two raters were required to have a discussion to ensure the same assessment approach was consistently applied between them. This process helped reduce the potential rater effect on the assessment outcome. To mitigate the halo effect, all sentences were randomized before the assignment (Myford and Wolfe 2003). Considering two separate scores were provided by raters, the inter-rater reliability was computed for each sub-scale. The two separate scores were combined into a composite score to represent the overall accuracy level of the interpreted sentence. The average scores assigned by the two assessors will be used for correlation analysis to explore the association between the sentiment score of interpreted speech and its level of accuracy.

Results

The sentiment scores of learners' rendition as compared to the reference

The present study used two sentiment analysis tools, *Liu & Hu* and *VADER*, to calculate the sentiment score of each interpreted sentence. As the two tools use different scoring scales, the analysis outcomes were converted into z-scores to normalise the values. The normalised sentiment scores were used to calculate the Cronbach's Alpha coefficient, which yielded a result of 0.772, surpassing the 0.7 benchmark. This result shows that the sentiment values calculated by these two tools were consistent with each other, confirming the reliability of these two tools.

To explore how the sentiment scores of learners' renditions differ from those of professional interpreters when they interpret the same source text, the study compared the sentiment scores of the learners to the reference across 83 sentences. A one-sample T-test was performed for each of the 83 sentences to compare the learners' sentiment scores to the reference. The tests revealed a statistically significant difference between the two groups in 61 percent and 65 percent of the comparison tests when sentiment values were calculated by *Liu & Hu* and *VADER*, respectively. This is shown in Figure 1. This result shows that over 60 percent of the sentences interpreted by learners did not convey the same amount of sentiment found in the professional interpreter's version. Since the professional interpreter's rendition was used as a gold standard, this finding may imply that learners' ability to convey the speaker's sentiment was only limited.

To further explore what contributed to learners' inadequate conveyance of the speaker's sentiment, this study conducted a manual accuracy analysis of the interpreted sentences with sentiment gaps. These are sentences for which the learners' renditions obtained statistically different sentiment scores compared to the reference. A total of 748 sentences were identified. To provide a more fine-grained analysis, each sentence was manually segmented into small chunks, with each

Table 1. Assessment criteria of accuracy

Aspects of accuracy	Criterion descriptors	Mark	Weight
1. Conveyance of propositional content	The interpreter maintains the propositional content of the utterance, that is, the semantic meaning on the surface or 'what' the speaker said. Instances of unjustified omission, addition, and distortion will lead to penalty points.	10	70%
2. Conveyance of pragmatic force	The interpreter maintains the pragmatic force of the utterance to convey the speaker's communicative intention and contextual information. This may include the speaker's speech style, sentiment, tone, use of figurative language, and illocutionary point. Instances of omission, addition, and distortion will lead to penalty points.	10	30%

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

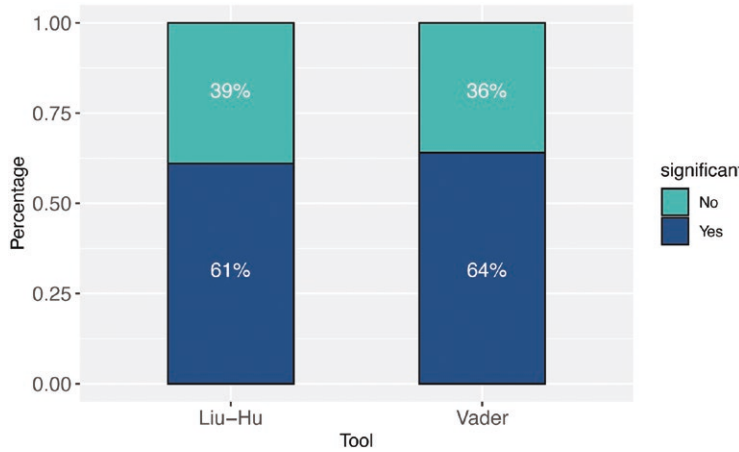

Figure 1 Percentage of interpreted sentences with different sentiment scores

chunk containing at least one noun phrase and one verb phrase. These small chunks were seen as the suitable processing segment for simultaneous interpreters to monitor before they start encoding (Goldman-Eisler 1972). After segmentation, each chunk in the sentence was manually coded to identify learners' failure to convey sentiment-related information units, which could be words or phrases. Three common error types, omission (OM), addition (AD), and distortion (DS) (Barik 1975; Lee 2008), were used as codes. For instance, if the coder found an unjustified addition of sentiment-related information, a code (AD) was added at the end of the chunk. The coding process involved three coders who hold Master's degree in translation and interpreting. Before the assignment, the coders were invited to attend a training session to be familiar with the coding process, the basic concept of interpreting accuracy, and examples of three types of errors pertinent to sentiment transfer. To ensure accurate coding, the coders finished one-third of the work independently each time and then discussed the results until an agreement was reached. Based on the coding results, three error types were quantified by counting the number of their occurrence in each sentence. The occurrence rate of each error type was calculated by dividing the frequency of each code by the total number of sentiment-related information units in that sentence.

Pearson's correlation coefficients were calculated to explore the relationships between the occurrence rate of each error type and the sentiment gap of one sentence. The results are summarised in Table 2. For the sentiment gaps calculated by *Liu & Hu*, no statistically significant correlation was found with any of the three error types. This result seems to indicate that *Liu & Hu* may not sufficiently capture the nuances of the sentiment-related information that learners failed to convey in their renditions. One possibility is that the sentiment aspects that this tool concentrates on may be less sensitive to the types of interpreting errors learners made. For the sentiment gaps calculated by *VADER*, statistically significant correlations were found in all types of errors. The prominently larger correlation efficiency shown between the sentiment gap and omission, as well as distortion, seems to suggest that learners' failure to fully convey the speaker's sentiment is likely to stem from unjustified omissions and distortions, highlighting the importance of these two error types in explaining sentiment gaps. This can be illustrated in Example 1 below. In addition, these findings also suggest that different sentiment analysis tools vary in their ability to detect sentiment gaps that are related to interpreting accuracy.

Example 1 below shows three learners' and one professional interpreter's rendition of the same source text. Under the analytical framework of opinion mining (Liu 2022), "任何人在任何情况下使用化学武器" (the use of chemical weapons by anyone under any circumstances) serves as the entity of the sentence. The phrase, "**坚决反对**" (firmly oppose), is the opinion that determines the

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

Table 2. Bivariate correlations between error types and sentiment gap

	OM%	AD%	DS%
Sentiment gap (*Liu & Hu*)	0.036	-0.022	0.043
Sentiment gap (*VADER*)	0.214***	-0.094***	-0.164***

Note. *** $P < 0.001$.

sentiment polarity. The professional interpreter provided an accurate rendition, accurately capturing the resolute stance of the Syrian government. However, Learner 1 interpreted the phrase as "go against", which is a distortion. It lacked the intensity conveyed by the original phrase, diminishing the strength of opposition expressed by the speaker. While Learner 2 accurately conveyed the essence of the sentiment phrase ("oppose"), the intensifier (firmly) was omitted, which softened the tone of the original utterance. Learner 3's rendition ("strongly oppose") appears to be the most accurate by keeping the semantic meaning of the source text and its strength.

Example 1

ST: 我们注意到叙利亚政府多次表示，叙方**坚决反对**任何人在任何情况下使用化武器…	**Learner 1**: We notice many times the Syria government iterates that they **go against** any usage of chemical weapon…
	Learner 2: We know that Syrian government said many times that they **oppose** anyone to use chemical weapons under any circumstances…
	Learner 3: We noticed that their governments has emphasized many times that they **strongly oppose** anyone to use chemical weapons under any circumstances…
	Professional: We know that the Syria government has announced on many occasions that they **firmly oppose** the use of chemical weapons by anyone under any circumstances…

Correlation between the sentiment score and level of accuracy

To investigate the association between a given interpreted speech's sentiment score and its level of accuracy as perceived by human raters, this study first calculated the sentiment gap between the learners and the reference. It is expected that the smaller the sentiment gap between the two, the higher the accuracy level should be for the learners' rendition. Before averaging the two raters' scores for analysis, inter-rater reliability was ensured via Cohen's kappa test for the propositional content score (kappa = 0.814, $P < 0.01$), the pragmatic force score (kappa = 0.767, $P < 0.01$), and the overall accuracy score (kappa = 0.824, $P < 0.01$). This result shows a high level of agreement between the two raters in each assigned sub-scale and the overall accuracy. Such a strong consistency may benefit from the initial training session, the clearly defined rubric, and the inter-rater calibration, which allowed the raters to develop a sufficient understanding of accuracy and apply the rubric consistently.

The level of accuracy was represented from three dimensions, namely accuracy of propositional content, accuracy of pragmatic force, and overall accuracy. As the sentiment scores were computed using two different tools, six pairs of Pearson's correlation were conducted. The results are shown in Table 3. Statistically significant negative correlations between the sentiment gap and the level of accuracy were found in all three dimensions and for both sentiment tools. Notably, *VALDER* exhibited a stronger degree of correlation between the two variables than *Liu & Hu*. This finding confirms the statistically significant correlation between the learner-reference gap of a given rendition and its accuracy level perceived by human raters, suggesting that sentiment

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

Table 3. Pearson's correlation between the sentiment gap and the level of accuracy

	Accuracy of propositional content	Accuracy of pragmatic force	Overall accuracy
Sentiment gap (Liu & Hu)	-0.168***	-0.176***	-0.172***
Sentiment gap (Vader)	-0.352***	-0.354***	-0.357***

analysis may be integrated into the interpreting quality assessment process as a valid indicator. Yet, the degree of correlation is moderate, which indicates that the predictive power of sentiment analysis as a standalone indicator to assess accuracy level is limited.

Discussion

Based on a corpus that consists of learners' simultaneous interpreting performance over a training period of 11 weeks and comparable professional interpreters' performance used as reference, this study explored how sentiment analysis can be used to assess accuracy in interpreting. Specifically, it investigated how the sentiment scores of learners' renditions vary from those of the reference, what factors cause the observed variations, and how the learner-reference sentiment gap of a given rendition correlates with its accuracy level as perceived by human raters.

Sentiment gaps between learners' output and reference

To begin with, statistically significant sentiment gaps were found for over 60 percent of the interpreted sentences in the corpus. As the professional interpreters' retentions were used as a reference due to their high level of accuracy, the existence of sentiment gaps indicates the learners' limited ability to convey the speaker's sentiment. Further analysis suggests that learners' failure to convey the speaker's intended sentiment largely resulted from their omissions and distortions of key sentiment words and their intensity. These findings support what was found in previous research on the effect of expertise on an interpreter's performance (Tang and Li 2016; Cheung 2016; Liu and Hale 2018; Stachowiak-Szymczak and Korpal 2019; Su and Li 2021; Hale et al., 2022a). According to Gile's (2009) tightrope hypothesis, simultaneous interpreters work close to saturation most of the time as they need to constantly deal with competing demands, such as keeping up with the speaker's pace and ensuring quality output. The interpreting errors found in learners' output seem to imply their struggle with real-time processing and decision-making. When facing the high cognitive load in interpreting, learners may have chosen to prioritise efficiency by trying to keep up with the speaker's pace. However, by doing so, learners had to sacrifice accuracy in capturing and conveying the nuanced sentiment expressed by the speakers. As for professional interpreters, due to years of practice and the systematic training they have received, they are likely to develop a more comprehensive understanding of accuracy and possess more cognitive resources to balance efficiency and quality (Stachowiak-Szymczak and Korpal 2019; Su and Li 2021; Hale et al., 2022a). Thus, they are more aware and capable of making informed decisions regarding sentiment representation to convey the speaker's intention and achieve accuracy.

Integrating sentiment analysis in the automated assessment of accuracy

This study also explored how the learner-reference sentiment gap of a given interpreted sentence correlates with its level of accuracy as perceived by human raters. The results revealed statistically significant correlations between the sentiment gap and the human-assigned accuracy score for both sentiment analysis tools. This finding shows that the sentiment polarity features of interpreted speech may be used to reflect certain aspects of its accuracy, confirming the validity of using lexicon-based sentiment analysis to assess accuracy. It provides corroborative evidence to previous research endeavors that attempted to assess quality based on linguistic or paralinguistic features of translation and interpreting output (Yu and van Heuven 2017; Liu 2021;

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

Ouyang, Lv and Liang 2021). This finding also lends support to the possibility of developing an automated computational approach to assess interpreting quality (Chung 2020; Han and Lu 2023; Lu and Han 2023). Integrating sentiment analysis into accuracy assessment has a sound theoretical foundation because accurate rendition includes the successful conveyance of a speaker's communicative intention (AUSIT 2012; Hale 2007), which includes the speaker's sentiments, emotions, attitudes, and opinions. Compared to the human rater's subjective assessment, sentiment analysis can be used as an objective measure and helps to ensure the consistent application of assessment criteria. Its advantages may also include instantaneity and cost-effectiveness, which are commonly cited advantages of automatic scoring (Lu and Han 2022).

Yet, it is important to note that while sentiment is a critical component of a message, it can hardly represent the full information contained in a message. In other words, sentiment analysis can only help determine whether the interpreter has successfully conveyed the semantic polarity of the source text. There are situations where two messages carry the same level of sentiment, but their semantic content is vastly different. Therefore, the predictive power of sentiment analysis is limited compared to other automated assessment approaches that attempt to provide a more comprehensive depiction of the semantic content of the rendition. For instance, Lu and Han (2023) found a strong correlation between BLEU, an automated metric for machine translation, and human-assigned scores. The results of these studies indicate that accuracy in interpreting is a complex construct (Cagigos 1990; Gile 1992; Hale 2004). An examination of a single linguistic aspect of interpreting output is not sufficient to determine its overall accuracy level. To ensure an adequate representation of accuracy from a computational linguistic perspective, automated tools should be able to automatically extract linguistic features that represent the multiple dimensions of accuracy. Moreover, sentiment analysis has its own limitations, such as its inability to capture the nuances of cultural references, non-verbal cues, and context-specific factors that contribute to accurate interpreting (Lei and Liu 2021; Liu 2022). These limitations may affect its accuracy and reliability. Therefore, sentiment analysis can only serve as a complementary tool to accuracy assessment rather than becoming the sole indicator of accuracy.

Implications for interpreting training

Furthermore, the findings of this study have practical implications for interpreting training. The notable disparity observed between learners and the reference in conveying the sentiment of source speeches highlights the need for learners to enhance their skills in accurately transferring the speaker's intention, including both its sentiment polarity and strength. In addition, using sentiment analysis as a complementary tool to assess accuracy, interpreting trainers can quickly identify whether learners accurately convey the intended sentiment of the speaker. This allows trainers to provide objective feedback in a timely manner. This feedback promotes learners' self-reflection and self-assessment, which helps to improve their understanding of the conception of accuracy and fosters a heightened sensitivity to emotional expression in language. This may motivate learners to develop relevant skills and strategies to maintain the pragmatic dimension of source speech content, contributing to the enhancement of their professional competence. Interpreting trainers may also consider integrating sentiment analysis in curriculum design to create interpreting exercises with different sentiment polarities for learners' practice.

Conclusion

The search for an automated approach to assess quality has always been a fascinating topic in translation and interpreting studies, garnering increasing academic attention (Yu and van Heuven 2017; Han and Lu 2023; Lu and Han 2023). The limited number of research in this line has shown that certain aspects of quality can be automatically assessed using objective measures, thus making the development of a fully automated tool a plausible objective. Situated within this context, the present study explored how sentiment analysis, a natural language processing technique, can be used to assess accuracy, a major indicator of interpreting quality. The results

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

largely confirmed the effectiveness of using sentiment analysis to examine learners' ability to achieve accuracy. Yet, the predictive power of sentiment analysis is only limited, which means it cannot be used as a standalone indicator to predict accuracy. Future research can explore the combination of sentiment analysis with other automatic assessment approaches to conduct a more comprehensive examination of interpreting quality.

The present study is not without its limitations. Firstly, it adopted one sentiment analysis method, namely the lexicon-based approach, which has its own limitations in accurately capturing the complexity of sentiment (Liu 2022). Secondly, the present study only tested one assessment scenario, including one rater type and one scoring method. Given the multifaceted nature of quality assessment, the results may be different when other rater types and scoring methods are involved (Liu 2013; Han 2018). Thirdly, this study focused on one specific interpreting setting, that is learner's simultaneous interpreting performance in a training context. Ongoing research efforts are much needed to test the application of other sentiment analysis methods, such as the machine learning-based approach, to examine interpreting outputs that involve different settings, language pairs, modes, and interpreters of varying qualifications. It would also be interesting to include more than one assessment scenario to enhance the robustness and reliability of the results (Lu and Han 2023).

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Notes on Contributors

Yujie Huang is a PhD student at the Hong Kong Polytechnic University. Her primary research interests include interpreting pedagogy, assessment, and evaluation, interpreting cognition, and interpreting technology.

Andrew K.F. Cheung specializes in empirical approaches to translation studies and interpreting studies. His research has been featured in scholarly journals such as Interpreting, Perspectives, Lingua, Babel, and the International Journal of Specialized Translation. He also serves as the Associate Editor of Babel, Translation Quarterly, and Humanities and Social Sciences Communications.

Kanglong Liu specializes in empirical approaches to translation studies, translation teaching, corpus-based translation research, and Hongloumeng research. His research has been featured in scholarly journals such as Target, Perspectives, Lingua, Language Sciences, International Journal of Specialized Translation, and System.

Han Xu is interested in conducting interdisciplinary studies to empirically investigate different aspects of interpreting and translation activity, such as issues related to quality, ethics, training and professionalism. Her research works are published in scholarly journals in the fields, such as Across Languages and Cultures, Lingua, Meta, Multilingua, Perspectives, Translation & Interpreting, Translation and Interpreting Studies, and Chinese Translators Journal.

Funding

This study was funded by The Hong Kong Polytechnic University (Projects No. P0043847, P0051009, I-8AK3).

Notes

[1] This rubric does not apply to a situation where a rendition is semantically accurate but loses most or all of the speaker's communicative intention. In this type of situation, the raters were instructed to concentrate on assessing whether the rendition had transferred the speaker's intention. Yet, this type of situation is rare in the present study.

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

References

Al-Shabi, M. A. (2020) 'Evaluating the Performance of the Most Important Lexicons Used to Sentiment Analysis and Opinions Mining', *International Journal of Computer Science and Network Security*, 20: 2–3.

AUSIT Code of Ethics and Code of Conduct. 2012. Retrieved from <https://ausit.org/wp-content/uploads/2020/02/Code_Of_Ethics_Full.pdf>, accessed 12 Feb. 2024.

Barik, H. C. (1975) 'Simultaneous Interpretation: Qualitative and Linguistic Data', *Language and Speech*, 18: 272–97. https://doi.org/10.1177/002383097501800310

Bonta, V., Kumaresh, N., and Naulegari, J. (2019) 'A Comprehensive Study on Lexicon Based Approaches for Sentiment Snalysis', *Asian Journal of Computer Science and Technology*, 8: 1–6.

Cagigos, C. 1990. *La fidélité en interprétation (Thèse de doctorat)*. Paris 3: Université Sorbonne Nouvelle.

Cheung, A. 2019. 'The hidden curriculum revealed in study trip reflective essays', in D.B. Sawyer, V. Enríquez Raído and F. Austermühl (Eds.) *The Evolving Curriculum in Interpreter and Translator Education*, pp. 393–408. Amsterdam/Philadelphia: John Benjamins.

Cheung, A. (2016) 'Paraphrasing Exercises and Training for Chinese to English Consecutive Interpreting', *FORUM / Revue internationale d'interprétation et de traduction / International Journal of Interpretation and Translation*, 14: 1–18. https://doi.org/10.1075/forum.14.1.01che

Chung, H. Y. (2020) 'Automatic Evaluation of Human Translation: BLEU vs. METEOR', *Lebende Sprachen*, 65: 181–205. https://doi.org/10.1515/les-2020-0009

Eng, T., Nawab, M. R. I., and Shahiduzzaman, K. M. (2021) 'Improving Accuracy of the Sentence-level Lexicon-based Sentiment Analysis Using Machine Learning', *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, 3307: 57–68.

Esuli, A. and F. Sebastiani. 2006. 'SENTIWORDNET: a publicly available lexical resource for opinion mining'. Proceedings of the 5th international conference on language resources and evaluation, pp. 417–22. Genoa, Italy: European Language Resources Association (ELRA).

Gamon, M. 2005. 'Pulse: mining customer opinions from free text'. *Advances in Intelligent Data Analysis VI: 6th International Symposium on Intelligent Data Analysis. Presented at the IDA 2005*, Berlin Heidelberg: Springer.

Gile, D. 2009. *Basic Concepts and Models for Interpreter and Translator Training. (Rev. ed.)*. Amsterdam: John Benjamins.

Gile, D. (1992) 'Predictable Sentence Endings in Japanese and Conference Interpretation', *The Interpreters' Newsletter*, Special Issue 1: 12–23.

Goldman-Eisler, F. (1972) 'Segmentation of Input in Simultaneous Translation', *Journal of Psycholinguistic Research*, 1: 127–40. https://doi.org/10.1007/BF01068102

Han, C. (2022) 'Interpreting Testing and Assessment: A State-of-the-Art Review', *Language Testing*, 39: 30–55. https://doi.org/10.1177/02655322211036100

Han, C. (2018) 'Using Rating Scales to Assess Interpretation: Practices, Problems and Prospects', *Interpreting*, 20: 59–95. https://doi.org/10.1075/intp.00003.han

Han, C., and Fan, Q. (2020) 'Using Self-Assessment as a Formative Assessment Tool in an English-Chinese Interpreting Course: Student Views and Perceptions of its Utility', *Perspectives*, 28: 109–25. https://doi.org/10.1080/0907676x.2019.1615516

Han, C., and Lu, X. (2023) 'Can Automated Machine Translation Evaluation Metrics be Used to Assess Students' Interpretation in the Language Learning Classroom?', *Computer Assisted Language Learning*, 36: 1064–87. https://doi.org/10.1080/09588221.2021.1968915

Hale, S. (2007) 'The Challenges of Court Interpreting: Intricacies, Responsibilities and Ramifications', *Alternative Law Journal*, 32: 198–202.

Hale, S., Goodman-Delahunty, J., Martschuk, N., and Lim, J. (2022a) 'Does Interpreter Location Make A Difference? A Study of Remote vs Face-to-face Interpreting in Simulated Police Interviews', *Interpreting*, 24: 221–53.

Hale, S., Goodman-Delahunty, J., Martschuk, N., and Doherty, S. (2022b) 'The Effects of Mode on Interpreting Performance in a Simulated Police Interview', *Translation and Interpreting Studies*, 17: 264–8. https://doi.org/10.1075/tis.19081.hal

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

Halliday, M. A. K., and Matthiessen, C. M. (2013) *Halliday's Introduction to Functional Grammar*. Milton Park, Abingdon, Oxfordshire: Routledge.

Han, C., and Shang, X. (2022) 'An Item-based, Rasch-calibrated Approach to Assessing Translation Quality', *Target*, 35: 63–96.

Herbert, J. 1952. *The Interpreter's Handbook: How to Become a Conference Interpreter*. Genève: Librairie de l'Université.

Hu, M. and B. Liu. 2004. 'Mining opinion features in customer reviews'. AAAI'04: Proceedings of the 19th National Conference on Artifical Intelligence, pp. 755–60.

Hutto, C., and Gilbert, E. (2014) 'VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text', *Proceedings of the International AAAI Conference on Web and Social Media*, 8: 216–25. https://doi.org/10.1609/icwsm.v8i1.14550

Jacobs, A. M., et al. (2020) 'Sentiment Analysis of Children and Youth Literature: Is There a Pollyanna effect?', *Frontiers in Psychology*, 11: 574746. https://doi.org/10.3389/fpsyg.2020.574746

Jiménez Ivars, A. (2020) 'Impartiality and Accuracy as a Case in Point While Interpreting in a Refugee Context', in *FORUM*, pp. 150–178. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Khan, M., Durrani, M., Ali, et al. (2016) 'Sentiment Analysis and the Complex Natural Language', *Complex Adaptive Systems Modeling* 4: 1–19. https://doi.org/10.1186/s40294-016-0016-9

Khoo, C. S., and Johnkhan, S. B. (2017) 'Lexicon-Based Sentiment Analysis: Comparative Evaluation of Six Sentiment Lexicons', *Journal of Information Science*, 44: 491–511. https://doi.org/10.1177/0165551517703514

Lee, J. (2008) 'Rating Scales for Interpreting Performance Assessment', *The Interpreter and Translator Trainer*, 2: 165–84. https://doi.org/10.1080/1750399x.2008.10798772

Lee, Y. H. (2005) 'Self-Assessment as an Autonomous Learning Tool in an Interpretation Classroom', *Meta*, 50: 5–6.

Lei, L. and D. Liu. 2021. *Conducting Sentiment Analysis*. Cambridge: Cambridge University Press.

Liu, B. (2022). *Sentiment Analysis and Opinion Mining*. New York: Springer Nature.

Liu, D., and Lei, L. (2018) 'The Appeal to Political Sentiment: An analysis of Donald Trump's and Hillary Clinton's Speech Themes and Discourse Strategies in the 2016 US Presidential Election', *Discourse, Context & Media*, 25: 143–52. https://doi.org/10.1016/j.dcm.2018.05.001

Liu, M. 2013. 'Design and analysis of Taiwan's interpretation certification examination'. D. Tsagari and R. van Deemter (eds) *Assessment Issues in Language Translation and Interpreting*, pp. 163–78. Frankfurt am Main: Peter Lang.

Liu, X. (2020) 'Pragmalinguistic Challenges for Trainee Interpreters in Achieving Accuracy: An Analysis of Questions and Their Interpretation in Five Cross-Examinations', *Interpreting. International Journal of Research and Practice in Interpreting*, 22: 87–116. https://doi.org/10.1075/intp.00035.liu

Liu, X., and Hale, S. (2018) 'Achieving Accuracy in a Bilingual Courtroom: The Effectiveness of Specialised Legal Interpreter Training', *The Interpreter and Translator Trainer*, 12: 299–321. https://doi.org/10.1080/1750399x.2018.1501649

Liu, Y. 2021. 'Exploring a corpus-based approach to assessing interpreting quality', in J. Chen and C. Han (eds) *Testing and Assessment of Interpreting: Recent Developments in China*, pp. 159–78. Springer, Singapore. https://doi.org/10.1007/978-981-15-8554-8_8

Liu, Y. 2023. *A sentiment Analysis of Oscar Wilde's Children's Stories in Chinese Translation (DALS dissertation)*. The Hong Kong Polytechnic University.

Lu, X., and Han, C. (2023) 'Automatic Assessment of Spoken-Language Interpreting Based on Machine-Translation Evaluation Metrics', *Interpreting. International Journal of Research and Practice in Interpreting*, 25: 4–10. https://doi.org/10.1075/intp.00076.lu

Martínez-Cámara, E., et al. (2014) 'Sentiment Analysis in Twitter', *Natural Language Engineering*, 20: 1–28. https://doi.org/10.1017/S1351324912000332

Mead, P. (2005) *Methodological Issues in the Study of Interpreters' Fluency*.

Myford, C. M., and Wolfe, E. W. (2003) 'Detecting and Measuring Rater Effects Using Many-facet Rasch measurement: Part I', *Journal of applied Measurement*, 4: 386–422.

Ouyang, L., Q. Lv, and J. Liang. 2021. 'Coh-Metrix Model-Based Automatic Assessment of Interpreting Quality', *New frontiers in translation studies*, 179–200. Berlin, Heidelberg, Dordrecht, and New York City: Springer International Publishing. https://doi.org/10.1007/978-981-15-8554-8_9

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025

Ozolins, U. 2015. 'Ethics and the Role of the Interpreter'. *The Routledge Handbook of Interpreting*, pp. 319–36. Milton Park, Abingdon, Oxfordshire: Routledge.

P Jain, A. and P. Dandannavar. 2016. *'Application of Machine Learning Techniques to Sentiment Analysis'. Presented at the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, IEEE. https://doi.org/10.1109/ICATCCT.2016.7912076, accessed 12 Feb. 2024.

Pöchhacker, F. 2001. 'Quality Assessment in Conference and Community Interpreting' *Meta* 46: 410–25. https://doi.org/10.7202/003847ar

Pöchhacker, F. 2004. *Introducing Interpreting Studies*. London Routledge.

Poria, S., et al. (2020) 'Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research', *IEEE Transactions on Affective Computing*, 14: 108–32. https://doi.org/10.1109/TAFFC.2020.3038167

Setton, R. and A. Dawrant. 2016. *Conference Interpreting: A Trainer's Guide*. Amsterdam: John Benjamins Publishing Company.

Setton, R., and Motta, M. (2007) 'Syntacrobatics: Quality and Reformulation in Simultaneous-with-text', *Interpreting*, 9: 199–230.

Stachowiak-Szymczak, K., and Korpal, P. (2019) 'Interpreting Accuracy and Visual Processing of Numbers in Professional and Student Interpreters: An Eye-tracking Study', *Across Languages and Cultures*, 20: 235–51. https://doi.org/10.1556/084.2019.20.2.5

Stewart, C., N. Vogler, J. Hu, J. Boyd-Graber, and G. Neubig. 2018. 'Automatic Estimation of Simultaneous Interpreter Performance'. *arXiv (Cornell University)*, pp. 662–6. Presented at the The 56th Annual Meeting of the Association for Computational Linguistics, January 1, Cornell University. https://doi.org/10.18653/v1/p18-2105

Su, W. (2019) 'Interpreting Quality as Evaluated by Peer Students', *The Interpreter and Translator Trainer*, 13: 177–89. https://doi.org/10.1080/1750399x.2018.1564192

Su, W., and Li, D. (2021) 'Exploring the Effect of Interpreting Training: Eye-Tracking English-Chinese Sight Interpreting', *Lingua*, 256: 103094. https://doi.org/10.1016/j.lingua.2021.103094

Taboada, M. (2016) 'Sentiment Analysis: An Overview From Linguistics', *Annual Review of Linguistics*, 2: 325–47. https://doi.org/10.1146/annurev-linguistics-011415-040518

Taboada, M., Brooke, J., Tofiloski, M., et al. (2011) 'Lexicon-based Methods for Sentiment Analysis', *Computational linguistics*, 37: 267–307.

Tang, F., and Li, D. (2016) 'Explicitation Patterns in English-Chinese Consecutive Interpreting: Differences Between Professional and Trainee Interpreters' *Perspectives*, 24: 235–55.

Tebble, H. (2012) 'Interpreting or Interfering', *Coordinating Participation in Dialogue Interpreting*, 1: 23–44.

Turner, B., Lai, M., and Huang, N. (2010) 'Error Deduction and Descriptors - A Comparison of Two Methods of Translation test assessment', *The International Journal of Translation and Interpreting Research*, 2: 11–23.

Wang, J., and Fang, J. (2019) 'Accuracy in Telephone Interpreting and On-Site Interpreting: A Comparative Study', *Interpreting International Journal of Research and Practice in Interpreting*, 21: 36–61. https://doi.org/10.1075/intp.00019.wan

Wen, J., and Lei, L. (2022) 'Linguistic Positivity Bias in Academic Writing: A Large-Scale Diachronic Study in Life Sciences Across 50 Years', *Applied Linguistics*, 43: 340–64. https://doi.org/10.1093/applin/amab037

Xu, H. (2021) 'Interprofessional Relations in Interpreted Lawyer–Client Interviews. An Australian Case Study', *Perspectives*, 29: 608–24. https://doi.org/10.1080/0907676x.2020.1769150

Xu, H. (2024) '"Please Make Sure We Don't Get This Interpreter Again" Australian Legal Aid Lawyers' Experience of Working With Interpreters', *Translation and Interpreting Studies*, 19: 257–76. https://doi.org/10.1075/tis.21070.xu

Xu, H., and Liu, K. (2024) 'The Impact of Directionality on Interpreters' Syntactic Processing: Insights From Syntactic Dependency Relation Measures', *Lingua*, 308: 103778. https://doi.org/10.1016/j.lingua.2024.103778

Yu, W., and van Heuven, V. J. (2017) 'Predicting Judged Fluency of Consecutive Interpreting From Acoustic Measures', *Interpreting*, 19: 47–68. https://doi.org/10.1075/intp.19.1.03yu

Downloaded from https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amaf026/8124526 by The Hong Kong Polytechnic University user on 26 August 2025