Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 9, 103627 (2026)
https://doi.org/10.29140/lea.2026.103627

Investigating the Predictive Power of Lexical, Syntactic, and Cohesive Features on Chinese-English Translation Quality: The Case of Test for English Majors Grade 8 (TEM8)

MINGWEI PAN[a]
HAOYUE YANG[b]

[a]*Shanghai International Studies University, China*
mwpan@shisu.edu.cn

[b]*Shanghai Jiayi Experimental High School, China*
1374893422@qq.com

Abstract

This study investigates the language features that may predict translation quality in TEM8 Chinese-English tasks and the magnitude of their predictive power. Drawing upon a self-compiled corpus comprising of 1,000 translation samples from 2022 and 2023 test administrations, 38 language feature indices spanning lexical, syntactic, and cohesive dimensions were extracted using Coh-Metrix (Graesser et al., 2004) and the NeoSCA (a refined iteration of the Second Language Syntactic Complexity Analyzer, namely L2SCA) (Tan, 2022). Quantitative analysis identified seven significant language feature predictors: four lexical and three syntactic. Notably, no cohesive features demonstrated statistically significant predictive capacity. Subsequent regression analysis revealed that five of these predictors collectively explained 17.1% of the variance in translation quality scores. These findings furnish empirical evidence for translation assessment and pedagogy, offering actionable insights for candidates and teachers to pinpoint specific linguistics domains warranting enhancement in TEM8 translation performance.

Keywords: Translation quality assessment (TQA), Chinese-to-English translation, Test for English Majors Grade 8 (TEM8), lexical feature, syntactic feature, cohesive feature

Data Availability Statement: *All relevant data are within this paper.*

Introduction

As global interconnectedness deepens, the ability to produce high-quality translation is increasingly recognized as vital for effective cross-cultural communication (Adil, 2020). This practical imperative is paralleled in the pedagogical sphere, where translation is viewed not merely as a skill-specific competency but as a potent vehicle for enhancing overall language proficiency (Cook, 2010). Within the Chinese specific context, this importance is further amplified, as fostering high-standard translation capability is deemed crucial for accurately articulating national narratives to the international community (Chen, 2022). Such prioritization is instantiated in standardized assessments, most notably the Test for English Majors Grade 8 (TEM8) in China, wherein the Chinese-English (C-E) translation task functions as a critical benchmark for gauging students' translation competence.

However, the field of translation assessment remains encumbered by persistent challenges of subjectivity and inconsistent criteria. A critical yet underexplored question concerns the empirical grounding of evaluative practices: specifically, which measurable language features constitute reliable predictors of translation quality. Situated within this broader endeavor to enhance assessment rigor, this study seeks to empirically identify the salient language features that predict translation quality in TEM8 C-E translation tasks and to precisely quantify their predictive power, thereby contributing to more objective and evidence-based evaluation practices.

Translation Quality Assessment

Translation Quality Assessment (TQA) emerges as a pivotal focus in translation studies. The outcome of TQA varies according to the diverse purposes of assessment (Melis & Hurtado Albir, 2001). Yang (2012) observed that while qualitative TQA is well-suited for evaluating published literary translations, quantitative TQA is more appropriate for large-scale proficiency and achievement language tests. For qualitative TQA, various scholars have put forward different models of qualitative assessment in translation (House, 2014; Nida & Taber, 1969; Reiss, 2004; Williams, 2004). Unlike qualitative TQA, quantitative TQA presents the assessment results in the form of scores (Yang, 2019). The representative examples of quantitative TQA include the model proposed by Colina (2008, 2009), the Diploma in Translation (DipTrans) exam (https://www.ciol.org.uk), employing an analytical scoring approach, and the American Translators Association (ATA) exam (https://www.atanet.org), which adopts an error deduction method in scoring. Within the Chinese context of English translation assessment, the translation section of the TEM8 exam differs from these exams by employing a mixed scoring approach, assessing translations both through error deduction and quantitative analytical scoring (see Appendix A for the TEM8 C-E translation rating scale). Similar to the DipTrans exam, the TEM8 translation task evaluates translations based on two dimensions: (1) translation fidelity, which accounts for 70% of the score, and (2) language quality, which constitutes 30% of the score. Additionally, the rating criteria describe errors in students' translations, and raters are required to score the translations according to these descriptions.

Having outlined the core approaches to TQA, namely qualitative and quantitative frameworks, it is important to note that recent advancements in computer science have driven a shift in this field. Many researchers have employed natural language processing tools and corpora to develop TQA models using automatic methods, significantly advancing the field of TQA (Zhu & Wang, 2020). According to Yang (2019), TQA that adopts automatic methods refers to the procedure in which translated texts are collected and organized into a corpus, after which computational programs are employed to automatically evaluate translation quality. It falls into two types of models (Yang, 2019). The reference translation-dependent model depends on a reference translation. To be more specific, researchers or raters examine the similarity between the meaning of the reference translation and the candidates'

translations, and concentrate on the fidelity of the translation. The more similar they are, the higher the score candidates will obtain. It offers objective, replicable scoring by quantifying semantic similarity to a benchmark. This model directly aligns with the core TQA criterion of fidelity, ensuring assessment relevance. However, this approach may rely too heavily on a single reference translation and may ignore multiple creative but accurate translations. In addition, it cannot adequately evaluate the naturalness and appropriateness of the language of the translated texts. High similarity to the reference translation does not necessarily mean the translated text is idiomatic and contextually appropriate. Conversely, the language feature-based model does not rely on a reference translation. Researchers or raters first find out the most representative language features of excellent translations. Scores increase commensurately with the incidence of these features. This paradigm emphasizes language features that distinguish excellent translations from mediocre ones, while foregrounding the inherent quality features of translations, such as accuracy, fluency, completeness, and appropriateness, which can reflect whether the translation is effective, appropriate, and comprehensible in real communicative situations. As it does not bind to a fixed reference, it accommodates translation creativity and contextual adaptability. Furthermore, the model also enables insights into why a translation is effective via language feature analysis, supporting formative feedback for translators. Nevertheless, the model requires extensive, high-quality corpora to identify valid, representative features of excellent translations. In addition, feature selection is subjective, so the overemphasis on surface-level traits may overlook deeper semantic or pragmatic qualities. Despite these shortcomings, many studies still analyze the relationship between language features and translation quality in view of the advantages of the language feature-based model (Jiang, 2013; Wang & Zhu, 2017; Zhu & Wang, 2020). While translation fidelity is essential, the present study focuses on the roles of language features in TQA.

Language Features in Assessing Translation Quality

Kyle and Crossley (2017) emphasized that language features are frequently used to assess learners' language performance and to create automatic scoring models. Although language features are less extensively studied in translation assessment than those in writing evaluation, the approach of analyzing language features is also applied in the development of quantitative TQA models, as evidenced by several studies (Jiang, 2016; Redelinghuys & Kruger, 2015; Wang et al., 2021; Zhu & Wang, 2020).

For example, the study of Redelinghuys and Kruger (2015) investigated whether there is a relationship between language features and translation expertise. The language features explored in this study included explicitation, which was measured by the ratio of that-omission and the frequency of conjunctive markers; simplification, standardized by type-token ratio (TTR), and the Flesch Reading Ease score; and normalization, measured through the ratio of contracted forms and the frequency of neologisms and loanwords. The study found that the level of translation expertise has an influence on the distribution and frequency of some of these features, most definitely as far as simplification is concerned, suggesting that these features could potentially be considered as indicators of translation expertise.

While Redelinghuys and Kruger (2015) adopted a broad perspective covering multiple macro-level text features, such as explicitation, simplification, and normalization, some other studies have delved into more specific, micro-level dimensions of one level of language features. For example, the study of Wang et al. (2021) mainly focused on the relationship between lexical features and translation quality. It developed a TQA framework based on lexical features such as fluency, lexical diversity, word frequency breadth, lexical difficulty, lexical density, and word semantics. It also explored the predictive power of these language features in the scoring model. They found significant correlations between these features and translation quality, underscoring their predictive value in scoring models.

Similarly, Jiang (2016) explored the predictive power of language features at the cohesive level. Primarily studying the role of cohesive devices in TQA, it employed the text analysis tool Coh-Metrix, and identified six cohesive indices that can effectively predict translation quality, including causal connectives incidence, temporal connectives incidence, the mean LSA (latent semantic analysis) similarity between adjacent sentences, the mean LSA similarity between all possible pairs of sentences in a paragraph, the mean LSA similarity between adjacent paragraphs, and the mean LSA similarity between all sentences. Significant differences were observed between the high- and low-level groups in both connectives and LSA.

Extending this line of inquiry, Zhu and Wang's (2020) study examined the relationship between language features and translation quality from three levels in detail. It explored how lexical, syntactic, and cohesive features influence translation quality. Lexically, they assessed both breadth (e.g., variation, frequency) and depth (e.g., semantic relations) of vocabulary knowledge. Syntactically, all 14 indices from Lu's (2010) L2SCA were analyzed, while cohesive features included connectives and latent semantic analysis. Their findings revealed genre-specific predictors. For narrative translations, among all examined indices, 23 significantly correlated with translation quality, with 11 of them forming an optimal multiple linear regression (MLR) model. For argumentative translations, 12 indices showed significance, 5 of which were selected to establish an optimal MLR model. Key predictors in argumentative translations involved lexical diversity (e.g., LDTTRa, LDVOCD), semantic properties (e.g., WRDHYPn, WRDHYPv), syntactic complexity (CT/T, CP/T), and cohesion markers (e.g., connectives and information structure indices).

In these studies, language features have proven to be significant indicators of translation quality and are worth studying.

Studies on the Translation Section of the Test for English Majors Grade 8 (TEM8)

When exploring the relationship between language features and translation quality, translations in large-scale exams serve as essential resources since they constitute valid evidence of candidates' proficiency. The Test for English Majors Grade 8 (TEM8) is one of the large-scale exams.

The C-E translation section of TEM8 evaluates students' ability to translate real-world Chinese materials into natural English. According to the official guidelines of TEM8 (http://tem.fltonline.cn/?p/74676), the test uses practical texts drawn from newspapers, magazines, and general literary works, with most texts being argumentative essays. The required translation speed is between 250 and 300 Chinese characters per hour. Moreover, TEM8 is a criterion-referenced test (Zou, 2012), which adopts a rating scale to score translated samples. As mentioned in the section *Translation Quality Assessment*, scores are calculated with 70% for translation fidelity and 30% for language quality (see Appendix A). This balanced approach ensures translations are both semantically correct and well-written. From this point of view, TEM8, as a prominent large-scale examination for English majors, provides an excellent context to investigate the predictive power of language features on translation quality. Pan and Zou (2020) demonstrated that TEM8 is widely recognized and used across various sectors of society. Their research indicates that some international universities accept TEM8 scores as equivalent to Test of English as a Foreign Language (TOEFL) or International English Language Testing System (IELTS) for postgraduate admissions. Similarly, the study found that several domestic universities use TEM8 scores to exempt students from postgraduate English courses, and certain civil service positions related to foreign affairs and middle school English teaching recruitment consider TEM8 certification as a basic requirement. In this context, developing a TQA model using automatic methods tailored to the TEM8 context represents a significant and worthwhile endeavor.

Research on TEM8 translation primarily focuses on error analysis. Most studies examine errors by comparing Chinese and English language features (e.g., Shao & Shao, 2012). A smaller number of studies investigate student strategies (e.g., Hai, 2003) or rater reliability (e.g., Chen, 2016). However, the field remains dominated by qualitative approaches, with limited use of quantitative or corpus-based methods. A notable exception is Wang (2009), who employed corpus tools to analyze 360 translation samples, identifying frequent interlanguage errors such as literal translation and Chinese-style English expressions, largely attributed to L1 negative transfer. Despite these efforts, significant research gaps persist. There is a scarcity of large-scale, quantitative studies utilizing corpora to explore language features of translations. Consistent with the methodological framework of translation competence research developed by Process in the Acquisition of Translation Competence and Evaluation (PACTE, 2000, 2003) and Transcomp (Göpferich, 2009, 2013), this study employs a corpus-based approach to analyze language features in TEM8 translations, to gain a more comprehensive understanding of how language features relate to translation quality. In summary, this study employed a quantitative method to examine the predictive power of lexical, syntactic, and cohesive features on translation quality. These features, derived from previous research and detailed in a following section, *Selection of Indices*, aim to address the following research questions (RQs):

RQ1: What specific lexical, syntactic, and cohesive features are effective predictors of TEM8 C-E translation quality? To what extent can they predict translation quality, respectively?

RQ2: Which of these predictors can form a multiple linear regression (MLR) model to predict TEM8 C-E translation quality jointly? To what extent can the model predict translation quality?

Methodology

Data Collection and Processing

The present study selected a sample of 500 TEM8 C-E translations from the 2022 and 2023 test administrations, respectively (the source texts and examples of translations are presented in Appendix B), totaling 1000 translations. The translation samples were drawn from a diverse pool of TEM8 candidates, representing a broad spectrum of universities across different regions and of various institutional levels. Furthermore, the present study adopted the official scores provided by TEM8. It should be noted that the TEM8 translations are scored on a 10-point scale as shown in Appendix A. The raw scores were later transformed into the final scores (15 points) by multiplying them by 1.5. Although final scores are reported to candidates, raw scores were used as the independent variable, as one index's capability of predicting scores does not vary just because the scores are linearly amplified. The present study employed a stratified random sampling approach, and the score distribution of the 500 translations each year followed a bell-shaped distribution, characterized by fewer translations with high and low scores, while the majority of translations had average scores. To be more specific, the translations awarded 5 and 6 points are the most frequent, while those getting 2 and 9 are the least. This distribution largely mirrors actual scoring patterns. However, another word of caution is that the TEM8 translations of 0 and 1 points were excluded from the data collection, given that most of them only have limited linguistic content with just one or half-translated sentences in broken language. These instances provide limited analytical value for examining the relationship between scores and language features, especially in terms of syntactic and cohesive features. The translations with a perfect score of 10 were also excluded due to their limited number. It should also be noted that translations with half-point scores were excluded from the study. The infrequent use of half-point ratings may result in uneven score distributions, potentially obscuring the monotonic relationship between language features and translation

quality. Meanwhile, TEM8 raters—senior experts in the field—underwent rigorous selection and calibration. The standard calibration process requires them to learn the official TEM8 rating scale and align their scoring through trial evaluations. In addition, a continuous monitoring system ensures the accuracy and consistency of scores by randomly selecting rated translations for verification. This safeguards the scores' reliability and validity.

Research Instruments

As noted earlier, automatic TQA typically evaluates language quality across three dimensions: lexical, syntactic, and cohesive features. To delve into the three aspects, the study employed two natural language processing tools: Coh-Metrix (Graesser et al., 2004) for analyzing lexical features and cohesive features, and NeoSCA (Tan, 2022) for analyzing syntactic features. As a refined iteration of L2SCA (Lu, 2010), NeoSCA started with the original codebase of L2SCA and has evolved into a separate project. It retains full functional equivalence with L2SCA for syntactic complexity analysis but adds the crucial advantage of Windows compatibility, whereas the original tool only operates on Mac OS and Linux.

Coh-Metrix, an online text analysis tool, is among the relatively broad and sophisticated automated textual assessment tools available to date (McNamara et al., 2014). It can analyze texts for various aspects, such as cohesion, language usage, and readability. It generates indices reflecting cohesion relations, world knowledge, as well as language and discourse characteristics (McNamara et al., 2014). In its latest iteration, Coh-Metrix 3.0 is equipped to analyze 106 indices from 11 key aspects. Coh-Metrix's syntactic complexity analysis is subject to significant limitations (Lu & Xu, 2016). In contrast, L2SCA, developed by Lu (2010), establishes its measurement indices of syntactic complexity directly from the L2 syntactic complexity studies. Lu (2010) selected 14 syntactic complexity indices from second language writing research (Ortega, 2003; Wolfe-Quintero et al., 1998) for automatic analysis, categorizing them into five types: length of production unit, sentence complexity, subordination, coordination, and particular structures.

Selection of Indices

Building on Zhu and Wang's (2020) research, this study adopts lexical and cohesive indices significantly correlated with argumentative translation scores, while also incorporating lexical density (Jiang, 2013) to examine the role of content-word ratio. Given that TEM8 source texts are predominantly argumentative and drawn from newspapers or magazines, cohesive features were also selected in alignment with Jiang (2016) to match text-type characteristics. For syntactic features, all 14 indices from L2SCA were included due to the limited existing research on syntactic complexity in translation assessment (Zhu & Wang, 2020). In total, 38 indices, including 16 lexical, 14 syntactic, and 8 cohesive indices, were employed to explore their predictive power for C-E translation quality in TEM8.

Lexical features were analyzed in terms of lexical diversity, lexical density, and lexical semantics. From the perspective of lexical diversity, the type-token ratio of all words (LDTTRa) and the lexical diversity of all words based on the VOCD measure (LDVOCDa) were selected. The indices that measure lexical density include noun incidence (WRDNOUN), verb incidence (WRDVERB), adjective incidence (WRDADJ), adverb incidence (WRDADV), and pronoun incidence (WRDPRO). Lexical semantics encompasses the meanings of individual words and the semantic relationships between words. These dimensions were operationalized through the following indices: the average polysemy values of all content words in a text (WRDPOLc), estimates of hypernymy for nouns (WRDHYPn), for verbs (WRDHYPv), and for the combination of both nouns and verbs (WRDHYPnv), the age in which a content word first appears in a child's vocabulary (WRDAOAc), a measure of how familiar a

word is to an adult (WRDFAMc), how concrete or non-abstract a word is (WRDCNCc), how easy it is to construct a mental image of the word (WRDIMGc), and the degree to which a word is related to other words (WRDMEAc).

In accordance with Lu (2010), syntactic complexity was operationalized across five dimensions, namely, length of production unit, sentence complexity, subordination, coordination, and particular structures. First, the dimension length of production unit includes three indices: mean length of clause (MLC), mean length of sentence (MLS), and mean length of T-unit (MLT). Second, sentence complexity is computed as the index sentence complexity ratio (C/S, defined as clauses per sentence). Third, the subordination dimension incorporates four indices: T-unit complexity ratio (C/T, defined as clauses per T-unit), complex T-unit ratio (CT/T), dependent clause ratio (DC/C, defined as dependent clauses per clause), and dependent clause per T-unit (DC/T). Fourth, coordination consists of three indices, including coordinate phrases per clause (CP/C), coordinate phrases per T-unit (CP/T), and sentence coordination ratio (T/S, defined as T-units per sentence). Fifth, the dimension particular structures are composed of complex nominals per clause (CN/C), complex nominals per T-unit (CN/T), and verb phrases per T-unit (VP/T).

Text cohesion was analyzed from two dimensions in the present study: latent semantic analysis (LSA) and connectives. Within the LSA dimension, four indices were employed, including mean LSA similarity between adjacent sentences (LSASS1), mean LSA similarity between all possible pairs of sentences in a paragraph (LSASSp), mean LSA similarity between adjacent paragraphs (LSAPP1), and average proficiency of how much given versus new information exists in each sentence in a text (LSAGN). Regarding connectives, four indices were used, involving the frequencies of causal connectives (CNCCaus), logical connectives (CNCLogic), additive connectives (CNCAdd), and temporal connectives (CNCTemp).

An inventory of these language features and their definitions is provided in Appendix C.

Data Analysis

The present study predominantly adopted a quantitative method to analyze the data. Quantitative analysis used Coh-Metrix and NeoSCA to extract language features, followed by simple linear regression analysis to identify predictors of TEM8 translation scores. Subsequently, multiple linear stepwise regression analysis was performed to examine their combined predictive effects. After establishing the simple and multiple linear regression models, residual analysis was also conducted to evaluate the rationality of the model and the validity of the results. The residual normality of these models was verified using the Kolmogorov-Smirnov test (K-S test). Residual is the difference between the observed value and the model's predicted value. If the K-S test yields a *p*-value greater than 0.05, it indicates that the model meets the assumption of residual normality. Residual homoscedasticity of these models was assessed using the Non-Constant Variance Score Test (ncvTest). Homoscedasticity means the variance of the residuals is roughly the same for all predicted values. If the ncvTest yields a *p*-value greater than 0.05, it demonstrates that the model meets the critical assumption of homoscedasticity, making it statistically reliable. Residual independence of the models was verified by Durbin-Warton Test (DW Test). The assumption of residual independence requires that the residual of one observation must not contain any predictive information about the residual of another. In DW tests, a DW value near 2 is ideal, confirming that the residuals are independent. For multiple linear regression analysis, besides the above-mentioned conditions, simple linear regression analysis had already been conducted to verify the linear relationship between the predictors and translation quality, and VIF was detected to avoid multicollinearity. All the regression analyses were implemented with R. In addition, typical examples were analyzed to support and enhance the robustness of the findings.

Results

Simple linear regression revealed that four lexical indices and three syntactic indices were significantly correlated with translation quality, while cohesive features show no significant correlation. The specific results are presented in the following sections.

Predictive Power of Lexical Features

Four lexical indices—WRDNOUN (noun incidence), WRDAOAc (mean age of acquisition for content words), WRDHYPn (mean value of hypernymy for nouns), and WRDHYPv (hypernymy for verbs)—were found to correlate significantly with translation quality. Residual normality, homoscedasticity, and independence were confirmed via K-S tests, ncvTests, and DW statistics respectively[1].

Results of the simple linear regression models are summarized in Table 1.

Table 1. *Results of simple linear regression models based on indices of lexical features*

Predictor	*b*	SE	β	*t*	*p*	95% CI	R^2
Model 1							.045
Intercept	7.766	0.336		23.138	< .001	[7.107, 8.425]	
WRDNOUN	–0.010	0.001	-–.212	-6.841	< .001	[–0.013, –0.007]	
Model 2							.038
Intercept	2.847	0.427		6.661	< .001	[2.009, 3.684]	
WRDAOAc	0.007	0.001	.194	6.258	< .001	[0.005, 0.009]	
Model 3							.062
Intercept	1.137	0.538		2.112	.035	[0.082, 2.192]	
WRDHYPn	0.705	0.087	.249	8.145	< .001	[0.535, 0.874]	
Model 4							.065
Intercept	2.819	0.326		8.661	< .001	[2.180, 3.458]	
WRDHYPv	1.901	0.228	.256	8.351	< .001	[1.454, 2.347]	

Note. N = 1,000. b = unstandardized coefficient; SE = standard error; β = standardized coefficient; CI = confidence interval.

WRDNOUN significantly correlates with translation quality. As shown in Table 2, Model 1 had an unstandardized coefficient of –0.010, indicating that a one-unit increase in WRDNOUN is associated, on average, with a -0.010 unit change in the score, controlling for other variables in the model. Since the aim is to interpret the relationship between the independent and dependent variables, reporting an unstandardized coefficient is appropriate. The same applies to Model 2-7. The model explained 4.5% of the variance (adj. $R^2 = 0.045$, $p < .001$). To illustrate this pattern empirically, consider two specific examples: Translation #769 (score = 7) contains 16 nouns out of 81 words, while #171 (score = 4) has 31 nouns out of 78 words. This contrast supports the observed negative effect of high noun incidence on translation quality.

[1] p-value of K-S tests: Model 1 = 0.378, Model 2 = 0.342, Model 3 = 0.177, Model 4 = 0.327; p-value of ncvTests: Model 1 = 0.559, Model 2 = 0.230, Model 3 = 0.059, Model 4 = 0.159; DW-value: Model 1 = 2.058, Model 2 = 2.016, Model 3 = 2.037, Model 4 = 1.986

Example 1

Cities have diven original **inhabitants** in the **wild** away and destroyed its ancient **scenery**. More **prosperity** in **order** has superseded the natural **scenery sights**. Currently, the **wild** is increasingly shrinking and decreasing. However, **people** shouldn't ignore it or disregard it; instead, they ought to find the most suitable **distance** to approach and accompany it. Being kind to the **wild** means to treat **human** himself kindly. We must be aware that it's never possible for **mankind** to defeat the **wild** by **cities**. (#769, score = 7)

Gentally, **nature area** and **city** are **contrast**. **City** makes the local **people** more out their **home**, and destroy the old **nature area sightsee**. More and more **florish** instead of the **nature sightsee**. Today, **nature area** become more and more small. **People** should't careless for **nature area**. We should found a better **metre** from **nature area** and **city**. Take **care** for **nature area**, equatlly take **care** for **people**. As we known, **people** is't win the **nature area** by **city**. (#171, score = 4)

WRDAOAc also correlates significantly with translation quality. Model 2 showed a positive unstandardized coefficient of 0.007, suggesting that when all other variables remain constant, for every one unit of increase in WRDAOAc, the score increases by 0.007 units. This illustrates that translations using words acquired later tend to achieve higher quality. This index accounted for 3.8% of variance (adj. $R^2 = 0.038$, $p < .001$). A concrete example can illustrate this relationship. For instance, Translation #754 (score = 7) uses more complex vocabulary (e.g., "approach" instead of "way"; "most appropriate" instead of "best"), consistent with later-acquired words leading to better quality. Specifically, although Translation #754 and #51 employ semantically equivalent vocabulary, Translation #754 uses more sophisticated lexical choices. While the WRDAOAc index (i.e., word age of acquisition) is not directly equivalent to word complexity or formality, it is closely associated with these two concepts: words acquired later in language learning are typically more academic, abstract, and formal, and thus perceived as more complex. This illustrates that translations characterized by higher WRDAOAc values tend to receive higher scores.

Example 2

.....However, human beings should find out the <u>most appropriate</u> <u>approach</u> to live in the wild rural places kindly and friendly, instead of ignoring them or
taking an indifferent attitude towards them..... (#754, score = 7)
.....but we should not ignore the kuangya, but to find the <u>best</u> <u>way</u> to stay with it..... (#51, score = 3)

WRDHYPn is positively correlated with translation quality, forming Model 3 with an unstandardized coefficient of 0.705, which indicates that when all other independent variables are held constant, a one-unit increase in WRDHYPn corresponds to a 0.705-unit increase in the score. It explained 6.2% of variance (adj. $R^2 = 0.062$, $p < .001$). In Example 3, Translation #754 (score = 7) uses a more specific hypernym ("residents") compared to #51 (score = 3), which uses a circumlocution ("people who originally lived in *kuangya*"). From the economic perspective, it is more appropriate to use the more specific noun "resident" in the translation, illustrating that lexical specificity is associated with higher scores.

Example 3

.....The arriving of the city, distroying the ancient landscape of the wild rural places, explicts the origional <u>resident</u>s there..... (#754, score = 7)
.....The city expels the <u>people who originally lived in kuangya</u>, destories the kuangya's old ancient view..... (#51, score = 3)

WRDHYPv is also positively correlated with translation quality. Model 4 had an unstandardized coefficient of 1.901, which means that with all other independent variables fixed, each one-unit rise in WRDHYPv is associated with a 1.901 increase in the dependent variable score. The model also explained 6.5% of variance (adj. $R^2 = 0.065$, $p < .001$). Example 4 demonstrates that Translation #796 (score = 7) uses specific verbs ("evolve", "reshape") to describe the ability of the traditional culture to update as the environment changes, while #114 (score = 3) uses a general verb ("influence"), which is unable to illustrate this ability vividly, supporting that verbal specificity improves translation quality.

> Example 4
> .....The traditional culture will evolve and reshape itself in the atmosphere of new generation while having impact on real life..... (#796, score = 7)
> .....And when she is influnced the truth, she's also influnced by the truth..... (#114, score = 3)

Predictive Power of Syntactic Features

Three syntactic indices, CP/T, CN/T, and CN/C, showed significant correlations with translation quality and were used to construct simple linear regression models (Models 5-7). CP/T means coordinate phrases per T-unit, CN/T refers to complex nominals per T-unit, and CN/C stands for complex nominals per clause. Residual normality, homoscedasticity, and independence were confirmed via K-S tests, ncvTests, and DW values respectively[2].

The model results are summarized in Table 2.

Table 2. *Results of simple linear regression models based on indices of syntactic features*

Predictor	*b*	SE	β	*t*	*p*	95% CI	R^2
Model 5							.043
Intercept	4.846	0.111		43.577	< .001	[4.628, 5.063]	
CP/T	0.955	0.142	.208	6.734	< .001	[0.677, 1.233]	
Model 6							.069
Intercept	4.352	0.144		30.294	< .001	[4.070, 4.634]	
CN/T	0.495	0.058	.263	8.611	< .001	[0.383, 0.608]	
Model 7							.038
Intercept	4.643	0.147		31.660	< .001	[4.355, 4.931]	
CN/C	0.504	0.080	.195	6.290	< .001	[0.347, 0.661]	

Note. N = 1,000. b = unstandardized coefficient; SE = standard error; β = standardized coefficient; CI = confidence interval.

CP/T positively correlates with translation quality. Model 5 yielded an unstandardized coefficient of 0.955, which means that, all other independent variables constant, a one-unit rise in CP/T brings a 0.955 increase in score. It explained 4.3% of the score variance (adj. $R^2 = 0.043$, $p < .001$). As illustrated in the following example, high-scoring Translation #999 used coordinated phrases (e.g., "impact modern Chinese and provide historic resources..."), which had richer meaning and more abundant

[2] p-value of K-S tests: Modal 5 = 0.094, Model 6 = 0.248, Model 7 = 0.244; p-value of ncvTests: Model 5 = 0.636, Model 6 = 0.145, Model 7 = 0.177; DW-value: Model 5 = 2.020, Model 6 = 2.012, Model 7 = 2.009

expressions than the phrase "provided historical experiences" in Translation #126 (score = 3) and thus conveyed the meaning of the original text more completely and appropriately.

> Example 5
> .....The Traditional Chinese culture were conducted by fruitful relics passing by our ancient generations, which constantly impact modern Chinese and provide historic resources and reality base for cultivate new cultures..... (#999, score = 9)
> .....It (the traditional Chinese culture) provided historical experiences for culture invation..... (#126, score = 3)

CN/T and CN/C are also positively related to translation quality. Model 6 had an unstandardized coefficient of 0.495, which shows that when all other independent variables remain unchanged, a one-unit increase in the CN/T is linked to a 0.495 increase in the dependent variable score. And it accounted for 6.9% of variance (adj. $R^2 = 0.069$, $p < .001$). Model 7 showed an unstandardized coefficient of 0.504, which displays that as all other independent variables are held constant, a one-unit increase in CN/C corresponds to a 0.504-unit increase in the dependent variable score, explaining 3.8% of variance (adj. $R^2 = 0.038$, $p < .001$). Example 6 illustrates that Translation #992 (score = 9) contains complex nominals (e.g., "a mighty river flowing...") within one of its clauses, while #126 (score = 3) used simpler sentences to express a similar meaning, lacking the richness and flexibility in the use of language. On the other hand, the excerpt from Translation #992 is a complete T-unit and has four complex nominals. In contrast, the excerpt from #126 contains two T-units, but none of them include complex nominals. On this account, Example 6 corroborates that higher-quality translations exhibit greater use of complex nominals.

> Example 6
> .....Much like a mighty river flowing for over 5000 years and never failing to maintain her abundance, Chinese traditional culture embodies time-honored conventions as well as future-oriented originality with her vigorous momentum to integrate the two and foster the new one..... (#992, score = 9)
> .....It just like a river. This river had been launched for five thousands years..... (#126, score = 3)

Absence of Significant Correlations in Cohesive Features

Simple linear regression analysis indicates that no cohesive features show significant correlations with translation quality in this study (see Table 3).

Table 3. *Results of simple linear regression models based on indices of cohesive features*

Predictor	*b*	SE	β	*t*	*p*	95% CI	R^2
LSASS1	–0.140	0.518	–.009	–0.271	.786	[–1.155, 0.875]	< .001
LSASSp	–0.626	0.517	–.038	–1.211	.226	[–1.641, 0.389]	.001
LSAPP1	0.661	0.790	.026	0.837	.403	[–0.889, 2.211]	.001
LSAGN	0.634	1.063	.019	0.597	.551	[–1.453, 2.721]	< .001
CNCCaus	0.016	0.004	.139	4.421	< .001	[0.009, 0.023]	.019
CNCLogic	0.010	0.003	.104	3.288	.001	[0.004, 0.017]	.011
CNCTemp	0.007	0.004	.049	1.562	.119	[–0.002, 0.015]	.002
CNCAdd	–0.005	0.003	–.053	–1.680	.093	[–0.010, 0.001]	.003

Note. N = 1,000 for all models. b = unstandardized coefficient; SE = standard error; β = standardized coefficient; CI = confidence interval.

Simple linear regression analysis shows that cohesive indices, including LSASS1, LSASSp, LSAPP1, LSAGN, CNCTemp, and CNCAdd, exhibit no significant correlation with translation quality. Although CNCCaus ($p < .001$) and CNCLogic ($p = .001$) exhibited significant correlations, the assumption of residual normality was violated (K-S test $p = .042$ and .005, respectively), and they were therefore excluded from further analysis.

The Multiple Linear Regression Model for Predicting Translation Quality

Following the assessment of individual indices, a MLR model was constructed using all significant language features to predict translation quality (see Model 8, Table 4). The linearity assumption was confirmed through preliminary regression analyses and multicollinearity was ruled out (all VIF < 10). Residual diagnostics confirmed normality, homoscedasticity, and independence, supporting the model's validity (Kutner et al., 2005)[3].

Table 4. *Result of the multiple linear stepwise regression model (Model 8)*

Predictor	*b*	SE	β	*t*	*p*	95% CI
Intercept	0.832	0.707		1.178	.239	[–0.555, 2.219]
WRDNOUN	–0.008	0.001	–.163	–5.536	< .001	[–0.011, -0.005]
WRDAOAc	0.003	0.001	.071	2.253	.024	[0.000, 0.005]
WRDHYPn	0.507	0.084	.179	6.071	< .001	[0.343, 0.671]
WRDHYPv	1.097	0.232	.148	4.736	< .001	[0.642, 1.552]
CN/T	0.322	0.057	.171	5.600	< .001	[0.209, 0.435]

Note. $N = 1{,}000$. $R^2 = .176$, Adjusted $R^2 = .171$. *b* = unstandardized coefficient; SE = standard error; β = standardized coefficient; CI = confidence interval. The overall regression was statistically significant, $F(5, 994) = 42.34$, $p < .001$.

Model 8 is selected as the optimal multiple linear regression model based on the criteria of the maximum adjusted coefficient of determination (adjusted R^2). It retains five predictors: four lexical indices (WRDNOUN, WRDAOAc, WRDHYPn, WRDHYPv) and one syntactic index (CN/T). Their standardized coefficients are −0.163, 0.071, 0.179, 0.148, and 0.171, respectively. Standardized coefficients are reported to facilitate comparison of effect magnitudes across predictors. Therefore, WRDHYPn is the strongest predictor, followed by CN/T, while WRDAOAc contributes less significantly.

The standardized regression equation is:

$$\textbf{Score} = -0.163 \times \textbf{WRDNOUN} + 0.071 \times \textbf{WRDAOAc} + 0.179 \times \textbf{WRDHYPn} + 0.148 \times \textbf{WRDHYPv} + 0.171 \times \textbf{CN/T}$$

The model explained 17.1% of the variance in translation quality (adj. $R^2 = 0.171$, $p < .001$).

Discussion

Interpreting the Simple Linear Regression Model: The Impact of Texts and Candidates

To address RQ1, simple linear regression analysis was conducted, identifying seven significant lexical and syntactic indices—WRDNOUN, WRDAOAc, WRDHYPn, WRDHYPv, CP/T, CN/T, and CN/C—that predict translation quality. No cohesive features showed significant correlations.

[3] VIF-value: WORDNOUN = 1.032, WRDAOAc = 1.198, WRDHYPn = 1.055, WRDHYPv = 1.170, CN/T = 1.123; K-S test p = 0.401, ncvTest p = 0.262, DW value = 2.046

At the lexical level, WRDAOAc, WRDHYPn, and WRDHYPv are positively correlated with quality, whereas WRDNOUN shows a negative correlation. These results partially align with prior research: WRDHYPn and WRDHYPv, which reflect semantic depth and lexical complexity, were also significant in Zhu and Wang (2020) and Wang et al. (2021). However, indices related to lexical diversity and frequency, which are significant in earlier studies, showed no predictive power in this study. This discrepancy likely reflects differences in text length: TEM8 translations averaged fewer than 100 words due to short source texts (129–142 characters), whereas Zhu and Wang (2020) used longer texts (~205 words), suggesting lexical diversity may require more extended discourse to emerge as a significant feature. The negative effect of WRDNOUN aligns with Wang and Zhu (2017), who also reported higher noun frequency correlating with lower quality in C-E translation. This may reflect less proficient candidates' over-reliance on high-frequency nouns due to limited grammatical flexibility, as seen in Example 1 in section *Predictive Power of Lexical Features*, where repetitive noun use replaced more complex structures such as pronouns, clauses, or non-finite verbs.

At the syntactic level, the indices CP/T, CN/T, and CN/C show a positive linear correlation with translation quality. Aligning with Zhu and Wang's (2020) findings, CP/T is also a significant indicator in this study. However, while CT/T (complex T-units per T-unit) was a key predictor in their research, it was not significant here. Conversely, CN/T and CN/C were significant predictors in the current study. A possible explanation for this discrepancy also lies in translation length. Zhu and Wang's (2020) texts were more than twice as long as those used here. The shorter translations in this study may have led raters to focus more on phrasal-level features, making indices such as CN/T and CN/C more predictive than sentence-level measures like CT/T.

At the cohesive level, no indices correlated significantly with translation quality, contradicting findings from Chi (2019), Jiang (2016), and Zhu and Wang (2020). This discrepancy may be attributed to four factors. First, the short length of the translation samples in this study (under 100 words) provided limited scope for cohesive devices to emerge, unlike the longer texts analyzed in prior research. Second, the TEM8 candidates may have adhered to a word-for-word approach from the loosely organized Chinese source text, lacking the skill to create explicit connections within one sentence. Or conversely, one plausible explanation is that all TEM8 candidates may have already become proficient in employing fundamental cohesive devices. As a result, no systematic disparity in quantitative measures, such as the frequency and density of these devices, is evident between translations with higher and lower scores. Third, candidates demonstrated limited ability to produce clearly discernible cohesion in their translations.

Consequently, raters had to rely more heavily on lexical and syntactic accuracy as the main evaluation criteria. This made cohesive indices less relevant to the results. Another possible reason may lie in the research instrument itself, Coh-Metrix. The cohesive indices in Coh-Metrix may be insensitive to the specific challenges of C-E translation. Coh-Metrix was developed for native English texts and only measures surface-level quantitative features, such as the frequency of connectives. However, these metrics cannot reflect the complex nature of C-E translation, which requires deep and abstract restructuring to follow English thinking and expression patterns. Candidates often actively employ methods, such as adding connectives and clarifying references, to achieve coherence. Yet Coh-Metrix only counts these surface features, without judging whether they are appropriate or effective. Thus, it is possible that the tool cannot adequately measure the actual cohesive competence demonstrated in cross-linguistic translation, explaining its weak correlation with translation quality.

In summary, this study's results partially align with prior research, with discrepancies attributable to variations in translation length, source text features, and candidates' proficiency.

Interpreting the Multiple Linear Regression Model: The Complexity of Translation and the Primacy of Lexis

The current study employed multiple linear regression to address RQ2. The results indicate that the five indices WRDNOUN, WRDAOAc, WRDHYPn, WRDHYPv, and CN/T collectively form the optimal MLR model, accounting for 17.1% of the variance.

To some extent, the limited explanatory power of the MLR model can be interpreted from the perspective of complexity in translation and translation assessment. As Richards (1953) noted, translation inherently involves complex cognitive processing. Thus, Model 8 reflects the multifaceted knowledge and skills candidates deployed in navigating the complex TEM8 translation task.

The assessment of translation is also complex. Like writing tests, translation tests also fall under the umbrella of behavioral tests (Xu & Ye, 2020). The evaluation of such tests constitutes an ill-structured task, implying that there is no definitive correct answer (DeRemer, 1998). To further elaborate, Fulcher et al. (2011) noted that the rating scale cannot encompass all language features present in translations, and the scale includes numerous descriptors that distinguish degrees. Consequently, raters may interpret these descriptors idiosyncratically. As the present model regresses language indices against rater-assigned scores, it captures only those features salient to raters' diverse evaluative criteria. Indices that are able to explain the remaining part of the language quality are still waiting to be discovered.

Moreover, since translations of TEM8 are the result of the complicated translation process, it is no wonder that they can be interpreted from various levels and indices. The MLR model incorporates indices from multiple linguistic levels rather than a single dimension. At the same time, even at the lexical level, the indices measure different aspects of words used in translations. For example, WRDNOUN is used to estimate the noun frequency, while WRDHYPn and WRDHYPv are used to assess the complexity of nouns and verbs in translation.

Notably, all four lexical indices were retained in the final model, compared with only one syntactic index. This suggests that lexical-level indices contribute more significantly to the model and exhibit better predictive power for translation quality in this context. One potential explanation for this is the limited length of the translations, which results in shorter sentences and less variation in syntactic features across translations of different quality. Consequently, syntactic features may not be as prominent in distinguishing translation quality. Another factor may be the limited proficiency of the candidates. As most of them are novice translators with little systematic training, they tend to use a word-for-word translation approach and lack the skill for restructuring sentences. This leads to translations that closely follow the sentence structure of the source text, resulting in minimal syntactic complexity differences between translations of varying quality levels. This inference aligns with the findings of Pan (2010), who found that TEM8 candidates' translations often closely mirror source text syntactic structures. Specifically, Pan (2010) found that 2004–2006 TEM8 candidates struggled to effectively deploy omission strategies. Consequently, it also provides evidence that students lack effective translation strategies and tend to rely heavily on a word-for-word approach in their translations.

In conclusion, the model can only explain 17.1% of the variance in scores, reflecting the complexity inherent in translation and its assessment. Lexical features contribute more significantly to the model compared to syntactic features, primarily due to the limited length of translations and the limited proficiency of candidates.

Conclusion

Major Findings

This study addressed two research questions concerning how language features predict TEM8 translation quality. For RQ1, simple linear regression revealed seven significant indices from lexical and syntactic levels: WRDNOUN (showing negative correlation), WRDAOAc, WRDHYPn, WRDHYPv, CP/T, CN/T, and CN/C (all positive). Each explained 3% to 7% of score variance. Notably, no cohesive indices correlated significantly with translation quality. For RQ2, multiple linear regression produced an MLR model of five indices (WRDNOUN, WRDAOAc, WRDHYPn, WRDHYPv, and CN/T) that collectively explained 17.1% of total variance. Overall, lexical features surpassed syntactic ones in predictive power, with word complexity being particularly significant, while cohesion proved irrelevant in this context.

Implications

From a theoretical perspective, this study advances the understanding of suitable indices for evaluating language quality in TEM8 translations. It develops an MLR model, comprising four lexical indices (WRDNOUN, WRDAOAc, WRDHYPn, and WRDHYPv) and one syntactic index (CN/T), collectively explaining 17.1% of score variance. These findings, which can be compared and discussed in similar future research, provide a valuable reference for studies in this area. Additionally, unlike most previous studies on TEM8 translation, which predominantly employed qualitative methods, this research relies more on quantitative analysis. This methodological attempt encourages future research to integrate statistical analysis with specific qualitative methods when exploring large-scale translation exams. Furthermore, while most existing studies on TEM8 translation exams concentrate on error analysis, this study shifts the focus to language features. By exploring the language quality of TEM8 from the perspective of language features, this study enriches the research from this perspective and offers a new reference point for future investigations. The study results presented in this study may also serve as a possible starting point, offering modest insights for future research aimed at building a more comprehensive theoretical foundation for automatic translation scoring.

In terms of practical application, this study offers several possible suggestions for translation pedagogy and insights for stakeholders involved in TEM8. For translation education, this study identifies some areas for improvement among candidates, particularly those with poor-quality translations. First, the study emphasizes the importance of not merely conveying the basic message but striving to achieve the precise meaning of the source text, which calls for more focus on the accuracy of word choice in translations. While general terms may offer a rough approximation of the source text's meaning, they fail to meet the standards of precision required for high-quality translations. Additionally, the findings suggest that translation pedagogy should emphasize syntactic restructuring to align with conventional English expression patterns. Unlike Chinese source texts, which often employ comma splices, English translations benefit from grammatical devices such as subordination and coordination. Therefore, translation pedagogy should prioritize these two dimensions. For test design, the findings suggest that increasing the length of source text may enhance the assessment of sentence complexity and cohesion. The disparity between previous research and the current study may stem from differences in source text length. Short source texts limit the ability to assess sentence complexity and the cohesion and coherence of translations effectively. Moreover, TEM8 rating scale explicitly reference syntactic complexity and cohesion, making longer source texts more conducive to a thorough evaluation of translation quality.

Limitations and Future Research

This study has three main limitations. First, the generalizability of the findings may be limited by the narrow scope of the analysis, which only includes data from two test administrations. Future research should incorporate translations from more years to strengthen findings. Second, the lack of supplementary qualitative data restricts our understanding of the reasoning behind linguistic choices. Future studies would benefit from mixed-methods approaches that include interviews or think-aloud protocols with candidates to gain deeper insights into translation processes and strategies. Third, the separate scores for fidelity and language quality were not available. Consequently, it was not possible to determine how much variance the current models could explain specifically for language quality. Future research could record fidelity and language quality scores separately to yield more precise and detailed results.

Full Disclosure of Generative Artificial Intelligence (GenAI) Use

The authors herein disclose the use of Generative Artificial Intelligence (GenAI) tools during the manuscript preparation process related to the present study.

1. GenAI Tool Information: The GenAI tool used was Doubao (12.2.0), developed by Beijing Chuntian Zhiyun Technology Co., Ltd.
2. Scope and Purpose of GenAI Use: GenAI was employed to assist in improving the academic expression and organization of English sentences, ensuring the clarity and coherence of the content. GenAI is not involved in core research design, data collection, statistical analysis, or result interpretation.
3. Limitations of GenAI Use and Authors' Responsibilities: The GenAI tool was used solely as an auxiliary tool, and all content generated by GenAI was thoroughly reviewed, verified, and revised by the authors. The authors bear full responsibility for any errors.

Funding

This work was supported by Shanghai International Studies University under Grant No. 2024DSYL008.

References

Adil, M. (2020). Exploring the role of translation in communicative language teaching or the communicative approach. *Sage Open*, *10*(2). https://doi.org/10.1177/2158244020924403

American Translators Association. (n.d.). *How the exam is graded*. https://www.atanet.org/certification/how-the-exam-is-graded/

Chen, X. L. (2022). Cultivating outstanding translation talents to promote China's international communication. *Foreign Language Education in China*, (2), 3–4.

Chen, Y. (2016). An examination of rater performance on the Chinese-to-English translation section of TEM8 in 2015. *Technology Enhanced Foreign Language Education*, (5), 77–82. https://doi.org/10.20139/j.issn.1001-5795.2016.05.015

Chi, Y. Q. (2019). An investigation into the cohesive explicitation of Chinese-English translation by English majors based on Coh-Metrix. *College English Teaching & Research*, (2), 3–9. https://doi.org/10.16830/j.cnki.22-1387/g4.2019.02.001

CIOL. (2022). *Qualification specification*. https://www.ciol.org.uk/sites/default/files/DipTrans-QualSpec-Apr2022.pdf

Colina, S. (2008). Translation quality evaluation: Empirical evidence for a functionalist approach. *The Translator, 14*(1), 97–134. https://doi.org/10.1080/13556509.2008.10799251

Colina, S. (2009). Further evidence for a functionalist approach to translation quality evaluation. *Target, 21*(2), 235–264. https://doi.org/10.1075/target.21.2.02col

Cook, G. (2010). *Translation in language teaching: An argument for reassessment*. Oxford University Press.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7–29.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29. https://doi.org/10.1177/0265532209359514

Göpferich, S. (2009). Towards a model of translation competence and its acquisition: The longitudinal study TransComp. In S. Göpferich, A. L. Jakobsen & I. M. Mees (Eds.). *Behind the mind: Methods, models and results in translation process research* (pp. 11–37). Samfundslitteratur Press.

Göpferich, S. (2013). Translation competence: Explaining development and stagnation from a dynamic systems perspective. *Target-International Journal of Translation Studies, 25*(1), 61–76. https://doi.org/10.1075/target.25.1.06goe

Graesser, A. C., D. S. McNamara, M. M. Louwerse & Z. Cai. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202. https://doi.org/10.3758/BF03195564

Hai, F. (2003). The translation process and reflections: A study on vocabulary strategies in Chinese-English translation of TEM8 (2002) candidates. *Chinese Translators Journal*, (1), 81–83.

House, J. (2014). *Translation quality assessment: Past and present* (1st ed.). Routledge. https://doi.org/10.4324/9781315752839

Jiang, J. L. (2013). An automatic approach to evaluating the linguistic quality of English-Chinese translations. *Modern Foreign Languages*, (1), 85–91.

Jiang, J. L. (2016). The application of Coh-Metrix in foreign language teaching and research. *Foreign Languages in China*, (5), 58–65. https://doi.org/10.13564/j.cnki.issn.1672-9382.2016.05.009

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill/Irwin.

Kyle, K. & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34*(4), 513–535. https://doi.org/10.1177/0265532217712554

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. F. & Xu, Q. (2016). L2 syntactic complexity analyzer and its applications in L2 writing research. *Foreign Language Teaching and Research*, (3), 409–420, 479–480.

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

Melis, N. M. & A. Hurtado Albir. (2001). Assessment in translation studies: Research needs. *Meta, 46*(2), 272–287. https://doi.org/10.7202/003624ar

Nida, E. & C. Taber. (1969). *The theory and practice of translation*. Brill.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

PACTE. (2000). Acquiring translation competence: Hypotheses and methodological problems of a research project. In A. Beeby, D. Ensinger & M. Presas (Eds.). *Investigating translation. Selected papers from the 4th international congress on translation, Barcelona, 1998* (pp. 99–106). John Benjamins.

PACTE. (2003). Building a translation competence model. In F. Alves (Ed.). *Triangulating translation: Perspectives in process oriental research* (pp. 43–66). John Benjamins.

Pan, M. W. (2010). A corpus-based study on the learners' application of the omission strategy in a Chinese-English translation test. *Foreign Languages Research*, (05), 72–77. https://doi.org/10.13978/j.cnki.wyyj.2010.05.013

Pan, M. W. & Z. Shen. (2020). Test for English majors in the new era: Challenges, solutions and future endeavors. *Technology Enhanced Foreign Language Education*, (02), 62–68. https://doi.org/10.20139/j.issn.1001-5795.2020.02.008

Redelinghuys, K., & Kruger, H. (2015). Using the features of translated language to investigate translation expertise: A corpus-based study. *International Journal of Corpus Linguistics, 20*(3), 293–325. https://doi.org/10.1075/ijcl.20.3.02red

Reiss, K. (2004). *Translation criticism: The potentials and limitations - Categories and criteria for translation quality assessment*. Shanghai Foreign Language Education Press.

Richards, I. A. (1953). Toward a theory of translating. In A. F. Wright (Ed.). *Studies in Chinese thought* (Vol. 55). Chicago University Press.

Shao, W. Y. & Shao, Z. H. (2012). On the translation of the psychological description text in light of the Chinese and English contrastive analyses: An evaluation of the Chinese-English translation test papers of TEM8(2012). *Foreign Languages and Literature*, (5), 125–129.

Tan, L. (2022). *NeoSCA* (version 0.0.55) [Computer software]. Github. https://github.com/tanloong/neosca

Wang, J. Q. & Zhu, Z. Y. (2017). A study on the automated assessment of Chinese-English translation competence. *Foreign Languages in China*, (2), 66–71. https://doi.org/10.13564/j.cnki.issn.1672-9382.2017.02.009

Wang, J. Q., Yu, X., & Wu, W. N. (2021). A study on translation quality assessment based on lexical measurement features. *Chinese Translators Journal*, (5), 113–120.

Wang, Q. (2009). Chinese-English interlanguage: A study on TEM8 translation corpus. *Shanghai Journal of Translators*, (2), 74–77.

Williams, M. (2004). *Translation quality assessment: An argumentation-centred approach*. University of Ottawa Press.

Wolfe-Quintero, K., S. Inagaki & H. Kim. (1998). *Second development in writing: Measures of fluency, accuracy and complexity*. University of Hawaii Press.

Xu, Y. & Ye, M. L. (2020). An investigation of rating strategies in the translation assessment: Based on CET-4. *Journal of China Examinations*, (6), 43–50. https://doi.org/10.19360/j.cnki.11-3303/g4.2020.06.007

Yang, Z. H. (2012). Quantitative assessment of translation quality: Models, trends and implications. *Foreign Languages Research*, (6), 65–69. https://doi.org/10.13978/j.cnki.wyyj.2012.06.013

Yang, Z. H. (2019). *Translation testing and assessment research*. Foreign Language Teaching and Research Press.

Zhu, Z. Y. & Wang, J. Q. (2020). Relationships among genre difference, language features and EFL learners' Chinese-English translation quality. *Foreign Languages in China*, (4), 58–68. https://doi.org/10.13564/j.cnki.issn.1672-9382.2020.04.008

Zou, S. (Ed.). (2012). *Language testing* (2nd ed.). Shanghai Foreign Language Education Press.

Appendix A

The Rating Scale of C-E Translation in TEM8

TEM8 汉译英评分标准

说明：该项目满分10分。根据考生的**译文忠实性和语言适切性**分别打分。
“译文忠实性”满分为7分，按考生的表现赋分： 0----1----2-----3-----4-----5-----6-----7
极差 不合格 合格 良好 优秀
“语言适切性”满分为3分，按考生的表现赋分：0-------1--------2--------3
极差 合格 良好 优秀
空白卷、仅写了一些与任务无关的字句、仅仅照抄指导语得**0**分。阅卷教师不用计算总分。

等级	操作性描述
优秀	能基本进行词语的正确翻译（突破汉语原文词语的表层含义，实现指称、语境、语体、搭配等多方面意义的对等）；有较强的翻译转移意识——能在不改变原意的基础上进行恰当的翻译转移（能避免汉语原文中的冗余信息，能补充汉语语句间的隐性逻辑关系，能对译句进行深层次的组构，译句能符合英语句子的信息架构和重量特点 ……）；没有语言上的失误，或仅有轻微失误（如介词、名词、代词、冠词等的误用）
良好	能一定地进行词语的正确翻译（突破汉语原文词语的表层含义，实现指称、语境、语体、搭配等多方面意义的对等）；基本能体现出翻译转移意识——在不改变原意的基础上进行一定的汉英翻译转移（如避免汉语原文中的冗余信息，补充汉语语句间的隐性逻辑关系，对译句进行深层次的组构，译句能符合英语句子的信息架构和重量特点 ……）；有些一般性语言错误（如谓语动词失误、时态错误、主谓一致性失误、词类错误、语序错误、搭配错误等）
合格	尚能进行词语的翻译，但多局限于指称意义的翻译，在语境、语体、搭配等意义的对等上有明显不足；句义尚且忠实，但缺乏翻译转移意识，基本不能在不改变原意的基础上进行恰当的翻译转移；有较多一般性语言错误或个别严重性语言失误（如包括破裂句、连写句在内的全句句法失误、从句失误等）
不合格	基本不能进行词语的翻译，意思偏离严重；没有翻译转移意识，机械式的对应严重，或对原文意思进行篡改，或有明显遗漏；有较多严重性语言失误
极差	不能进行词语的翻译，没有翻译转移意识，完全局限于机械式的对应，或对原文意思进行明显篡改，或有大量遗漏；有相当多的严重性语言失误

It should be noted that the rating scale for the TEM8 is formulated in Chinese, as this enables Chinese raters to achieve a more accurate and in-depth understanding of the scale.

Appendix B

Source Texts and Examples of Translations in the Years 2022 and 2023

The source text of the year 2022:
旷野与城市，从根本上讲，是对立的。
城市驱散了旷野原有的住民，破坏了旷野古老的风景，越来越多地以井然有序的繁华，取代我行我素的自然风光。
今天，旷野日益退缩着，但人们不应忽略旷野，漠视旷野，而要寻觅出与其相亲相守的最佳间隙。善待旷野就是善待人类自身。要知道，人类永远不可能以城市战胜旷野。

Examples of translations in the year 2022:

Countryside and city, actually, contrast. City City destroyed countryside old scene. More and more order instead freedom scene. Today, countryside is decrease ultimately, but, person don't should Furthermore, person seek. (#16, score = 2)

For endness, field is contrast to city. Many people who lived in the city have been escaped. The city destory old seen that in field before. Now, the field become small every day. But people shouldn't ingort field. Be kind to field is kind to human. (#45, score = 3)

Actually, ground and cities are contrast. Urbanization causes the decrease of local habitants and changes original views of ground. There are more lines of buildings instead of natural sights. Nowadays, with the reducing of ground, people should try to find ways to stay with nature. It's not a good choice without attention and reaction. Existence of ground much relates to humans neither, in the past or future, people can never be the winner. (#211, score = 4)

Cities drive the original residents and destroy the antique scenery of the wild. More and more prosperities which are in order take place of the free scene of nature. Nowadays, the wild decrease day by day. But people shouldn't ignore and scorn the wild. People should try to find out the best gap to live together with the wild. Treating the wild well equals to treating ourselves well. People should know that they can never beat the wild with cities. (#301, score = 5)

Cities not only excluded original residents from the remorse countryside, but also destroyed old and beautiful sceneries of the remorse countryside. Consequently, more well-ordered abundance substitues tedious views of the countryside.

Nowadays, countryside sceneries gradually fade day by day. However, people should not ignore the remorse countryside and regard it indifferently, instead, people should discover best gaps of relations so as to people's loving and protection. Kindly treating the remorse countryside is treating human-beings themselves kindly. We all need to know that Humans will never depend on cities to conquer the remorse countryside. (#501, score = 6)

Cities have diven original in habitants in the wild away and destroyed its ancient scenery. More prosperity in order has superseded the natural scenary sights.

Currently, the wild is increasingly shrinking and decreasing. However, people shouldn't ignore it or disregard it; instead, they ought to find the most suitable distance to approach and accompany it. Being kind to the wild means to treat human himself kindly. We must be aware that it's never possible for mankind to defeat the wild by cities. (#769, score = 7)

Cities expelled indigenous inhabitants of vast lands and devastated the old landscape of vast lands which resulted in a phenomenon that the orderly prosperity increasingly replaces the free natural landscape. Today, the area of vast lands is gradually shrinking. But we human beings should not neglect or overlook vast lands. Thus, we have to find the best point where we live in harmony with it. Being kind to the lands means treating ourselves well. We're supposed to know that human beings will never manage to replace vast lands with cities. (#861, score = 8)

Cities chased away the native inhabitants of the wild nature, destroying the old scenery and replacing the undisturbed natural beauty with orderly urban prosperity. Nowadays, the wild nature is shrinking. Instead of ignoring wild nature or casting a cold glance at it, we humans should find a way to coexist with it in harmony. Taking care of the wild nature is taking care of humans ourselves. After all, we can never conquer nature with cities. (#973, score = 9)

The source text of the year 2023:

中国传统文化是我们先辈传承下来的丰厚遗产。她无时无刻不在影响今天的中国人，为我们开创新文化提供历史根据和现实基础。传统文化在影响现实的同时，也必然在新时代的氛围中发生蜕变。中国传统文化犹如一条奔腾了五千年的永不干涸的大河，她亦旧亦新，不断吐故纳新，持续创新。

Examples of translations in the year 2023:

China tradition is our acent wealthy. She always inflence China people now and creat culture to provide. Tradition culture inflence now world meanwhile change in new world. China tradition culture like run river. She always beautiful and creat happiness. (#30, score = 2)

Chinese traditional cultural

Chinese people has been affected by Chinese traditional cultural for long time. It provided historical experiences for culture invation. Nowdays was affected by it. In the meantime, it will chang in the future. It just like a river. This river had been launched for five thousands years. (#126, score = 3)

China tradional culture is the fully hiretativa which we get frome our profesor. She is infulence our daily life, frome now and on, give the history ideal and realitice function, when the tradional culture infulence the realitic sociaty, it also created in the new history atmosphere. China tradional culture just like a never dirty river which go through five thousheds years. She contined tradional and creative, continued to creative forever. (#223, score = 4)

The traditional culture of China now has a great influence on each Chinese people all the time and also provide us historical evidence and social bases from creating the new culture. When the traditional culture of China effects on reality, at the same time the fact traditional culture of China has changed in the period of new century is undoubtable. The traditional culture of China is like a never dry river driving for five thousand years. It is involved in old culture and new culture, but constantly still give up the out-of-time part and is willing to accept the excellent part and creat the new part. (#401, score = 5)

She is influencing the Chinese people today all the time, and providing historical and real base for the innovation of our culture. When she is making a difference on the reality, she will, at the same time, renew herself in the new times. Like a long-lasting great river which has running for five thousand years, Chinese traditional culture is both old and new, and she will continue to abandon the old-fashioned things and embrace the new things, and keep the innovation go on and on. (#601, score = 6)

She continuously influences Chinese people today, providing historical roots and real-world basis for us bringing out new cultures. Traditional culture definitely has her own change and growth in this new era, while keeping to influence the reality. Traditional Chinese culture is old yet new and conventional yet innovative, constantly evolving and progressing like a river running over five thousand years with unfailing energy. (#834, score = 7)

It has been having influence on Chinese people today all the time, offering historical base and realistic foundation to us for developing new cultures. While influencing the reality, traditional cultures will be bound to make changes and improvements under the atmosphere of new times. Chinese traditional culture is like a never-empty spectacular river which has never stopped for the past five thousand years, integrating the old things with the new things, constantly assimilating new ideas while abandoning old thoughts and keeping the ability of innovation. (#925, score = 8)

Chinese traditional culture is the rich property, which be given from our ancestors. It constantly influences our Chinese in every aspects today and also provides us with the solid foundation and historical evidence to creat new cultures. The traditional culture is not only influencing the reality but also changing constantly under the new era. Chinese tradition culture is a 5000-year river, consistently flowing in our country, which includes all the things whether they are new or not. It will remove the dated things, receive the advanced things and constantly innovate the culture. (#996, score = 9)

Appendix C

Inventory of Language Features and their Definitions

Aspect		Index	Definition
Lexical features	Lexical diversity	LDTTRa	Type-token ratio for all words
		LDVOCDa	Vocd for all words
	Lexical density	WRDNOUN	The number of nouns per 1000 words
		WRDVERB	The number of verbs per 1000 words
		WRDADJ	The number of adjectives per 1000 words of the text
		WRDADV	The number of adverbs per 1000 words
		WRDPRO	The number of pronouns per 1000 words of the text
	Lexical semantic	WRDPOLc	Average polysemy for content words
		WRDHYPn	Hypernymy for nouns
		WRDHYPv	Hypernymy for verbs
		WRDHYPnv	Hypernymy for a combination of both nouns and verbs
		WRDAOAc	When a content word is learnt by children
		WRDFAMc	A rating of how familiar a content word seems to an adult
		WRDCNCc	How concrete or non-abstract a content word is
		WRDIMGc	How easy it is to construct a mental image of the content word
		WRDMEAc	The degree to which a content word is associated with other words
Syntactic complexity features	Length of production unit	MLC	Mean length of clause
		MLS	Mean length of sentence
		MLT	Mean length of T-unit
	Sentence complexity	C/S	Sentence complexity ratio
	Subordination	C/T	T-unit complexity ratio
		CT/T	Complex T-unit ratio
		DC/C	Dependent clause ratio
		DC/T	Dependent clauses per T-unit
	Coordination	CP/C	Coordinate phrases per clause
		CP/T	Coordinate phrases per T-unit
		T/S	Sentence coordination ratio
	Particular structures	CN/C	Complex nominals per clause
		CN/T	Complex nominals per T-unit
		VP/T	Verb phrases per T-unit

Aspect		Index	Definition
Cohesive features	Latent semantic analysis	LSASS1	LSA similarity between adjacent sentences
		LSASSp	LSA similarity between all possible pairs of sentences in a paragraph
		LSAPP1	LSA similarity between adjacent paragraphs
		LSAGN	LSA Given-New
	Connectives	CNCCaus	Causal connectives
		CNCLogic	Logical connectives
		CNCAdd	Additive connectives
		CNCTemp	Temporal connectives