Can *Peter Pan* Survive MT? A Stylometric Study of LLMs, NMTs, and HTs in Children's Literature Translation

Delu Kong[1,2] **and Lieve Macken**[2]

[1]School of Foreign Studies, Tongji University, Shanghai, 200092, China
[2]Language and Translation Technology Team, Ghent University, Ghent, 9000, Belgium
Correspondence: kongdelu2009@hotmail.com

Abstract

This study focuses on evaluating the performance of machine translations (MTs) compared to human translations (HTs) in English-to-Chinese children's literature translation (CLT) from a stylometric perspective. The research constructs a Peter Pan corpus, comprising 21 translations: 7 human translations (HTs), 7 large language model translations (LLMs), and 7 neural machine translation outputs (NMTs). The analysis employs a generic feature set (including lexical, syntactic, readability, and n-gram features) and a creative text translation (CTT-specific) feature set, which captures repetition, rhythm, translatability, and miscellaneous levels, yielding 447 linguistic features in total.

Using classification and clustering techniques in machine learning, we conduct a stylometric analysis of these translations. Results reveal that in generic features, HTs and MTs exhibit significant differences in conjunction word distributions and the ratio of 1-word-gram-一样, while NMTs and LLMs show significant variation in descriptive words usage and adverb ratios. Regarding CTT-specific features, LLMs outperform NMTs in distribution, aligning more closely with HTs in stylistic characteristics, demonstrating the potential of LLMs in CLT.

1 Introduction

With the advent of LLMs[1] , various evaluations have been made within user and researcher communities (Jiao et al., 2023; Castilho et al., 2023; Enis and Hopkins, 2024). A consensus has been reached: LLMs appear to be useful for handling texts that use highly formulaic language, such as contracts, technical documents, and web pages. However, when it comes to translating highly creative texts, such as literary works, it remains a highly challenging task (Kocmi et al., 2024).

Among all text types, literary texts serve as a formidable "bastion" (Toral and Way, 2014) that challenge the performance of MT engines. Compared to informational texts, literary texts place greater emphasis on aesthetic creation. They are characterized by intricate linguistic structures, rich metaphors, and deep cultural nuances, interwoven with the styles of different nations, cultures, eras, and even individual authors (Hadley et al., 2022). Therefore, literary translation not only facilitates the cross-linguistic transmission of ideas but also needs to recreate the artistic charm of the original work. As some scholars have pointed out, the innovative nature of literary texts "almost implies an inherent degree of resistance to automation" (Ruffo, 2022, p. 18), which also creates tricky challenges for MT developers seeking effective solutions.

As a significant branch of literary translation, children's literature translation (CLT for abbreviation) inspires limitless imagination in young readers worldwide. The choice to research CLT arises from its distinctive combination of literary artistry and creative expression: (1) it is crucial for cultural exchange and dissemination of children literature worldwide; (2) children's literature often features relatively unique expressions, making it ideal for assessing MT systems' semantic and cultural handling; (3) evaluating MT performance is vital to prevent mistranslations or hallucinations that could mislead young readers early on.

2 Related work

2.1 Stylometric investigation in CLT studies

Over the past decade, the field of CLT studies has expanded significantly. Scholars note that the

© 2025 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

[1]MT is used as a superordinate term encompassing both NMT and LLM translation, while NMT and LLM can also be treated as distinct categories based on the underlying generation engines.

first two decades of the 21st century have been a "blooming period" for research on translating children's books (Fornalczyk-Lipska, 2022). Once a marginal topic, it is now recognized that CLT works become a crucial part in young readers' literary experiences worldwide (Van Coillie and McMartin, 2020). Puurtinen (2003) was among the pioneering scholars who employed a corpus-based methodology to investigate CLT. Her study examines translation-specific features in Finnish children's literature, identifying a high frequency of non-finite structures, a relative absence of colloquial expressions as potential hallmarks of translated texts.

In recent years, the study of CLT has witnessed more refined outputs within the framework of stylometry or quantitative stylistics[2]. Čermáková (2018) uses corpus-based method to analyse the feature of repetition in CLT, showing that translators often find repetition uncomfortable and tend to compensate for it by using synonymy. For the Chinese-English language pair, Zhang et al. (2019) examine increased lexical-grammatical explicitness in Chinese CLT, and show a higher frequency of personal pronouns compared to non-translated texts, likely due to cross-linguistic influence from the source language. Also, Zhao et al. (2022) explore how narrative space is transferred in CLT from Chinese works to English translated versions, and show that selective appropriation (patterns of omission and addition designed to suppress, accentuate or elaborate particular aspects of a narrative) is the most used strategy.

The aforementioned studies have established a foundation for applying stylometric analysis to CLT. However, compared to general translation studies, where broad features such as Type-token-ratio (TTR), PoS-tags, and word frequencies are commonly analyzed (see Päpcke et al., 2022; Ding, 2024; Ploeger et al., 2024), CTT-features are more specifically utilized in CLT studies and are less frequently examined in broader translation research. CLT requires attention to more specific aspects, including repetition, metaphor, and rhythmic patterns. To address this, the present study incorporates both generic stylometric features and those specifically tailored to CLT, which are presented in Section 3.2.

[2]In this paper, stylometry is used interchangeably with quantitative stylistics, both referring to approaches that emphasize feature engineering and statistical analysis of textual style. In contrast, stylistics more broadly involves interpretive analysis of how language produces meaning, literary effects, and context-sensitive nuances.

2.2 MT in creative texts

MT's application in creative texts raises debates on its impact on the accuracy of cultural and creative expression. Even before the advent of NMT, Toral and Way (2014) raised the question of the usefulness of MT for literature. The very next year, Toral and Way (2015) demonstrated that a statistical machine translation system adapted to literary texts outperformed generic baseline systems, and could be further incorporated into the literary translator's workflow.

The NMT system advancement has sparked growing interest among scholars in applying MT to literary texts. Toral and Way (2018) evaluate NMT for literary texts, specifically novels, and find that NMT achieves an 11% relative improvement in BLEU scores compared with statistical MT. Yet if we look at literary translators' attitudes, the story is different. Studies show more experienced literary translators prefer translating from scratch, and might be resistant in using NMTs for more freedom (Moorkens et al., 2018; Way et al., 2023). From a reader perspective, Guerberof-Arenas and Toral (2024) observed that, in a case study of different translations, the MT version, compared to HT and PE, received the lowest ratings for narrative understanding and attentional focus, with structural and lexical issues contributing to a disrupted and confusing reading experience.

This raises the question: why are MT outputs often not well received by translators and readers? One major issue lies in stylistic limitations. Daems (2022) points out that MT systems struggle with capturing stylistic nuances, humour, and contextual or cultural subtleties—elements widely recognized as central to the nature of literary texts. Farrell (2018) and Taivalkoski-Shilov (2018) both emphasize that MT or post-edited output may result in homogenization and normalization in the target texts, trends that run counter to the diverse and creative nature of literary texts. Several studies investigate the stylistic difference between HTs and NMTs of literary works. For example, Jiang and Niu (2022) show that NMTs might be less coherent in discourse and less consistent in lexical choices. A more recent study also shows that NMT outputs can be distinguished from the HT counterparts based on sentence length distribution (Rybicki, 2025). These studies inspired us to follow a similar vein and examine whether LLM outputs differ from those of NMT systems.

2.3 Research gaps and questions

Summarizing the above studies, despite the solid foundation of CLT research, stylometric methods remain underexplored. Additionally, while existing studies of MT on literary works predominantly focus on general literary works, children's literature has received comparatively less attention. This gap underscores the need for research that accounts for the unique linguistic characteristics of CLT. Furthermore, the translation capabilities of LLM systems have gained significant interest; however, most research in this domain remains centered on NMT engines. The effectiveness of LLMs in CLT needs further investigation.

Based on these gaps, our study focuses on CLT among three translation groups, namely HTs, NMTs and LLMs. We construct a large dataset by incorporating a total of 21 English to Chinese translations of *Peter Pan*, and establish a more comprehensive feature set with a sub-category tailored for CTT. We adopt the research design of Daems et al. (2017) and Lynch and Vogel (2018), and collect experimental results and draw salient features, and then compare inter- and intra-group performance.

We address the following research questions:

- RQ1: Do MTs differ from HTs in CLT on generic textual features and CTT-specific features?
- RQ2: Do LLM outputs differ from NMT outputs in CLT on generic textual features and CTT-specific features?
- RQ3: How do salient features illustrate the differences among HTs, NMTs, and LLMs in CLT?

3 Methodology

3.1 Dataset

The corpus used in this study is based on J.M. Barrie's 1911 novel *Peter and Wendy*, the classic children's version of the Peter Pan story[3]. The character of Peter Pan was first introduced in Barrie's 1902 novel *The Little White Bird*. The 1911 novel has been widely cherished by children worldwide for its vivid imagination and the thrilling adventures of Peter Pan. The HTs are selected from **7** Chinese translations published by reputable publishers, spanning a wide time range (1929–2020) to ensure representativeness across different eras.

For MTs, the translated texts are categorized into two groups: NMTs and LLMs, with **7** engines selected for each category to produce a total of **14** MTed versions. In the NMT category, while some providers have reportedly begun integrating LLM technologies into their traditional NMT systems, e.g. DeepL[4], this study specifically aims to examine "pure" NMTs. To achieve this goal, models that explicitly state using an NMT engine are prioritized. For providers offering multiple MT engine versions, the NMT variant is selected whenever possible. If no version details were disclosed, the default engine is used. All NMTs are accessed via API. Details can be viewed in the online supplementary materials.

For LLMs, we select state-of-the-art commercial generic models, such as ChatGPT and Claude, as well as open-sourced DeepSeek, and MT-tailored LaraTranslate, and both open-sourced and MT-tailored Unbabel-TowerInstruct. Large-parameter deep reasoning models such as GPT-o1 and DeepSeek-R1, though strong in contextless multilingual translation, are not adopted in our framework due to their high inference cost, slower processing speed, and tendency to generate rambling outputs in Chinese (Chen et al., 2025). These factors significantly increase computational complexity and reduce overall efficiency in translation tasks. Consequently, we prioritize LLM models optimized for faster, more direct translation processes.

The MT process for LLMs involves prompt engineering, adhering to practices outlined in Andrew Ng's course[5] and the CRISPE framework[6]. This ensures a structured and standardized prompt design, consistently applied across all engines during translation. The complete prompt used in our experiments is provided in Appendix A for reference. Furthermore, we distinguish between MT engines developed by Chinese and non-Chinese enterprises, as this distinction may influence translation strategies and quality, particularly given that Chinese is the target language.

All texts underwent rigorous preprocessing, in-

[3]The 1911 novel uniquely includes the final chapter "When Wendy Grew Up", which does not appear in play or later abridged versions. This chapter is present in all Chinese translations used in this study. Although some translations do not explicitly state the source text, the inclusion of this chapter indicates that all were ultimately based on the 1911 novel.

[4]https://www.smartling.com/blog/how-accurate-is-deepl

[5]https://learn.deeplearning.ai/courses/chatgpt-prompt-eng

[6]https://github.com/mattnigh/ChatGPT3-Free-Prompt-List

Type	Translator	Abbr.	Engine	Acquisition	Token	Type	Sent.	Year
Source	-	-	-	E-book	47,978	5,334	3,334	2005
HTs	Liang	HTL	-	OCR	47,627	5,108	3,229	1929
	Yang & Gu	HYG	-	E-book	50,222	5,798	3,500	1991
	Ren	HTR	-	E-book	51,271	4,968	3,424	2006
	Ma	HTM	-	E-book	47,777	5,358	3,662	2011
	Sun	HSU	-	E-book	56,674	5,872	3,673	2017
	Shi	HSH	-	E-book	50,451	5,220	3,571	2018
	Huang	HTH	-	E-book	51,233	5,477	3,937	2020
NMTs	DeepL	NDL	Classic	API	46,731	4,980	2,850	2025 Feb.
	GoogleTrans	NGT	v2	API	46,891	4,887	3,268	
	MicrosoftTrans	NMS	-	API	46,854	4,711	3,150	
	AmazonTrans	NAZ	-	API	46,097	4,599	3,206	
	BaiduTrans*	NBD	-	API	45,697	4,514	3,205	
	YoudaoTrans*	NYD	-	API	47,817	4,690	3,393	
	NiuTrans*	NNT	-	API	46,296	4,541	3,197	
LLMs	ChatGPT	LCG	4o	Web	50,958	5,965	3,980	2025 Feb.
	Claude	LCL	3.5-sonnet	Web	46,503	5,355	3,426	
	Gemini	LGM	1.5-flash	API	48,174	5,489	3,474	
	Kimi*	LKM	v1	API	57,320	5,221	3,869	
	DeepSeek*	LDS	v3	OpenSource	45,951	4,865	3,421	
	TowerInstruct	LTI	7b-v0.2	OpenSource	46,097	4,869	3,124	
	LaraTrans	LLT	Creative	Web	48,576	4,709	3,176	
Total	-	-	-	-	**1,025,217**	**107,196**	**71,735**	-

Table 1: Overview of the datasets used in this study. Pure NMT version is selected if the provider offers version options (such as NDL and NGT). LaraTranslate provides a "creative text" style option but, despite claiming to use LLM technology, does not support custom prompts. Engines marked with * are developed by Chinese enterprises. The total count excludes the source text.

cluding cleaning, denoising, part-of-speech (PoS) tagging, and dependency (Dep) parsing. Since Chinese lacks explicit word boundaries, prior word segmentation is necessary. To achieve SOTA performance, we utilize the Language Technology Platform (LTP)[7], a comprehensive natural language processing toolkit (Che et al., 2021). LTP's deep learning model (Base2) is used for word segmentation, PoS tagging, and syntactic analysis, achieving reported accuracies of 99.18%, 98.69%, and 90.19% for these tasks, respectively.[8]

The final *Peter Pan* translation corpus exceeds over one million tokens. A detailed overview of the dataset is provided in Table 1.

3.2 Feature set

Based on the principles to construct feature set for stylometric analysis (Volansky et al., 2013)[9] and referring to previous research (see Huang and Liu, 2009; Lynch and Vogel, 2018; Toral, 2019; De Clercq et al., 2021), the following section presents the feature set applied in this study. A brief feature summary is in Table 3 (in Appendix B). All together, we've employed 447 features in this study. It should be noted that all features are represented as ratios or weighted measures to mitigate the influence of sample size differences and ensure comparability across texts.

3.2.1 Generic textual features

The generic textual features are designed to capture text characteristics from a holistic perspective. "Generic" means that these features are commonly used in other types of stylometric studies, not constrained on CLT studies. They are divided into four linguistic dimensions: lexical, syntactical, readability, and N-gram features.

Lexical features reflect word-level characteristics on lexical diversity, density, and richness, such as TTR. Additionally, PoS tag features, extracted using the LTP platform[10], capture the distribution of word classes, such as nouns and verbs.

Syntactic features focus on overall sentence structure and syntactic patterns, including average sentence length. Dependency tags are used

[7] https://github.com/HIT-SCIR/ltp

[8] https://github.com/HIT-SCIR/ltp/blob/main/README.md

[9] Although the quoted research focuses on translationese study, their proposed principles are highly relevant to construct a well-balanced and interpretable feature set for studying linguistic features of translated texts.

[10] https://ltp.ai/docs/appendix.html#id2

to further analyze syntactic roles, such as Mean Dependency Distance and the ratio of head and node.

Readability features include nine readability metrics proposed by Lei et al. (2024)[11], which evaluate lexical, syntactic, and semantic variability to assess a text's difficulty and comprehensibility for the target audience. Moreover, four concreteness features are included, measuring lexical concreteness based on the work of Xu and Li (2020).

N-gram features utilize N-word-grams and N-PoS-grams, where N ranges from 1 to 3, to capture locally constrained phrasal patterns. These features are extracted by comparing the target corpus with the LCMC reference corpus[12]. To maintain consistency, LCMC was re-tagged using the same LTP tools to ensure a uniform PoS-tag set.

3.2.2 CTT-specific features

CTT-specific features refer to characteristics specifically designed for creative text translation. In this study, we incorporate insights from previous Chinese research on CLT. These features are tailored to the linguistic and stylistic elements of CLT and are categorized into four subcategories:

Repetition features are a widely recognized in CLT, as a narrative strategy, often used to capture young readers' attention (Mastropierro, 2022). In this study, repetition features include AA-pattern (two-character repetition), AAA-pattern (three-character repetition), and ABAB-pattern structures.

Rhythm features examine phonetic patterns within the text, as previous research has identified distinct rhythmic patterns in CLT (Cooper, 1989). This study captures key rhythm features, such as rhyme proportion, vowel balance, and tonal alternation.

Translatibility features[13] assess language transfer and translation completeness between the source and the target texts through five key aspects: completeness, foreignness, code-switching, abbreviation, and untranslatable elements. Notably, completeness identifies untranslated English phrases longer than three words, while foreignness measures the ratio of English to Chinese characters.

The miscellaneous category includes five unique linguistic features that could not satisfactorily fall into the previous classifications but are frequently observed in CLT. These include the proportion of onomatopoeia, the usage of the Chinese-specific "-er suffix"[14], and the frequency of sentence-final particles, among others.

We put more detailed explanations and examples in the supplementary materials.

3.3 Algorithms

3.3.1 Feature selection

To streamline the experiment, mitigate feature noise, and enhance efficiency, we adopt a feature selection strategy based on chi-square (χ^2) ranking in both classification and clustering tasks. Features are prioritized according to their χ^2 values, with the top 30 being retained. If a given category contains fewer than 30 features, all available ones are preserved.

3.3.2 Classification experiment

The classification experiment follows a hierarchical structure based on different feature set levels. Initially, classification is performed separately for each feature sub-level. The experimental setup consists of the following comparison groups: (1) HTs vs. MTs, where MTs encompass both NMTs and LLMs; (2) HTs vs. NMTs and LLMs separately; (3) LLMs vs. NMTs; and (4) intra-group classification within the NMTs and LLMs categories.

To evaluate classification performance, five classifiers are employed: Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest. The average performance across these classifiers is reported. The SVM model is configured with a linear kernel, while the remaining classifiers use their default settings. Following the methodology outlined by Rahman et al. (2024), the effectiveness of the ensemble classifier is assessed using Accuracy (ACC) score. All classification tasks, except intra-group classifications within the NMTs and LLMs groups, are binary classification tasks.

[11] https://github.com/leileibama/AlphaReadabilityChinese

[12] https://www.lancaster.ac.uk/fass/projects/corpus/LCMC/

[13] We categorize translatability features under the CTT-specific level, considering that CLT is particularly sensitive to language shifts. This sensitivity is especially pronounced when the target audience is mainly children. The interweaving of English elements within Chinese text might influence their reading experience.

[14] The "-er suffix" in Chinese is a linguistic phenomenon where the suffix "儿" (ér) is added to the end of a word, often altering its pronunciation and adding a diminutive meaning. This is widely used in spoken and literary Chinese. For example, 花 (huā, "flower") becomes 花儿 (huār), and 鸟 (niǎo, "bird") becomes 鸟儿 (niǎor).

3.3.3 Clustering experiment

The clustering experiment employs the k-means algorithm to categorize data. Rather than predefining the number of clusters (k), it is determined based on performance evaluation. Euclidean distance serves as the similarity metric for clustering, and the feature set is derived from the Top-k features identified in the earlier analysis.

To assess clustering effectiveness, the Adjusted Rand Index (ARI) is used as the primary evaluation metric. ARI quantifies the alignment between the clustering results and ground truth labels while accounting for chance, providing an objective measure of clustering quality (Warrens and van der Hoef, 2022). Beyond numerical evaluation, interactive clustering visualizations are generated using Python's Plotly library.

In addition to k-means clustering, hierarchical clustering is introduced as a complementary approach. This method utilizes the stylo package in R, incorporating the top 100 most frequent words and Eder's delta (Eder, 2015) as key metrics. By integrating multiple clustering strategies, we aim to strengthen the robustness of our analysis and enhance the overall reliability of the experimental findings.

4 Results

4.1 Classification

Level	Sub-level	HTs -MTs	HTs -NMTs	HTs -LLMs	NMTs -LLMs
Generic textual features	Lexical	0.8673	0.9149	0.8526	0.7439
	Syntactical	0.8001	0.8443	0.7128	0.6752
	Readability	0.6645	0.7056	0.5538	0.6178
	N-gram	0.8674	0.8803	0.8486	0.7724
CTT-specific features	Repetition	0.6274	0.6800	0.5778	0.5436
	Rhythm	0.6650	0.5564	0.6444	0.5593
	Translatibility	0.6883	0.7930	0.6091	0.6067
	Miscellaneous	0.7378	0.7521	0.6393	0.6172
All	-	0.9149	0.9376	0.8896	0.7464

Table 2: Classification results across different feature levels and comparison groups. "Generic textual features" encompass general linguistic attributes, while "CTT-specific features" focus on aspects relevant to creative text translation. "All" represents the combined performance when all features are used together.

Table 2 presents the results across feature levels and groups. It should be noted that a high ACC score indicates a clear distinction between the examined categories, while a low ACC score suggests greater similarity. We generalize two tendencies:

First, from group comparison perspective, HTs consistently achieve the highest performance across different pairwise comparisons, with HTs-NMTs reaching the highest accuracy (0.9376) and HTs-MTs following closely (0.9149). When LLMs are involved, accuracy drops, as seen in HTs-LLMs (0.8896) and NMTs-LLMs (0.7464). This suggests that distinguishing between HTs and other MTed translations (NMTs and LLMs) is relatively easier, while differentiating within the same category (intra-group) is much harder, as in HTs (0.6785), NMTs (0.5965), and LLMs (0.5917).

Second, from feature categories perspective, among the generic textual features, lexical and N-gram features contribute the most to classification, with the highest ACC across different translation types (e.g., HTs-NMTs: 0.9149 and 0.8803, respectively). Readability exhibits the lowest performance, as in HTs-LLMs (0.5538). Regarding CTT-specific features, translatability and miscellaneous features contribute notably in differentiating HTs from other groups, while repetition and rhythm drop sharply in ACC scores. Overall, generic features perform better than CTT-specific features in classification, and using all features leads to the best performance.

Figure 1 (in Appendix D) presents a pairwise classification heatmap to provide a visualized plot and a fine-grained classifying result. It reveals three main results:

First, for HTs, the classification accuracy between HTs and other groups (NMTs and LLMs) is generally high. Within the HTs group, the highest ACC exceeds 0.90. The highest intra-group accuracy is observed in the HTL-HTH pair (0.98), also HTL achieves the highest average ACC (0.97) compared with all other samples. This may be due to the significant temporal gap, as HTL is the earliest translation among HTs. Conversely, the HSU-HTM and HSU-HTR pairs (both 0.76) exhibit the lowest accuracy, with HSU having the lowest average ACC (0.87). The greater similarities among these translations might be newer versions drawing references from previously published ones.

Second, for NMTs, ACC in distinguishing NMTs-LLMs is relatively lower than that of NMTs-HTs, meaning NMTs are much more similar to LLMs in style. The lowest inter-group ACC is observed in NGT-LDS (0.59). The score within the NMTs group is considerably lower than that within the HTs group. The highest intra-group accuracy is

NAZ-NYD (0.94), while the lowest is NBD-NGT (0.44). Also, in terms of average ACC, we found that NAZ (0.92) and NDL (0.9) achieve relatively higher compared with the rest NMT engines, while NGT (0.79) is the lowest.

Third, within the LLM group, the highest intra-group ACC is observed in LCG-LLT (0.99), and LKM-LLT (0.65) exhibits the lowest. Still, on average ACC, LCG (0.96) and LCL (0.93) have the relative highest score compared with other LLM engines, while LDS (0.83) has the lowest one. No significant differences are found between MT systems developed by Chinese companies and those by international companies.

4.2 Clustering

Figure 2 presents the clustering results using K-means (left) and hierarchical clustering (right). The K-means clustering, with an ARI score of 0.4873, demonstrates a clear separation between HTs and NMTs, as the two groups are positioned far apart. However, LLMs exhibit a more complex distribution, which cannot be clustered into a distinct group. Notably, three LLMs (LCG, LCL, LGM) cluster closer to HTs, suggesting that their translations share more similarities with human translations. Meanwhile, a subset of LLMs (LKM, LDS, LLT, LTI) aligns more closely with NMTs.

The hierarchical clustering (right) supports these findings, displaying relatively stable and well-separated clusters for HTs and NMTs. In contrast, LLMs show a more dispersed pattern, with samples integrating into both HT and NMT clusters. This reveals that LLMs exhibit heterogeneous translation characteristics, with some models leaning towards human-like translation styles and others resembling NMT outputs. Both clustering results reinforce observations drawn from previous classification experiments.

5 Discussion

5.1 Overview from generic features

This section discusses these differences from the perspective of generic features and provides an overview on their variance.

5.1.1 Ratio of conjunctions

From Table 4, we can see that conjunction words and N-grams contribute greatly to the separation of HTs and MTs. For the ratio of conjunction words (Figure 3), HTs differ significantly from MTs, with MTs exhibiting a higher ratio (ANOVA F = 91.10, $p < 0.0001$)[15]. Among conjunction words, the 1-word-gram-然后 (English: then) serves as a good example, since it is particularly prominent, and exhibits a similar trend on its over-usage in MT outputs. It means that MTs tend to rely more on explicit logical connectors, exhibiting a certain tendency toward "explicitation" (Zhang et al., 2019), whereas HTs demonstrate greater flexibility in expression and are not strictly bound by the logical transitions of the source text. However, no significant difference is observed between NMTs and LLMs in terms of conjunction (p = 0.06).

5.1.2 Ratio of 1-word-gram-一样

Another feature that distinguishes HTs and MTs apart is the 1-word-gram-一样 (same). This word frequently follows another word to form a Chinese phrase "像...一样 (same as...)" which conveys a simile meaning. As shown in Figure 4, MTs exhibit a significantly higher frequency of "一样" compared to HTs ($p < 0.0001$), indicating that MTs tend to produce more explicit comparative structures. Particularly NMTs use "像...一样" more frequently. Hence, MTs are more constrained by source text structures and produce similar patterns, leading to potential 'homogenisation' (Daems et al., 2024) in lexical terms with less variation in figurative expressions. But for LLMs, we see that LCG (ChatGPT) and LCL (Claude) use relatively less "像...一样". Given that the principle of fidelity to source texts should be generally maintained in translation, the reduction in "像...一样" likely reflects a shift in how figurative meaning is expressed in the LLM translations, rather than a loss of figurative content.

Drawing on actual concordance as an example (see Figure 8), we observe that in HTs, there are only two occurrences of "像...一样", while other variations, such as "像...似的" and "宛如", are also used creatively. In contrast, all seven NMT outputs employ the same fixed expression, whereas LLMs exhibit a mix, with three outputs using the same phrase.

5.1.3 Ratio of descriptive and adverbial words

Figure 5 illustrates two important features that show significant differences between NMTs and LLMs. For the proportion of descriptive words (ratio_dscrptW), texts translated by LLMs exhibit

[15]To determine significant differences, we first conduct a normality test on the data. If the data met the normality assumption, we apply ANOVA; otherwise, we use the non-parametric Kruskal-Wallis test.

a significantly higher usage of descriptive expressions than NMTs ($p < 0.0001$). Since descriptive words are generally regarded as enhancing textual vividness and specificity by providing richer contextual details, their higher occurrence in LLMs suggests more expressive and stylistically nuanced outputs than NMTs. Second, the proportion of adverbs (ratio_adverb) across HTs, NMT, and LLM indicates significant differences among these three systems ($p < 0.0001$). The trend suggests that HTs employ adverbs more frequently than MTs, where LLM restores some adverb usage compared to NMT.

5.2 Zooming into CTT-specifc features

If we narrow down our analysis to a more specific CLT perspective, we can see from Table 4 that several CTT-specific features also stand out in distinguishing different translation groups.

5.2.1 Ratio of foreignness

To begin with, the left column of Figure 6 presents the foreignness feature at the level of translatability. By definition, this feature quantifies the ratio of English words that appeared in the translated text. HTs exhibit zero occurrence of the retained foreign words. In contrast, MTs, particularly NMTs, demonstrate a notably higher ratio. Interestingly, LLMs display a substantially lower foreignness ratio compared to NMT, approaching HT-like tendencies.

Most untranslated cases are names (Such as "其次是Slightly" in NBD; "Tink确实又开始四处乱窜" in NGT), and idiomatic expressions ("他与他们分道扬way了" in NAZ). Although this is less common in LLMs, some expressions are still translated incompletely ("他们 perfectly safe，不是吗？" in LGM), a type of error also pointed out by Macken (2024). For child readers, minimizing source-language element leakage in CLT is a way both to improve acceptability and to mitigate "cultural colonialism" , since children's limited experience may necessitate a higher degree of adaptation than adult fiction (Lathey, 2015, p. 38). In this regard, LLMs demonstrate an advantage over NMTs.

5.2.2 Ratio of er-suffix

For the "ratio_er_suffix" feature, as shown in the middle column of Figure 6, the upper graph indicates that HTs employ significantly more "-er suffixes" than MTs, while in the lower graph, LLMs exhibit a higher ratio compared to NMTs.

The "-er suffix" serves as an important figurative expression in CLT, as it often conveys a colloquial, playful, or affectionate tone. It is commonly used in northern Chinese dialects, particularly in Beijing (see Chen, 2000; Fu, 2022). Such phonetic modifications cater to children's cognitive development and linguistic preferences, making texts more engaging and accessible. Based on this, HTs remain the most effective in preserving "-er suffix" in CLT. However, prompt-tuned LLMs demonstrate a stronger ability to capture "-er suffix" features compared to NMTs, suggesting that LLMs are more aligned with HTs in this aspect.

5.2.3 Ratio of repetitive expression

Repetitive expressions, such as AA (e.g., "热热的" warm and cozy), AAA (e.g., "慢慢慢" very slowly), and ABAB (e.g., "很久很久" a very long period), are a prominent stylistic feature in children's literature. The ratio_AA[16] feature in the right column of Figure 6 indicates that HTs employ significantly more repetition than MTs, while LLMs outperform NMTs in preserving this pattern. Research has shown that repetitive structures enhance readability, reinforce linguistic patterns, and facilitate memory retention for young readers, making texts more engaging and accessible (Tannen, 1989; Hickmann, 2003). Given these, the results suggest that LLMs better capture the stylistic and cognitive functions of repetition in CLT than NMTs, making them more aligned with HTs.

5.3 Some further remarks on the LLM translations

5.3.1 On LCG and LCL

Figure 1 shows that two LLM-based engines, LCG (ChatGPT) and LCL (Claude) exhibit significant distinctions in pair-wise classification compared to HTs and other MTs. By examining the feature importance list in the classification logs, we observe that LCG demonstrates exceptionally high divergence in the *Average Number of Children per Node* feature (see Figure 7), exceeding the values of other texts by approximately 1.5 times. Specifically, HT averages around 20, while LCG reaches approximately 35.

This feature reflects the syntactic dependency tree structure, where a higher value suggests LCG favors a flatter syntactic structure rather than a deeply nested one. LCG appears to prioritize paral-

[16]It should be noted that the AA pattern mentioned here excludes fixed proper nouns, such as "妈妈 mom".

lelism and broader phrase expansion. Corpus analysis (see online supplementary material) further corroborates this pattern where LCG tends to segment sentences more frequently, breaking complex structures into multiple shorter clauses deliberately, perhaps in order to retain readability suitable to children's levels.

Subsequently, we observed that LCL exhibits a significantly higher deviation in the ratio_quote (quotation mark) feature, reaching approximately twice the value of other engines (see Figure 7). While other LLMs maintain an average ratio of 0.13, LCL reaches approximately 0.25. Although *Peter Pan* is a children's novel rich in dialogue, such an unusually high occurrence of quotation marks appears atypical.

Upon inspecting the corpus, we found that other LLMs use directional Chinese quotation marks, whereas LCL employs non-directional English quotation marks. Our quotation-matching process was designed to recognize left quotation mark and non-directional mark, but not for right quotation mark. This explains why LCL's quotation mark count is nearly double that of other models. However, in formal writing, Chinese translations should adhere to standard typographic conventions, using directional Chinese quotation marks. In this regard, LCL's handling of punctuation is less consistent with formal Chinese writing norms compared to other engines.

5.3.2 On LDS, LTI and LLT

In this section, we discuss in more detail the recent open-source engine LDS (DeepSeek) and MT-tailored engine LLT (LaraTranslate), and both MT-tailored and open-sourced engine LTI (Unbabel Tower_instruct).

The three engines exhibit a high degree of confusion with other MT engines in classification tasks, with LDS and NGT achieving classification accuracies of only 0.59 and 0.66 with NBD, respectively. While these engines (particularly LDS) have gained significant attention recently, a more critical evaluation reveals that their performance in MT tasks shows no substantial improvement compared to other LLM-based engines.

Among MT-tailored LLMs, LLT and LTI fail to reach a satisfactory level in CTT-specific features with notably poorer performance than other commercial LLM models (see Figure 7). Specifically, they exhibit lower usage of the "-er suffix" and fewer repetitive expressions. Despite LLT's popularity and its advertised "creative translation" capabilities, our analysis finds little evidence of enhanced stylistic performance in its output. Additionally, during our testing, LLT produced a significant number of hallucinations, and its trans_foreignness score was the highest.

Some marketing claims about certain LLMs should be critically evaluated rather than taken at face value. While some models may perform well in general machine translation tasks, their effectiveness in specialized domains—such as children's literature—requires thorough empirical validation.

6 Conclusion

This study investigates the extent to which MTs diverge from HTs in CLT from a stylometric perspective, focusing on both generic textual features and CTT-specific features.

For RQ1, our findings confirm that MTs exhibit significant differences from HTs across both feature sets. Looking at generic text features, MTs deviate from HTs in conjunction word distributions and the ratio of 1-word-grams, where MTs tend to favor literal translation strategies, with a stronger influence from the source text. Looking at CTT-specific features, MTs generally fail to reproduce stylistic elements crucial in CLT, such as repetition and "-er suffix", further reiterating the challenges of automated translation in preserving literary expressiveness.

For RQ2, NMTs and LLMs diverge notably in both generic and CTT-specific stylistic features. While LLMs and NMTs exhibit significant differences in descriptive word usage and adverb ratios, LLMs show greater alignment with HTs. LLMs also outperform NMTs in "-er suffix" usage, AA-pattern repetition, and foreignness. Thus, they are better and more effective at capturing stylistic patterns in CLT than NMTs.

For RQ3, our analysis reveals distinct stylistic differences among HTs, NMTs, and LLMs. HTs still act as the gold standard in stylistic expressions, while NMTs produce more rigid and less engaging outputs. LLMs strike a balance, with greater stylistic fluidity than NMTs and approximating HT-like translation patterns. However, performance varies across LLMs. ChatGPT and Claude exhibit stronger stylistic consistency, whereas open-sourced or MT-tailored models show no clear advantages over the rest.

Limitations and future work

This study has several limitations that future research should address:

First, in terms of dataset selection, this study is limited to a single literary work, *Peter Pan*, which, while representative, may not fully capture the diversity of children's literature. Future research should expand the dataset to include translations of varied genres and styles to improve generalizability. Moreover, it should be noted that the training data of LLMs may contain human translations of *Peter Pan*, potentially influencing the results and blurring the boundaries between human and LLM translations.

Second, methodologically, this study primarily relies on quantitative stylometric analysis, offering a broad "distant reading" of translation patterns. However, it lacks in-depth qualitative analysis to explain why certain stylistic deviations occur between HTs, LLM and NMT outputs. Future work could incorporate qualitative case studies or human evaluations to better understand how MT outputs impact the reading experience of child audiences and whether stylistic deficiencies could be mitigated through prompt engineering or fine-tuning techniques.

Lastly, the feature set in this study remains lexical and syntactic-centric, with limited exploration of semantic and discourse-level attributes. Some feature overlaps were also observed, which could introduce redundancy in classification tasks. Future work should incorporate feature correlation analysis and dimensionality reduction methods (e.g., PCA) to refine the feature set and explore network-based approaches for a more holistic view of stylistic variations.

Acknowledgments

We gratefully acknowledge support from the China Scholarship Committee for the Visiting PhD Project (Num. 202406260211). We also thank the three anonymous reviewers for their constructive feedback.

Supplementary material

The supplementary material used in this study is uploaded onto Github https://github.com/DanielKong1996/CLT_MTsummit

Sustainability statement

This study primarily utilizes commercial MT engines' APIs during the translation acquisition phase. Due to the nature of these proprietary systems, accurately estimating the associated carbon footprint is challenging. Additionally, the classification and clustering experiments conducted during the machine learning phase of this research require relatively low computational resources. All experiments were performed on a personal laptop, ensuring minimal energy consumption. As a result, the overall environmental impact of this research is expected to be low.

References

Sheila Castilho, Clodagh Mallon, Rahel Meister, and Shengya Yue. 2023. Do online machine translation systems care for context? What about a GPT model? Tampere, Finland. European Association for Machine Translation (EAMT).

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-LTP: an open-source neural language technology platform for Chinese. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–49, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, and Min zhang. 2025. Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis. *Preprint*, arXiv:2502.11544.

Matthew Y. Chen. 2000. *Tone sandhi: patterns across Chinese dialects*, volume 92. Cambridge University Press.

B. Lee Cooper. 1989. Rhythm 'n' rhymes: character and theme images from children's literature in contemporary recordings, 1950–1985. *Popular Music & Society*.

Joke Daems. 2022. Dutch literary translators' use and perceived usefulness of technology: the role of awareness and attitude. In *Using Technologies for Creative-text Translation*. Routledge.

Joke Daems, Orphée De Clercq, and Lieve Macken. 2017. Translationese and post-editese: how comparable is comparable quality? *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 16:89–103.

Joke Daems, Paola Ruffo, and Lieve Macken. 2024. Impact of translation workflows with and without MT on textual characteristics in literary translation. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 57–64, Sheffield, United Kingdom. European Association for Machine Translation.

Orphée De Clercq, Gert De Sutter, Rudy Loock, Bert Cappelle, and Koen Plevoets. 2021. Uncovering

machine translationese using corpus analysis techniques to distinguish between original and machine-translated French. *Translation Quarterly*, (101):21–45.

Guoqi Ding. 2024. Triangulating text relationship in literary retranslation: The Great Gatsby in Chinese. *Digital Scholarship in the Humanities*, (39):849–863.

Maciej Eder. 2015. Rolling stylometry. *Digital Scholarship in the Humanities*, 31(3):457–469.

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: advancing low-resource machine translation with Claude. *arXiv preprint*.

Michael Farrell. 2018. Machine translation markers in post-edited machine translation output. In *Proceedings of the 40th Conference Translating and the Computer*, pages 50–59.

Anna Fornalczyk-Lipska. 2022. Repetitive or innovative? Children's literature in translation as the main focus of B.A. and M.A. theses. *Journal of Research in Higher Education*, 6:38–51.

Boer Fu. 2022. Contrast preservation in mandarin R-suffixation: a comparative study of beijing and liaoning dialects. *Glossa: a Journal of General Linguistics*, 7(1).

Ana Guerberof-Arenas and Antonio Toral. 2024. To be or not to be. *Target. International Journal of Translation Studies*, 36(2):215–244.

James Luke Hadley, Kristiina Taivalkoski-Shilov, Carlos S. C. Teixeira, and Antonio Toral. 2022. *Using Technologies for Creative-Text Translation*. Taylor & Francis.

Maya Hickmann. 2003. *Children's discourse: person, space, and time across languages*. Cambridge University Press, Cambridge, UK.

Wei Huang and Haitao Liu. 2009. Application of quantitative characteristics of Chinese genres in text clustering [In Chinese]. *Computer Engineering and Applications*, 45(29):25–27, 33.

Yue Jiang and Jiang Niu. 2022. A corpus-based search for machine translationese in terms of discourse coherence. *Across Languages and Cultures*, 23(2):148–166.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: the LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Gillian Lathey. 2015. *Translating children's literature*. Routledge, London.

Lei Lei, Yaoyu Wei, and Kanglong Liu. 2024. AlphaReadabilityChinese: a tool for the measurement of readability in Chinese texts and its applications. *Foreign Languages and Their Teaching*, 46(1):83–93.

Gerard Lynch and Carl Vogel. 2018. The translator's visibility: detecting translatorial fingerprints in contemporaneous parallel translations. *Computer Speech & Language*, 52:79–104.

Lieve Macken. 2024. Machine translation meets large language models: evaluating ChatGPT's ability to automatically post-edit literary texts. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 65–81, Sheffield, United Kingdom. European Association for Machine Translation.

Lorenzo Mastropierro. 2022. The avoidance of repetition in translation: a multifactorial study of repeated reporting verbs in the Italian translation of the harry potter series. In *Advances in Corpus Applications in Literary and Translation Studies*, pages 138–157. Routledge.

Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262.

Esther Ploeger, Huiyuan Lai, Rik Noord, and Antonio Toral. 2024. Towards tailored recovery of lexical diversity in literary machine translation.

T. Puurtinen. 2003. Genre-specific features of translationese? Linguistic differences between translated and non-translated finnish children's literature. *Literary and Linguistic Computing*, 18(4):389–406.

Simon Päpcke, Thomas Weitin, Katharina Herget, Anastasia Glawion, and Ulrik Brandes. 2022. Stylometric similarity in literary corpora: non-authorship clustering and*Deutscher novellenschatz*. *Digital Scholarship in the Humanities*, 38(1):277–295.

Md. Mostafizer Rahman, Ariful Islam Shiplu, and Yutaka Watanobe. 2024. CommentClass: a robust ensemble machine learning model for comment classification. *International Journal of Computational Intelligence Systems*, 17(1):184.

Paola Ruffo. 2022. Collecting literary translators' narratives: towards a new paradigm for technological innovation in literary translation. In *Using Technologies for Creative-text Translation*. Routledge.

Jan Rybicki. 2025. Can machine translation of literary texts fool stylometry? *Digital Scholarship in the Humanities*, page fqaf010.

Kristiina Taivalkoski-Shilov. 2018. Ethical issues regarding machine(-assisted) translation of literary texts. *Perspectives*, 27(5):689–703.

Deborah Tannen. 1989. *Talking voices: repetition, dialogue, and imagery in conversational discourse*. Cambridge University Press, Cambridge, UK.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. *arXiv preprint arXiv:1907.00900*.

Antonio Toral and Andy Way. 2014. Is machine translation ready for literature. In *Proceedings of Translating and the Computer 36*, London, UK. AsLing.

Antonio Toral and Andy Way. 2015. Machine-assisted translation of literary text: a case study. *Translation Spaces*, 4(2):240–267.

Antonio Toral and Andy Way. 2018. What level of quality can Neural Machine Translation attain on literary text? In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: from Principles to Practice*, Machine Translation: Technologies and Applications, pages 263–287. Springer International Publishing, Cham.

Jan Van Coillie and Jack McMartin, editors. 2020. *Children's literature in translation: texts and contexts*. Leuven University Press.

V. Volansky, N. Ordan, and S. Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Matthijs J. Warrens and Hanneke van der Hoef. 2022. Understanding the adjusted rand index and other partition comparison indices based on counting object pairs. *Journal of Classification*, 39(3):487–509.

Andy Way, Andrew Rothwell, and Roy Youdale. 2023. Why literary translators should embrace translation technology. *Tradumàtica Tecnologies De La Traducció*, pages 87–102.

Xu Xu and Jiayin Li. 2020. Concreteness/abstractness ratings for two-character chinese words in MELD-SCH. *PLOS ONE*, 15(6):e0232133.

Xiaomin Zhang, Haidee Kotze (Kruger), and Jing Fang. 2019. Explicitation in children's literature translated from English to Chinese: a corpus-based study of personal pronouns. *Perspectives*, 28(5):717–736.

Meijuan Zhao, Lay Hoon Ang, Sabariah Md Rashid, and Florence Haw Ching Toh. 2022. Translating narrative space in children's fiction bronze and sunflower from Chinese to English. *Sage Open*, 12(1):21582440211068498.

Anna Čermáková. 2018. Translating children's literature: some insights from corpus stylistics. *Ilha Do Desterro*, 71:117–133.

A LLM prompt

The engineered prompt is originally drafted in Chinese, as follows:

你是一位专精于儿童文学翻译与创作的资深译者，你的任务是将 J.M. Barrie 的《Peter Pan》翻译成富有创造力的中文版本。这不仅是一次普通的翻译，而是一种"创译"——你的目标是保留原文的幻想色彩和情感基调，同时使译文更加符合中文儿童的阅读习惯，充满趣味性和文学感染力。

请遵循以下创译原则：

- 情感与想象力再现：保持原著的童话氛围，使译文生动、富有画面感，如必要可调整句式，使其更具表现力。

- 符合儿童语言习惯：采用简练、口语化、充满韵律感的表达方式，避免生硬直译。可适当使用拟声词、叠词、押韵句式等。

- 文化适配：调整可能不易理解的文化元素，使其符合中文语境，但同时保留原作的神秘感和奇幻感。

- 角色个性化语言：确保各个角色的独特特点能在译文中得以体现。

- 增添文学趣味：适当运用修辞（比喻、拟人、夸张等），调整句式，使故事更具节奏感，增强朗读时的感染力。

请翻译以下《彼得潘》片段，并确保符合上述原则: [原文]

English Translation:

You are a seasoned translator and writer specializing in children's literature. Your task is to create a highly creative Chinese adaptation of J.M. Barrie's *Peter Pan*. This is not merely a literal translation but a transcreation - your goal is to preserve the whimsical essence and emotional tone of the original while making the text more engaging and accessible for Chinese children ensuring it is rich in imaginative appeal and literary charm

Guiding Principles for Transcreation:

- Emotional and Imaginative Recreation: Maintain the fairy-tale atmosphere of the original, making the translation vivid and evocative. Feel free to adjust sentence structures to enhance expressiveness.

- Child-Friendly Language: Use concise, rhythmic, and conversational expressions, avoiding rigid literal translation. Incorporate onomatopoeia, reduplication, rhyming phrases, and other playful linguistic elements as appropriate.

- Cultural Adaptation: Modify cultural references that may be difficult for Chinese readers to grasp, ensuring they fit the Chinese linguistic and

cultural context while preserving the mystical and fantastical essence of the original.

- Character-Specific Speech: Ensure that each character's unique personality and speech style are well reflected in the translation.

- Enhanced Literary Appeal: Utilize figurative language (e.g., metaphors, personification, hyperbole) and varied sentence structures to enrich the storytelling, improve readability, and enhance the rhythm and emotional impact, particularly when read aloud.

Please translate the following excerpt from Peter Pan while adhering to these principles: [text]

B Feature set in summary

A summary list of features used in the study is in Table 3.

C Selected features used in experiments

A summary list of significant features used in different experiments is in Table 4.

D Supplementary figures

Figure 1 illustrates pair-wised classification results among HTs, NMTs, and LLMs groups.

Figure 2 shows the clustering results.

Figure 3 illustrates ratio of conjunction words, and the ratio of the 1-word-gram-然后 between HTs vs. MTs, and NMTs vs. LLMs.

Figure 4 presents ratio of 1-word-gram-一样 between HTs vs. MTs, and NMTs vs. LLMs, and the distribution ratio of "像……一样" in all texts.

Figure 5 illustrates generic features between HTs vs. MTs, and NMTs vs. LLMs.

Figure 6 shows CTT-specific features between HTs vs. MTs, and NMTs vs. LLMs.

Figure 7 mainly illustrates the distribution of six key features among seven LLMs.

Figure 8 shows an example of the phrase "像...一样" drawn from actual concordance.

Feature level	Sub level	Total	Feature instances
Generic textual features	Lexical	69	STTR, MTLD, noun, verb, content words, idioms ...
	Syntactical	26	WordsPerSent, QuestionSent, MDD, AvgChildrenPerNode...
	Readability	16	lexical_richness, Concreteness score, AvgConcrete ...
	N-grams	309	N_word_gram, N_PoS_gram, N = 1 - 3 ...
CTT-specific features	Repetition	7	ratio_AA, ratio_ABAB, ratio_AAA, ratio_AABB...
	Rhythm	10	Open Syllable Ratio, Rhyme Density, Rhyme Ratio ...
	Translatability	5	completeness, foreignness, code_switching, untranslatable ...
	Miscellaneous	5	ratio_onomatopoeia, ratio_StrongModifier, ratio_er_suffix ...

Table 3: Summary of features used in this study. Due to space constraints, only representative feature instances are listed, with "..." indicating additional features available in the full list (see supplementary materials). Feature names shortened for formatting when necessary.

Groups	Generic features	Specific features
HTs vs. MTs	Ratio_conjunction	foreignness_ratio
	word_1gram_然后	switching_ratio
	word_1gram_一样	ratio_er_suffix
	word_1gram_它们	ratio_StrSentMdfyr
	ratio_adverb	ratio_aabb
	ratio_ContentWords	ratio_AA
NMTs vs. LLMs	word_1gram_没有	code_switching_ratio
	pos_2gram_ws	foreignness_ratio
	ratio_dscrptW	ratio_StrSentMdfyr
	ratio_adverb	ratio_AA
	POB	-
	ratio_prep	-

Table 4: Summary of significant features used in different experiments. 6 features in each sub-categories are selected from top-40 salient features, with "-" mark representing no more features in this category are found in the top-40. Feature names shortened for formatting when necessary.

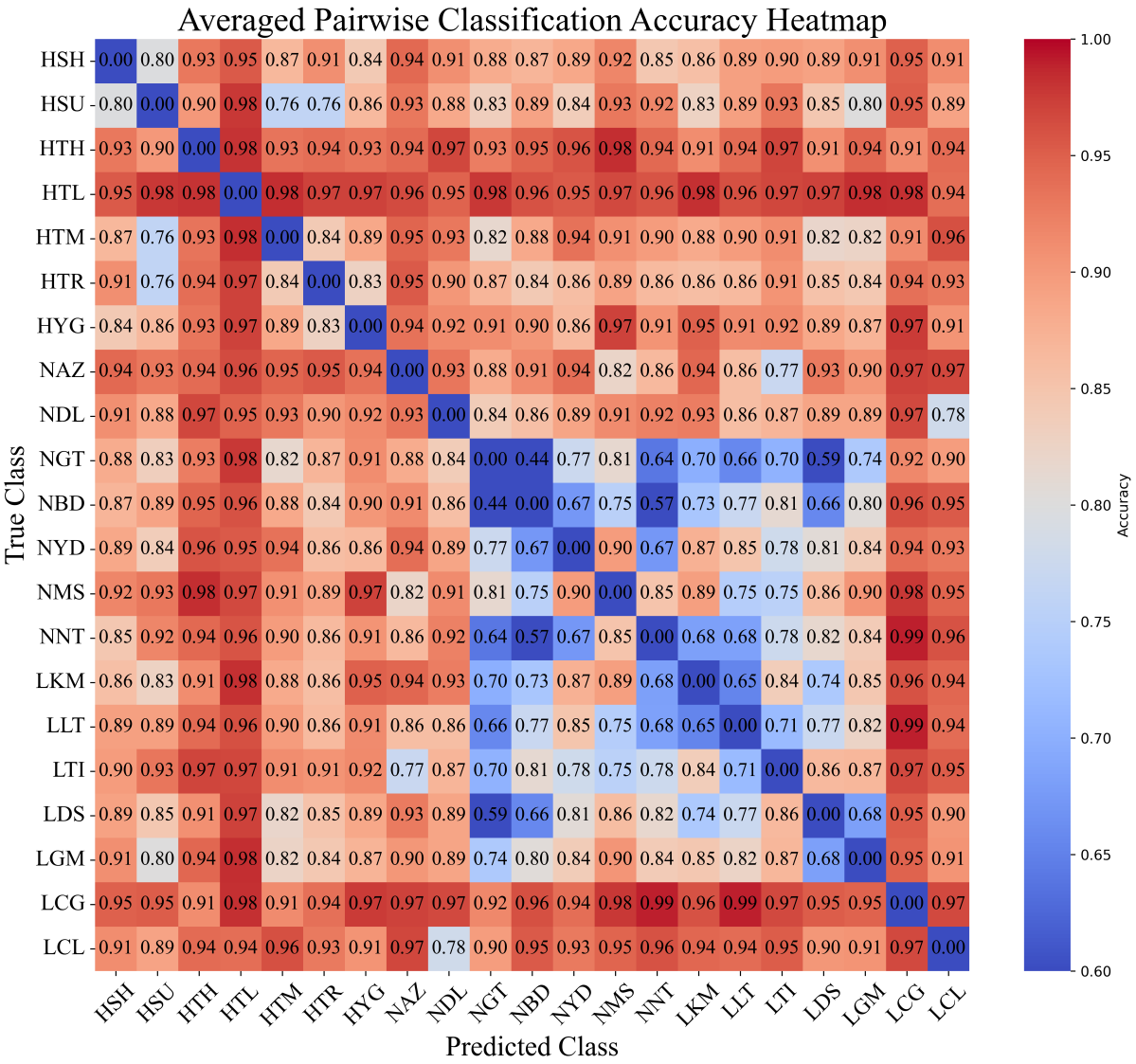


Figure 1: Pair-wise comparison of different MT engines based on five averaged classifiers and top-30 salient features

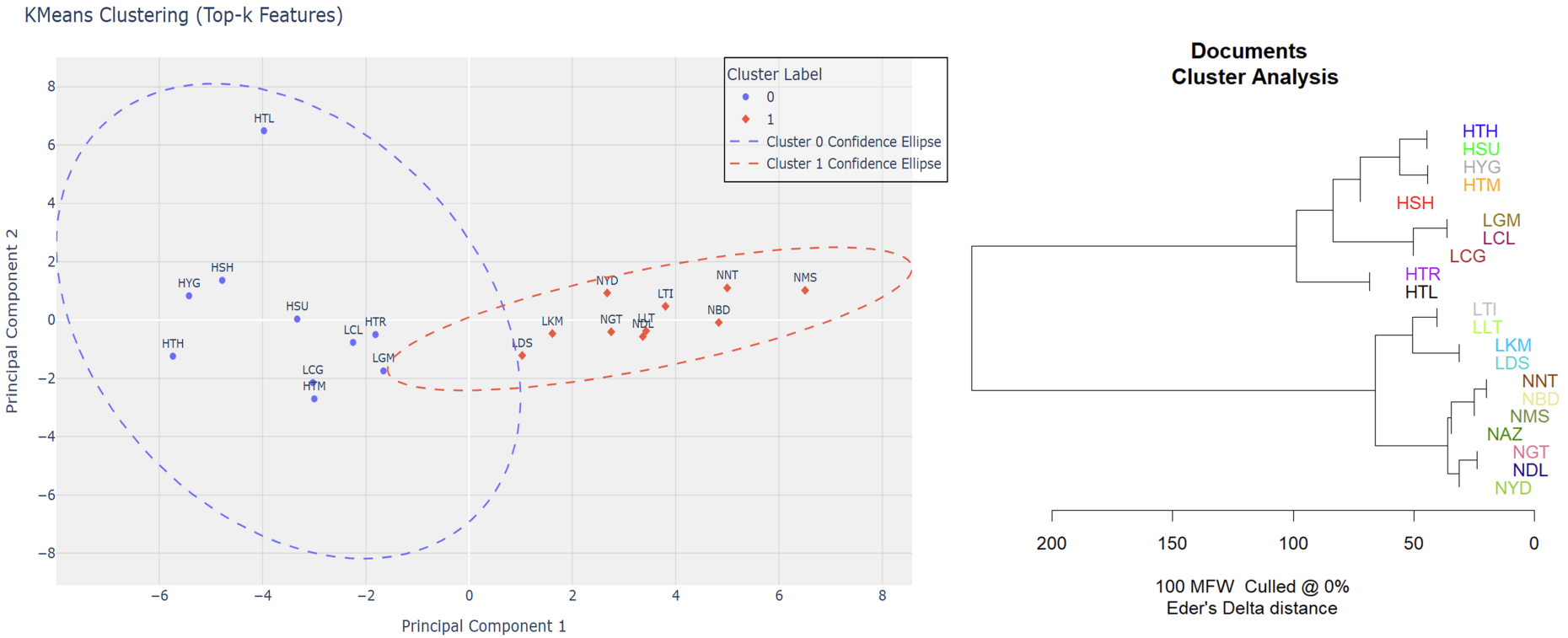


Figure 2: Left: using Top-30 features in K-means clustering with $k = 2$, where the true labels are set as HTs = 0 and MTs = 1. ARI is 0.4873, and the confidence ellipse is set at 0.7. Right: hierarchical clustering using the Stylo package with 100 MFWs and Eder's delta as the distance measure.

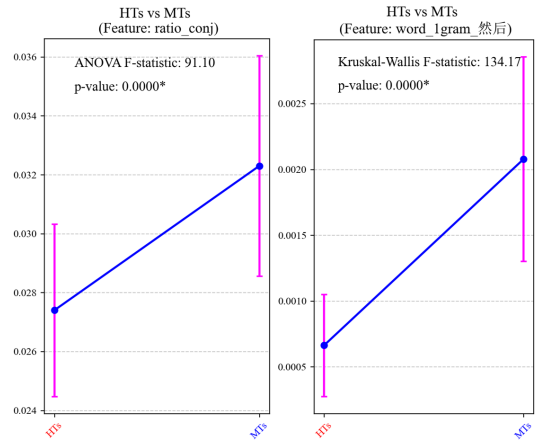


Figure 3: Generic differences between HTs and MTs. The left panel compares ratio of conjunction words, while the right panel examines ratio of the 1-word-gram-然后

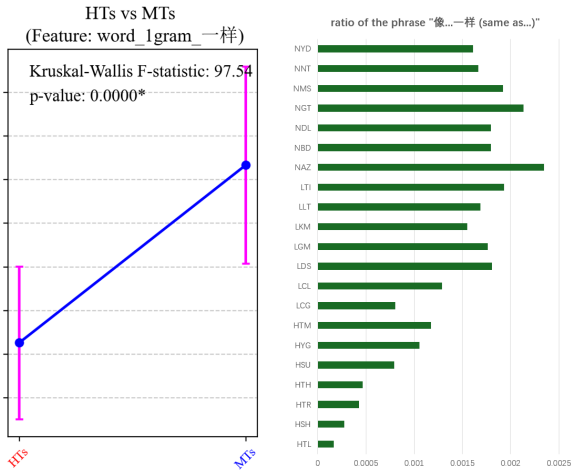


Figure 4: The left panel compares ratio of 1-word-gram-一样, while the right panel shows the distribution ratio of "像…一样" in all texts.

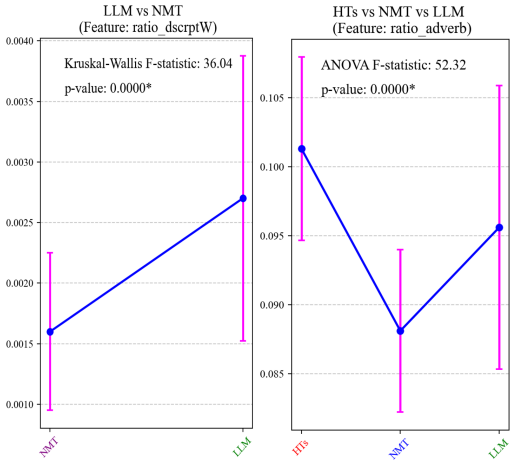


Figure 5: Generic differences between HTs and MTs. The left panel compares ratio of descriptive words, while the right panel examines ratio of the adverbs among HTs, NMTs and LLMs.

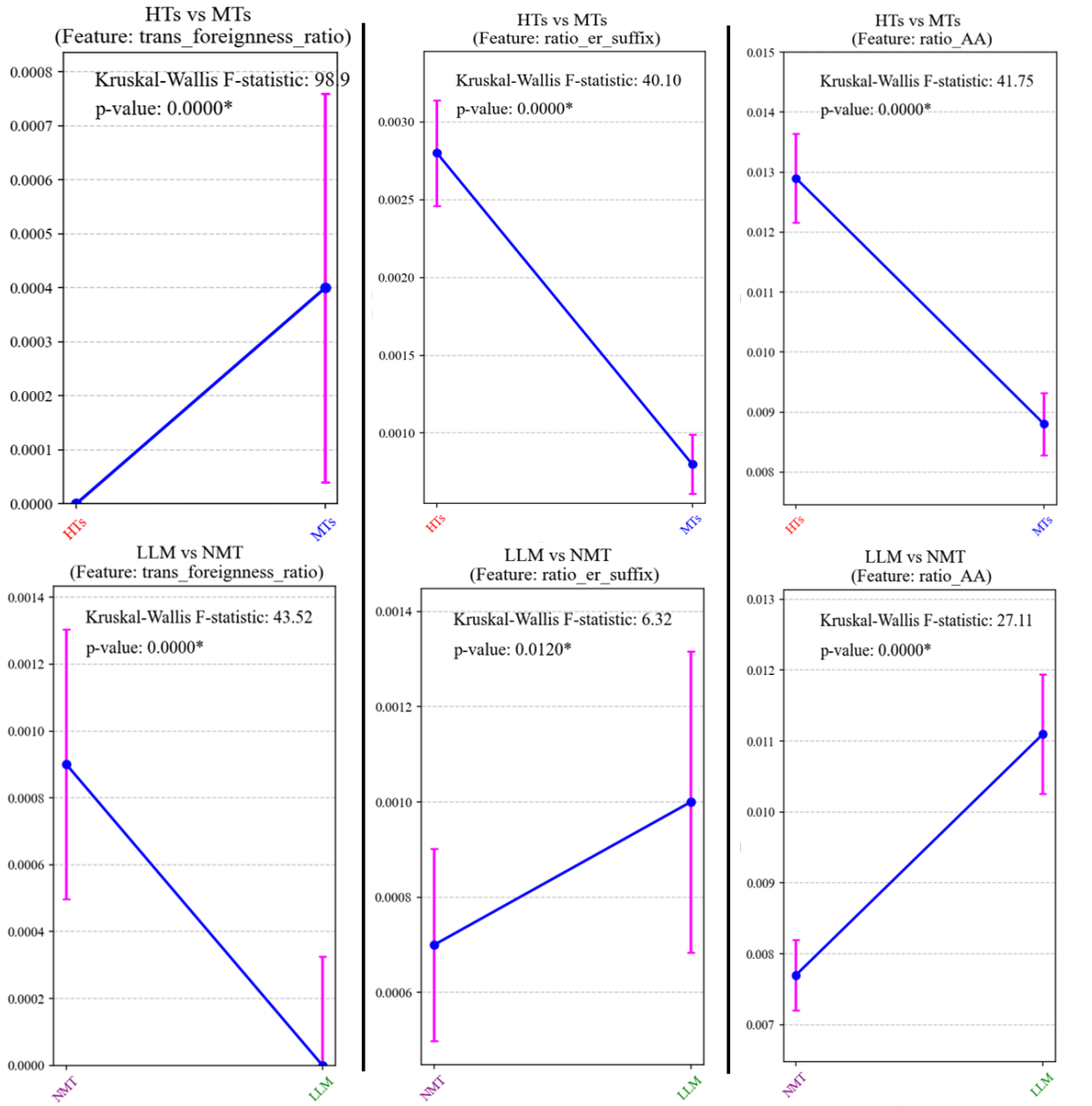


Figure 6: Feature distribution comparisons between different translation groups. The left column presents the differences in *trans_foreignness_ratio*; the middle column shows variations in *ratio_er_suffix*; and the right column illustrates the differences in AA-pattern repetitive expression.

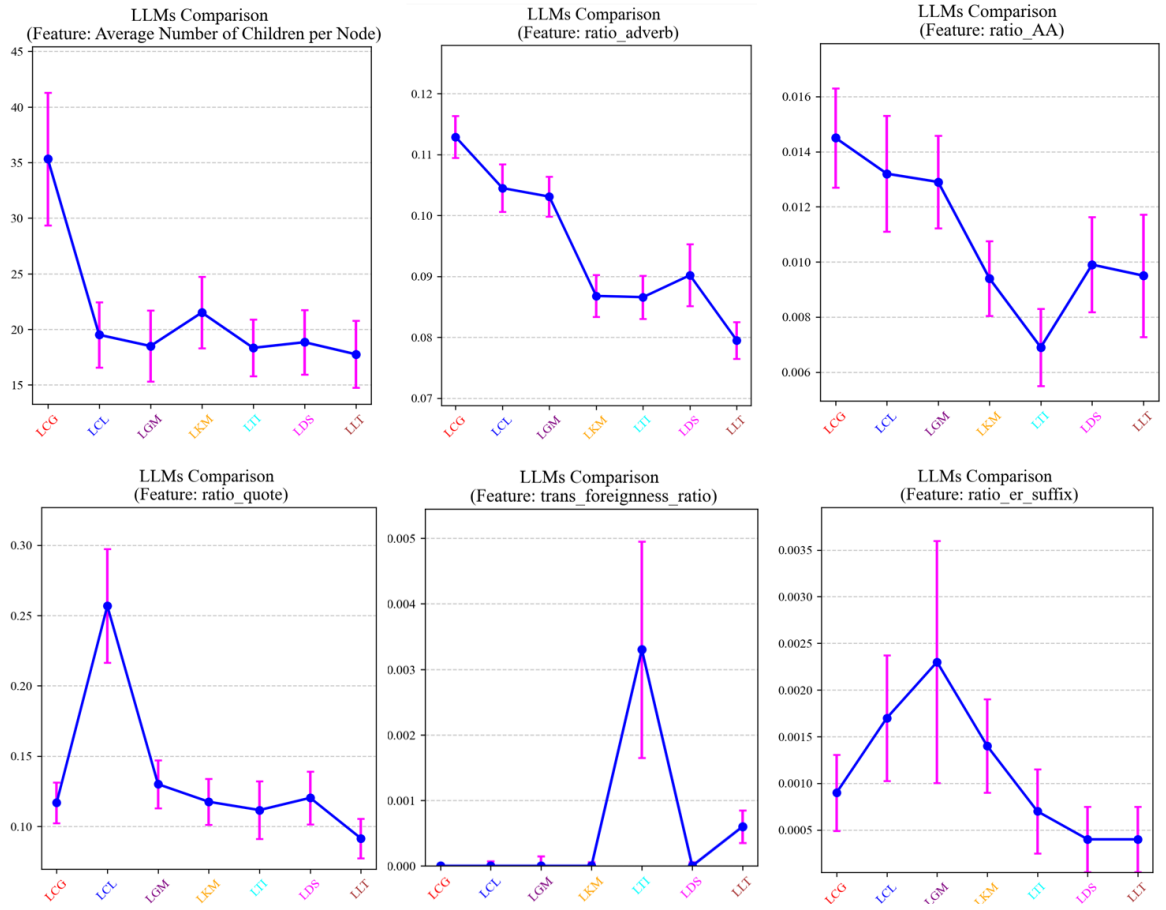


Figure 7: Comparison of key linguistic features across seven different LLMs. The top-left plot shows differences in *Average Number of Children per Node*; the top-middle *ratio_adverb*; the top-right *ratio_AA*. The bottom-left plot presents differences in *ratio_quote*; the bottom-middle *trans_foreignness_ratio*; the bottom-right *ratio_er_suffix*.

Origin: He was accompanied by a strange light, no bigger than your fist, which darted about the room like a living thing.
NAZ: 他 身边 有 一 道 奇怪 的 光芒 ， 比 你 的 拳头 还 大 ， 它 像 活物 一样 在 房间 里 飞来 飞 去 ， CRLF
NYD: 他 身边 有 一 道 奇怪 的 光 ， 没有 你 的 拳头 大 ， 像 个 活物 一样 在 房间 里 窜 来 窜 去 ， CRLF
NNT: 伴随 着 他 的 是 一 道 奇怪 的 光 ， 没有 你 的 拳头 大 ， 像 一个 活物 一样 在 房间 里 飞来 飞去 ， CRLF
NMS: 他 身边 有 一 道 奇怪 的 光 ， 不 比 你 的 拳头 大 ， 像 个 活物 一样 在 房间 里 飞来 飞去 ， CRLF
NGT: 他 身边 还 伴 有 一 道 奇怪 的 光 ， 不 比 拳头 大 ， 像 活物 一样 在 房间 里 飞来 飞去 ， CRLF
NDL: 伴随 着 他 的 是 一 束 奇怪 的 光 ， 比 拳头 大 不 了 多少 ， 像 活物 一样 在 房间 里 飞来 飞 去 ， CRLF
NBD: 他 身边 有 一 道 奇怪 的 光 ， 比 你 的 拳头 还 小 ， 像 活物 一样 在 房间 里 飞来 飞去 ， CRLF
LCL: 他 身边 跟着 一 道 奇异 的 光 ， 不 比 你 的 拳头 大 ， 在 房间 里 像 个 活物 似的 飞来 飞 去 。 CRLF
LCG: 他 并非 独自 一 人 ， 身旁 还 跟着 一 道 奇异 的 光亮 ， 只 有 拳头 大小 ， 却 像 活物 般 在 屋里 飞快 地 穿梭 。 CRLF
LTI: 他 身边 还有 一 股 奇怪 的 光芒 ， 大小 不 过 手掌 一个 ， 它 在 房间 里 来回 穿梭 ， CRLF
LLT: 他 身边 闪 着 一 道 奇怪 的 光 ， 不 比 你 的 拳头 大 ， 像 个 活生生 的 东西 一样 在 房间 里 飞来 飞 去 ， CRLF
LKM: 他 身边 伴随 着 一 束 奇异 的 光芒 ， 那 光束 不 比 你 的 拳头 大 ， 它 在 房间 里 像 活物 一样 四处 飞 蹿 ， CRLF
LGM: 他 身边 带 着 一 盏 奇怪 的 光 ， 不 比 你 的 拳头 大 ， 在 房间 里 像 个 活物 一样 飞来 飞 去 ， CRLF
HYG: 伴随 着 他 的 ， 还有 一 团 奇异 的 光 ， 那 光 还 没有 你 的 拳头 那么 大 ， 它 像 一个 活物 在 房间 里 四处 乱 飞 。 CRLF
HTR: 他 带有 一 种 奇怪 的 光 。 不 比 你 的 拳头 大 ， 像 个 活 的 东西 那样 在 房间 里 窜 来 窜 去 ； CRLF
HTM: 跟 他 在 一起 的 还有 一 道 奇异 的 光 ， 跟 你 的 拳头 差不多 大 ， 像 一个 活物 一样 在 房间 里 蹿动 ， CRLF
HTL: 陪 着 他 来 的 还有 一 颗 奇异 的 火亮 ， 不见得 比 你 的 头 大 ， 像 是 活 东西 似的 在 屋里 乱 飞 ， CRLF
HTH: 男孩 身边 有 一 团 奇怪 的 光 ， 比 拳头 大 不 了 多少 ， 那 团 光 在 房间 里 横冲直撞 ， 像 是 个 活物 。 CRLF
HSU: 紧跟着 他 的 ， 是 一 团 怪异 的 光 ， 一 团 还 没有 拳头 大 的 光 ， 像 个 有 生命 的 活物 一样 ， 在 房间 里面 四处 乱 飞 。 CRLF
HSH: 与 他 相伴 的 还有 一 道 奇妙 的 光 ， 那 光 并 不 比 你 的 拳头 大 ， 宛如 一个 精灵 在 房间 里 横冲直撞 ， CRLF

Figure 8: Actual concordance in the corpus of the feature “像……一样”. Highlights by matching regular expression “像[一-龟]+一样”.

This file is accessed from Delu Kong's Repo, you may find the official version from ACL anthology web.
Latest version will also be upload onto https://github.com/DanielKong1996/CLT_MTsummit