

Translation Style Variation across Registers in Machine Translation and Large Language Model Outputs

A Corpus-Based Study of English-to-Chinese Translation

Shuyue Wang

School of Foreign Languages, Northeast Normal University, Changchun, China

Gang Cheng

School of Foreign Languages, Northeast Normal University, Changchun, China

Abstract This study examines translation style variation across registers in English Chinese outputs from TencentMT, DeepSeek, and ArkLLM. Using an aligned corpus of public speech, legal-diplomatic texts, and news, it compares length ratio, mean sentence length, and connective density through mixed-effects modelling and qualitative analysis. ArkLLM generally produced longer outputs, while TencentMT produced shorter sentences and used more explicit connectives. Register also shaped these patterns, showing that AI translation style varies across systems and communicative contexts.

Keywords Machine translation. Large language models. Translation style. Register variation. English-to-Chinese translation.

Summary 1 Introduction. – 2 Related Work. – 3 Data and Methods. – 4 Results. – 5 Discussion. – 6 Conclusions.



Peer review

Submitted 2026-04-30
Accepted 2026-08-03
Published 2026-10-12



Open access

© 2026 Wang, Cheng | 4.0

1 Introduction

The rapid development of neural machine translation and large language models has changed the conditions under which translated texts are produced, evaluated, and revised.¹ Dedicated machine translation systems are now widely used in professional and non-professional translation settings, while large language models are increasingly adopted as flexible multilingual text-generation tools. Their growing use raises a question that is important for both machine translation research and translation studies: when different AI systems translate the same source text, do they merely differ in quality, or do they also construct different kinds of translated discourse?

Mainstream research on machine translation has been organised around quality evaluation. Automatic metrics such as BLEU, BERTScore, and COMET have enabled large-scale comparison of system outputs and have helped researchers estimate adequacy, fluency, and semantic similarity to human references (Papineni et al. 2002, 311-13; Zhang et al. 2020, 1; Rei et al. 2020, 2685). Recent studies of large language models have also examined whether GPT-style systems can match or outperform dedicated machine translation systems across language pairs and domains (Jiao et al. 2023, 1-2; Hendy et al. 2023, 1-2). These approaches remain valuable, but they do not fully explain how translations are textually organised. A system output may be semantically acceptable while still differing from another output in length, sentence segmentation, discourse cohesion, or degree of explicitness.

This distinction is central to corpus-based translation studies. Translated texts are not only carriers of transferred meaning; they are also linguistic products with recurrent patterns that can be observed, counted, and interpreted (Baker 1993, 235; Laviosa 2002, 21-3). Research on translationese and translational variation has shown that translated language may display systematic tendencies in simplification, explicitation, normalisation, or source-language

1 Data Availability Statement: the source corpora used in this study are publicly available under their respective access and reuse conditions. The analysis scripts, variable definitions, connective list, and representative examples may be made available in a public repository. The full derived segment-level dataset is not publicly deposited because it contains source and reference materials subject to the reuse conditions of the original corpora. Further information about the analysed subset may be made available by the corresponding author upon reasonable request. Funding: this research received no external funding. Conflict of Interest Statement: the author declares no conflict of interest. Statement on AI Tools: AI translation systems were used as research objects in this study. No AI-generated output was used as empirical evidence unless it was included in the analysed dataset. The author remains responsible for the research design, analysis, interpretation, and manuscript content.

interference (Chesterman 2004, 8; Volansky, Ordan, Wintner 2015, 103-12). With the rise of AI translation, a related question emerges: do machine translation systems and large language models produce their own textual profiles, and are these profiles stable across different registers?

Register is central to this question. Linguistic choices are shaped by communicative purpose, audience, medium, and discourse convention (Biber 1988, 28-46; Biber, Conrad 2009, 31-49). Public speech, legal-diplomatic institutional writing, and journalistic news differ not only in topic but also in textual organisation. Public speech often relies on rhythm, interpersonal orientation, repetition, and explicit discourse markers. Legal-diplomatic institutional texts tend to favour formulaic expressions, long syntactic units, and conventional patterns of obligation or procedure. News texts are shaped by information density, participant reference, attribution, and journalistic readability. A translation system that appears expansive in one register may be following formulaic constraints in another; a lower density of connectives may indicate more implicit cohesion in one context but weakened logical marking in another.

English-to-Chinese translation provides a particularly informative context for examining these issues because differences in information packaging, cohesion, and sentence organisation are especially visible between the two languages. English often uses explicit syntactic links and embedded structures, while Chinese may rely more heavily on clause order, punctuation, contextual inference, and paratactic relations. Translating from English into Chinese therefore frequently involves decisions about whether to preserve source sentence boundaries, split dense structures, reduce explicit connectives, or expand information for target-language readability. These decisions are not merely surface-level choices; they shape the style and discourse texture of the translated Chinese text.

This study investigates translation style variation in English-to-Chinese AI translation across three registers: public speech, legal-diplomatic institutional texts, and journalistic news. It compares one dedicated machine translation system, TencentMT, with two large language model-based systems, DeepSeek and ArkLLM. The analysis focuses on three measurable textual features: length ratio against human reference translation, mean sentence length, and connective density per 1,000 Chinese token-like units. These variables are not treated as direct measures of translation quality. They are used as indicators of how systems manage textual volume, sentence-level packaging, and explicit discourse cohesion.

The study is based on an aligned corpus subset in which each English source segment was translated by all three systems. This repeated-measures design makes it possible to compare systems on the same source materials while controlling for source-segment

effects through mixed-effects modelling. Quantitative results are then interpreted with qualitative examples, so that statistical patterns can be connected to actual translation behaviour such as compression, expansion, segmentation, formulaic convergence, and source-boundary preservation.

The article argues that AI translation systems are not stylistically neutral. They do not simply produce more or less accurate versions of the same target text; they construct Chinese translations through different textual strategies. TencentMT tends to produce compact outputs with more explicit connective marking, ArkLLM tends to produce longer and more expansive translations, and DeepSeek often occupies an intermediate position while remaining closer to ArkLLM in sentence organisation. These tendencies are further shaped by register: public speech allows greater variation in rhetorical segmentation, legal-diplomatic institutional discourse narrows stylistic freedom through formulaic constraints, and the selected news examples reveal informational expansion and possible source-boundary effects.

By shifting attention from quality ranking to register-sensitive translation style, this study contributes to AI translation evaluation and corpus-based translation studies. It shows that system comparison should consider not only whether a translation is acceptable, but also what kind of translated discourse a system produces. This is especially important for professional translation and post-editing, where translators do not work with abstract quality scores but with concrete texts that must be revised, adapted, and made appropriate for specific communicative situations.

2 Related Work

2.1 Translation Quality, Translationese, and Textual Style

Machine translation evaluation has long been supported by automatic metrics such as BLEU, BERTScore, and COMET, which provide quantitative estimates of similarity or quality-related performance rather than qualitative analyses of textual organisation (Papineni et al. 2002, 311-13; Zhang et al. 2020, 1; Rei et al. 2020, 2685-6). These metrics are useful for benchmarking, but they are less suited to explaining how translated texts are organised as discourse.

Corpus-based translation studies offer a complementary perspective by examining systematic linguistic patterns in translated texts (Baker 1993, 235; Laviosa 2002, 23-7). Work on translationese and translation universals has discussed tendencies such as simplification, explicitation, normalisation, and source-language

interference (Chesterman 2004, 8; Volansky, Ordan, Wintner 2015, 105-7). These tendencies are not deterministic rules, but they provide a way to study translation as patterned language use.

This perspective is particularly relevant for machine-generated translation, where outputs may exhibit systematic textual tendencies beyond conventional quality scores. Toral (2019, 273-6) shows that post-edited MT may retain or intensify certain translationese features rather than simply becoming indistinguishable from human translation. Vanmassenhove et al. (2021, 2203-4) propose the notion of “machine translationese”, suggesting that MT output may display algorithmically shaped regularities and reduced linguistic diversity. Such work shifts attention from isolated errors to broader textual profiles. The present study follows this direction by examining AI translation style through measurable features of textual volume, sentence packaging, and explicit cohesion.

Translation style is therefore not treated as an impressionistic judgement of elegance, but as a set of systematic textual patterns. It refers to systematic variation in how systems construct target-language texts. Length ratio, mean sentence length, and connective density do not replace quality evaluation, but they make visible dimensions of system behaviour that overall metrics may obscure.

2.2 LLM Translation and Document-level MT Evaluation

Neural machine translation has substantially improved automatic translation performance. Sequence-to-sequence architectures, attention mechanisms, and Transformer-based models have enabled MT systems to generate more fluent and contextually plausible target texts than earlier statistical systems (Bahdanau, Cho, Bengio 2015, 1-3; Vaswani et al. 2017, 5998-6000; Koehn 2020, 67-88). Dedicated MT systems are usually trained and optimised for translation tasks, often using large parallel corpora and task-specific decoding procedures. Their output behaviour is therefore shaped by translation-oriented objectives and the conventions embedded in parallel training data.

Large language models create a different translation setting because translation is embedded within broader text-generation capabilities. They are general-purpose text-generation systems that perform translation through instruction following rather than through a task-specific MT interface alone. Studies of ChatGPT and GPT-style models suggest that LLMs can produce competitive translations in many language pairs and domains, although their performance varies according to prompt, language direction, domain, and evaluation setting (Jiao et al. 2023, 1-2; Hendy et al. 2023, 1-2). LLM translation should therefore not be understood simply as a better or worse version of dedicated MT. It may involve different

behaviours, including paraphrase, elaboration, restructuring, and register adjustment.

This distinction matters for document-level and discourse-sensitive evaluation. Sentence-level metrics may fail to capture coherence, reference tracking, discourse relations, and information flow (Läubli, Sennrich, Volk 2018, 4794-5; Maruf, Saleh, Haffari 2021, 2-6). Even when a translation is locally accurate, it may still preserve source sentence boundaries too rigidly, disrupt target-language rhythm, or alter discourse organisation. These issues are especially important when machine output is used for post-editing, where translators must revise concrete texts rather than abstract quality scores.

The present study therefore compares not only a dedicated MT system with LLM-based systems, but also two LLM-based systems with each other. This design avoids treating LLM translation as a homogeneous category. DeepSeek and ArkLLM may share some broad features as LLM-based systems, but they may differ in expansion, segmentation, connective marking, and sensitivity to register. Such differences are important for understanding how AI systems construct translated Chinese discourse.

2.3 Register, Cohesion, and English-to-Chinese Translation

Register provides the interpretive framework for the study. In register analysis, linguistic patterns are shaped by communicative situation, discourse function, medium, and audience (Biber 1988, 28-46; Biber, Conrad 2009, 31-49). Texts belonging to different registers do not simply discuss different topics; they organise information differently. Public speech, legal-diplomatic institutional writing, and journalistic news differ in sentence length, formulaicity, repetition, discourse markers, information density, and preferred modes of textual cohesion.

This perspective is necessary for AI translation analysis because the same textual feature may carry different meanings across registers. A long sentence in an institutional resolution may reflect legal-diplomatic convention, while a long sentence in public speech may weaken oral rhythm. A high density of connectives may support audience guidance in a speech, but may sound overly explicit in a news report. A shorter output may indicate efficient compression in one context and loss of rhetorical nuance in another. Length ratio, sentence length, and connective density must therefore be interpreted as register-sensitive stylistic features rather than universal indicators of quality.

English and Chinese may differ in their preferred ways of organising information and expressing discourse relations. English often relies on explicit syntactic links, subordination, and connective markers,

whereas Chinese can allow relations to be inferred through clause order, punctuation, topic continuity, and contextual logic. Cohesive devices contribute to textual connectedness (Halliday, Hasan 1976, 1-9), and shifts in cohesion and coherence have long been recognised as important in translation (Blum-Kulka 1986, 17-21). In English-to-Chinese AI translation, connective density can therefore reveal whether systems tend to preserve explicit relations, reduce them, or reformulate them through more implicit target-language resources.

Sentence segmentation is another central issue. English source segments, especially in legal-diplomatic and news texts, often contain dense syntactic structures with embedded modifiers, reported speech, or multiple events. Chinese translation may preserve the source boundary, split the information into shorter units, or reorganise clauses according to target-language readability. In public speech, segmentation also carries rhetorical and performative meaning. A system that merges short discourse units into a long sentence may preserve propositional content but alter the rhythm and audience orientation of the source.

The literature therefore suggests that AI translation should be studied not only as a quality-ranking task, but also as textual production under register constraints. This study addresses that issue by comparing TencentMT, DeepSeek, and ArkLLM on the same aligned English source segments across three registers. The analysis examines how systems differ in length ratio, mean sentence length, and connective density, and how these differences are shaped by public speech, legal-diplomatic institutional discourse, and journalistic news.

3 Data and Methods

3.1 Corpora and Translation Systems

The study uses an aligned English Chinese dataset drawn from three corpora representing distinct registers: TED2020 for public speech, UNPC for legal-diplomatic institutional discourse, and WMT20_News for journalistic news. The datasets were obtained from publicly available parallel-corpus resources: TED2020 through OPUS and related TED subtitle data (Tiedemann 2012, 2214-18; Reimers, Gurevych 2020, 4512-14), UNPC through the United Nations Parallel Corpus (Ziems, Junczys-Dowmunt, Pouliquen 2016, 3530-4), and WMT20_News through the WMT20 news translation setting (Barrault et al. 2020, 1-5). The selection was motivated by register diversity rather than by corpus size equivalence. Because the analysis compares systems on an aligned subset rather than on

raw corpus frequencies, differences in the original corpus sizes were controlled through the segment-level comparative design.

The analysis is conducted at the source-segment level. Source segments were retained according to the segmentation available in the corpus files, and no additional reconstruction of paragraph boundaries was performed. Each source segment was submitted independently for translation.

Table 1 Corpus distribution of the final aligned dataset

Corpus	Register	Source segments	System outputs
TED2020	Public speech	169	507
UNPC	Legal-diplomatic institutional texts	57	171
WMT20_News	Journalistic news	63	189
Total	—	289	867

The final aligned dataset contains 289 English source segments and 867 Chinese system outputs. The unit of comparison is the aligned source segment: each segment serves as a repeated observation across TencentMT, DeepSeek, and ArkLLM. Each source segment was translated by all three systems, resulting in three comparable outputs for each source item. The dataset was therefore balanced at the system level, although the number of segments differed across corpora. This design ensured that system comparisons were made on the same source material rather than on partially overlapping subsets.

Three systems were compared: TencentMT, DeepSeek, and ArkLLM. TencentMT was treated as a dedicated machine translation system, while DeepSeek and ArkLLM were treated as LLM-based systems. Only source segments for which all three systems produced valid Chinese translations were retained. This criterion was necessary because some LLM outputs were incomplete or invalid and could not be included in a balanced comparison.

The human Chinese reference translations were the reference texts distributed with the respective corpora or shared-task datasets. Their institutional status differs across resources; for example, UNPC contains official United Nations translations, while TED2020 and WMT20_News provide reference translations associated with their respective corpus resources. In this study, these references were used to calculate relative length. They were not treated as the only possible correct translations or as a complete standard of translation quality. Their function was to provide a stable comparison point for measuring compression and expansion in system outputs.

For the LLM-based systems, each source segment was processed independently in a fresh interaction context, with no previous outputs or source segments included. The translation instruction required a

complete Simplified Chinese translation only, without commentary or explanation; stage directions in parentheses were translated when present. This setting preserved the segment-aligned design and made the outputs directly comparable across systems. The complete prompt and preprocessing procedure are provided in Appendix A.

3.2 Textual Variables

The analysis focuses on three textual variables: length ratio, mean sentence length, and connective density. These variables were selected as indicators of translation style rather than direct measures of translation quality. Together, they capture textual volume, sentence-level packaging, and explicit discourse cohesion.

Length ratio measures the length of a system output relative to the corresponding human reference translation:

$$\text{length ratio} = \text{system output token-like units} / \text{reference translation token-like units}$$

A value below 1 indicates that the system output is shorter than the reference, while a value above 1 indicates that it is longer. This variable was used to identify compression and expansion tendencies. Because ratio values were right-skewed, length ratio was log-transformed in the statistical model.

Mean sentence length measures how much information a system packages into each Chinese sentence unit:

$$\text{mean sentence length} = \text{translation token-like units} / \text{translation sentence count}$$

This variable was not interpreted as a direct measure of quality. A long sentence may reflect formulaic legal-diplomatic convention in UNPC, but it may indicate weakened oral rhythm in public speech or possible source-boundary preservation in news translation. Mean sentence length was calculated within each system output; no one-to-one sentence correspondence between source, output, and reference sentences was assumed.

Connective density measures the number of occurrences of predefined Chinese connective expressions per 1,000 token-like units:

$$\text{connective density} = (\text{number of connective occurrences} / \text{number of translation token-like units}) \times 1,000$$

This variable was used as an indicator of explicit cohesive marking. A

higher connective density does not necessarily indicate better cohesion. It signals a stronger tendency to mark discourse relations overtly. A lower connective density may reflect more implicit target-language cohesion, but it may also indicate under-marked logical relations in local contexts. The connective list is provided in Appendix A.

3.3 Statistical Modelling and Qualitative Analysis

The dataset has a repeated-measures structure because each English source segment appears three times, once for each system output. These observations cannot be treated as independent: the same source segment constrains length, sentence organisation, lexical choices, and connective use across all three outputs. Mixed-effects models were therefore fitted with source segment as a random intercept, following standard practice for clustered linguistic data (Baayen, Davidson, Bates 2008, 390-5; Bates et al. 2015, 1-6).

Models were fitted as linear mixed-effects models using restricted maximum likelihood estimation. Fixed-effect significance was assessed using asymptotic Wald z tests. For the length-ratio model, the dependent variable was log-transformed to reduce skewness. ArkLLM and TED2020 were used as reference levels.

The main models were specified as follows:

$$\log(\text{length_ratio}) \sim \text{system} + \text{corpus} + (1 \mid \text{sample_segment_id})$$

$$\text{translation_mean_sentence_len} \sim \text{system} + \text{corpus} + (1 \mid \text{sample_segment_id})$$

$$\text{connectives_per_1000} \sim \text{system} + \text{corpus} + (1 \mid \text{sample_segment_id})$$

In the model formulas, `translation_mean_sentence_len` and `connectives_per_1000` correspond to mean sentence length and connective density, respectively.

Under this coding, system coefficients estimate differences between DeepSeek or TencentMT and ArkLLM, while corpus coefficients estimate differences between UNPC or WMT20_News and TED2020.

The choice of `sample_segment_id` as the random intercept reflects the structure of the data. The strongest dependency comes from the fact that the same source segment was translated by all three systems. Model diagnostics also showed high intraclass correlation coefficients at the segment level, confirming that source-segment variation accounted for a substantial share of the variance in the dependent variables. Full model outputs, including random-effect variance, ICC values, and interaction models, are provided in Appendix B.

Interaction models with system $\times$ corpus terms were fitted to examine whether system effects varied across registers. Under the same Wald-test framework, the interaction terms did not show statistically reliable effects. The main-effects models were therefore retained as a more parsimonious summary of the observed system and register tendencies. The models were used to identify systematic textual tendencies rather than to rank systems by translation quality.

Quantitative modelling was complemented with qualitative examples. The examples were selected to illustrate compression, expansion, segmentation, formulaic convergence, connective use, and source-boundary preservation. They were not treated as isolated errors, but as cases that made the statistical patterns interpretable. Additional examples are provided in Appendix C.

4 Results

4.1 Quantitative Results

The mixed-effects models show systematic differences among the three systems after controlling for register and source-segment clustering. For length ratio, ArkLLM produced the longest translations overall. DeepSeek was significantly shorter than ArkLLM, with an estimated coefficient of -0.027 , corresponding to an approximate 2.7% decrease in relative length. TencentMT showed a stronger shortening effect, with an estimated coefficient of -0.080 , corresponding to an approximate 7.7% decrease compared with ArkLLM.

Table 2 Mixed-effects model for log-transformed length ratio

Predictor	Estimate	SE	z	p
Intercept: ArkLLM / TED2020	-0.062	0.025	-2.51	0.012
DeepSeek vs ArkLLM	-0.027	0.009	-3.20	0.001
TencentMT vs ArkLLM	-0.080	0.009	-9.45	<0.001
UNPC vs TED2020	0.142	0.049	2.92	0.003
WMT20_News vs TED2020	0.074	0.047	1.57	0.116

UNPC translations were significantly longer than TED2020 translations, while WMT20_News did not differ significantly from TED2020. The source-segment ICC was 0.903, indicating that length ratio was strongly conditioned by the source segment.

For mean sentence length, TencentMT produced significantly shorter sentence units than ArkLLM. DeepSeek did not differ significantly from ArkLLM. Register effects were substantial: UNPC

and WMT20_News both had significantly longer sentence units than TED2020, with the effect especially strong for UNPC.

Table 3 Mixed-effects model for mean sentence length

Predictor	Estimate	SE	z	p
Intercept: ArkLLM / TED2020	25.70	2.99	8.59	<0.001
DeepSeek vs ArkLLM	-0.55	0.56	-0.97	0.333
TencentMT vs ArkLLM	-2.34	0.56	-4.14	<0.001
UNPC vs TED2020	39.45	5.92	6.66	<0.001
WMT20_News vs TED2020	13.60	5.71	2.38	0.017

The source-segment ICC was 0.970, showing that sentence length was highly dependent on the syntactic and informational structure of the source segment. Nevertheless, the significant TencentMT effect indicates that system-specific sentence packaging remained observable after source-segment variation was controlled.

For connective density, TencentMT used significantly more explicit connectives than ArkLLM, with an estimated increase of 1.51 connective occurrences per 1,000 token-like units. DeepSeek again did not differ significantly from ArkLLM. TED2020 had higher connective density than both UNPC and WMT20_News, reflecting the discourse character of public speech.

Table 4 Mixed-effects model for connective density

Predictor	Estimate	SE	z	p
Intercept: ArkLLM / TED2020	6.40	1.00	6.42	<0.001
DeepSeek vs ArkLLM	0.55	0.60	0.91	0.361
TencentMT vs ArkLLM	1.51	0.60	2.50	0.012
UNPC vs TED2020	-4.06	1.86	-2.18	0.029
WMT20_News vs TED2020	-3.80	1.79	-2.12	0.034

The ICC for connective density was 0.711, lower than those for length ratio and mean sentence length. This suggests that connective use was less tightly determined by the source segment and left more room for system-level variation.

Taken together, the three models indicate that ArkLLM showed the most expansive output profile in terms of length, TencentMT showed the most compact profile in terms of length and sentence packaging, and DeepSeek showed an intermediate pattern in output length. TencentMT was not only shorter in overall length; it also produced shorter sentence units while using more explicit connectives. This combination points to a compact but overtly marked cohesive profile.

4.2 System Profiles across Registers

The quantitative results suggest three system profiles. ArkLLM produced longer translations and longer sentence units than TencentMT, while using fewer explicit connectives. Its output profile appears more expansive in terms of length and less dependent on the counted connective items. DeepSeek showed an intermediate pattern in output length, while remaining closer to ArkLLM in sentence organisation and connective density. TencentMT displayed the most compact output profile in terms of length and sentence packaging, while showing higher connective density.

Register shaped how these profiles were realised. UNPC appeared to constrain the systems toward long formulaic structures, reducing visible divergence in sentence organisation. TED2020 allowed greater variation in compression, segmentation, and rhetorical organisation. The qualitative examples from WMT20_News provided evidence of system-specific expansion and possible source-boundary preservation.

The system $\times$ corpus interaction terms were not statistically significant, suggesting that the main direction of system differences was broadly stable across registers. Their textual interpretation, however, remains register-dependent. A long sentence in UNPC may reflect legal-diplomatic convention, while a long sentence in public speech may weaken rhetorical rhythm. Higher connective density may support audience guidance in speech, but may sound more overtly marked in news translation.

4.3 Qualitative Examples

Representative examples were examined to clarify how the quantitative patterns are realised in actual translation. Full source texts, human references, and system outputs are provided in Appendix C.

In the TED2020 example, the source segment is a song-lyric passage composed of short line-like units. This example represents an extreme case of segmentation divergence rather than the average TED2020 pattern. TencentMT produces a highly compressed and partly distorted output. DeepSeek preserves most semantic material but merges the verse-like units into one long sentence. ArkLLM produces a similarly long output but preserves line-by-line segmentation more clearly. This example illustrates how segmentation differences may affect rhetorical organisation in performance-oriented discourse. The contrast is not only about length, but also about rhythm, rhetorical structure, and the reconstruction of spoken or performed discourse.

The UNPC example presents a different pattern. The source is a formal UN resolution clause. All three systems retain a long integrated sentence structure and follow the formulaic logic of legal-diplomatic institutional discourse. The differences are mainly lexical or formulaic rather than structural. This convergence is consistent with the quantitative finding that UNPC was associated with longer, more formulaic sentence organisation and narrower visible system variation.

The selected WMT20_News examples suggest both system-specific expansion and possible source-boundary preservation. In one news segment concerning dog control legislation, ArkLLM produces the longest translation by expanding participant reference and inserting the cohesive expression “此举”. TencentMT remains more compact and uses the awkward expression “狗控制法”, while DeepSeek occupies an intermediate position. Another news example shows that the human reference splits a dense English lead into two Chinese sentences, whereas TencentMT, DeepSeek, and ArkLLM all preserve it as one sentence. This suggests that even LLM-based systems may follow segment-level input boundaries when translating English news into Chinese.

These examples illustrate the statistical results. Public speech reveals differences in compression, integration, and rhetorical segmentation. Legal-diplomatic institutional discourse appears to constrain system variation and produce formulaic convergence. The selected news examples suggest both informational expansion and possible source-boundary preservation.

5 Discussion

The findings show that AI translation systems differ not only in quality-related performance but also in the textual profiles they produce. ArkLLM, DeepSeek, and TencentMT translated the same English source segments into Chinese, yet their outputs differed systematically in length control, sentence packaging, and explicit cohesive marking. These differences indicate that each system constructs translated Chinese discourse through relatively stable textual strategies.

This supports a shift from quality ranking to translation style analysis. Automatic metrics remain useful for estimating adequacy, fluency, or similarity to reference translations, but they do not fully explain how a translation is organised as discourse. A semantically acceptable output may still change rhetorical rhythm, compress information, preserve source sentence boundaries, or increase explicit connective marking. For translation studies and post-editing practice, these features matter because translations are used as texts in specific communicative settings.

TencentMT provides an illustrative case of the relationship between output compression and explicit cohesive marking. It produced shorter translations and shorter sentence units than ArkLLM, but used more explicit connectives. This suggests a compact but overtly marked cohesive profile. ArkLLM, by contrast, produced longer outputs with lower connective density, suggesting a more expansive profile that relies less on the counted connective items and more on fuller information rendering or structural continuity. DeepSeek was slightly more compact than ArkLLM in length ratio, but closer to ArkLLM in mean sentence length and connective density.

The comparison between DeepSeek and ArkLLM also shows that LLM-based translation should not be treated as a homogeneous category. Existing studies have shown that LLMs can perform competitively in translation tasks, but the present findings suggest that model-specific textual behaviour also needs to be examined (Jiao et al. 2023, 1-2; Hendy et al. 2023, 1-2). Different LLMs may vary in how expansive, source-bound, cohesive, or register-sensitive their translations are.

Register mediates these system tendencies. The same surface feature can carry different meanings across registers. In UNPC, long sentence units are consistent with legal-diplomatic institutional convention, where formulaic expression and syntactic integration are expected. In TED2020, a long sentence may weaken oral rhythm and reduce the rhetorical force of short discourse units. In news translation, long sentence units may indicate excessive preservation of English source boundaries, especially when a dense English lead would be more readable in Chinese if split into shorter sentences.

The connective-density results further show why discourse-level indicators require cautious interpretation. A higher density of connectives does not automatically mean better cohesion, and a lower density does not automatically mean weaker coherence. In English-to-Chinese translation, discourse relations may be expressed explicitly through connectives, implicitly through clause order and punctuation, or through broader information structure. TencentMT's higher connective density may support explicit local cohesion, but it may also produce a more source-oriented or overtly marked style in some contexts. ArkLLM's lower connective density may reflect more implicit Chinese discourse organisation, but in local cases it may also risk under-marking logical relations.

The selected news examples point to a possible shared tendency across systems: source-boundary preservation. Even the LLM-based systems sometimes retained the English source segment as a single Chinese sentence when the human reference split it into more than one sentence. This suggests that segment-aligned input may influence output organisation. The finding echoes calls for document-level and discourse-sensitive MT evaluation, where coherence, segmentation,

and information flow cannot be fully captured by sentence-level comparison (Läubli, Sennrich, Volk 2018, 4794-5; Maruf, Saleh, Haffari 2021, 2-6).

These findings have practical implications. Translators selecting AI translation tools may need to consider not only accuracy but also the textual behaviour of each system. A compact and explicitly cohesive output may be useful for some informational tasks, but less suitable for preserving rhetorical rhythm. An expansive LLM output may improve readability or clarify participant relations, but it may require editing when concision or strict institutional formulation is needed. AI translation style therefore emerges from the interaction of system behaviour, source segmentation, language-pair features, and register constraints.

6 Conclusions

This study examined translation style variation in English-to-Chinese AI translation across public speech, legal-diplomatic institutional discourse, and journalistic news. It compared TencentMT, DeepSeek, and ArkLLM using length ratio, mean sentence length, and connective density as indicators of textual volume, sentence-level organisation, and explicit discourse cohesion. Mixed-effects models were used to control for source-segment clustering, and qualitative examples were used to interpret the statistical patterns.

The results show that the three systems produced distinct textual profiles with respect to the selected indicators. ArkLLM generated the longest outputs overall, indicating a more expansive output profile in terms of length. TencentMT produced significantly shorter translations and shorter sentence units, but used more explicit connectives than ArkLLM. DeepSeek showed an intermediate pattern in output length, while remaining closer to ArkLLM in sentence organisation and connective density. These patterns indicate that AI translation systems do not merely differ in quality; they construct Chinese translations through different strategies of compression, expansion, segmentation, and cohesive marking.

Register appeared to shape these tendencies. Public speech allowed visible differences in rhetorical segmentation and discourse rhythm. Legal-diplomatic institutional discourse appeared to constrain systems toward long formulaic structures. The qualitative news examples provided evidence of system-specific expansion and possible source-boundary preservation. AI translation evaluation should therefore not treat texts as register-neutral. The same textual feature may be appropriate in one register and problematic in another.

The study contributes to AI translation research by shifting attention from system ranking to register-sensitive translation style. It also shows that LLM-based translation is not a uniform category. Different LLM-based systems may produce different textual profiles, and these differences matter for translation practice and post-editing.

The study has limitations. The final aligned subset contains 289 source segments, which ensures comparability across systems but limits the scale of the analysis. Only three systems were examined, so the findings should not be generalised to all MT or LLM-based systems. The selected variables capture important textual features, but they do not provide a full account of translation quality, terminology accuracy, pragmatic appropriateness, or reader reception.

Future research could expand the corpus size, include more systems and language pairs, and compare sentence-level prompting with document-level prompting. It would also be useful to combine textual indicators with professional translator judgements in order to examine how observed stylistic differences affect post-editing effort and target-register acceptability.

The central implication is that AI translation systems may produce systematically different textual profiles. Dedicated MT and LLM-based systems produce different forms of translated Chinese discourse, and these forms are shaped by register. A discourse-oriented and register-sensitive approach is therefore necessary for understanding how current AI translation systems behave in real translation contexts.

Appendix A. Translation Prompt and Preprocessing

A.1 Translation prompt

For the LLM-based systems, each English source segment was translated independently using the following instruction. The original Chinese prompt and an English gloss are provided for reproducibility.

Original prompt: 请将以下英文完整翻译为简体中文。只输出译文，不要解释。请翻译所有英文内容；括号中的舞台提示也要翻译为中文。

English gloss: Please translate the following English text completely into Simplified Chinese. Output only the translation and do not provide explanations. Translate all English content; stage directions in parentheses should also be translated into Chinese.

Each source segment was submitted separately. No previous source segments, human reference translations, or earlier system outputs were included in the prompt context. This procedure was adopted to keep all outputs comparable at the aligned segment level.

A.2 Inclusion and Exclusion Criteria

Only source segments for which all three systems produced valid Chinese outputs were retained in the final aligned subset. Outputs were excluded when they were empty, incomplete, or consisted mainly of non-translation or meta-commentary. Occasional untranslated fragments within otherwise identifiable translation outputs were retained as part of the system output.

The final dataset contained 289 source segments and 867 system outputs.

A.3 Token-Like Units

Chinese translation length was measured in token-like units. The same counting procedure was applied to human reference translations and all system outputs. The purpose of token-like counting was not to establish a theoretically definitive unit of Chinese segmentation, but to provide a consistent basis for comparing relative length, sentence length, and connective density across systems.

A.4 Sentence Segmentation

Chinese sentence units were identified using major sentence-final punctuation marks, including 。, !, ?, and equivalent full-stop punctuation where applicable. Short rhetorical units separated by line breaks or semicolons were checked manually in qualitative examples when segmentation was relevant to interpretation.

A.5 Connective List

Connective density was calculated using a predefined list of Chinese connective expressions. The list included common items expressing additive, contrastive, causal, conditional, concessive, sequential, and summarising relations.

Relation type	Examples
Additive	而且, 并且, 此外, 另外, 同时
Contrastive	但是, 然而, 不过, 相反
Causal	因此, 所以, 因而, 由于, 因为
Conditional	如果, 只要, 除非
Concessive	尽管, 虽然, 即使
Sequential	然后, 随后, 接着, 首先, 最后
Summarising	总之, 综上, 由此可见

The connective list was used consistently across corpora and systems. Connective density was interpreted as a measure of overt cohesive marking rather than as a direct indicator of translation quality.

Appendix B. Full Model Outputs

The following tables report REML-estimated mixed-effects models. p-values are based on asymptotic Wald z tests for fixed effects.

B.1 Length Ratio Model

Model specification:

$\log(\text{length_ratio}) \sim \text{system} + \text{corpus} + (1 \mid \text{sample_segment_id})$

Reference levels: ArkLLM and TED2020.

Predictor	Estimate	SE	z	p
Intercept: ArkLLM / TED2020	-0.062	0.025	-2.51	0.012
DeepSeek vs ArkLLM	-0.027	0.009	-3.20	0.001
TencentMT vs ArkLLM	-0.080	0.009	-9.45	<0.001
UNPC vs TED2020	0.142	0.049	2.92	0.003
WMT20_News vs TED2020	0.074	0.047	1.57	0.116

Random-effect variance for sample_segment_id: 0.0971

Residual variance: 0.0105

ICC: 0.903

B.2 Mean Sentence Length Model

Model specification:

translation_mean_sentence_len~system+corpus+(1|sample_segment_id)

Reference levels: ArkLLM and TED2020.

Predictor	Estimate	SE	z	p
Intercept: ArkLLM / TED2020	25.70	2.99	8.59	<0.001
DeepSeek vs ArkLLM	-0.55	0.56	-0.97	0.333
TencentMT vs ArkLLM	-2.34	0.56	-4.14	<0.001
UNPC vs TED2020	39.45	5.92	6.66	<0.001
WMT20_News vs TED2020	13.60	5.71	2.38	0.017

ICC: 0.970

B.3 Connective Density Model

Model specification:

connectives_per_1000 ~ system + corpus + (1 | sample_segment_id)

Reference levels: ArkLLM and TED2020.

Predictor	Estimate	SE	z	p
Intercept: ArkLLM / TED2020	6.40	1.00	6.42	<0.001
DeepSeek vs ArkLLM	0.55	0.60	0.91	0.361
TencentMT vs ArkLLM	1.51	0.60	2.50	0.012
UNPC vs TED2020	-4.06	1.86	-2.18	0.029

Predictor	Estimate	SE	z	p
WMT20_News vs TED2020	-3.80	1.79	-2.12	0.034

ICC: 0.711

B.4 Interaction Models

Interaction models with system × corpus terms were fitted for all three dependent variables. Under the same Wald-test framework, the interaction terms did not show statistically reliable effects. The main-effects models were therefore retained for the main analysis because they provided a more parsimonious summary of the observed system and register tendencies.

Interaction model specification:

dependent variable ~ system * corpus + (1 | sample_segment_id)

The absence of significant interaction terms suggests that the main direction of system differences was broadly stable across TED2020, UNPC, and WMT20_News. The interpretation of these differences, however, remained register-dependent.

Appendix C. Qualitative Examples

Table C1 TED2020 example: rhetorical segmentation

Text type	Text
Source	(Music) ♪ Go ahead and congratulate yourself ♪ ♪ Give yourself a hand, the hand is your hand ♪ ♪ And the eye that eyes itself is your eye ♪ ♪ And the ear that hears itself is near ♪ ♪ ‘Cause it’s your ear, oh oh ♪ ♪ You’ve done the impossible now ♪ ♪ Took yourself apart ♪ ♪ You made yourself invulnerable ♪ ♪ No one can break your heart ♪ ♪ So you wear it out ♪ ♪ And you wring it out ♪ ♪ And you wear it out ♪ ♪ And you break it yourself ♪ ♪ Breaking your own, break it yourself ♪ ♪ Breaking your own, break it yourself ♪ ♪ Breaking your own ♪ (Applause) Thanks.
Human reference	(音乐) ♪ 前进吧, 祝贺自己吧 ♪ ♪ 帮自己一把, 用自己的手 ♪ ♪ 眼睛, 用自己的眼睛 ♪ ♪ 耳朵听到它自己就在附近 ♪ ♪ 因为它是你的耳朵, 哦 哦 ♪ ♪ 你现在做到了本不可能的事 ♪ ♪ 把自己拆开 ♪ ♪ 你让自己不再受伤害 ♪ ♪ 没人能让你伤心 ♪ ♪ 你把它耗尽 ♪ ♪ 你把它扭干 ♪ ♪ 你把它耗尽 ♪ ♪ 你自己把它击碎 ♪ ♪ 突破自己, 自己击碎它 ♪ ♪ 突破自己, 自己击碎它 ♪ ♪ 突破自己 ♪ (掌声) 谢谢。
TencentMT	(音乐) ~哦哦~~ out BID你把它磨坏了吗? 你自己打破它了吗? 打破你自己的, 你自己打破它吗? 打破你自己的, 你自己打破它吗? (掌声) 谢谢。
DeepSeek	(音乐) ♪ 来吧, 给自己一点嘉许 ♪ ♪ 为自己鼓个掌, 那手就是你的手 ♪ ♪ 那只注视自己的眼睛就是你的眼睛 ♪ ♪ 那只倾听自己的耳朵就近在咫尺 ♪ ♪ 因为那是你的耳朵, 哦哦 ♪ ♪ 你已完成了不可能的事 ♪ ♪ 将自己拆解开来 ♪ ♪ 你让自己变得无懈可击 ♪ ♪ 没有人能伤透你的心 ♪ ♪ 所以你把它耗尽 ♪ ♪ 你把它拧干 ♪ ♪ 你把它耗尽 ♪ ♪ 你亲手将它击碎 ♪ ♪ 击碎你自己的, 亲手击碎它 ♪ ♪ 击碎你自己的, 亲手击碎它 ♪ ♪ 击碎你自己的 ♪ (掌声) 谢谢。
ArkLLM	(音乐声) ♪ 尽管恭喜你自己吧 ♪ ♪ 给自己鼓个掌, 这手是你的手 ♪ ♪ 凝视着自身的眼睛, 是你的眼睛 ♪ ♪ 聆听着自身的耳朵, 近在咫尺 ♪ ♪ 因为那是你的耳朵, 哦哦 ♪ ♪ 你如今做到了不可能的事 ♪ ♪ 将自己彻底拆解 ♪ ♪ 你让自己变得无懈可击 ♪ ♪ 没人能伤你的心 ♪ ♪ 于是你把它耗到枯竭 ♪ ♪ 把它拧到干透 ♪ ♪ 你把它耗到枯竭 ♪ ♪ 亲手将它打碎 ♪ ♪ 打碎你自己的, 亲手打碎 ♪ ♪ 打碎你自己的, 亲手打碎 ♪ ♪ 打碎你自己的 ♪ (掌声) 谢谢。
Comment	TencentMT is highly compressed and partially distorted. DeepSeek preserves semantic content but merges the verse-like units into one long sentence. ArkLLM preserves more line-level segmentation and therefore maintains a more speech-like or performance-oriented structure.

Table C2 UNPC Example: Legal-diplomatic Institutional Formulaicity

Text type	Text
Source	5. Requests the Secretary-General to bring the present resolution to the attention of Member States, organs, bodies and other components of the United Nations system, and intergovernmental and non-governmental organizations and to disseminate it on the widest possible basis;
Human reference	5. 请秘书长提请会员国、联合国系统各机关、机构和其他部门以及政府间组织和非政府组织注意本决议，并尽可能广泛地予以分发；
TencentMT	5. 请秘书长提请会员国、联合国系统机关、机构和其他部门以及政府间组织和非政府组织注意本决议，并尽可能广泛地传播；
DeepSeek	5. 请秘书长将本决议提请各会员国、联合国系统各机关、机构及其他组成部分以及政府间组织和非政府组织注意，并尽可能广泛地予以分发；
ArkLLM	5. 请秘书长提请会员国、联合国系统各机关、机构和其他组成部分以及政府间组织和非政府组织注意本决议，并尽可能广泛地分发本决议；
Comment	All systems retain the long formal structure expected in legal-diplomatic institutional discourse. The differences are mainly lexical or formulaic: TencentMT uses “传播”, DeepSeek adds “各” before “会员国”, and ArkLLM omits “予以” while repeating “本决议”.

Table C3 WMT20_News Example: Informational Expansion

Text type	Text
Source	The Scottish Parliament’s Post-Legislative Scrutiny Committee has been hearing evidence from victims and parents of children who have been attacked or killed by dogs, prompting the Government to review dog control laws.
Human reference	苏格兰议会的立法后监察委员一直在听取被狗袭击或，害的受害者和儿童父母的证词，敦促政府审查养狗法。
TencentMT	苏格兰议会立法后审查委员会一直在听取受害者和被狗袭击或，害的儿童父母的证据，促使政府审查狗控制法。
DeepSeek	苏格兰议会立法后审查委员会一直在听取遭狗攻击或致死儿童的受害者及家长的证词，促使政府重新审视犬只管控法律。
ArkLLM	苏格兰议会立法后审查委员会一直在听取遭犬只袭击或致死的受害者，以及子女遭犬只袭击或身亡的家长的证词，此举促使政府对犬只管控法律展开审查。

Text type	Text
Comment	ArkLLM expands participant reference through “子女遭犬只袭击或身亡的家长” and inserts “此举”, producing a more explicitly packaged news sentence. TencentMT remains more compact but uses the awkward expression “狗控制法”. DeepSeek is more idiomatic than TencentMT but less expansive than ArkLLM.

Table C4 WMT20_News Example: Source-Boundary Preservation

Text type	Text
Source	The 28-year-old woman was skydiving with a group of seven people around 2 p.m. Thursday, California Highway Patrol PIO Ruben Jones told CNN.
Human reference	这名 28 岁的女子与另外七个人在下午 2 点左右跳伞。星期四，加州公路巡邏队公共信息官鲁本·琼斯告诉美国有线电视新闻网。
TencentMT	加州公路巡邏队 PIO Ruben Jones 告诉 CNN，周四下午 2 点左右，这名 28 岁的女子与一行七人一起跳伞。
DeepSeek	加州公路巡警发言人鲁本·琼斯告诉美国有线电视新闻网，这名 28 岁的女子周四下午 2 点左右与一行七人一起进行跳伞活动。
ArkLLM	加州公路巡警局公共信息官鲁本·琼斯告诉美国有线电视新闻网，周四下午 2 点左右，这名 28 岁的女子正与一个 7 人小组一同跳伞。
Comment	The human reference splits the English lead into two Chinese sentences. All three systems preserve the source segment as a single Chinese sentence in this example, suggesting possible source-boundary preservation under segment-level input.

Bibliography

- Baayen, R.H.; Davidson, D.J.; Bates, D.M. (2008). "Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items". *Journal of Memory and Language*, 59(4), 390-412. <https://doi.org/10.1016/j.jml.2007.12.005>.
- Bahdanau, D.; Cho, K.; Bengio, Y. (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". *International Conference on Learning Representations (ICLR 2015)*.
- Baker, M. (1993). "Corpus Linguistics and Translation Studies: Implications and Applications". Baker, M.; Francis, G.; Tognini-Bonelli, E. (eds), *Text and Technology: In Honour of John Sinclair*. Amsterdam; Philadelphia: John Benjamins, 233-50. <https://doi.org/10.1075/z.64.15bak>.
- Barrault, L. et al. (2020). "Findings of the 2020 Conference on Machine Translation (WMT20)". *Proceedings of the Fifth Conference on Machine Translation. Online: Association for Computational Linguistics*, 1-55. <https://doi.org/10.18653/v1/2020.wmt-1.1>.
- Bates, D. et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". *Journal of Statistical Software*, 67(1), 1-48. <https://doi.org/10.18637/jss.v067.i01>
- Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CB09780511621024>.
- Biber, D.; Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CB09780511814358>.
- Blum-Kulka, S. (1986). "Shifts of Cohesion and Coherence in Translation". House, J.; Blum-Kulka, S. (eds), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*. Tübingen: Gunter Narr, 17-35.
- Chesterman, A. (2004). "Hypotheses about Translation Universals". Hansen, G.; Malmkjær, K.; Gile, D. (eds), *Claims, Changes and Challenges in Translation Studies: Selected Contributions from the EST Congress, Copenhagen 2001*. Amsterdam; Philadelphia: John Benjamins, 1-13. <https://doi.org/10.1075/btl.50.02che>.
- Halliday, M.A.K.; Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Hendy, A. et al. (2023). "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation". *arXiv preprint*, arXiv:2302.09210. <https://doi.org/10.48550/arXiv.2302.09210>.
- Jiao, W. et al. (2023). "Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine". *arXiv preprint*, arXiv:2301.08745. <https://doi.org/10.48550/arXiv.2301.08745>.
- Koehn, P. (2020). *Neural Machine Translation*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781108608480>.
- Laviosa, S. (2002). *Corpus-Based Translation Studies: Theory, Findings, Applications*. Amsterdam: Rodopi.
- Läubli, S.; Sennrich, R.; Volk, M. (2018). "Has Machine Translation Achieved Human Parity? A Case for Document-Level Evaluation". *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics*, 4791-6. <https://doi.org/10.18653/v1/D18-1512>.
- Maruf, S.; Saleh, F.; Haffari, G. (2021). "A Survey on Document-Level Neural Machine Translation: Methods and Evaluation". *ACM Computing Surveys*, 54(2), article 45, 1-36. <https://doi.org/10.1145/3441691>
- Papineni, K. et al. (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation". *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia: Association for Computational Linguistics, 311-18. <https://doi.org/10.3115/1073083.1073135>.

- Rei, R. et al. (2020). "COMET: A Neural Framework for MT Evaluation". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Online: Association for Computational Linguistics, 2685-702. <https://doi.org/10.18653/v1/2020.emnlp-main.213>.
- Reimers, N.; Gurevych, I. (2020). "Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Online: Association for Computational Linguistics, 4512-25. <https://doi.org/10.18653/v1/2020.emnlp-main.365>.
- Tiedemann, J. (2012). "Parallel Data, Tools and Interfaces in OPUS". *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: European Language Resources Association, 2214-18.
- Toral, A. (2019). "Post-edited: an Exacerbated Translationese". *Proceedings of Machine Translation Summit XVII: Research Track*. Dublin: European Association for Machine Translation, 273-81.
- Vanmassenhove, E.; Shterionov, D.; Gwilliam, M. (2021). "Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation". *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, 2203-13. <https://doi.org/10.18653/v1/2021.eacl-main.188>.
- Vaswani, A. et al. (2017). "Attention Is All You Need". *Advances in Neural Information Processing Systems*, 30, 5998-6008.
- Volansky, V.; Ordan, N.; Wintner, S. (2015). "On the Features of Translationese". *Digital Scholarship in the Humanities*, 30(1), 98-118. <https://doi.org/10.1093/lc/fqt031>.
- Zhang, T. et al. (2020). "BERTScore: Evaluating Text Generation with BERT". *International Conference on Learning Representations (ICLR 2020)*.
- Ziemski, M.; Juncys-Dowmunt, M.; Pouliquen, B. (2016). "The United Nations Parallel Corpus v1.0". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association, 3530-4.