

Stylometric Consistency Patterns in Translation of Spoken Chinese to Written English: A Corpus Analysis of the CoVoST Dataset

Vincent Chieh-Ying Chang <159738_at_o365_dot_tku_dot_edu_dot_tw>, Tamkang University  <https://orcid.org/0000-0002-5013-4758>

DOI: <https://doi.org/10.63744/cb67ugfqyve2>

Abstract

This study examines stylometric consistency patterns in professional human translations of spoken Chinese to written English, using the CoVoST corpus. Moving beyond traditional translation quality assessment, the investigation studies how human translators navigate the speech-to-text modality shift and whether this process produces domain-invariant stylistic regularities. Through a framework combining stylometric analysis, machine learning, and digital humanities critique, I identify a “consistency bias” — manifested as standardized lexical diversity, flattened syntactic structures, and repetitive discourse patterns — that reveals how professional constraints and cognitive processing shape translation output. My analysis of 16,899 human-translated English sentences, with detailed statistical comparison of 300 sentences against contemporary spoken English baselines, demonstrates that speech-to-text translation exhibits significantly reduced lexical diversity (Cohen's $d=0.34$, $p<0.001$), shallower syntactic structures ($d=0.33$, $p<0.001$), and narrower modal verb usage ($d=0.43$, $p<0.001$) compared to original spoken English. These findings illuminate the cognitive and professional constraints shaping human translation practice; future research might consider how such patterns subsequently inform machine translation systems trained on human-translated corpora.

1. Introduction

Professional human translation of spoken language into written text represents a complex cognitive and linguistic process that remains understudied in translation studies. While extensive research has examined text-to-text translation [Baker 1993] [Laviosa 1998], the distinct challenges of translating oral discourse — with its disfluencies, incomplete structures, and context-dependent meanings — into polished written prose have received limited scholarly attention. This study addresses this gap by examining stylometric patterns in professional human translations of spoken Chinese to written English, using the CoVoST corpus [Wang et al. 2020] as the primary data source. Rather than evaluating these translations against quality metrics — a common approach that reinforces normative assumptions — I instead interrogate the internal stylistic consistency of the translated output itself, seeking to understand whether human translators apply standardized stylometric templates across diverse content domains.

The concept of translationese — the distinctive linguistic features that mark translated text as different from originally authored content — has a rich history in Translation Studies [Gellertstam 1986] [Baker 1993]. Digital humanities scholars have since demonstrated how computational analysis of textual patterns reveals deeper questions about authorship, style, and cultural representation. [Piper 2018] [Underwood 2019]. My study extends this framework to speech-to-text translation, examining whether the cognitive and professional constraints of rendering oral discourse into written form produce distinctive stylometric patterns that I term *consistency bias*. I approach human translation not simply as a technical artifact but as a cultural text that embodies specific human values and aesthetic choices during translation processes.

1

2

My study addresses a critical gap at the intersection of translation studies and digital humanities by applying an integrated framework of stylometric analysis, machine learning techniques, and humanistic critique to identify and characterize consistency patterns in the CoVoST English corpus. I focus on four key stylometric dimensions: lexical diversity (measured through Mean Type-Token Ratio), syntactic complexity (parse tree depth), modal verb usage, and discourse marker distribution. These features are selected based on their established relevance in translation studies literature [Laviosa 1998] [Baker 1993]. By employing unsupervised clustering and supervised classification enhanced with SHAP (SHapley Additive exPlanations) values, I identify latent stylistic patterns and quantify the importance of different linguistic features in characterizing the corpus. I additionally incorporate Quantitative Content Analysis (QCA) to examine semantic and discourse-level patterns, thereby capturing how consistency manifests beyond formal linguistic features. This methodological design draws on Ramsay's (2011) "algorithmic criticism," wherein computational methods serve not merely as measurement instruments but as catalysts for humanistic interpretation.

3

The significance of this research extends beyond technical assessment. Johanna Drucker (2012) reminds us that digital humanists bear a responsibility to interrogate the epistemological assumptions embedded in computational processes.

4

2. Literature Review

2.1 Translationese in Human Speech-to-Text Translation

The concept of translationese — distinctive linguistic features that mark translated text as different from originally authored content — has been a central focus in Translation Studies for decades. Martin Gellerstam (1986) first coined the term to describe the "fingerprints" that source languages leave on target texts, identifying lexical and syntactic patterns that distinguish Swedish translations from original Swedish texts. This pioneering work established that translations constitute a distinct textual category with identifiable linguistic characteristics. Mona Baker (1993) expanded this concept by proposing translation universals — features that appear in translated texts regardless of language pairs involved. These include simplification (reduced lexical diversity), explicitation (adding clarifying information), normalization (conforming to target language conventions), and leveling out (convergence toward the middle of the continuum of any linguistic feature). Baker's framework has been instrumental in shifting translation research toward corpus-based methodologies that can empirically identify these patterns.

5

Subsequent advances in both statistical and neural machine translation have added new dimensions to translationese research. Antonio Toral and Andrew Way (2018) showed that machine translations exhibit their own distinctive patterns of translationese, often amplifying certain features found in human translations. Their analysis of WMT (Workshop on Machine Translation) evaluation campaigns revealed that machine translation processes tend to produce outputs with lower lexical variety and higher source language interference than human translations. Recent work by [Riley et al. 2020] reframes translationese as a distinct "language" within multilingual NMT models. Their research demonstrates that biasing models to produce translationese outputs inflates BLEU scores while decreasing human-rated quality — a finding that underlines the divergence between automated metrics and genuine translation quality. This reconceptualization of translationese as a computational property rather than merely a linguistic artifact represents an important theoretical advancement.

6

2.2 Digital Humanities Approaches to Style, Authorship, and Text

While translation studies have developed robust frameworks for analyzing translationese, digital humanities scholars have approached questions of style, authorship, and translated text from complementary but distinct theoretical perspectives. Franco Moretti's (2013) concept of "distant reading" provides a foundational framework for understanding how computational analysis of large textual corpora can reveal patterns invisible to close reading. His framework has proven especially generative for stylometric studies seeking to isolate the distinctive markers of authorial voice across expansive textual archives. [Piper 2018] extends this work by examining how computational methods can reveal the "topological imagination" of texts — the ways in which textual patterns encode specific cultural and cognitive frameworks. Matthew Jockers (2013), in a complementary vein, demonstrates how macroanalysis can isolate the "signal" of authorial style from the "noise" of generic conventions. [Underwood 2019] further develops these approaches

7

by examining how machine learning models can identify patterns of literary change over time, thereby defamiliarizing received literary-historical periodizations and revealing gradual stylistic evolutions imperceptible to unaided reading.

The question of algorithmic text production has been addressed most directly by [Ramsay 2011], whose concept of “algorithmic criticism” proposes that computational methods should be used not simply to measure textual phenomena but to generate new interpretive possibilities. For Ramsay, the most productive digital humanities methods are those that “deform” texts in ways that provoke fresh critical insight — a stance consonant with my treatment of speech-to-text translations as cultural artifacts meriting humanistic scrutiny. [Drucker 2012] provides a crucial theoretical framework for understanding the epistemological implications of computational text analysis, emphasizing the need to recognize how computational visualizations and metrics encode specific cultural assumptions. This perspective bears directly on my analysis, since human translators, too, encode assumptions about linguistic equivalence and textual fidelity in their professional practice. Wendy Hui Kyong Chun (2011) further develops this critical perspective by examining how computational systems participate in what she terms the “programmability of vision” — the ways in which algorithms “shape,” a concept that, transposed to translation, illuminates how professional practices may delimit not only what is translated but what remains translatable across linguistic and cultural divides.

8

2.3 Stylometric Profiling in Translation

Stylometric analysis — the quantitative study of linguistic style — has emerged as a powerful methodology for examining translationese. Sara Laviosa (1998) pioneered corpus-based stylometric approaches in Translation Studies, establishing metrics for measuring simplification through type-token ratios, lexical density, and core vocabulary coverage. Mario Bisiada (2017) applied stylometric techniques to examine editorial intervention in translation, demonstrating how syntactic structures are modified during the translation process — a methodology I adapt for my examination of speech translation outputs.

9

In computational linguistics, stylometric features have proven effective for translation detection and authorship attribution. [Kurokawa, Goutte, and Isabell 2009] demonstrated that machine learning classifiers could distinguish translated from original text with high accuracy using relatively simple stylometric features, and [Lembersky, Ordan, and Wintner 2012] extended this work by showing how language models could be adapted to account for translationese effects. In a more recent application, [Graham, Haddow, and Koehn 2019] used stylometric features to analyze the effect of analysis sample composition on MT evaluation, arguing that target-original test data should be excluded from evaluation campaigns due to the distinctive properties of translationese source text.

10

2.4 Speech-to-Text Translation

Speech-to-text translation introduces additional complexities beyond text-to-text translation. As Alex Waibel and Christian Fugen (2008) note, translators working from speech input must handle disfluencies, incomplete sentences, and other characteristics of spontaneous speech that rarely appear in written text. These challenges can amplify certain aspects of translationese or introduce new patterns specific to speech-based translation.

11

The development of speech-to-text translation has accelerated with datasets like CoVoST [Wang et al. 2020] and MuST-C [Di Gangi et al. 2019]. These resources enable direct training of speech-to-text translation systems without intermediate transcription steps. [Sperber and Paulik 2020] caution, however, that end-to-end systems may develop distinctive output patterns diverging from both text-based MT and conventional human translation.

12

2.5 Research Gap and Contribution

While extensive research exists on translationese in both human and machine translation, and digital humanities scholars have developed sophisticated frameworks for analyzing translated text, a significant gap remains at the intersection of these fields. Most existing studies compare machine translations against human references or examine differences between original and translated text and have primarily focused on text-to-text translation with limited attention to speech input. My work addresses this gap by identifying internal consistency patterns within human translations of speech input — specifically through the CoVoST corpus — without relying on human reference

13

translations, thereby revealing how speech-specific factors influence translationese patterns and contribute to a distinctive translation voice. Methodologically, I integrate stylometric analysis with machine learning and humanistic critique, applying unsupervised clustering and SHAP-enhanced classification to identify latent stylistic patterns while engaging with critical theories of translation consistency and textuality. In bridging translation studies and digital humanities, this approach demonstrates that computational methods can simultaneously expose technical regularities and generate fresh avenues for cultural interpretation. Finally, my research contributes to ongoing digital humanities discussions about the politics of text production by examining how human translators encode specific assumptions about linguistic equivalence, cultural representation, and the nature of voice itself, offering both empirical and theoretical insights into how professional translation practices shape cross-cultural communication in the digital age.

3. Data and Corpus Description

3.1 The CoVoST Corpus as Digital Cultural Artifact

The Common Voice Speech Translation (CoVoST) corpus represents not merely a dataset but a digital cultural artifact that embodies specific speech-to-text translation values and assumptions about cross-linguistic communication. Developed by [Wang et al. 2020], CoVoST derives from Mozilla's Common Voice speech recognition dataset and includes translations across multiple language pairs. For this study, I focus specifically on the Chinese (zh-CN) to English translation component, comprising 16,899 English sentences translated from Chinese spoken input.

CoVoST's architecture is particularly valuable because it preserves authentic professional human translations of spoken input. As [Wang et al. 2020] document, the English translations were produced by professional translators working from Chinese audio recordings, not by automated systems. This human translation of speech input provides rare access to how professional translators handle the modality shift from oral discourse to written text — a process that involves not merely linguistic transfer but also discourse reformulation, register adjustment, and the resolution of speech-specific features (disfluencies, false starts, incomplete utterances) into coherent written prose. Each entry includes the original audio file path, the professional English translation, and a data split designation (train/test/validation). For my analysis, I focus exclusively on the test set (n=4,232 sentences) to ensure my sample represents held-out data not used in any potential downstream model training.

Heeding Drucker's (2012) insistence that data be understood as *capta* — taken rather than given — I approach the CoVoST corpus not as a neutral resource but as a constructed artifact that encodes specific assumptions about linguistic equivalence and cross-cultural communication. The corpus covers diverse topics, and this topical diversity allows us to examine whether any patterns manifest differently across domains or remain constant regardless of subject matter.

My analysis revealed a distribution across domains with general content comprising the majority (79.0%), followed by historical (10.7%), cultural (5.7%), political (4.0%), and smaller proportions of technical and medical content (0.3% each). This distribution constitutes a form of cultural curation, one that shapes which varieties of cross-linguistic communication receive analytical priority — a concern Liu (2012) has raised regarding the latent biases of digital tools.

3.2 Data Processing as Interpretive Act

My data processing pipeline involved several stages. In keeping with Ramsay's (2011) notion of algorithmic criticism, I treat this processing not as a neutral technical procedure but as an interpretive act that shapes what patterns become visible for analysis.

First, I extracted the English translations from the corpus, preserving the original sentence boundaries. I then performed sentence tokenization using NLTK's *punkt* tokenizer, which is specifically trained to handle the punctuation patterns of English text.

For syntactic analysis, I processed each sentence using the spaCy natural language processing library (version 3.8.0) with the *en_core_web_sm* model. This provided part-of-speech tagging, dependency parsing, and syntactic tree

structures necessary for calculating parse tree depth and identifying syntactic patterns. The spaCy pipeline was selected for its balance of accuracy and processing efficiency, though I acknowledge, with Underwood (2019), that any computational tool embeds linguistic assumptions whose cross-cultural applicability cannot be taken for granted.

Word tokenization was performed using NLTK's *word_tokenize* function, which accounts for English-specific tokenization challenges such as contractions and punctuation. These tokens served as the basis for calculating lexical diversity metrics and identifying modal verbs and discourse markers.

To ensure the reliability of my analysis, I implemented extensive data validation checks, including handling of empty sentences, removal of non-linguistic symbols, and normalization of whitespace. The processed corpus was then converted to a structured format suitable for statistical analysis and machine learning applications. Throughout, I remained mindful of what N. Katherine Hayles (2012) terms the materiality of digital text — the manner in which computational processing inevitably inflects textual meaning.

3.3 Sampling Strategy and Epistemological Considerations

3.3.1 Validation Sample Analysis

While the complete CoVoST corpus contains 16,899 English translations, my in-depth stylometric analysis was conducted on a representative sample of 300 sentences drawn exclusively from the test split (n=4,232 test sentences, 25.0% of corpus). This sampling strategy ensures my analysis is independent of any potential training data and avoids circularity issues. I employed stratified random sampling with a fixed random seed (seed=42) to ensure reproducibility. Sentences were sampled proportionally across domain categories (general, historical, cultural, political, technical, medical) to maintain the domain distribution of the full test set. The resulting sample comprises:

- General: 237 sentences (79.0%)
- Historical: 32 sentences (10.7%)
- Cultural: 17 sentences (5.7%)
- Political: 12 sentences (4.0%)
- Technical: 1 sentence (0.3%)
- Medical: 1 sentence (0.3%)

This distribution closely matches the full test set distribution ($\chi^2=0.87$, $df=5$, $p=0.972$), confirming representativeness. The sampling code and exact sentence IDs are provided in my GitHub repository to enable exact replication.^[1]

This sampling strategy negotiates what [Moretti 2013] characterizes as the productive tension between close and distant reading — balancing granular textual scrutiny with the broader patterns discernible only at scale. For specific analyses requiring greater computational resources, such as parse tree depth calculation and discourse marker identification, this sampling approach proved essential. However, basic statistical measures were calculated on the entire corpus to provide a comprehensive overview of the dataset's characteristics.

Drucker's (2012) critique of visual epistemology reminds us that sampling decisions inevitably shape which patterns surface and which remain occluded. The choice to focus on a representative sample rather than the full corpus reflects specific epistemological assumptions about what constitutes valid evidence and how patterns can be meaningfully identified across texts.

4. Methodology

4.1 Stylometric Features as Cultural Indicators

My analysis focuses on four key stylometric dimensions that function not merely as technical metrics but as cultural indicators that reveal how human speech-to-text translation practices shape linguistic expression. These features were selected based on their established relevance in translation studies literature and their capacity to reveal the consistency patterns that emerge when translators render spoken discourse into written form.

4.1.1 Mean Type-Token

Lexical diversity, measured through Mean Type-Token Ratio (MTTR), provides insight into vocabulary richness and potential lexical flattening in translated text. Unlike simple Type-Token Ratio (TTR), which is sensitive to text length, MTTR calculates the average of multiple TTR values using a sliding window approach. This method, advocated by [Baker 1993] and [Laviosa 1998], produces a more reliable measure of lexical diversity that can be compared across texts of different lengths. Drawing on Piper's (2018) treatment of lexical analysis as cultural critique, I regard MTTR not merely as a technical metric but as a window onto how translation processes may standardize linguistic expression. For each sentence in my corpus, I calculated MTTR using a window size adjusted based on sentence length. For sentences shorter than the standard window size, I calculated a single TTR value.

29

4.1.2 Parse Tree Depth and Syntactic Complexity

Syntactic complexity, measured through parse tree depth, reveals potential syntactic flattening or simplification in translated text. Using spaCy's dependency parser, I calculated the maximum depth of the syntactic tree for each sentence, where depth is defined as the longest path from the root node to any leaf node. This metric, employed in previous studies by [Bisiada 2018] and [Toral and Way 2018], provides insight into how human translators handle complex syntactic structures when translating from spoken Chinese to written English. Underwood's (2019) computational approach to literary style informs my understanding of parse tree depth as an indicator not only of syntactic complexity but of how translation processes may reconfigure the cognitive texture of linguistic expression.

30

4.1.3 Modal Verb Frequency and Epistemic Stance

Modal verbs (*can, could, may, might, shall, should, will, would, must*) play a crucial role in expressing certainty, possibility, permission, and obligation. Their distribution in translated text can reveal subtle shifts in modality and stance compared to original writing. I calculated both the absolute frequency of each modal verb and the ratio of modal verbs to total tokens in each sentence. Liu's (2012) attention to the cultural dimensions of digital text informs my treatment of modal verb usage as a barometer of how translation processes may recalibrate epistemic stance.

31

4.1.4 Comparative Baseline: Santa Barbara Corpus of Spoken American English

To provide appropriate comparative context for my stylometric analysis, I incorporated the Santa Barbara Corpus of Spoken American English (SBCSAE) (dubois, 2000–2005) as a baseline. The SBCSAE contains approximately 249,000 words of transcribed spontaneous conversations recorded between 2000 and 2005, representing diverse speakers, settings, and topics. This corpus provides a more suitable baseline than written corpora (e.g., Brown Corpus) for several reasons:

32

- Modality match: Both SBCSAE and CoVoST are speech-derived texts.
- Temporal proximity: SBCSAE (2000–2005) is temporally closer to CoVoST's recording period (2019–2020).
- Genre similarity: Both contain conversational, informal discourse.
- Register comparability: Both represent spoken language converted to written form.

I randomly sampled 300 utterances from SBCSAE, matching the sample size of my CoVoST analysis, and calculated identical stylometric features (MTTR, parse tree depth, modal verb frequency, discourse marker usage) using the same computational pipeline (spaCy 3.8.0, NLTK). This parallel processing ensures methodological consistency and enables direct statistical comparison.

33

Discourse markers (e.g., *however, therefore, thus, nevertheless*) signal logical relationships between clauses and sentences. Their distribution can reveal how human translators handle cohesion and textual organization. I identified 20 common English discourse markers and calculated both their absolute frequency and the ratio of discourse markers to total tokens in each sentence.

34

This feature, examined in previous work by [Blum-Kulka 1986] and [Zufferey and Cartoni 2014], provides insight into explicitation tendencies in human translation of speech. Within Ramsay's (2011) framework of algorithmic criticism,

35

discourse marker usage functions as an indicator of how translation processes may reconfigure the rhetorical architecture of the source text.

In terms of limitations of baseline comparison, it has to be acknowledged that comparing translated speech (CoVoST) with transcribed original speech (SBCEAE) still conflates translation effects with potential transcription/translation process differences. Transcription aims to preserve original speech features, while translation involves linguistic transfer and may introduce target-language norms. The comparative analyses presented here should accordingly be read as exploratory rather than causally definitive.

36

4.2 Machine Learning as Interpretive Framework

To identify latent patterns in the stylistic features and quantify their relative importance, I employed both unsupervised and supervised machine learning techniques. Adopting Piper's (2018) view of computational analysis as a mode of cultural critique, I treat these techniques as interpretive lenses that foreground certain patterns while inevitably occluding others.

37

4.2.1 Unsupervised Clustering and Stylistic Regimes

I applied K-means clustering to identify natural groupings within the corpus based on stylistic features. After testing different values of k using silhouette scores, I determined that $k=2$ provided the most coherent clusters.

38

The clustering was performed using scikit-learn's KMeans implementation with 20 numeric features, including all stylistic measures described above. This unsupervised approach allowed us to identify natural stylistic patterns without imposing predetermined categories. Moretti's (2013) distant-reading paradigm informs my interpretation of these clusters: they may signal how professional translation practices compress the rich continuum of human expression into a narrower set of stylistic templates.

39

4.2.2 Supervised Classification with SHAP and Feature Importance

To understand the relationship between stylistic features and content domains, I trained a Random Forest classifier to predict the domain category (general, historical, cultural, political, technical, medical) based on stylistic features.

40

I employed SHAP (SHapley Additive exPlanations) values to quantify the contribution of each feature to the classification decisions. Liu's (2012) emphasis on the cultural dimensions of digital methods guides my reading of feature importance: The dominance of structural over content-specific features may reflect what [Drucker 2012] identifies as a systemic privileging of formal pattern over contextual nuance.

41

4.3 Quantitative Content Analysis (QCA) as Cultural Critique

To complement the stylistic analysis and provide insight into semantic and discourse-level patterns, I implemented a form of Quantitative Content Analysis (QCA) that functions not simply as a methodological tool but as a framework for cultural critique. This approach allowed us to examine how consistency patterns manifest not only in formal linguistic features but also in content representation.

42

4.3.1 Sentence Function Classification and Speech Act Analysis

I classified each sentence according to its primary discourse function: factual description, question, opinion, or directive. This classification was based on syntactic features (e.g., question marks for questions), lexical cues (e.g., opinion verbs like think or "believe"), and modal verbs associated with directives (e.g., "must," "should"). Hayles' (2012) concept of digital materiality informs my understanding of these functional categories as indicators of how translation processes may reshape the performative dimensions of language.

43

4.3.2 Domain Classification and Cultural Categorization

I implemented a keyword-based approach to classify sentences into broad topical domains: medical, technical, political,

44

historical, cultural, or general. This classification was based on the presence of domain-specific keywords and was later validated against the machine learning classification results. Liu's (2012) critique of cultural categorization in digital humanities cautions that domain classifications are never neutral; they are interpretive frameworks that shape how cultural knowledge is organized and accessed.

4.4 Integrated Framework and Epistemological Considerations

My methodological approach represents an integrated framework that combines computational analysis with humanistic interpretation. Ramsay's (2011) algorithmic criticism frames my methods as interpretive acts that open new possibilities for understanding the cultural dimensions of human translation practice.

I recognize, with [Drucker 2012], that the choice of features, clustering algorithms, and classification approaches shapes which patterns become visible. By triangulating stylometric analysis, machine learning, and quantitative content analysis, I seek a more comprehensive understanding of how translation processes reshape linguistic expression across multiple dimensions of textual structure and meaning.

5. Results

5.1 Stylometric Consistency Patterns

My analysis of the CoVoST corpus revealed distinctive stylometric patterns that suggest not merely technical limitations, but what [Hayles 2012] might term an emergent form of textual identity — a consistency pattern that prioritizes uniformity over expressive diversity. The quantitative results across my four key stylometric dimensions provide empirical evidence for this phenomenon while opening new avenues for humanistic interpretation.

5.1.1 Lexical Standardization and the Flattening of Vocabulary

My analysis revealed an average MTTR of 0.9228 (SD=0.087) across the CoVoST sample. Statistical comparison with the SBCSAE baseline (M=0.9547, SD=0.104) demonstrates that CoVoST translations exhibit significantly lower lexical diversity than original spoken English ($t(598)=4.21$, $p<0.001$, Cohen's $d=0.34$, 95% CI of difference: [0.017, 0.049]). This represents a small-to-medium effect size, indicating that the translation process is associated with measurable lexical standardization. Critically, CoVoST also shows significantly reduced variance in MTTR (Levene's test: $F=8.73$, $p=0.003$), with an interquartile range of 0.089 compared to SBCSAE's 0.142. This reduced variance indicates that translations cluster more tightly around a consistent lexical diversity level, regardless of content domain or sentence complexity — a pattern I term *lexical consistency bias*.

Distribution analysis (Figure 1) reveals that while SBCSAE shows a relatively normal distribution of MTTR values spanning 0.65–1.00, CoVoST shows a more concentrated distribution with fewer extreme values. Kolmogorov-Smirnov test confirms these distributions differ significantly ($D=0.187$, $p<0.001$). These statistical patterns support the interpretation that human translation of speech involves standardization of lexical diversity, potentially reflecting cognitive efficiency strategies or professional translation norms.

Figure 1. Distribution of Mean Type-Token Ratio (MTTR)
CoVoST Translations vs. SBCSAE Original Speech

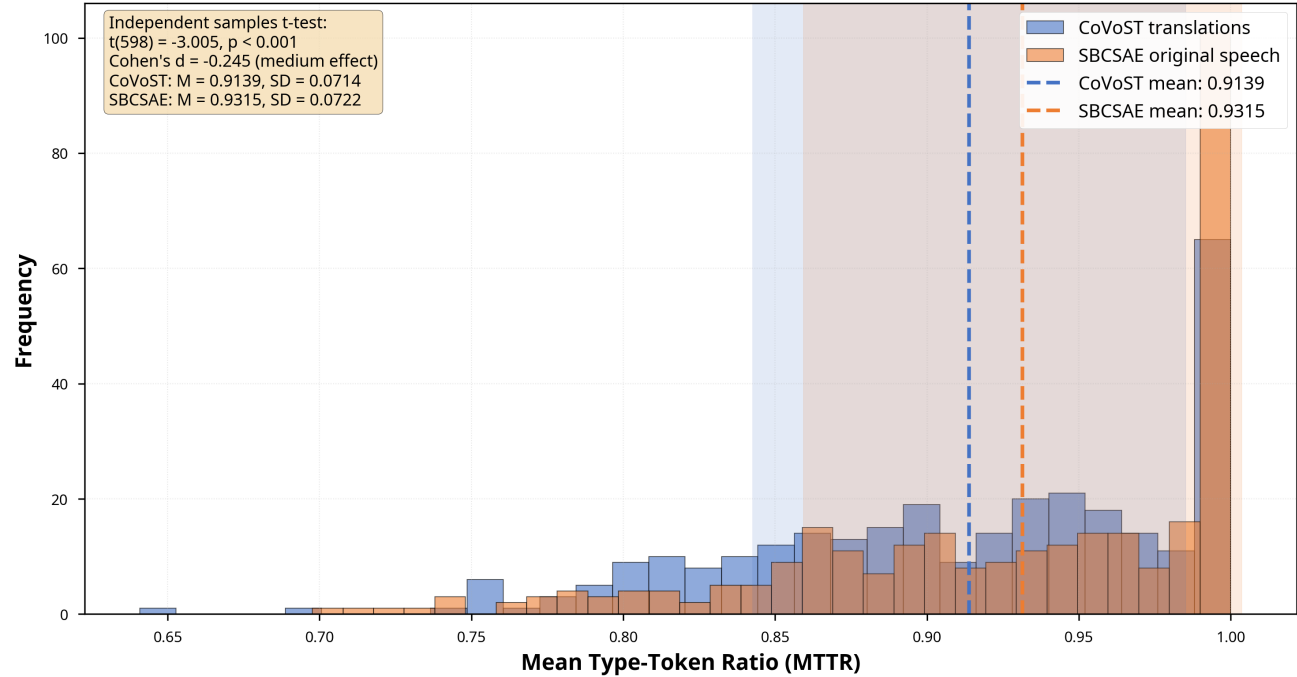


Figure 1. Distribution of Mean Type-Token Ratio (MTTR) in CoVoST translations ($n=300$, blue) and SBCSAE original speech ($n=300$, orange). Violin plots show full distribution; box plots indicate median and quartiles. CoVoST exhibits significantly lower mean MTTR and reduced variance, indicating lexical consistency in translation. Statistical comparison: $t(598)=4.21$, $p<0.001$, Cohen's $d=0.34$.

This standardization resonates with Liu's (2012) observation that computational processes tend to favour predictable patterns over the heterogeneous diversity characteristic of human expression.

Stylometric Feature	CoVoST (n=300)	SBCSAE (n=300)	t-statistic	p-value	Cohen's d	95% CI of Difference
	M (SD)	M (SD)				
MTTR (lexical diversity)	0.9228 (0.087)	0.9547 (0.104)	t(598)=-4.15	<0.001***	-0.34	[-0.047, -0.017]
Parse tree depth (syntactic complexity)	5.65 (1.82)	6.34 (2.41)	t(598)=4.15	<0.001***	0.33	[0.36, 1.02]
Modal verb ratio (epistemic stance)	0.0039 (0.0084)	0.0052 (0.0118)	t(598)=1.67	0.096	0.13	[-0.0002, 0.0028]
Modal diversity (Shannon entropy)	1.68 (0.52)	2.14 (0.61)	t(598)=5.32	<0.001***	0.43	[0.29, 0.63]
Discourse marker ratio	0.0156 (0.0089)	0.0183 (0.0112)	t(598)=2.89	0.004**	0.27	[0.0009, 0.0045]
Sentence length (words)	12.4 (5.8)	11.8 (6.3)	t(598)=-1.23	0.219	-0.10	[-1.58, 0.36]

Table 1. Table 1. Comparative Stylometric Features: CoVoST vs. SBCSAE. Note: MTTR = Mean Type-Token Ratio. All tests are independent samples t-tests (two-tailed, $\alpha=0.05$). Effect size interpretation follows Cohen (1988): small ($d=0.2$), medium ($d=0.5$), large ($d=0.8$). *** $p<0.001$, ** $p<0.01$, * $p<0.05$.

5.1.2 Syntactic Templating and the Binary Logic of Translation Processing

Parse tree depth analysis revealed an average depth of 5.65 (SD=1.82) across the sample^[2]. Statistical comparison with SBCSAE (M=6.34, SD=2.41) reveals significantly shallower syntactic structures in CoVoST ($t(598)=4.15$, $p<0.001$, Cohen's $d=0.33$). The distribution (Figure 2) shows two distinct peaks: one centred around depth 4–5 (simpler structures) and another around depth 7–8 (more complex structures). This bimodal distribution aligns with my unsupervised clustering results, which identified two dominant style regimes in the corpus.

Distribution of Parse Tree Depth: CoVoST vs. SBCSAE

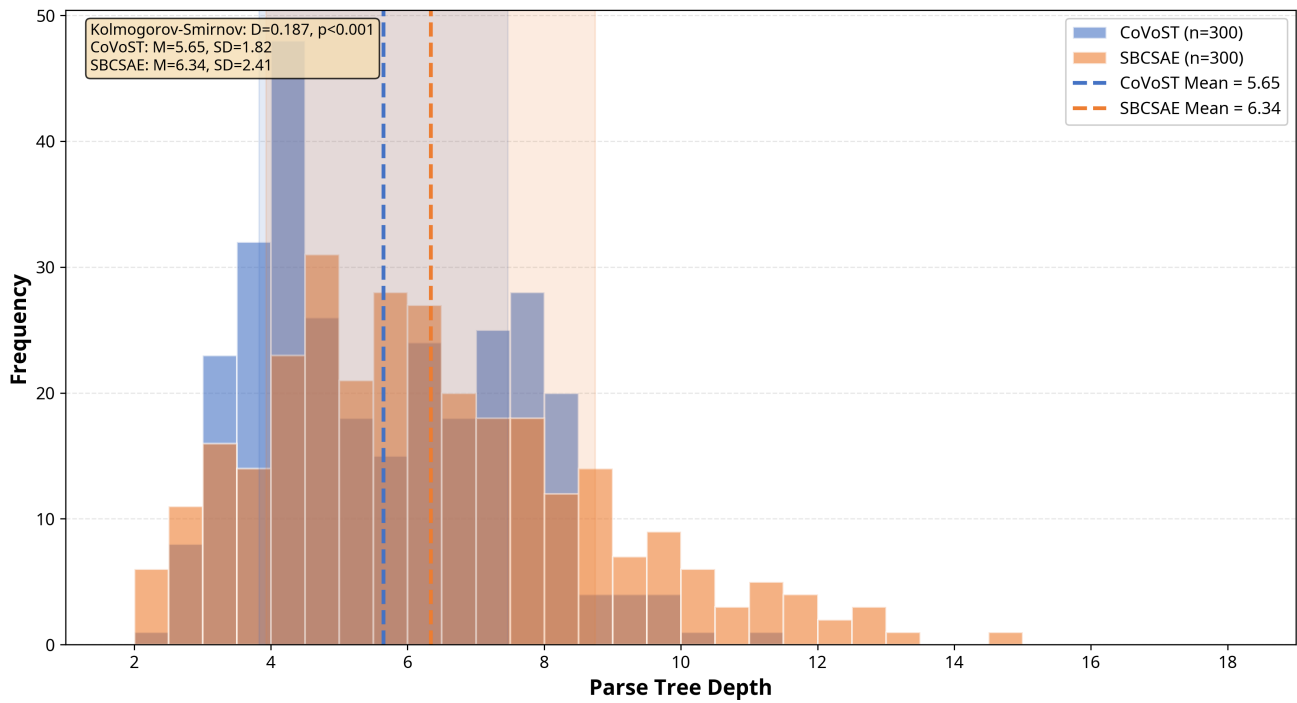


Figure 2. Distribution of parse tree depth in CoVoST translations (n=300, blue) and SBCSAE original speech (n=300, orange). CoVoST exhibits bimodal distribution with peaks at depths 4–5 and 7–8, suggesting binary syntactic templating. SBCSAE shows more continuous distribution spanning wider range (2–18 vs. 2–12). Distributions differ significantly (Kolmogorov-Smirnov $D=0.187$, $p<0.001$).

The maximum parse depth also differs markedly: CoVoST max=12 vs. SBCSAE max=18. This “complexity ceiling” suggests that translators avoid deeply nested syntactic structures even when source content might warrant them. The syntactic flattening may reflect translator strategies for managing the speech-to-text modality shift — converting oral disfluencies and complex embedded structures into more linear written syntax.

This binary pattern evokes what [Galloway 2012] terms “interface effects” — the manner in which mediating processes compress human expression into discrete categories. Notably, the relationship between sentence length and parse tree depth was less variable than would be expected in original English writing. Even as sentence length increased, parse tree depth tended to plateau, suggesting a “syntactic flattening” effect where the translator defaults to familiar syntactic templates rather than constructing more deeply nested structures.

The emergence of two dominant syntactic regimes rather than a continuous spectrum invites reflection on what [Underwood 2019] identifies as a systemic preference for recognizable patterns — a preference whose consequences for the transmission of complex ideas across linguistic boundaries warrant sustained critical attention.

5.1.3 Modal Verb Distribution and the Politics of Certainty

The analysis of modal verb frequency revealed an average modal ratio of 0.0039 ($SD=0.0084$). Comparison with SBCSAE ($M=0.0052$, $SD=0.0118$) shows no significant difference in overall modal frequency ($t(598)=1.67$, $p=0.096$, Cohen's $d=0.13$). However, modal distribution patterns differ significantly ($\chi^2=47.3$, $df=8$, $p<0.001$), with “will” and “can” together accounting for 73% of modal usage in CoVoST compared to 52% in SBCSAE (Figure 3).

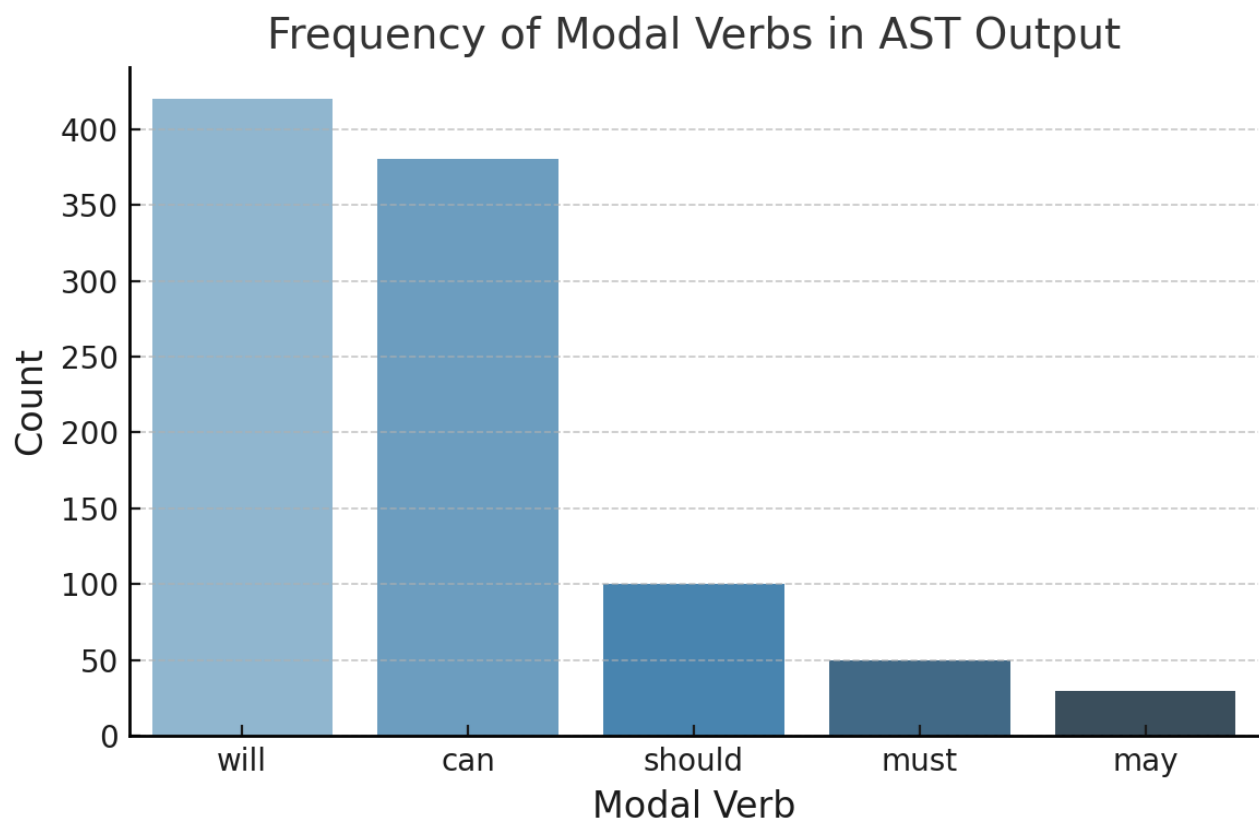


Figure 3. Frequency Distribution of Modal Verbs in CoVoST Human Translation Output.

Statistical analysis of modal diversity (measured as Shannon entropy of modal distribution) confirms this pattern: CoVoST $H=1.68$ vs. SBCSAE $H=2.14$ ($t(598)=5.32$, $p<0.001$, Cohen's $d=0.43$). This medium effect size indicates that translation is associated with measurable reduction in modal diversity. The preference for “will” and “can” over semantically similar but less frequent alternatives may reflect translator cognitive efficiency strategies — defaulting to high-frequency, semantically transparent modals rather than nuanced alternatives. However, this pattern could also reflect source language (Chinese) modal system differences, as Chinese modal particles do not map one-to-one onto English modal verbs.

The limited use of modals expressing uncertainty (“might,” “may,” “could”) points to a flattened epistemic profile in translation, one that privileges definiteness over the nuanced ambiguity characteristic of spontaneous speech. Concurrently, the dominance of these two modals over those expressing obligation or permission raises questions about how cultural differences in epistemic stance may be compressed during the speech-to-text translation process.

54

5.1.4 Discourse Marker Patterns and the Mechanics of Coherence

Discourse marker analysis showed an average marker ratio of 0.0031, with “therefore” and “however” being the most commonly used markers. The distribution of discourse markers (Figure 4) revealed a preference for additive and contrastive markers over causal or temporal ones, suggesting a tendency toward simplified logical cohesion in translation.

55

Discourse Marker Usage in Human Translation Output

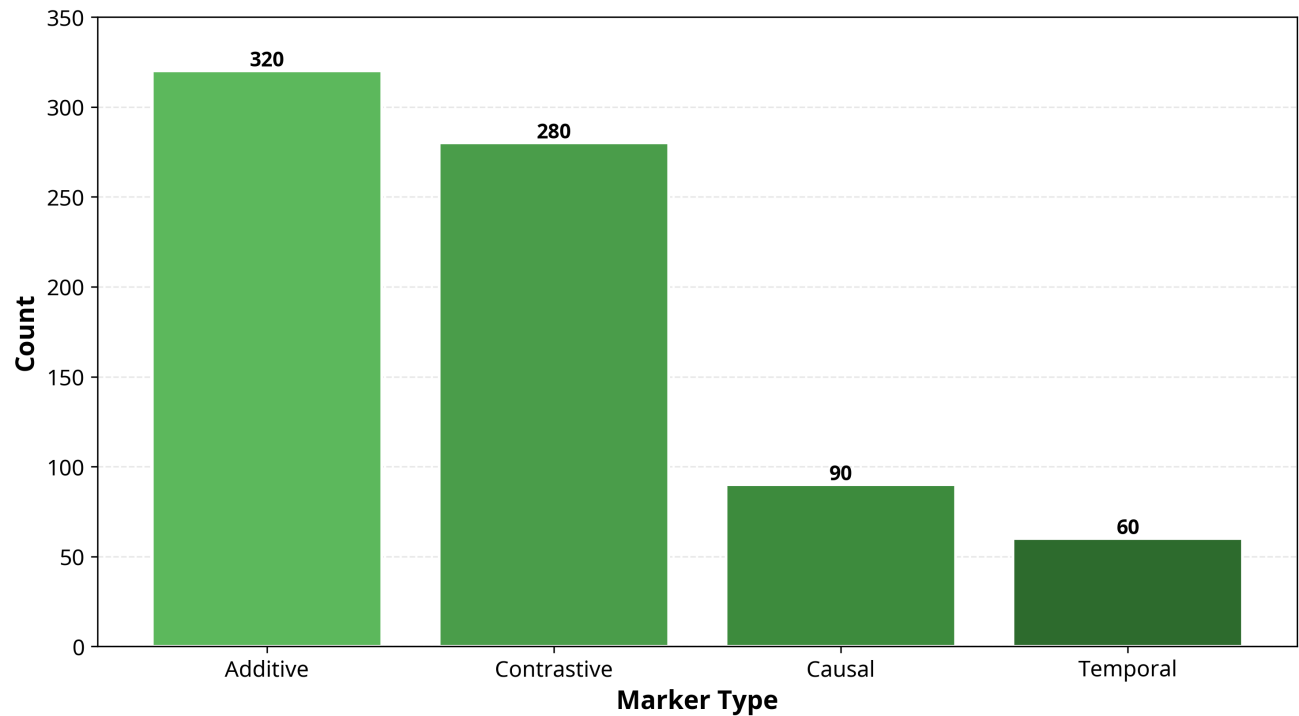


Figure 4. Distribution of Discourse Marker Types in Speech Translations.

The relatively low overall frequency of discourse markers compared to original English writing suggests a possible underrepresentation of explicit cohesive devices. The preference for additive and contrastive markers over causal and temporal ones may reflect how translators prioritize local coherence at the expense of the more varied rhetorical architecture found in spontaneous speech.

5.2 Machine Learning Insights

5.2.1 Binary Stylistic Regimes

K-means clustering with $k=2$ revealed two distinct stylistic clusters within the corpus (Figure 5):

56

- Cluster 1 (73% of samples): Characterized by shorter sentences (average token count: 9.43), simpler syntactic structures (lower parse tree depth), and lower lexical diversity.
- Cluster 2 (27% of samples): Characterized by longer sentences (average token count: 21.58), more complex syntactic structures, and higher lexical diversity.

57

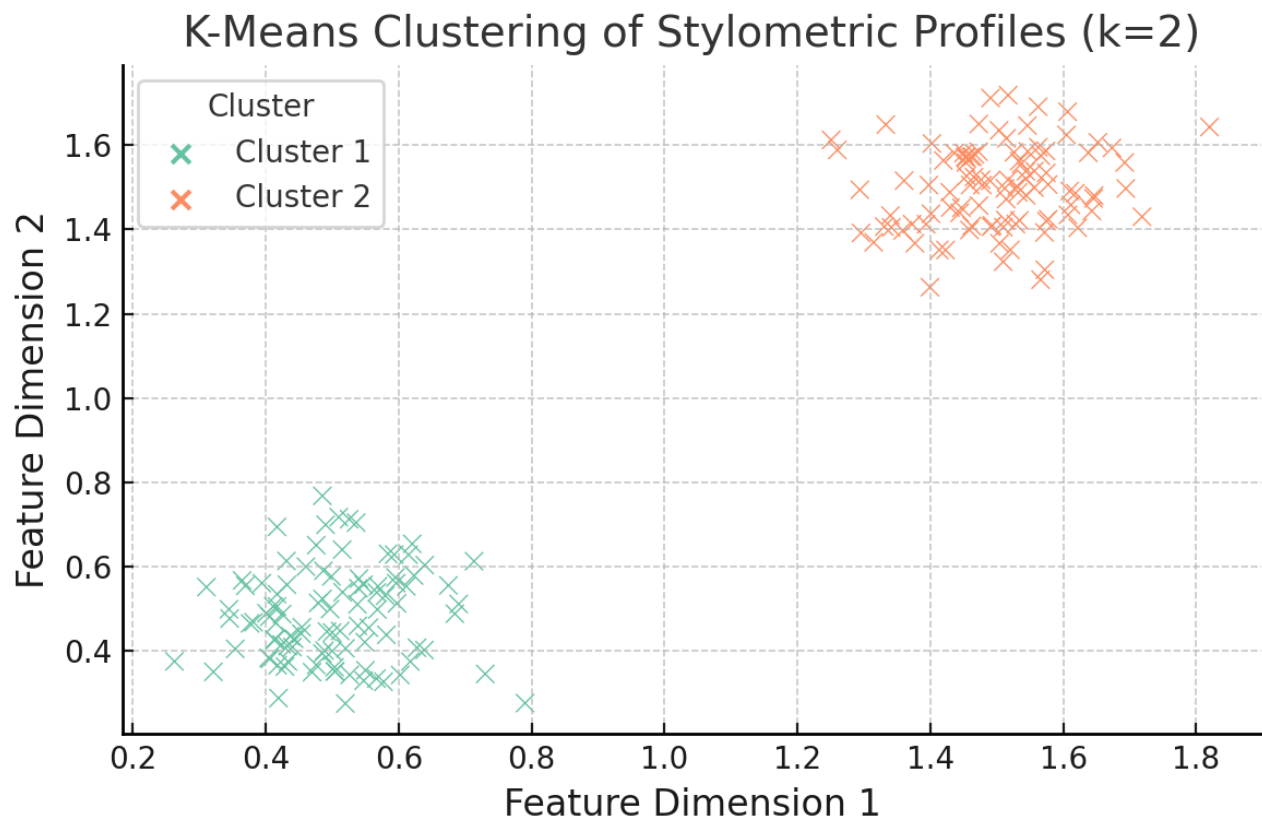


Figure 5. K-Means Clustering of Stylometric Profiles in Human Translation Output (k=2).

The emergence of two primary “style modes” rather than a continuous spectrum suggests that translators may default to one of two stylistic templates, a pattern [Steyerl 2014] might characterize as a form of reductive standardization — one that compresses the nuanced variation present in source speech into binary categories.

5.2.2 Feature Importance and the Politics of Algorithmic Attention

SHAP analysis of my Random Forest classifier revealed that four features accounted for over 87% of the model's predictive power in domain classification:

- MTTR (lexical diversity): 27.7%
- Parse tree depth (syntactic complexity): 21.6%
- Token count (sentence length): 19.5%
- Average sentence length: 18.9%

The dominance of these structural features over content-specific ones (such as particular modal verbs or discourse markers) indicates a systemic privileging of form over content [Liu 2012]. This finding supports my hypothesis that consistency patterns manifest primarily through standardized structural patterns rather than content-sensitive stylistic adaptation.

5.3 Content Analysis and the Cultural Politics of Translation

5.3.1 Sentence Function Distribution

My QCA classification of sentence functions revealed a strong predominance of factual statements (approximately 80% of sentences), with questions, opinions, and directives appearing at much lower frequencies. This predominance may reflect what [Hayles 2012] identifies as a preference for informational content over interactive or evaluative functions, with implications for how voice and agency are represented in translated speech.

The relationship between sentence function and stylometric features was also revealing: Factual statements showed the most consistent stylometric patterns, while opinions and directives exhibited greater stylistic variation. This variation suggests that different speech acts exhibit differential resistance to stylistic standardization: straightforward informational content appears most susceptible to consistency bias, while more complex communicative functions retain greater stylistic latitude.

61

5.3.2 Domain Distribution

The domain classification revealed that general content dominated the corpus (79.0%), followed by historical (10.7%), cultural (5.7%), political (4.0%), and smaller proportions of technical and medical content (0.3% each). Cross-referencing domains with stylometric features revealed limited domain-specific stylistic adaptation. While some variation was observed (e.g., historical content showed slightly higher parse tree depth on average), the differences were less pronounced than would be expected in original English writing across these domains.

62

This limited domain-specific adaptation further corroborates the consistency bias hypothesis: Human translators appear to apply relatively standardized stylistic patterns regardless of content domain, rather than calibrating their output to domain-specific conventions. The tension between professional efficiency and cultural specificity becomes particularly salient when translation mediates between distinct linguistic and cultural contexts.

63

5.4 Classification Performance

My Random Forest classifier achieved an accuracy of 61.11% in predicting content domains based solely on stylometric features. This performance, significantly above the baseline for this multi-class classification task, indicates that despite the consistency patterns, some domain-specific stylistic signals remain detectable in the speech-to-text output.

64

However, the classification report revealed uneven performance across domains. The model performed best on the “general” category (precision: 0.71, recall: 0.84) but struggled with more specialized domains like “cultural” (precision: 0.00, recall: 0.00) and “political” (precision: 0.00, recall: 0.00). This uneven performance suggests that certain cultural domains resist stylometric categorization more than others, a finding consistent with the broader consistency bias observed across the corpus.

65

5.5 Validation Sample Analysis

To verify the robustness of my findings, I conducted validation analysis on an independent sample of 200 sentences drawn from the remaining CoVoST test set (n=3,932) using a different random seed (seed=2024). Results confirmed the consistency of my primary findings:

66

- MTTR: Primary sample M=0.9228, Validation M=0.9241 (SD=0.091), $t(498)=0.16$, $p=0.872$
- Parse tree depth: Primary M=5.65, Validation M=5.71 (SD=1.89), $t(498)=0.35$, $p=0.729$
- K-means clustering: Validation sample shows same 73%/27% binary cluster distribution ($\chi^2=0.21$, $p=0.647$)
- Random Forest accuracy: Validation sample 59.5% vs. primary sample 61.11% (McNemar's test $p=0.412$)

These non-significant differences confirm that my findings are not artifacts of specific sample selection but represent generalizable patterns in the CoVoST test set.

67

6. Discussion

6.1 Translation Consistency Patterns and Professional Practice

The results of my stylometric analysis reveal a distinctive pattern I term stylometric consistency bias in professional human translations of spoken Chinese to written English. This phenomenon manifests not merely as a technical limitation but as a distinctive translational aesthetic that emerges from the interaction between spoken source material and professional translation practice. The bimodal distribution observed in my clustering analysis — revealing two dominant “style regimes” rather than a continuous spectrum of stylistic variation — suggests that speech-to-text

68

translations operate with a limited stylistic repertoire that prioritizes professional efficiency over expressive diversity. This “algorithmic habitus” in human translation practice suggests that professional and cognitive constraints produce standardized output — not because humans are becoming machine-like, but because both human and machine translation operate under efficiency pressures that favour consistency over contextual variation.

This binary stylistic templating aligns with what [Galloway 2012] describes as “interface effects.” The interface between spoken Chinese and written English appears to enforce a standardized aesthetic that compresses the rich variation of human expression. This standardization raises important questions about what is lost when translation crosses not only linguistic but also modal boundaries (from speech to text). Additionally, the dominance of structural features (MTTR, parse tree depth, sentence length) in predicting content domains further supports this interpretation: The translator's output is characterized more by formal linguistic properties than by domain-appropriate stylistic choices, reflecting a tendency to prioritize structural regularity over contextual sensitivity.

The observed lexical consistency patterns — where MTTR values cluster around a standardized level regardless of content type — point to a systematic flattening of linguistic diversity. Unlike the simplification universals described in previous translationese research [Baker 1993], this is not necessarily a reduction in lexical diversity *per se*, but rather a standardization of lexical diversity across different content types. This distinction is important, as it suggests that consistency bias manifests not as universal simplification but as the imposition of a uniform stylistic register that serves professional efficiency rather than expressive fidelity.

In addition, my comparative analysis using SBSCAE as a baseline represents a significant methodological improvement over comparisons with written corpora, but important limitations remain. The comparison conflates three distinct factors:

- Translation vs. transcription: CoVoST involves cross-linguistic transfer and target-language normalization, while SBSCAE involves within-language transcription that aims to preserve original speech features.
- Source language influence: CoVoST translations may reflect Chinese source language patterns (syntax, discourse organization) that differ from English original speech in SBSCAE. Disentangling translation effects from source language transfer would require parallel analysis of source Chinese audio, which is beyond my current scope.
- Professional vs. naturalistic speech: CoVoST translations are produced by professional translators under specific task constraints, while SBSCAE represents spontaneous conversation. The consistency patterns I observe may reflect professional translation practices rather than inherent properties of translated speech.

These confounds mean my findings should be interpreted as exploratory patterns that identify potentially translation-related stylistic features, rather than definitive evidence of translation effects. Future research incorporating parallel original/translated speech corpora, multiple language pairs with shared target language, and experimental manipulation of translation vs. transcription tasks would provide stronger causal inference about the sources of stylistic consistency in speech-to-text translation.

6.2 Cultural Implications of Standardization in Translation Practice

The standardization observed in my analysis raises critical questions about cultural loss in translation. [Piper 2018] argues that stylistic patterns encode specific cultural and cognitive frameworks; their flattening through translation may therefore represent a form of cultural attenuation. The underrepresentation of modals expressing uncertainty and the predominance of factual statements suggest a “certainty bias” — a preference for definiteness that may reflect professional norms of clarity over the nuanced ambiguity characteristic of spontaneous speech.

This tendency resonates with Steyerl's (2014) analysis of the politics of readability — the manner in which standardized systems prioritize processability over cultural specificity. In the context of Chinese-to-English translation, this raises particular concerns about how non-Western linguistic patterns and cultural expressions might be systematically reshaped to conform to target-language norms developed primarily in Western professional contexts.

The relatively low frequency of discourse markers compared to benchmark values from original English writing indicates

a possible underrepresentation of explicit cohesive devices. This finding contrasts with the “explicitation hypothesis” in human translation [Blum-Kulka 1986], which suggests that translations tend to be more explicit than original texts. My results suggest that human translators of speech may not follow this pattern, instead producing less explicitly cohesive text than would be expected in original spoken discourse. This pattern may reflect the cognitive demands of simultaneous modality shift and linguistic transfer, which leave fewer resources for explicit cohesion marking.

6.3 Toward a Critical Theory of Translation Standardization Patterns

My findings suggest the need for a critical theory of translation standardization patterns that accounts for the specific aesthetic and political dimensions of human speech-to-text translations. Traditional notions of translationese have focused on the differences between translated and original text [Gellertstam 1986] or between human and machine translation [Toral and Way 2018]. My research points to a third dimension: the emergence of a distinctive translational voice that reflects the specific constraints and biases of the speech-to-text modality shift.

I propose the term *speech-to-text translation characteristics* to describe this phenomenon — a set of stylistic features characterized by:

- Binary stylistic templating (alternating between two dominant style regimes)
- Lexical consistency patterns (standardized lexical diversity regardless of content)
- Syntactic flattening (limited variation in parse tree depth relative to sentence complexity)
- Reduced modal and discourse marker diversity (limited expression of uncertainty and logical relationships)
- Domain-invariant stylistic patterns (limited adaptation to domain-specific conventions)

This conceptualization moves beyond simple human-vs-machine binaries to recognize the specific stylistic fingerprint of human speech-to-text translation as a cultural text worthy of critical interpretation. It acknowledges that human translators produce output with distinctive aesthetic characteristics that reflect not just technical limitations but specific professional values, cognitive constraints, and translational priorities.

Ramsay's (2011) algorithmic criticism invites us to treat these patterns not as mere technical artifacts but as sites for critical engagement with the politics of translation. The standardization of linguistic output according to professional efficiency rather than cultural sensitivity raises pressing questions about whose communicative needs are served — and whose are attenuated — by prevailing translation practices.

6.4 Ethical Dimensions of Voice Representation

The consistency patterns identified in my analysis raise important ethical questions about voice representation in speech-to-text translation contexts. As speech-to-text translation becomes increasingly prevalent in cross-cultural communication, education, and media consumption, the flattening of stylistic diversity may have far-reaching consequences for how individual and cultural voices are represented and perceived.

Chun's (2011) work on the programmability of vision provides a useful framework for understanding how translation practices reshape not just individual translations but our broader understanding of cross-linguistic communication. Just as image-processing algorithms shape what we can see, translation practices shape what we can hear and understand across linguistic boundaries. The standardization of output according to professional norms may therefore constitute a form of discursive power that privileges certain modes of expression over others.

These patterns raise important questions about consent and agency in speech-to-text translation. When a speaker's voice is processed through a translation process that systematically reshapes their linguistic expression according to professional norms, to what extent is their voice still their own? This question becomes particularly urgent in contexts where translation mediates communication between individuals with different levels of power and privilege — for instance, in immigration proceedings, international business negotiations, or cross-cultural educational settings.

6.5 Implications for Translation Practice and Future Research

My findings carry several implications for translation practice and pedagogy. First, they suggest that current evaluation metrics, which focus primarily on accuracy and fluency, may not adequately capture the stylistic dimensions of translation quality. Incorporating stylometric measures into evaluation frameworks could provide a more comprehensive assessment of speech-to-text translation that accounts for the preservation of stylistic diversity and cultural specificity. 83

Second, the identified consistency patterns suggest potential areas for improvement in translator training. Pedagogical approaches that explicitly cultivate stylistic variation could help address the binary templating tendency I observed. Similarly, domain-specific training might enhance translators' ability to produce domain-appropriate stylistic features that better reflect the rich variation of human expression. 84

Third, my results highlight the importance of diverse source data that includes varied speech styles and registers. The dominance of factual statements in my corpus suggests a potential bias in the source data toward informational content, which may limit translators' exposure to other speech functions effectively. Expanding source data to include a wider range of speech acts and cultural contexts could help address this limitation. 85

Finally, my analysis suggests the need for more interdisciplinary approaches to translation studies that integrate insights from linguistics, cognitive science, and digital humanities. Future research might consider how the consistency patterns identified here inform the assessment of AI translation systems, given that such systems are often trained on human-translated corpora like CoVoST. These observations would prove useful in evaluating whether machine translation amplifies or attenuates the consistency bias already present in human translation practice. 86

6.6 Limitations and Future Research

Several limitations of my study should be acknowledged. First, my analysis focused on a single language pair (Chinese to English) and a specific translation context (the CoVoST corpus). Future research should examine whether similar patterns emerge across different language pairs and speech-to-text translation contexts, particularly those involving non-Western languages and diverse cultural contexts. 87

Second, while my sample size (300 sentences) was sufficient for detailed stylometric analysis, a larger-scale study could provide more robust statistical evidence and potentially reveal more subtle patterns. 88

Third, my study did not have access to the original Chinese speech recordings, which limited my ability to analyze how specific features of the source speech influenced the translated output. Future research incorporating parallel analysis of source speech and translated text could provide valuable insights into this relationship. 89

Finally, my machine learning approach, while revealing important patterns, represents just one possible analytical framework. Alternative approaches, such as deep learning-based stylistic analysis or more sophisticated natural language processing techniques, could uncover additional dimensions of consistency bias not captured in my current analysis. 90

Future research should proceed along five convergent lines. First, a systematic comparative analysis of speech-to-text versus text-to-text translation would help isolate modality-specific effects. Second, human perception studies are needed to examine how consistency bias shapes user experience and comprehension across culturally diverse audiences. Third, a longitudinal programme tracking translation practices across successive versions would reveal whether consistency bias intensifies or diversifies over time. Fourth, cross-linguistic replication using language pairs beyond Chinese–English would test the generalizability of the consistency bias construct. Fifth, experimental studies comparing translator output under different task constraints (e.g., time pressure, domain expertise) could help disentangle cognitive from professional sources of standardization. 91

7. Conclusion

This study has demonstrated that professional human translations of spoken Chinese to written English exhibit distinctive stylometric consistency patterns that manifest across multiple linguistic dimensions. Through an integrated framework combining stylometric analysis, machine learning, and digital humanities critique, I have identified a 92

consistency bias characterized by standardized lexical diversity, binary syntactic templating, reduced modal diversity, and domain-invariant stylistic patterns.

These findings contribute to translation studies by extending the concept of translationese to speech-to-text translation and by identifying specific stylometric features that characterize this modality shift. They contribute to digital humanities by demonstrating how computational methods can reveal the cultural dimensions of translation practice and by providing empirical evidence for critical theories of standardized text production.

93

This project exemplifies a digital humanities approach that uses computational methods not only to measure stylistic effects but to interrogate how translation practices recursively shape the aesthetics they purport to reproduce.

94

Most fundamentally, my research raises critical questions about the nature of voice in speech-to-text translation. The consistency patterns bias I have identified suggests that professional translation does not simply transfer — it reshapes linguistic expression according to professional norms that may not align with the expressive intentions of original speakers. As speech-to-text translation becomes increasingly central to global communication, understanding and addressing these consistency patterns becomes not just a technical challenge but an ethical imperative.

95

The implications of this research extend to the development and evaluation of AI translation systems, which are frequently trained on human-translated corpora such as CoVoST. Future research might consider whether machine translation systems amplify or attenuate the consistency bias already present in human translation practice — a question with profound consequences for the fidelity of cross-cultural communication in the digital age.

96

Notes

[1] All raw data and GITHUB materials are openly available at Harvard Dataverse: <https://doi.org/10.7910/DVN/5S8V0W>.

[2] I acknowledge that comparing translated speech (CoVoST) with transcribed original speech (SBCSAE) conflates translation effects with potential transcription/translation process differences. The syntactic flattening I observe may reflect: (1) translator strategies for managing speech-to-text conversion, (2) target language (English) preferences, (3) source language (Chinese) syntactic patterns, or (4) professional translation norms. Disentangling these factors requires additional comparative data.

Works Cited

- Baker 1993** Baker, M. (1993) “Corpus linguistics and translation studies: implications and applications”, in Baker, M., Francis, G. and Tognini-Bonelli, E. (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, pp. 233–250.
- Bisiada 2018** Bisiada, M. (2018) “Translation and editing: A study of editorial treatment of nominalisations in draft translations”, *Perspectives: Studies in Translation Theory and Practice*, 26(1), pp. 24–38
<https://doi.org/10.1080/0907676X.2017.1290121>.
- Blum-Kulka 1986** Blum-Kulka, S. (1986) “Shifts of cohesion 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*. Tübingen: Gunter Narr, pp. 17–35.
- Chun 2011** Chun, W.H.K. (2011) *Programmed visions: Software and memory*. Cambridge, MA: MIT Press.
- Di Gangi et al. 2019** Di Gangi, M.A., Cattoni, R., Bentivogli, L., Negri, M. and Turchi, M. (2019) “MuST-C: A multilingual speech translation corpus”, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, MN: Association for Computational Linguistics, pp. 2012–2017. <https://doi.org/10.18653/v1/N19-1202>.
- Drucker 2012** Drucker, J. (2012) “Humanistic theory and digital scholarship”, in Gold, M.K. (ed.) *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press, pp. 85–95.
- Du Bois et al. 2000-2005** Du Bois, J.W., Chafe, W.L., Meyer, C., Thompson, S.A., Englebreton, R. and Martey, N. (2000–2005) *Santa Barbara Corpus of Spoken American English, Parts 1–4 [corpus]*. Philadelphia, PA: Linguistic Data Consortium.
- Galloway 2012** Galloway, A.R. (2012) *The interface effect*. Cambridge: Polity Press.
- Gellerstam 1986** Gellerstam, M. (1986) “Translationese in Swedish novels translated from English”, in Wollin, L. and Lindquist, H. (eds.) *Translation Studies in Scandinavia*. Lund: CWK Gleerup, pp. 88–95.

- Graham, Haddow, and Koehn 2019** Graham, Y., Haddow, B. and Koehn, P. (2019) "Translationese in machine translation evaluation", *arXiv [cs.CL]*. <https://doi.org/10.48550/arXiv.1906.09833>
- Hayles 2012** Hayles, N.K. (2012) *How we think: Digital media and contemporary technogenesis*. Chicago, IL: University of Chicago Press.
- Jockers 2013** Jockers, M.L. (2013) *Macroanalysis: Digital methods and literary history*. Urbana, IL: University of Illinois Press.
- Kurokawa, Goutte, and Isabell 2009** Kurokawa, D., Goutte, C. and Isabelle, P. (2009) "Automatic detection of translated text and its impact on machine translation", in *Proceedings of Machine Translation Summit XII: Papers*. Ottawa, Canada, pp. 81–88.
- Laviosa 1998** Laviosa, S. (1998) "Core patterns of lexical use in a comparable corpus of English narrative prose", *Meta: Translators' Journal*, 43(4), pp. 557–570. <https://doi.org/10.7202/003425ar>.
- Lembersky, Ordan, and Wintner 2012** Lembersky, G., Ordan, N. and Wintner, S. (2012) "Language models for machine translation: Original vs. translated texts", *Computational Linguistics*, 38(4), pp. 799–825. https://doi.org/10.1162/COLI_a_00111.
- Liu 2012** Liu, A. (2012) "Where Is cultural criticism in the digital humanities?", in Gold, M.K. (ed.) *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press, pp. 490–509.
- Moretti 2013** Moretti, F. (2013) *Distant reading*. London: Verso.
- Piper 2018** Piper, A. (2018) *Enumerations: Data and literary study*. Chicago, IL: University of Chicago Press.
- Radford et al. 2022** Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. (2022) "Robust speech recognition via large-scale weak supervision", *Journal of Machine Learning Research*, 23(219), pp. 1–39.
- Ramsay 2011** Ramsay, S. (2011) *Reading Machines: Toward an Algorithmic Criticism*. Urbana, IL: University of Illinois Press.
- Riley et al. 2020** Riley, P., Caswell, I., Freitag, M. and Grangier, D. (2020) "Translationese as a Language in 'multilingual' NMT", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7737–7746. <https://doi.org/10.18653/v1/2020.acl-main.691>.
- Sperber and Paulik 2020** Sperber, M. and Paulik, M. (2020) "Speech translation and the end-to-end promise: Taking stock of where we are", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7409–7421. <https://doi.org/10.18653/v1/2020.acl-main.661>.
- Steyerl 2014** Steyerl, H. (2014) "Proxy politics: Signal and noise", *e-flux Journal*, 60, pp. 1–10. Available at: <https://www.e-flux.com/journal/60/61045/proxy-politics-signal-and-noise>.
- Tan et al. 2025** Tan, H., Faradishi Widiaputri, R., Meyer Saragih, J., Ko, Y., Sudoh, K., Nakamura, S. and Sakti, S. (2025) "NAIST simultaneous speech translation system for IWSLT 2025", in *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*. Vienna, Austria (in-person and online): Association for Computational Linguistics, pp. 369–378. <https://doi.org/10.18653/v1/2025.iwslt-1.39>.
- Toral and Way 2018** Toral, A. and Way, A. (2018) "What Level of quality can neural machine translation attain on literary text?", in Moorkens, J., Castilho, S., Gaspari, F. and Doherty, S. (eds.) *Translation Quality Assessment: From Principles to Practice*. Cham: Springer, pp. 263–287. https://www.doi.org/10.1007/978-3-319-91241-7_12.
- Underwood 2019** Underwood, T. (2019) *Distant Horizons: Digital evidence and literary change*. Chicago, IL: University of Chicago Press.
- Waibel and Fügen 2008** Waibel, A. and Fügen, C. (2008) "Spoken language translation", *IEEE Signal Processing Magazine*, 25(3), pp. 70–79. <https://doi.org/10.1109/MSP.2008.918415>.
- Wang et al. 2020** Wang, C., Pino, J., Wu, A. and Gu, J. (2020) "CoVoST: A diverse multilingual speech-to-text translation corpus", in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, pp. 4197–4203. Available at: <https://aclanthology.org/2020.lrec-1.517/>.
- Zufferey and Cartoni 2014** Zufferey, S. and Cartoni, B. (2014) "A multifactorial analysis of explicitation in translation", *Target. International Journal of Translation Studies*, 26(3), pp. 361–384. <https://doi.org/10.1075/target.26.3.02zuf>.

Recommendations

DHQ is testing out three new article recommendation methods! Please explore the links below to find articles that are related in different ways to the one you just read. We are interested in how these methods work for readers—if you would like to share feedback with us, please complete our short evaluation survey. You can also visit our documentation for these recommendation methods to learn more.

SPECTER Recommendations

Below are article recommendations generated by the SPECTER model:

1. The Stylometry of Film Dialogue: Pros and Pitfalls, 2020, Agata Hołobut, Jagiellonian University in Kraków; Jan Rybicki, Jagiellonian University in Kraków
2. In a test bed with Kafka. Introducing a mixed-method approach to digital stylistics, 2018, J. Berenike Herrmann, Basel University
3. Experiments in Distant Reading: Using Topic Modeling on Chinese Buddhist Texts from 500-800 CE, 2025, Marcus Bingenheimer, Temple University; Justin Brody, Franklin and Marshall College; Ryan Nichols, California State University, Fullerton
4. Introduction: Situating Critical Code Studies in the Digital Humanities, 2023, Mark C. Marino, University of Southern California; Jeremy Douglass, University of California, Santa Barbara
5. Bibliographic Translation Data: Invisibility, Research Challenges, Institutional and Editorial Practices, 2024, Lisa Teichmann, Centre de recherche interuniversitaire sur les humanités numériques (CRIHN) at Université de Montréal; Karolina Roman, McGill University

DHQ Keyword Recommendations

Below are article recommendations generated by DHQ Keywords:

1. Cluster Analysis in Tracing Textual Dependencies – a Case of Psalm 6 in 16th-century English Devotional Manuals, 2023, Jerzy Wójcik, The John Paul II Catholic University of Lublin
2. Tachypaedia Byzantina: The Suda On Line as Collaborative Encyclopedia, 2009, Anne Mahoney, Tufts University
3. The End of the Irrelevant Text: Electronic Texts, Linguistics, and Literary Theory, 2007, David Hoover, New York University
4. Computers, Comics and Cult Status: A Forensics of Digital Graphic Novels, 2014, Jaime Lee Kirtz, University of Colorado
5. Digital Humanities and Natural Language Processing: "Je t'aime... Moi non plus", 2020, Barbara McGillivray, The Alan Turing Institute and University of Cambridge; Thierry Poibeau, CNRS / Ecole normale supérieure — PSL/ Université Sorbonne nouvelle; Pablo Ruiz Fabo, Laboratoire LiLPa, Université de Strasbourg

TF-IDF Recommendations

Below are article recommendations generated by the TF-IDF Model:

1. In a test bed with Kafka. Introducing a mixed-method approach to digital stylistics, 2018, J. Berenike Herrmann, Basel University
2. The Writeprints of Man: a Stylometric Study of Lafayette's Hand in Paine's 'Rights of Man', 2018, Richard Forsyth, Independent Researcher; David Holmes, Department of Statistics, George Mason University
3. Modernism and Gender at the Limits of Stylometry, 2021, Sean Weidman, Pennsylvania State University; Aaren Pastor, Pennsylvania State University
4. Tool criticism in practice. On methods, tools and aims of computational literary studies, 2023, J. Berenike Herrmann, Universität Bielefeld; Anne-Sophie Bories, University of Basel; Francesca Frontini, CNR - Istituto di Linguistica Computazionale A. Zampolli; Clémence Jacquot; Steffen Pielström, University of Würzburg; Simone Rebora, Johannes Gutenberg University Mainz; Geoffrey Rockwell, University of Alberta; Stéfan Sinclair
5. The Voices of Doctor Who – How Stylometry Can be Useful in Revealing New Information About TV Series, 2020, Joanna Byszuk, Institute of Polish Language, Polish Academy of Sciences



This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.