Distinguishing ChatGPT-Generated Translation and Neural Machine Translation from Human Translation: A Linguistic and Statistical Approach

Abstract: The growing popularity of neural machine translation (NMT) and generative artificial intelligence represented by ChatGPT underscores the need for a deeper understanding of their distinct characteristics and inter-relationships, an area that remains underexplored. This study aims to bridge the gap by investigating the distinguishability of ChatGPT-generated translation and NMT from human translation, and exploring the extent to which ChatGPT-generated translation and NMT align with or diverge from patterns of human translation (HT). To achieve these objectives, we employ machine learning algorithms, multidimensional analysis (MDA), and distance computation based on a customized corpus comprising diplomatic translations. The results reveal a clear distinction among these translation varieties, as evidenced by the high accuracy of our machine learning algorithms. Additionally, it is observed that ChatGPT-produced translations exhibit greater similarity to NMT than to HT across most MDA dimensions. This finding is further supported by distance computation, which indicated that translations generated by ChatGPT bear resemblance to both HT and NMT, but are closer to NMT. Furthermore, we identify a distinct form of machine translationese associate with the Simplification and Explicitation translation universal hypotheses.

Keywords: neural machine translation; ChatGPT; machine learning; multidimensional analysis; distance

1. Introduction

In the dynamic intersection of translation studies and computational linguistics, the emergence of large language models exemplified by ChatGPT, and NMT engines such as Google Translate and DeepL, has propelled a reevaluation of traditional translation paradigms. Recent advancements have significantly enhanced the capabilities of AI-driven translation technologies, bringing them closer to human-level proficiency (Hu and Li 2023; Hendy *et al.* 2023). However, the extent to which machine translations resemble the patterns and characteristics of HT remains a subject of considerable academic inquiry (Gaspari *et al*. 2015).

Since its release, ChatGPT has emerged as a prominent subject in discussions surrounding AI and its societal implications, primarily due to its efficacy in addressing practical challenges, including translation across diverse registers. While some scientists contend that ChatGPT has possessed the capability to generate language resembling human expression and accomplish intricate language-related tasks, skeptics emphasize the fundamental disparities between the output of language models and human language.

Furthermore, despite ChatGPT's ability to perform translation tasks, it significantly differs from NMT engines across various aspects. While ChatGPT is designed as a versatile language model capable of handling a wide range of language-related tasks, NMT systems are specifically tailored for translation purposes. From a technical perspective, ChatGPT adopts a generative decoder-only architecture, lacking the encoder component that is integral to NMT systems. Additionally, NMT systems primarily rely on parallel corpora for training, whereas ChatGPT is trained on monolingual corpora in multiple languages. Given these notable distinctions, further investigation is warranted to systematically explore how these differences may impact the style, patterns, and characteristics of their respective translation outputs (Hendy *et al.* 2023).

Another interest of this study is to investigate how ChatGPT-generated translation and NMT align with or diverge from established hypotheses of Translation Universals (TUs) and thereby explore the potential existence of machine translationese. Existing corpus-based studies on TUs have predominantly focused on translations performed by human translators, with only a limited number of studies delving into the realm of machine translation (Lapshinova-Koltunski 2015; Zhang and Toral 2019). Consequently, a significant knowledge gap exists regarding whether machine translations, devoid of the cognitive mechanisms inherent in human translators, conform to certain TUs. By drawing from existing TU hypotheses, we can expand the limited knowledge of advanced translation technologies in the AI era and deepen our understanding of machine translationese. In this study, particular emphasis is placed on the Explicitation and Simplification hypotheses, as they have received the most extensive investigation and can be readily operationalized using computational methods. The Simplification hypothesis pertains to the tendency for translated texts to exhibit simpler syntax and vocabulary compared to their source texts (Laviosa 1998), while the Explicitation hypothesis refers to the inclination to make implicit information explicit in the translation (Blum-Kulka 1986).

Our study is centered on the linguistically rich and contextually complex translations of Spokesperson's Remarks. These texts, characterized by a blend of spontaneity in response and formality in presentation, serve as valuable resources to probe the styles and patterns generated by machine and human translators. The dual nature of these remarks offers an ideal testing ground to examine how different translation sources handle the intricacies of language, and how they strike a balance between conversational fluidity and structured formality.

Specifically, this study is guided by the following research questions (RQs):

RQ1: Are ChatGPT-generated translations, NMT, and HT distinguishable from each other?

RQ2: What are the distinctive linguistic patterns of the three translation varieties?

RQ3: Are translations from ChatGPT more similar to HT or NMT?

RQ4 How are the Simplification and Explicitation hypotheses manifest in these translation varieties?

We employ a multi-feature methodology by examining a group of relevant features at lexical, syntactic, and textual levels simultaneously. Specifically, we adopt widely applied machine learning methods, including classification and clustering techniques, together with Biber's (1988) multidimensional analysis (MDA) to answer the first two research questions. Previous studies have already demonstrated the feasibility and necessity of analyzing translation through a simultaneous examination of multiple features (Hu *et al*. 2019; Kruger and van Rooy 2016; Kruger and van Rooy 2018). This approach also corresponds to a "new, updated research agenda" for translation studies (De Sutter and Lefer 2020, p. 6), which calls for an interdisciplinary scope, a multimethodological framework, and an in-depth understanding of the multidimensional structure of translation (Calzada Pérez and Sánchez Ramos 2021). For the third question, we resort to the technique of distance calculation and dimension-reduced visualization to reveal the similarities among the three translation varieties as measured by linguistic features.

2. Literature Review

2.1. Comparative studies of NMT, ChatGPT, and HT

Unlike traditional NMT systems that are constrained by the source language and its encoded representation, ChatGPT can generate translations in a more flexible manner, exhibiting more lexical diversity, syntactic variations, and textual adjustments (Hendy *et al*., 2023). This can lead to fluent and context-aware but potentially less accurate translations, especially in cases where strict adherence to the source language is crucial (Hendy *et al.* 2023; Peng *et al.* 2023).

Numerous studies have already compared NMT and HT from various perspectives. Most of them focus on literary translation, aiming to identify the differences between the two translation approaches (Kuo 2018; Frankenberg-Garcia 2022; Hu and Li 2023). For instance, Kuo (2018) examined the use of function words in machine-translated Chinese and in original Chinese, and discovered an overuse of function words in MT. Frankenberg-Garcia (2022) conducted a comparative lexical analysis of literary works translated by NMT and HT, revealing that HT to exhibited more explicitation, idiomaticity, register awareness, and risk aversion. In a comparison between Shakespearian plays translated by DeepL and human translators, Hu and Li (2023) observed a certain degree of creativity in NMT. More pertinent to the current study is Sheng and Kong (2023), which examined the machine-translated Chinese political document in contrast to human translation, and found NMT to lack the subjectivity and flexibility of professional translators. These previous studies have provided valuable insights into the characteristics of NMT and HT. However, given the emergence and widespread adoption of ChatGPT, it is essential to expand the scope of comparison to include ChatGPT, so as to stay up-to-date with the rapid advancements in AI-powered language technology.

There are also abundant studies comparing the translation quality of advanced NMT engines and ChatGPT using automated metrics and human evaluation (Hendy *et al.* 2023; Raunak *et al.* 2023). Their findings indicated that for high-resource language pairs, such as English and French, ChatGPT and GPT-4 could exhibit state-of-the-art translation capabilities, rivaling or outperforming the mainstream NMT systems. However, for low-resource language pairs and in highly domain-specific fields, ChatGPT still lagged behind NMT systems (Karpinska and Iyyer 2023). However, since ChatGPT is mainly trained on high-resource languages, we are not fully aware of its competence in understanding and translating a middle-resource language like Chinese. Moreover, most of the previous assessments are conducted on publicly available corpora from OPUS or WMT, which leaves other registers to be underexplored.

2.2. Multi-feature methods in translation studies

Multi-feature analyses are commonly employed in corpus-based translation studies to explore the simultaneous effects of multiple relevant properties at lexical, syntactic, and textual levels. By considering multiple linguistic properties together, these methods provide a macroscopic view of various linguistic phenomena that cannot be captured by a single feature alone.

These methods are often applied in studies of translator attribution and translation stylistics in literary works. For instance, Rybicki and Heydel (2013) attempted to attribute the correct translator in a collaborative translation that was completed by more than one translator. Mohamed *et al*. (2023) used machine learning algorithms to attribute Arabic translations of well-known literary books, aiming to identify which translator translated what texts. Wang and Li (2011) compared two Chinese translations of Ulysses using parallel and comparable corpora. They analyzed keywords, lexical features, and syntactic features, concluding that translator fingerprints could be identified. Similarly, Fang and Liu (2023) conducted a comparative study on three Chinese translations of *Alice's Adventure in Wonderland.* Their findings suggested that translator style was visible and could be identified through multi-feature analyses.

Based on either customized or balanced corpora, multi-feature methodology also serves as a powerful tool for studies interested in distinguishing translational texts from non-translational ones. This line of research treats translated language as a distinctive type of language, often referred to as the “third language” (Duff 1981, p.12), the “third code” (Frawley 2000, p.253), “constrained language" (Kruger and van Rooy 2016), or "translationese" (Newmark 1991). A representative study was Hu *et al*. (2019), which adopted a multi-feature statistical model adapted from MDA to examine differences between translated English and original English across registers. Kruger and van Rooy (2016) used MDA to investigate the relationships between translated English and L2 varieties of English, and whether their shared features could be explained by constrained bilingual language production. Treating translation as a special language variety because of contact with other languages, Kruger and van Rooy (2018) explored the influence of translation as well as other language varieties, register differences, and their combined effect in linguistic variation using MDA and a regression model. They identified register as the most significant factor in explaining linguistic variations. More related to the current study is Calzada Pérez and Sánchez Ramos (2021), which only focused on one specific register, the parliamentary speech, to investigate linguistic variation by drawing on the identified dimensions in Biber (1988).

These earlier studies have already delved into the characteristics of translated language from various perspectives, showing that multi-feature methods can be used to investigate linguistic variations across text types, registers and varieties of language. As neural machine translation and generative AI gain widespread adoption in recent years, scholars have paid more attention to identify their distinctive characteristics and patterns. Our study thus treats NMT and ChatGPT-generated translation as two different varieties of human translation, and aims to find their similarities and differences as exhibited in their respective translation output.

3. Methodology

The methodology of this study features a combination of various computational techniques. Specifically, machine learning algorithms, MDA, and distance computation are employed to examine the relationships among ChatGPT-generated translations, NMT, and HT. The objective is to explore their distinguishability, distinctive characteristics, and relative proximity to one another. Five major procedures are involved: (1) corpus building and text processing; (2) feature extraction; (3) applying machine learning classification algorithms; (4) implementing MDA; (5) calculating and visualizing the distance among the three translation varieties; (6) discussing machine translationese in relation to the Simplification and Explicitation TU hypotheses

3.1. Corpus building and text processing

The customized corpus in this study consists of three sub-corpora: (1) English translation made by institutional translators (Human_Trans); (2) English machine translation by Google Translate (Machine_Trans); (3) English translation generated by ChatGPT (ChatGPT_trans). Their source texts are 147 pieces of spokesperson's remarks between 2018 and 2022. Each remark contains questions proposed by foreign reporters and answers from the Chinese spokespersons centering several foreign affairs at a range of press conferences. Questions in these materials are asked in English, which are answered by the spokespersons in Chinese and then translated into other languages by institutional translators. Both the Chinese source texts and human translations underwent adjustment in wordings and contents, as well as corrections of speaking errors to be in line with the requirements of government websites.

We choose these textual materials out of three considerations: (1) data availability, as all the textual materials are readily accessible from the official website[1]of the Ministry of Foreign Affairs of the People's Republic of China; (2) high-quality reference, since the human translation is performed by professional institutional translators working for the government and can serve as reliable reference; (3) complexity, because diplomatic discourse needs to demonstrate spontaneity and formality simultaneously.

To build the sub-corpus NMT_trans, the source texts are translated on the text level into English by Google Translate. The sub-corpus ChatGPT_trans consists of English translations by ChatGPT, or the gpt-3.5-turbo model from OpenAI. Both NMT and translations by ChatGPT are conducted on 20 October 2023. Basic information of the sub-corpora used in this study is listed in Table 1.

Table 1 Basic information of the sub-corpora

Sub-corpus	Tokens	Number of texts
Source	59,505	147
Human_Trans	38,697	147
Machine_Trans	40,675	147
ChatGPT_Trans	41,162	147

Since our objective is to uncover distinctive patterns inherent in the three translation varieties, it is crucial to mitigate the external influence of stylistic differences attributed to spokespersons, text lengths, and contents. To be specific, as each spokesperson's remark may vary in length and topic, using their translations directly as text samples for frequency calculation could lead to misleading results.

To address this concern, we resorted to a technique called rolling stylometry (Eder, 2016) to process the textual data. This approach involved the following steps: First, we concatenated all the translated texts in each sub-corpus into a single file. Next, we split the three concatenated files into equal-sized blocks of 5000 words. To ensure overlap and continuity between samples, we set the moving window size as 500 words. As a result, each sample, except the first and last ones, overlaps with its preceding and following samples. The remaining contents less than 5000 words were discarded. This process guaranteed that each sample is representative of its translation variety, and is adequate for feature extraction. We demonstrate the operation of rolling stylometry in Figure 1. Following this procedure, we randomly sampled 50 samples for each sub-corpus. These samples will be repetitively used to address our three research questions.

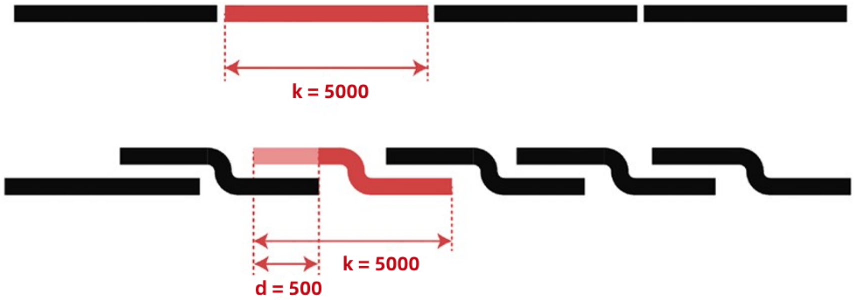

Figure 1 An illustration of rolling stylometry adapted from (Eder 2016)

3.2. Feature extraction

Considering the uncertainty of which features may be significant to distinguish different translation varieties, we followed Biber (1988) and Hu *et al*. (2019) to incorporate as many standardized features as possible in the initial stage. We used MAT (Nini 2019) and MFTE (Le Foll and Shakir 2023) to tag and extract 121 features in total. MFTE was developed as an extension of MAT to incorporate semantic tags from Biber *et al.* (1999) and Biber (2006). MFTE evaluated its performance in comparison to MAT and presented the reliability report[2]. The output is a csv file of normalized frequency counts. It will be used in the following machine learning experiments and MDA to address RQ1 and RQ2. The entire list of features and their descriptions can be found in Appendix A.

3.3. Machine learning

To examine whether the three translation varieties are distinguishable, we applied five supervised machine learning classifiers: Linear SVM, Random Forest, Multi-Layer Perceptron (MLP), AdaBoost, and Naïve Bayes, to perform a three-way classification task. We split our samples into a training set comprising 120 samples and a testing set containing 30 samples. The five classifiers were only trained on the training set without being exposed to the testing set, and thus we were able to examine whether they can classify the unseen samples into the correct translation variety. If we see high accuracy, recall, and F1 scores, it means that the extracted features hold great discriminatory power to separate different translation varieties.

3.4 Multidimensional Analysis

To gain an in-depth understanding of the linguistic patterns of the three translation varieties, we employed MDA (Biber 1988) to analyze linguistic variation as reflected by the co-occurrence of extracted features. Based on the statistical technique of factor analysis, MDA is a "bottom-up, data-driven" method (Thompson *et al*. 2017) that considers registers, dimensions of co-occurring linguistic features, and text functions comprehensively.

While we could base our analysis directly on the dimensions identified in Biber (1988), we decided to conduct a factor analysis from scratch, with an aim to identify context and register-specific factors in diplomatic translation. The procedure involves the following six steps: (1) selecting statistically significant linguistic features; (2) determining the number of factors based on scree plot; (3) performing factor extraction and factor rotation; (4) retrieving factor loadings; (5) interpreting the meaning of each factor; (6) comparing dimension scores of samples from HT_trans, NMT_trans, and ChatGPT_trans.

To avoid the multicollinearity issue, we first examined whether or not each feature was statistically significant to distinguish the three translation varieties. This was realized by conducting non-parametric Kruskal-Wallis H test. To avoid Type 1 error caused by multiple comparisons, we used Bonferroni correction. We also computed the pairwise correlations among the statistically significant features to exclude those exhibiting strong correlations with other features.

In addition, we carried out KMO Measure of Sampling Adequacy and Bartlett's Test of Sphericity to examine the feasibility of our data for factor analysis. After that, we calculated eigenvalues, proportions of variance, and cumulative variance of each factor to determine the number of factors.

To enhance the interpretability of factors, we performed the Varimax factor rotation. This method enforces orthogonality between factors, meaning that the factors are uncorrelated with each other. This simplifies the interpretation by ensuring that each factor represents a unique and independent dimension of the linguistic features.

We also calculated the dimension scores of all the text samples within each factor by summing up the scores of positive features and subtracting those of negative features. We then standardized these scores using z-transformation, and used boxplots to display their relative positions along each dimension.

3.5. Calculation of the pairwise Euclidean distances and visualization with t-SNE

To investigate whether translations by ChatGPT were closer to HT or NMT, we calculated the pairwise Euclidean distance among these three translation varieties using the z-transformed dimension scores, which resulted in three distance matrices. To represent the distance distribution, we used t-SNE (t-Distributed Stochastic Neighbor Embedding) for the purpose of visualization. T-SNE is a nonlinear dimensionality reduction technique that models the pairwise similarities between data points in a high-dimensional space, and maps them to a lower-dimensional space, where the similarities are preserved as much as possible. By using this technique, we can visualize the distances among the three translation varieties in a reduced-dimensional space, and gain an intuitive understanding of the proximity of ChatGPT-generated translations to HT and NMT.

4. Analysis of Findings

4.1. Classification results

Table 2 presents the outcomes of the five classifiers on the test set. Overall, these results show the effectiveness of the classifiers in accurately classifying our samples into their respective classes. Notably, the Random Forest classifier and MLP classifier achieved full accuracy, precision, recall, and F1-scores for all classes, suggesting that HT, NMT, and ChatGPT samples are perfectly distinguishable. Linear SVM and AdaBoost gained full points on all the metrics for HT samples classification, but made a few mistakes on NMT and ChatGPT samples. Likewise, the Naïve Bayes classifier exhibited slightly lower scores for the NMT and ChatGPT class, but displayed perfect scores for the HT class, indicating the presence of a distinct boundary between HT and the other two translation varieties. However, it also suggests that NMT and ChatGPT-produced translations may share some commonalities that cause the classifiers to misclassify them into the wrong categories.

Table 2 Results of machine learning classifiers with linguistic features

Classifiers	Accuracy	Class	Precision	Recall	F1-score	Support
Linear SVM	0.97	ChatGPT	0.91	1.00	0.95	10
		HT	1.00	1.00	1.00	9
		NMT	1.00	0.91	0.95	11
Random Forest	1.00	ChatGPT	1.00	1.00	1.00	10
		HT	1.00	1.00	1.00	9
		NMT	1.00	1.00	1.00	11
MLP	1.00	ChatGPT	1.00	1.00	1.00	10
		HT	1.00	1.00	1.00	9
		NMT	1.00	1.00	1.00	11
AdaBoost	0.97	ChatGPT	1.00	1.00	1.00	10
		HT	1.00	0.89	0.94	9
		NMT	0.92	1.00	0.96	11
Naïve Bayes	0.90	ChatGPT	0.77	1.00	0.87	10
		HT	1.00	1.00	1.00	9
		NMT	1.00	0.73	0.84	11

4.2. Interpreting dimensions as a result of co-occurring features

Following the implementation of Kruskal-Wallis H, we found that 74 out of 121 linguistic features exhibited statistical significance The complete list of these features is in the Appendix. As shown in Table 3, their frequency matrix demonstrates a KMO score exceeding the commonly employed threshold of 0.5 and thus ensures sampling adequacy. Additionally, Bartlett's test yields a large chi-square value with a *p* value below 0.001. Therefore, we can safely proceed for the following analyses. However, as shown in Figure 2, ToVDSR and LDE are found to be highly correlated to other features, rendering the whole feature matrix a “singular matrix”. In this case, factor analysis could not be conducted. To solve this issue, we made the decision to exclude these two features from the feature set, resulting in a remaining set of 72 features.

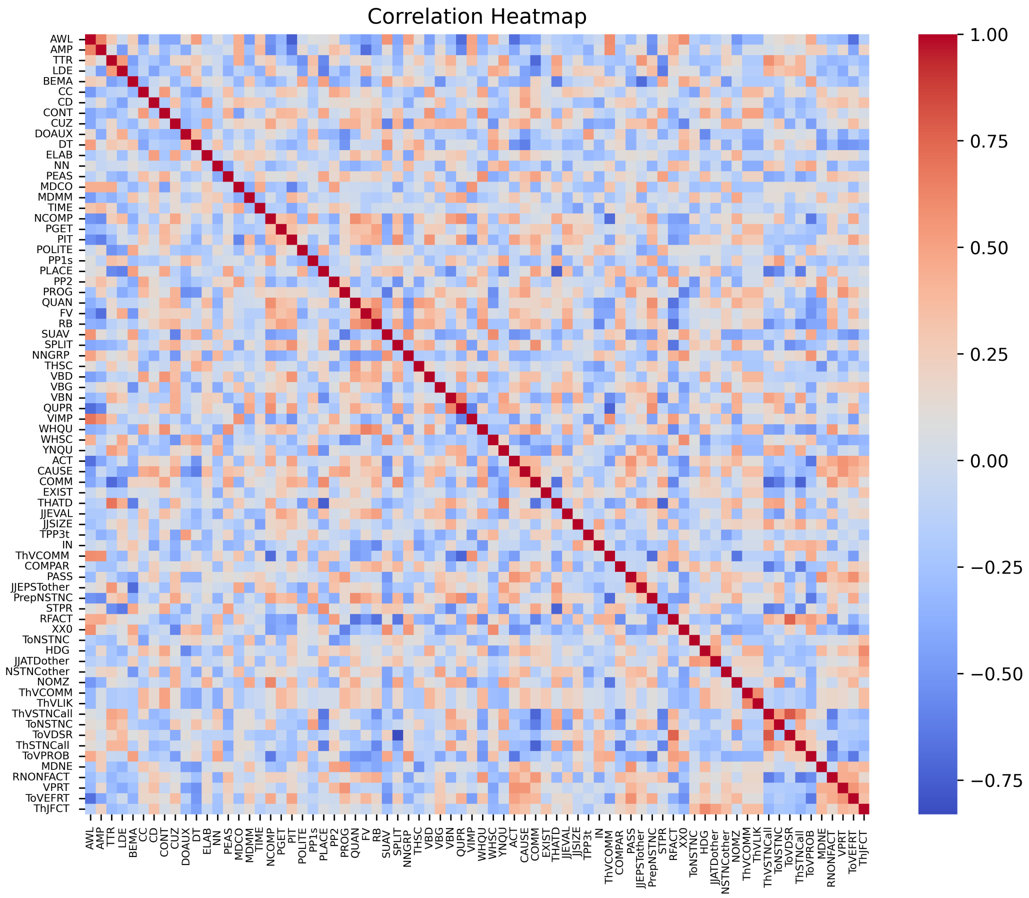

Figure 2 Correlation heatmap among the 74 statistically significant features

Table 3 Results of KMO and Bartlett’s test

KMO measure of sampling adequacy		0.709
Bartlett's test of sphericity	Chi-square	35663.45
	Significance	< 0.001

The scree plot in Figure 3 illustrates the eigenvalues associated with each factor, arranged in descending order. In Table 4, we present the corresponding eigenvalues, proportions of variance, and cumulative variance explained by each factor. Notably, the first five factors collectively account for almost 70 percent of the total variance.

To determine the optimal number of factors, we manually evaluated the linguistic features within each dimension when considering a range of 4 to 10 factors. After careful examination, we concluded that the five-factor solution yields the most meaningful and interpretable outcomes. As a result, we have chosen to set the number of factors to five.

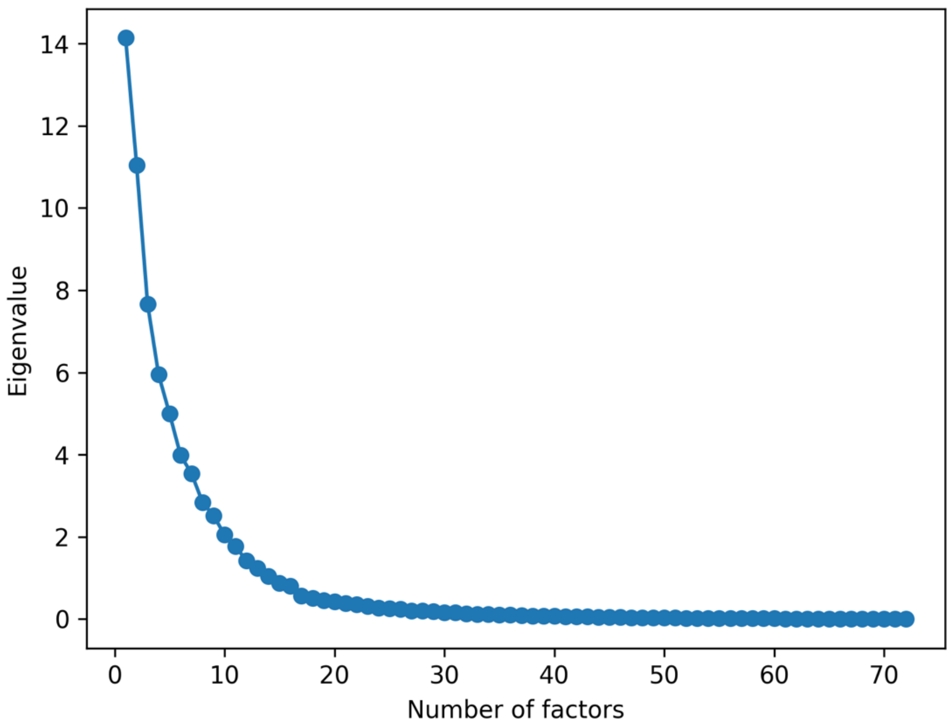

Figure 3 Scree plot: The plot shows the eigenvalue of the corresponding factor number. It helps determine the suitable number of factors to retain, as it identifies the point where the eigenvalues sharply decrease, indicating diminishing returns in terms of explained variance.

Table 4 Eigenvalue, proportion of variance and cumulative variance explained

Factor number	Eigenvalue	% of variance explained	% of cumulative variance explained
0	14.0873	19.57%	19.57%
1	11.0070	15.29%	34.85%
2	7.6040	10.56%	45.41%
3	5.8799	8.17%	53.58%
4	4.9208	6.83%	60.42%
5	3.9331	6.46%	66.88%
6	3.4670	4.82%	70.69%
7	2.7698	3.85%	74.54%
8	2.4456	3.40%	77.94%
9	1.9909	2.77%	80.70%
10	1.6816	2.34%	83.04%

We only analyze features with loadings exceeding 0.30 or falling below -0.30. From Table 5 we can see that, in Dimension 1, we have identified a total of 30 such features. Notably, there exists a relatively larger cluster of features with positive loadings. The high frequency of nouns, noun compounds, nominalizations, and factive verbs indicates a high information density. High average word length and type/token ratio are often associated with rich vocabulary and elaborated language. Additionally, perfect aspect, past tense, passives, politeness markers, *it* pronoun reference, split auxiliaries, and infinitives can be indicators of a formal and polite tone.

For high-loading negative features, first person and second person pronouns are usually used in interactive discourse (Biber 1988) and conversational scenarios. Modals *may* and *might, to* clauses preceded by stance nouns, together with private verbs serve as indicators of personal stance and attitude. Progressive aspect, time references, and place references are often used to denote concrete and specific events. Altogether, Dimension 1 can be interpreted as a dimension of precise information delivery with dense information and a formal tone at one end, and a rather interactive discourse in a relatively informal style at the other end.

Table 5 Features with positive and negative loadings in Dimension 1

Dimension	Positive and negative features	
Dimension 1	Positive	PASS (0.86), VBD (0.79), NN (0.76), NOMZ (0.71), AWL (0.71),

H (2, 150) = 31.43, p < 0.001, η^2= 0.605		NCOMP (0.69), WHSC 0.64), RFACT (0.62), TTR (0.55), IN (0.51), PEAS (0.47), THSC (0.41), DOAUX (0.34), POLITE (0.32), QUPR (0.31), PIT (0.30), SPLIT (0.30)
	Negative	VPRT (-0.79), PP1s (-0.77), ToNSTNC (-0.68), ELAB (-0.61), THAHD (-0.57), MDMM (-0.53), PP2 (0.49), PROG (-0.44), RB (-0.40), PRIV (-0.35), TIME (-0.33), PLACE (-0.31)

The z-transformed scores of text samples in the three translation varieties show that translations generated by ChatGPT contain more negative features in Dimension 1 compared with HT and NMT (Figure 4). This suggests that ChatGPT-generated translations are characterized by an engaged and interactive style with a relatively informal tone, whereas HT and NMT translations showcase a greater degree of formality and sophisticated language use. Their difference can be illustrated by Example 1.

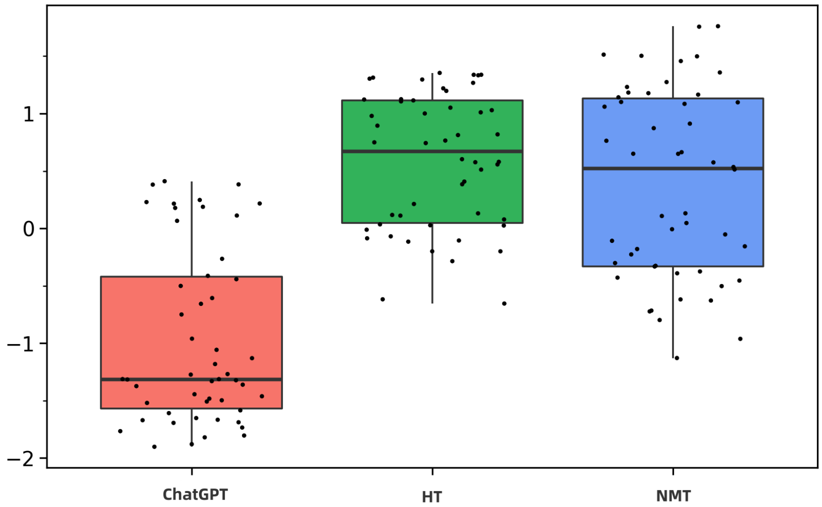

Figure 4 z-transformed scores along Dimension 1

Example 1:

HT: It is reported that Indian Prime Minister Narendra Modi visited the so-called "Arunachal Pradesh" on February 9th. […] It has been verified that there were eight Chinese citizens on board, including one from the Hong Kong SAR.

NMT: It is hoped that all parties in Myanmar will proceed from the fundamental and long-term interests of the country and the nation, resolve emerging problems by peaceful means under the constitutional and legal framework, and continue to advance the process of democratic transition in the country in an orderly manner.

ChatGPT: Really afraid of chaos in the world! We can't help but ask these lawmakers, are you "legislators" or "lawbreakers"? […] You better mind your own business. Hong Kong doesn't need you to worry about it.

Dimension 2 comprises a total of 25 features, consisting of 16 positive features and 9 negative features (see Table 6). Noticeably, a wide range of positive-loading features in this factor are used to express stance (e.g, *that* complement clauses preceded by a stance adjective or verb, *will* and *shall* modals, modal *could*, amplifiers, negation, attitudinal adjectives, stance nouns, and hedges). Features such as direct WH-questions, communicative verbs often serve to engage with the others, delivering one's own or asking for other people's opinions. While amplifiers flag heightened emotions, hedges and concessive conjunctions can mitigate the intensity of attitudes. The most distinctive characteristic of these negative-weighted features is the prevalence of various types of verbs, including activity verbs, finite verbs, facilitation and causative verbs, and non-finite *ed* verb forms. Modals such as *will*, *shall*, and *could* all indicate future action. Overall, this factor can be identified as one that distinguishes stance-oriented expressions from action-focused language.

Table 6 Features with positive and negative loadings in Dimension 2

Dimension 2 H (2, 150) = 27.38, $p < 0.001$, $\eta^2 = 0.447$	Positive	ThSTNCall (0.87), COMM (0.85), THSC (0.81), ThVCOMM (0.74), MENTAL (0.71), THSC (0.70), MDWS (0.69), AMP (0.64), MDCO (0.59), WHQU (0.57), XX0 (0.47), PrepNSTNC (0.43), JJATDother (0.41), NSTNCother (0.39), HDG (0.36), CONC (0.31)
	Negative	ACT (-0.79), FV (-0.74), CAUSE (-0.63), ToVEFRT (-0.55), ToVDSR (-0.53), VBN (-0.46), CUZ (-0.43), RP (-0.38), CC (-0.32)

Based on Figure 5, it is evident that translations generated by ChatGPT exhibit a considerable degree of similarity with NMT translations, while displaying significant differences from HT translations in Dimension 2. Human translators tend to employ fewer stance-related expressions and rely more on verbs, conjunctions, and coordinators. In contrast, translations produced by ChatGPT and NMT contain a higher frequency of linguistic features associated with the direct expression of stance and attitudes (e.g., negation, attitudinal adjectives, stance verbs, *will* and *shall* modals, and modal *could*).

Example 2 is provided to illustrate their distinctions. We can see that the human translator used three consecutive verb phrases to describe the efforts taken by the Algerian President to strengthen China-Algeria relations. In contrast, ChatGPT and NMT resorted to attitudinal adjectives “appropriate,” “necessary,” and “important,” attitudinal adverb “resolutely,” as well as the negation device “no” to convey a strong sense of stance taking.

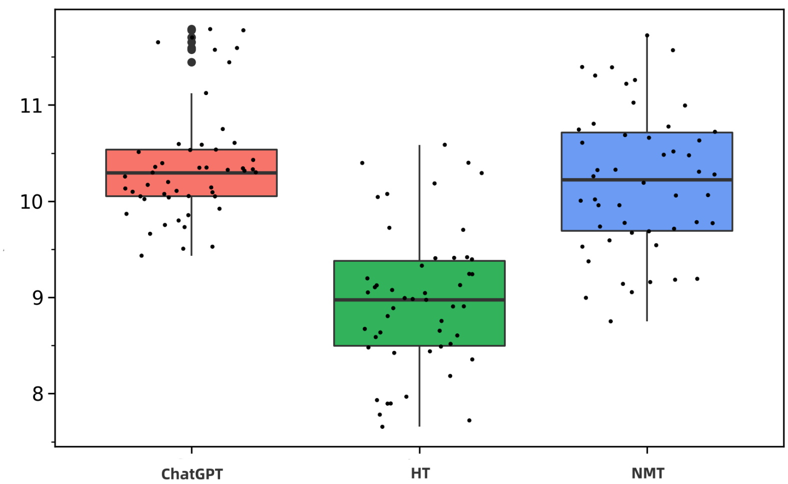

Figure 5 z-transformed scores along Dimension 2

Example 2:

HT: During his term as Algerian President, he actively promoted the development of China-Algeria relations, deepened bilateral friendly cooperation and enhanced the friendship between the two peoples.

NMT: No force can stop the progress of the Chinese people and the Chinese nation. The most important criterion for judging whether the Chinese power situation is good or not is whether the Chinese people are satisfied.

ChatGPT: China will resolutely take appropriate and necessary countermeasures according to the development of the situation.

In Dimension 3, there are 12 features with positive loadings above the threshold of 0.3 (see Table 7), including numbers, causal conjunctions, subordinator *that* omission, factive adverbs, yes/no questions, nouns referring to group, *that* subordinate clauses preceded by factive adjectives, existential or relationship verbs, nouns referring to human, non-finite *ed* verb forms, progressive aspect, proper nouns, and get-passives. These features can be interpreted as devices to describe factual information or actual events that have already happened in an explicit and concrete manner.

The six features with negative loadings include likelihood verbs, to clauses preceded by verbs of probability, non factive adverbs, *that* subordinate clauses preceded by likelihood adjectives, elaborating conjunctions, and determiners. The frequent co-occurrence of these features can be indicators of likelihood, possibility, and inference of future events. Taking both the positive and negative features into consideration, we can explain Factor 3 as distinguishing between factual description and inferential conjecture.

Table 7 Features with positive and negative loadings in Dimension 3

Dimension 3 H (2, 150) = 5.21, $p = 0.07$	Positive	CD (0.73), CUZ (0.71), THATD (0.64), RFACT (0.57), ThJFCT (0.51), EXIST (0.44), NNGROUP (0.41), VBN (0.33), PROG (0.31), NNP (0.31), PGET (0.30)
	Negative	VLIKother (-0.70), ToVPROB (-0.48), RNONFACT (-0.42), ThJLIK (0.40), ELAB (-0.36), DT (-0.34)

From Figure 5, it is evident that both translations generated by ChatGPT and NMT lean towards the positive end of Dimension 3, indicating a slight preference for fact-oriented and information-focused language. In contrast, HT is more inclined to the negative end, suggesting that human translators use more expressions indicative of uncertainty and possibility. However, their differences do not amount to significance.

From Example 3, we can see that translations by ChatGPT and NMT on the positive pole characterize frequent use of numbers, exact dates, and other factual information, while HT on the negative pole tends to convey likelihood (e.g., modal *could*) rather than absolute facts.

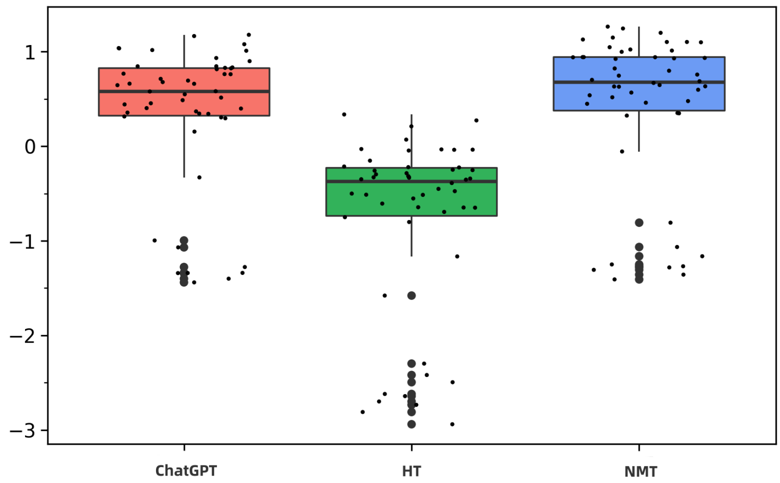

Figure 6 z-transformed scores along Dimension 3

Example 3:

HT: Should it choose to go further down the wrong path, it could expect more countermeasures from China.

NMT: From 2010 to 2018, the Uyghur population in Xinjiang rose from 10,171,500 to 12,718,400, an increase of 2,546,900, an increase of 25.04%, which was not only higher than the 13.99% increase in the entire Xinjiang population, but also significantly higher than the **2%** increase of the Han population.

ChatGPT: According to reports, Tanzania's National Electoral Commission released official results of the presidential election on October 30 showing incumbent President John Pombe Joseph Magufuli winning another term with 84.3 percent of the vote.

Dimension 4 contains a total of 12 features, with 9 of them being positively weighted and 3 being negatively weighted (see Table 8). The positively loaded features include evaluative adjectives, epistemic adjectives without a that clause after, non-finite verb *ing* forms, that subordinate clauses other than relatives, size-related adjectives, comparatives, *be* as main verbs, reference to more than one non-interactant and single *they* reference, and *that* subordinate clauses preceded by communicative verbs. These features are associated with judgement and evaluative meanings, as well as comparison between entities or individuals

The three negatively weighted features consist of stranded prepositions, *that* complement clauses not preceded by a stance adjective or verb, as well as auxiliary. They can be related to rather complicated sentence structure and non-evaluative discourse. We thus categorize this factor as a dimension characterizing a contrast between evaluative discourse and non-evaluative discourse.

Table 8 Features with positive and negative loadings in Dimension 4

Dimension 4 H (2, 150) = 21.67, p < 0.001, η^2= 0.425	Positive	JJEVAL (0.64), JJEPSTother (0.55), VBG (0.51), THSC (0.46), JJSIZE (0.40), COMPAR (0.37), BEMA (0.37), TPP3t (0.31), ThVCOMM (0.30)
	Negative	STPR (-0.58), FV (-0.43), CAUSE (-0.36)

Figure 7 shows that ChatGPT-produced translations are more similar to NMT in Dimension 4. Both the two exhibit a higher frequency of positive features compared to HT translations. Their differences are clearly observable in Example 4, which comprises translations from the same source text. We can see that ChatGPT uses the evaluative adjectives “inappropriate,” “very bad,” and “unkind” and *be* as main verbs in its translation, to the effect that the evaluation is direct and intense. Similarly, NMT uses the epistemic adjective “factual,” and the evaluative adjectives including “appropriate,” “very bad” and “unkind”. On the other hand, the human translator chooses different expressions such as “neither in line with the facts nor out of place,” “go in the opposite way,” and “not a gesture of goodwill” to convey the same meaning. These choices lead to a reduced intensity of evaluation compared to ChatGPT and NMT.

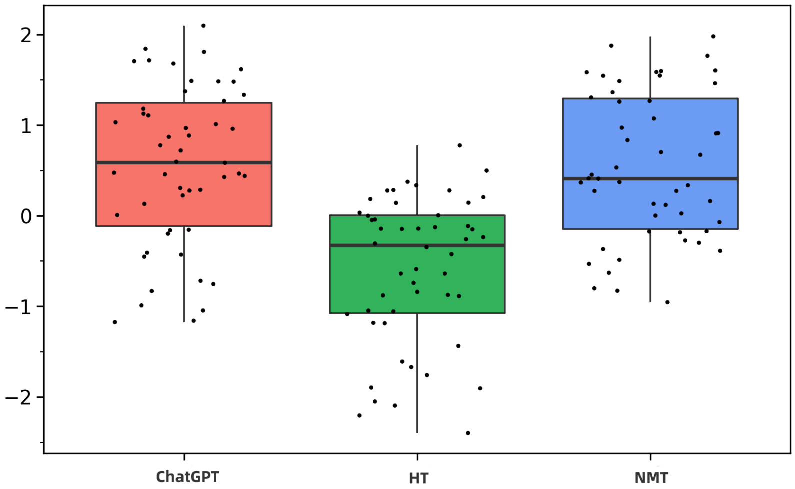

Figure 7 z-transformed scores along Dimension 4

Example 4:

HT: In sharp contrast, certain US officials' words and actions are neither in line with the facts nor out of place. Just as the WHO recommended against travel restrictions, the US rushed to go in the opposite way. Certainly not a gesture of goodwill. (HT)

NMT: In contrast, the words and deeds of the US side are neither factual nor appropriate. The World Health Organization called on countries to avoid travel restrictions, but before the words fell, the United States did the opposite, with a very bad head. It's so unkind.

ChatGPT: In contrast, the words and actions of the US side not only do not conform to the facts, but are also inappropriate. The World Health Organization called on countries to avoid implementing travel restrictions, but before the words had even settled, the United States went against this and set a very bad precedent. That's really unkind.

Dimension 5 is the smallest in size, containing only five features with positive loadings (see Table 8). Among them, the most prominent one is suasive verbs, which is often used for the purpose of persuasion. Similarly, necessity modals convey the speaker’s strong personal conviction about a particular situation, and indicate a sense of obligation. They are often used to provide advice or make recommendations. Verbal contractions is typically seen in spoken discourse. Their co-occurrence alongside group nouns and present tense suggests that Factor 5 mainly revolves around the demonstration of strong will.

Table 9 Features with positive and negative loadings in Dimension 5

Dimension 5 H (2, 150) = 19.82, p < 0.001, η^2= 0.376	Positive	SUAV (0.53), MDNE (0.49), CONT (0.42), NNGRP (0.33), VPRT (0.33)

Figure 9 shows that HT scores the highest in this dimension, while NMT and ChatGPT-produced translations score lower. Also, the score distribution of HT is much more concentrated, suggesting that the dimensional characteristic is generally more noticeable in HT, which can be illustrated by Example (13).

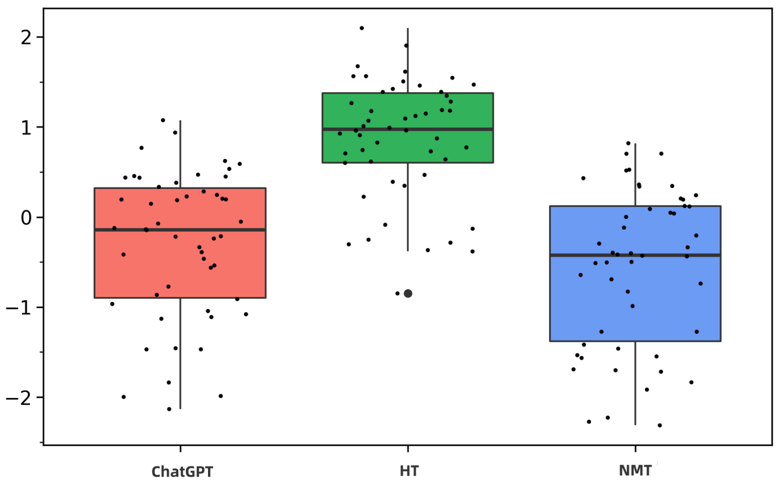

Figure 8 z-transformed scores along Dimension 5

Example 5:

HT: It must be pointed out that Hong Kong's prosperity and stability are in line with the interests of all parties, including the United States. […] The Chinese government and Chinese people are firmly resolved in safeguarding national sovereignty, security and development interests.

In summary, our factor analysis yields five meaningful dimensions. Four of them (Dimension 1, 2, 4, and 5) are able to distinguish ChatGPT-produced translations, NMT and HT with statistical significance. In Dimension 1 we observe a contrast between precise information in a formal style and interactive discourse with an informal tone. Dimension 2 in general differentiates expressions indicative of stance from language related to action taking. Dimension 4 characterizes a contrast between evaluative discourse and non-evaluative discourse, and Dimension 5 centers the demonstration of strong will. Measured by dimension scores, translations from ChatGPT seem to be more casual, explicit in stance-taking, and evaluative. NMT is mostly similar to ChatGPT-generated translations, but tends to be more formal, while HT features more formality and strong will but less evaluation.

4.3. Calculating and visualizing distances

Based on the z-transformed dimension scores, the calculation of mean Euclidean distances between samples from ChatGPT and HT, ChatGPT and NMT, as well as HT and NMT are 4.233, 2.875, and 4.670 respectively. This implies that NMT is closest to ChatGPT-produced translations, and farthest from HT in a general sense. These findings are largely consistent with the results in Section 4.2. The dimension-reduced t-SNE visualization is shown in Figure 9.

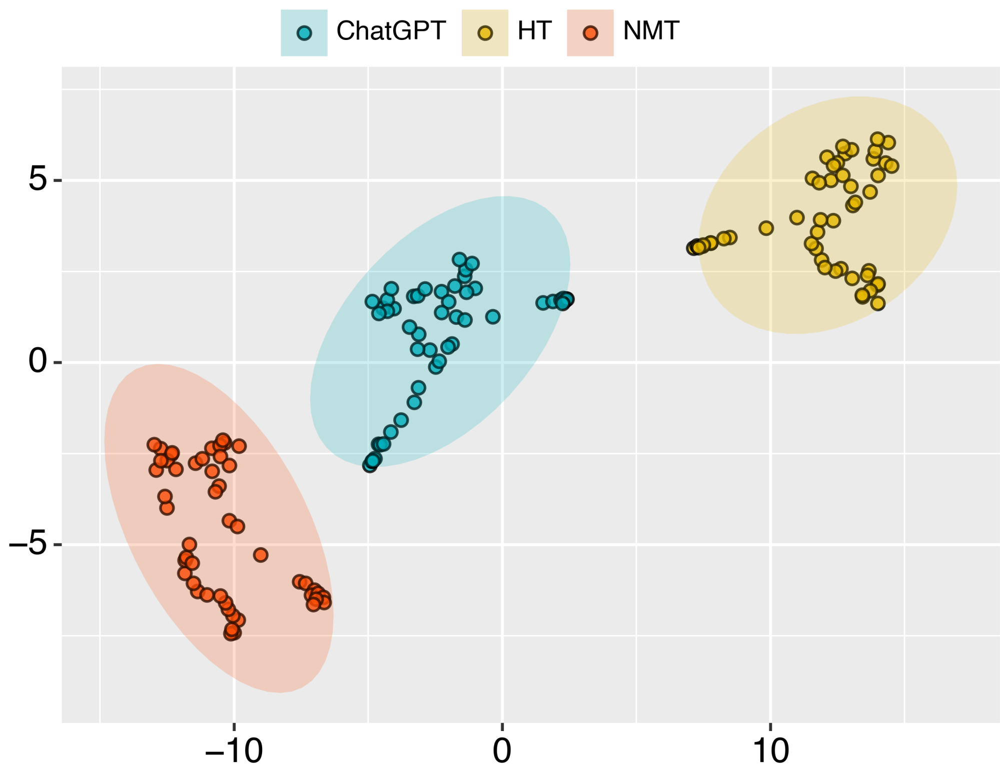

Figure 9 t-SNE visualization of distance distributions

5. Discussion and conclusion

Our first research question is whether translations by ChatGPT, NMT, and HT are distinguishable from each other. Results of five supervised classification algorithms show that distinctions among the three translation varieties are easily identifiable, though there were still cases where NMT and translations from ChatGPT were misclassified. We also found that HT was constantly classified correctly by all the supervised algorithms, suggesting that HT seems to be more easily identifiable compared with the other two translation varieties. Based on these findings, we were confirmed that each translation variety possessed its own characteristics, but coarse-grained classification tasks were insufficient to unveil their exact differences.

This thus led us to the second research question, which delved into the distinctive characteristics of ChatGPT-generated translations, NMT, and HT. We drew on Biber's (1988) MDA to conduct a fine-grained stylistic analysis by identifying and analyzing dimensions consisting of co-occurring features. We found that in four out of five dimensions, the z-transformed dimension scores of ChatGPT-produced translations were very close to those of NMT, as supported by their relative locations in the boxplots (see Figure 6, 7, 8, 9) and the text examples. In particular, we found that evaluative and attitudinal expressions were frequently observed in ChatGPT and NMT translations, but were less prevalent in HT. A possible reason is that human translators, who work for and represent the Chinese government, are proposed to be cautious in their language use, and adhere to the common practices in diplomatic translation. In other words, human translators are more risk-adverse (Pym 2005) than NMT engines and ChatGPT. This is explainable, since both NMT engines and ChatGPT are in essence complicated neural networks, which resemble human brain but still lack the capacity to think as humans do. Therefore, their translations offer a semblance of "common sense" (Lee 2023), but lack the cultural sensitivity, linguistic flexibility, adaptability, and awareness of translation norms exhibited by human translators. As proposed by Pym (2015), an important technique in translators' repertoire is "text tailoring," where translators can change the content of the source text to better serve the purposes of the translated text. However, the current cutting-edge translation technologies, be it ChatGPT or NMT, are still unable to acquire such competency as making judgement and adaptation according to contexts and communicative needs.

A follow-up question addressed in this study is whether translations by ChatGPT are closer to NMT or HT. In line with our expectation, both the calculation of Euclidean distance and t-SNE visualization demonstrated that ChatGPT-generated translations were closer to MT, while HT was distant from both. The longest distance was observed between MT and HT. Similar observation was found in Frankenberg-Garcia (2022), which offered a comprehensive analysis of the lexical differences between HT and NMT. The author found human translators to be superior in idiomaticity, the use of translation strategies, conveying register, and handling communication breakdowns. Karpinska and Iyyer (2023) showed that paragraph-level translations by ChatGPT were more aligned with high-quality human translation, exhibiting reduced mistranslations, grammatical errors, and stylistic inconsistencies as compared to Google Translate.

What, then, is the relationship between the typical features of the three translation varieties and the TU hypotheses? To answer this question, we computed the average normalized frequencies of features commonly associated with specific TU hypotheses (see Table 10). Notably, translations produced by ChatGPT and NMT exhibit higher frequencies in features linked to explicitness (NNP, NN, DEMO, CC, ELAB) and linguistic complexity (AWL, LDE, TTR, NOMZ), while demonstrating lower frequencies in features indicative of simplified language use (THATD, CONT). This suggests that MT does not align with the Simplification hypothesis but adheres to the

Explicitation hypothesis, showcasing a distinct pattern of machine translationese that we term "Sophisticated Explicitation." These findings provide additional empirical support in line with Luo and Li (2022), which also fails to confirm the existence of the simplified language use. Our results are also consistent with Jiang and Niu (2022) and Lapshinova-Koltunski (2015), both of which observe that NMT systems employ more connectives and coordinators that indicate increased explicitness. However, our results contradict with Kuo (2019), which did not confirm the Explicitation hypothesis.

We can explore the causes of machine translationese from two perspectives. From a cognitive standpoint, since the translation process involves effortful code-switching, translators may resort to simplified language use to relieve their cognitive crutch (Jiang and Niu 2022; Kruger 2018). In contrast, machines are not subject to the same cognitive mechanisms as human translators, rendering the purpose of simplification irrelevant in machine translation. Technically, the training of AI models revolves around statistical computation rather than thinking and comprehension. As a result, machine translation may rely on surface-level linguistic devices like coordinators and connectives to reflect textual connections. Human translators, on the other hand, may employ more implicit and semantic means to convey textual coherence, which cannot be captured through the feature computation employed in this study.

Table 10 Mean normalized frequencies of features associated with the Explicitation and Simplification TU hypothesis in the three translation varieties

TU	Features	ChatGPT	NMT	HT
Explicitation	Proper nouns (NNP)	12.78	12.43	12.37
	Prepositions (IN)	14.17	14.18	14.11
	Demonstrative pronouns and articles (DEMO)	0.58	0.66	0.55
	Coordinators (CC)	5.91	6.07	5.82
	Elaborating conjunctions (ELAB)	0.06	0.05	0.03
Simplification	Average word length (AWL)	5.48	5.35	5.30

Lexical density (LDE)	0.59	0.61	0.57
Lexical diversity (TTR)	0.52	0.50	0.52
Subordinator *that* omission (THATD)	0.02	0.02	0.06
Verbal contractions (CONT)	0.01	0.00	0.02
Nominalization (NOMZ)	2.61	2.55	2.45

One implication of our analyses is that, even though NMT and generative intelligence represented by ChatGPT have made huge advances, there is still a marked gap between their translations and HT. There is thus a call for further investigations into the factors that contribute to the uniqueness and distinction of HT, encompassing not only formal linguistic properties but also other aspects. One possible way to improve the translation performance of ChatGPT and NMT is to identify the distinctive characteristics of top-notch translations, and then incorporate these characteristics into the training process of advanced AI-powered language models. Future research can focus on enhancing the adaptability and cultural awareness of automated translation tools, by creating translation technologies that combine the efficiency and speed of NMT and ChatGPT with the cultural sensitivity, linguistic flexibility, and domain expertise exhibited by professional human translators. This is crucial for the future development of NMT and artificial general intelligence (AGI) as a whole. Human translators, on the other hand, can make informed decisions when integrating these advanced translation technologies into their workflow.

Lastly, we should acknowledge that this study is restricted in register and scope. Our investigation was conducted only in the field of Chinese-to-English diplomatic translation, and the relatively limited corpus size necessarily makes our analysis selective. Nevertheless, the findings may provide valuable insights into the characteristics of and relationship among ChatGPT-generated translations, HT, and MT. Moreover, the study also offers a set of methods and tools for future exploration with similar focuses.

Notes

1. https://www.fmprc.gov.cn/fyrbt_673021/
2. https://github.com/elenlefoll/MultiFeatureTaggerEnglish/blob/main/Introducing_the_MFTE_v3.0.pdf.

Funding

This work was supported by the National Institutes of Health [AA123456 to C.S., BB765432 to M.H.]

References

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., and Quirk, R. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins Publishing Company.

Blum-Kulka, S. (1986). 'Shifts of ccohesion and coherence in translation'. In House, J. and Blum-Kulka, S. (eds.) *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies,* pp. 17–35. Tübingen: Narr.

Calzada Pérez, M., and Sánchez Ramos, M. d. M (2021) 'MDA analysis of translated and non-translated parliamentary discourse'. In Ji, M. and Oakes, M. P. (eds.) *Corpus Exploration of Lexis and Discourse in Translation,* pp. 26-55. London: Routledge. https://doi.org/10.4324/9781003102694-2

Chen, P., Guo, Z., Haddow, B. and Heafield, K. (2023) 'Iterative translation refinement with large language models', *arXiv*: 2306.03856[Cs]. https://arxiv.org/pdf/2306.03856.pdf

De Sutter, G., and Lefer, M. A. (2020) 'On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach', *Perspectives*, 28(1): 1–23. https://doi.org/10.1080/0907676X.2019.1611891

Duff, Alan. 1981. *The third language: Recurrent problems of translation into English.* Oxford: Pergamon.

Eder, M. (2016) 'Rolling stylometry', *Digital Scholarship in the Humanities,* 31(3): 457-469. https://doi.org/10.1093/llc/fqv010

Fang, Y., and Liu, H. (2023) 'Seeing various adventures through a mirror: Detecting translator's stylistic visibility in Chinese translations of Alice's Adventure in Wonderland', *Digital Scholarship in the Humanities*, 38(1): 50-65. https://doi.org/10.1093/llc/fqac024

Frankenberg-Garcia, A. (2022) 'Can a corpus-driven lexical analysis of human and machine translation unveil discourse features that set them apart?', *Target*, 34(2): 278-308. https://doi.org/10.1075/target.20065.fra

Frawley, W. (2000) 'Prolegomenon to a theory of translation', In L. Venuti (ed.) *The translation studies reader,* pp. 250-263. London and New York: Routledge.

Gaspari F, Almaghout H, and Doherty S. (2015) 'A survey of machine translation competences: Insights for translation technology educators and practitioners', *Perspectives*, 23(3): 333–358. https://doi.org/10.1080/0907676X.2014.979842

He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S. and Wang, X. (2023) 'Exploring Human-Like Translation Strategy with Large Language Models', *arXiv*: 2305.04118[Cs]. https://arxiv.org/pdf/2305.04118.pdf

Hendy, A., Abdelrehim, M.G., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y., Afify, M. and Awadalla, H.H. (2023) 'How good are GPT models at machine translation? A comprehensive evaluation', *arXiv*: 2302.09210[Cs]. https://arxiv.org/pdf/2302.09210.pdf

Hu, K., Li, X. (2023) 'The creativity and limitations of AI neural machine translation A corpus-based study of DeepL's English-to-Chinese translation of Shakespeare's plays', *Babel*, 69 (4): 546-563. https://doi.org/10.1075/babel.00331.hu

Hu, X., Xiao, R., and Hardie, A. (2019) 'How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis', *Corpus Linguistics and Linguistic Theory*, 15(2): 347-382. https://doi.org/10.1515/cllt-2014-0047

Karpinska, Marzena, and Mohit Iyyer. (2023) 'Large language models effectively leverage document-level context for literary translation, but critical errors persist', In *Proceedings* of *the Eighth Conference on Machine Translation (WMT23)*, Singapore, 6-7 December 2019, pp.419–451. https://doi.org/10.18653/v1/2023.wmt-1.41

Kruger, H., and Rooy, B. (2018) 'Register variation in written contact varieties of English', *English World-Wide*, 39 (2): 214-242. https://doi.org/10.1075/eww.00011.kru

Kruger, H. (2018) 'That again: A multivariate analysis of the factors conditioning syntactic explicitness in translated English', *Across Languages and Cultures*, 20(1): 1–33. https://doi.org/10.1556/084.001

Kruger, H., and van Rooy, B. (2016) 'Constrained language', *English World-Wide*, 37(1), 26-57. https://doi.org/10.1075/eww.37.1.02kru

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

Kuo, C. (2018) ‘Function words in statistical machine-translated Chinese and original Chinese: A study into the translationese of machine translation systems’, *Digital Scholarship in the Humanity,* 34(4): 752-771. https://doi.org/10.1093/llc/fqy050

Lapshinova-Koltunski, E. (2015) ‘Variation in translation: Evidence from corpora’, In Fantinuoli, C. and Zanettin, F. (eds.) *New Directions in Corpus-Based Translation Studies,* pp. 81–99. Language Science Press.

Laviosa, S. (1998) ‘Core patterns of lexical use in a comparable corpus of English narrative prose’, *Meta*, 43(4): 557-570.

Le Foll, E. and Shakir, M. (2023). MFTE Python (Version 1.0) [Computer software]. https://github.com/mshakirDr/MFTE

Lee, T. K. (2023) ‘Artificial intelligence and posthumanist translation: ChatGPT versus the translator’, *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2023-0122

Lu, Q., Qiu, B., Ding, L., Xie, L., and Tao, D. (2023) ‘Error analysis prompting enables human-like translation evaluation in Large Language Models: A case study on ChatGPT’, *arXiv*: 2303.13809[Cs]. https://arxiv.org/pdf/2303.13809.pdf

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

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

Mohamed, E., Sarwar, R., and Mostafa, S. (2023) ‘Translator attribution for Arabic using machine learning’, *Digital Scholarship in the Humanities,* 38(2): 658-666. https://doi.org/10.1093/llc/fqac054

Newmark, Peter. (1991). *About Translation*. Clevedon: Multilingual Matters.

Nini, A. (2019) ‘The Muli-Dimensional Analysis Tagger’, In Berber Sardinha, T. and Veirano Pinto, M. (eds) *Multi-Dimensional Analysis: Research Methods and Current Issues,* pp. 67-94. London; New York: Bloomsbury Academic.

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y. and Tao, D. (2023) ‘Towards Making the Most of ChatGPT for Machine Translation’, *arXiv*: 2303.13780[Cs]. https://arxiv.org/pdf/2303.13780.pdf

Raunak, V., Menezes, A., Post, M. and Hassan. H. (2023) ‘Do GPTs Produce Less Literal Translations?’, In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* Toronto, Canada, 9-14 July 2023*,* pp. 1041–1050. https://doi.org/10.18653/v1/2023.acl-short.90

Rybicki, J., and Heydel, M. (2013) ‘The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish’, *Literary and Linguistic Computing,* 28(4): 708-717. https://doi.org/10.1093/llc/fqt027

Sheng, A., and Kong, Y. (2023) ‘Neural machine translation and human translation’, *Babel,* 69(4): 483-498. https://doi.org/10.1075/babel.00332.she

Thompson, P., Hunston, S., Murakami, A., and Vajn, D. (2017) ‘Multi-Dimensional Analysis, text constellations, and interdisciplinary discourse’, *International Journal of Corpus Linguistics*, 22 (2): 153-186. https://doi.org/10.1075/ijcl.22.2.01tho

Wang, Q., and Li, D. (2011) ‘Looking for translator's fingerprints: a corpus-based study on Chinese translations of Ulysses’, *Literary and Linguistic Computing*, 27(1): 81-93. https://doi.org/10.1093/llc/fqr039

Pym, A. (2015) ‘Translating as risk management’, *Journal of Pragmatics*, 85, 67–80

Pym, A. (2005) ‘Text and risk in translation’, In Sidiropoulou, M. and Papaconstantinou, A. (eds.) *Choice and difference in translation. The specifics of transfer*, pp. 27–42. Athens: University of Athens.

Appendix A

121 features extracted

Category	Feature (tag)
General text properties	total number of words (Words), average word length (AWL), lexical diversity (TTR), lexical density (LDE), finite verbs (FV)
Adjectives	Attributive adjectives (JJAT), predictive adjectives (JJPR),
Adverbials	frequency references (FREQ), place references (PLACE), time references (TIME), other adverbs (RB)
Determinatives	s-genitives (POS), determiners (DT), quantifiers (QUAN), numbers (CD), demonstrative pronouns and articles (DEMO)
Discourse organizations	elaborating conjunctions (ELAB), coordinators (CC), causal conjunctions (CUZ), concessive conjunctions (CONC), conditional conjunctions (COND), discourse/pragmatic markers (DMA), filled pauses and interjections (FPUH), direct WH-questions (WHQU), question tags (QUTAG), yes/no question (YNQU), *that* relative clauses (TRHC), *that* subordinate clauses (other than relatives) (THSC), subordinator *that* omission (THATD), WH subordinate clauses (WHSC)
Lexis	Total nouns (including proper names) (NN), noun compounds (NCOMP), hashtags (HST), superlatives (SUPER), comparatives (COMPAR), nominalization (NOMZ)
Negation	Negation (XX0)
Prepositions	Prepositions (IN)
Pronouns	Reference to the speaker/ writer (PP1S), Reference to the speaker/ writer and others (PP1P), reference to addressee(s) (PP2), *it* pronoun reference (PIT), any personal pronoun not included in the other categories (PPOther), single, male third person reference (PP3m), single, female third person reference (PP3f), reference to more than one non-interactant and single *they* reference (TPP3t), quantifying pronouns (QUPR)
Stance-taking devices	Politeness markers (POLITE), amplifiers (AMP), downtoners (DWNT), emphatics (EMPH), hedges (HDG)
Stative forms	existential *there* (EX), *be* as main verb (BEMA)
Verb features	Verbal contractions (CONT), particles (RP), be-passives (PASS), get-passives (PGET, g*oing to* constructions (GTO), past tense (VBD), non-finite verb-*ing* forms (VBG), non-finite *ed* verb forms (VBN), imperatives (VIMP), present tense (VPRT), perfect aspect (PEAS), progressive aspect (PROG), *have got* constructions (HGOT)
Verb semantics	*do* auxiliary (DOAUX), necessity modals (MDNE), modal *can* (MDCA), modal *could* (MDCO), modals *may* and might (MDMM), *will* and *shall* modals (MDWS), modal *would* (MDWO), *be able to* (ABLE), activity verbs (ACT), aspectual verbs (ASPECT), suasive verbs (SUAV), facilitation and causative verbs (CAUSE), communication verbs (COMM), existential or relationship verbs (EXIST), mental verbs (MENTAL), private verbs (PRIV), public verbs (PUBV), Seem/appear (SMP), occurrence verbs (OCCUR),

	communicative verbs in other contexts (VCOMMother), factive verbs in other contexts (VFCTother), likelihood verbs in other contexts (VLIKother)
Adjectives semantics	attitudinal adjectives without a clause after (JJATDother), adjectives related to color (JJCOLR), epistemic adjectives without a *that* clause after (JJEPSTother), evaluative adjectives (JJEVAL), relational adjectives (JJREL), relational adjectives (JJREL), size related adjectives (JJSIZE), time related adjectives (JJTIME), topical adjectives (JJTOPIC)
Adverb semantics	attitudinal adverbs (RATT), factive adverbs (RFACT), adverbs of likelihood (RLIKELY), non factive adverbs (RNONFACT)
Noun semantics	Nouns abstracted and process (NNABSPROC), nouns cognitive (NNCOG), nouns concrete (NNCONC), nouns group (NNGRP), nouns human (NNHUMAN), nouns place (NNPLACE), nouns quantity (NNQUANT), nouns technical (NNTECH), nominalization (NOMZ), proper nouns (NNP), stance nouns without prepositions (NSTNCother)
Syntax	*that* subordinate clauses (other than relatives) preceded by attitudinal adjectives (ThJATT), *that* subordinate clauses (other than relatives) preceded by adjectives of evaluation (ThJEVL), *that* subordinate clauses (other than relatives) preceded by likelihood adjectives (ThJLIK), *that* subordinate clauses (other than relatives) preceded by adjectives of evaluation (ThJEVL), that subordinate clauses (other than relatives) preceded by factive nouns (ThNATT), *that* subordinate clauses (other than relatives) preceded by factive nouns (ThNFCT), *that* subordinate clauses (other than relatives) preceded by attitudinal verbs (ThVATT), *that* subordinate clauses (other than relatives) preceded by communicative verbs (ThVCOM), *that* subordinate clauses (other than relatives) preceded by factive verbs (ThVFCT), *that* subordinate clauses (other than relatives) preceded by likelihood verbs (ThVLIK), mental/attitudinal verbs in other contexts (VATTother), *to* clauses preceded by ability adjectives (ToJABL), *to* clauses preceded by certainty adjectives (ToJCRTN), *to* clauses preceded by adjectives of ease (ToJEASE), *to* clauses preceded by factive adjectives (ToJFCT), *to* clauses preceded by evaluative adjectives (ToJEVAL), *to* clauses preceded by verbs of desire (ToVDSR), *to* clauses preceded by verbs of effort (ToVEFRT), *to* clauses preceded by mental verbs (ToVMNTL), *to* clauses preceded by verbs of probability (ToVPROB), *to* clauses preceded by verbs of speech (ToVSPCH), *WH* subordinate clauses preceded by attitudinal verbs (WhVATT), *WH* subordinate clauses preceded by communicative verbs (WhVCOM), *WH* subordinate clauses preceded by factive verbs (WhVFCT), *WH* subordinate clauses preceded by likelihood verbs (WhVLIK), *To* clauses preceded by stance nouns (ToNSTNC), prepositions preceded by stance nouns (PrepNSTNC), split auxiliaries and infinitives (SPLIT), stranded propositions (STPR)

Appendix B

Significant features tested under Kruskal-Wallis H

Features	***p* value**
AWL	<0.001
AMP	<0.001
TTR	<0.001
LDE	<0.001
BEMA	<0.001
CC	<0.001
CD	<0.001
CONT	<0.001
CUZ	<0.001
DOAUX	<0.001
DT	0.012
ELAB	<0.001
NN	<0.001
PEAS	<0.001
MDCO	<0.001
MDMM	<0.001
TIME	<0.001
NCOMP	<0.001
PGET	<0.001
PIT	<0.001
POLITE	<0.001
PP1s	<0.001
PLACE	<0.001
PP2	<0.001
PROG	0.009
QUAN	<0.001
FV	<0.001
RB	<0.001
SUAV	<0.001
SPLIT	<0.001
NNGRP	<0.001
THSC	<0.001
VBD	<0.001
VBG	<0.001
VBN	<0.001
QUPR	0.016
VIMP	<0.001
WHQU	<0.001
WHSC	<0.001
YNQU	<0.001
ACT	<0.001
CAUSE	<0.001
COMM	<0.001
EXIST	<0.001
THATD	0.005
JJEVAL	<0.001
JJSIZE	<0.001
TPP3t	<0.001
IN	<0.001
ThVCOMM	<0.001
COMPAR	<0.001
PASS	<0.001
JJEPSTother	<0.001
PrepNSTNC	<0.001
STPR	<0.001
RFACT	0.013
XX0	<0.001
ToNSTNC	<0.001
HDG	<0.001
JJATDother	<0.001
NSTNCother	<0.001
NOMZ	<0.001
ThVCOMM	<0.001
ThVLIK	<0.001
ThVSTNCall	<0.001
ToNSTNC	<0.001
ToVDSR	<0.001
ThSTNCall	0.011
ToVPROB	<0.001
MDNE	<0.001
RNONFACT	<0.001
VPRT	<0.001
ToVEFRT	<0.001
ThJFCT	<0.001