Automated Information Extraction and Template Filling from Client Style Guides

Leonor Graça
TransPerfect, Lisbon, Portugal
leonor.graca.int@unbabel.com

Vera Cabarrão
TransPerfect, Lisbon, Portugal
vera.cabarrao@unbabel.com

Helena Moniz
University of Lisbon, Portugal
CLUL, Lisbon, Portugal
INESC-ID, Lisbon, Portugal
helena.moniz@edu.ulisboa.pt

Abstract

Style guides are a centrepiece of professional translation workflows. Yet, their integration into automatic pipelines remains underexplored. This paper presents exploratory work on information extraction from client style guides and application to a templated style guide, developed to be a system prompt. This template is then applied during an LLM-based translation to automatically produce outputs that are compliant to client's requirements. The study focused on seven language pairs (LP), evaluating the automatic extraction, and translation quality and compliance with the style guide. The extraction demonstrated reliable performance across languages and file formats. Translation quality and adherence were evaluated using human preference annotation, comparing two Tower models (Tower Zen 9B and Tower+ 72B). The results indicate a modest advantage for Tower+, but with mutual acceptability in certain instances. These findings establish a viable semi-automatic framework for style guide integration in translation workflows, and motivate further investigation across broader domains, clients, and LPs.

1 Introduction

Language is never neutral. Every decision in a translation task carries meaning, even more so in a professional translation setting. Inconsistent usage of terminology can erode brand trust. An ambiguous tone can lead to miscommunications between the customer and the seller. Improper formatting of legal sections could lead to liability issues. The absence of clear rules, beyond language guidelines, forces the translator to make decisions without context, leading to inconsistent results.

Style guides exist to prevent exactly that. They are often described as reference documents, but that quite undersells them. Style guides are one of the most crucial tools in a translation company. When looking at different perspectives, from translators themselves to language providers (American Translators Association, 2019; Lingualinx, 2024; Ghosal, 2024), they all agree that without clear guidelines that determine tone, grammar, terminology, formatting, and audience definition, translators are left to make calls that should already have been resolved, leading to inconsistent messaging, costly revisions, and even legal consequences. ISO (2015) proves this point, creating a standard for translation processes. It is precisely the presence of guidelines that eliminate the guesswork by giving every linguist, reviewer, translator, etc., the same rulebook to work from.

However, for all their importance, they come with one significant challenge. As they are provided by the client, no two are alike, arriving with different formats, lengths, structures, and types of information. Some run for dozens of pages with detailed examples, while others are a page of bullet points. This inconsistency is difficult enough to handle with human translators. But, as the industry turns to Large Language Models (LLM), this becomes more acute. As powerful as they might be, LLMs tend to struggle in instruction-following (Heo et al., 2025; Young et al., 2025), even more so when parsing these style guides, having one of

the most valuable assets in the translation workflow becoming one of its biggest blind spots.

To circle around this, the research aims to develop an automated framework for generating machine-readable style guides that can be integrated directly into LLM-assisted translation workflows, enhancing translation quality through real-time application of client-specific style guides as system prompts while simultaneously improving operational efficiency by automating the style guide implementation process across diverse client requirements and linguistic specifications.

2 Literature Review

A translation style guide is a document that incorporates relevant information for translation tasks. Through these, translators can ensure consistency in format and language throughout all brand documentation (Aranbure et al., 2018).

However, a different challenge arises. Recent industry data underlines the rapid transformation toward AI-assisted workflows. The Business Research Company (2026) projects a Compound Annual Growth Rate (CAGR) of 25.2% for AI in translation, forecasting market expansion to $2.94 billion in 2025. Furthermore, CSA Research reports a $24.6 billion decline in the translation market since 2019, representing a 45% drop, suggesting industry restructuring alongside technological adoption. The question is no longer whether to adopt them, but rather how to integrate them effectively into established workflows.

Despite their capabilities, LLMs still exhibit notable limitations, particularly in adhering to instructions. Research has proven that LLMs struggle when processing complex and large guidelines (Jang et al., 2022; Jaroslawicz et al., 2025), suggesting that directly inputting extensive client style guides would likely result in unsatisfactory results. Furthermore, these style guides are developed for the human mind, and do not follow prompt engineering strategies, begging the question, how can we make LLMs follow style guides?

With that, a primary work was built upon this. A study conducted by the author (Graça, 2025), prior to this, presented initial research on how to properly develop style guides that are readable by an LLM, through its application as a system prompt. This was done with the development of a three-step methodology: a General Template — a one-fits-all approach, any document from any domain could be translated using it; a Domain Template — a style guide tailored for one specific domain in which any document within it could be translated using it; and a Document Template — one style guide for each type of document, within a domain.

The results showed its effectiveness in improving translation quality and compliance, proposing the Domain Template for simpler documents and the Document Template for more complex ones as the best approach. However, this was tailored to be more manual, as all the style guides are filled-in by a human before application. Additionally, it misses the specifications that certain clients have. Domains have generic rules. However, certain clients may wish to follow different parameters.

As such, the fourth-step of this methodology was thought out, the Client Template, to fit the needed industry use case, and tailor this system prompt approach to client specifications.

3 Methodology

Building upon Graça (2025), this research added to the framework by creating a fourth-step. This methodology brought two additions to the previous work, automating the process through a script integrated with the GPT API, while making it adaptable to the needs of individual clients. Data was extracted from the client's style guide and applied to a template structure, automatically creating an LLM-readable style guide that follows the clients needs and requirements. This template was later applied as a system prompt during translation, through TowerLLM.

3.1 Information Extraction and Application

3.1.1 Template

The previous work mentioned in Section 2 led to the development of AI-readable style guides and, subsequently, to its templated structure. As the study intends to be generally applicable to all clients, the initial step consisted of an adaptation of the General Template for this use-case.

However, this created a new challenge. The previous templates were filled-in by a linguist and/or translator. But, as the purpose was to automatise the first step, it was necessary to turn the template into a prompt to be adapted by the model. As such, the template itself was adapted through prompt engineering, adding instructions to the relevant sections, e.g. prompt *"add the relevant rules in the style guide for formal/informal usage and tone of*

voice." added to the section **Tone and Formality**. In this way, the model should incorporate the rules regarding formality and tone of voice in this section while employing the extracted data.

In order to create an appropriate template, it was necessary to analyse different client style guides so as to identify key elements. An aspect to consider was the importance of a section on one style guide, that was irrelevant for another. A marketing client may require the section **Humor**, while a legal client would not. Not all categories are intended for all clients. Nevertheless, testing revealed that the best approach was to have a prompt filter the necessary sections based on the client.

Furthermore, it was determined what rules must be unchangeable, e.g. *"Apply punctuation consistently, following the rules of the [Target Language]."*. For certain sections, without a fixed prompt, the translation would not adhere to the language norms. In other words, the model already knows the generic grammar rules, as such those are not added to the template. However, without that prompt in the template, the translation will only apply the exceptions, and lose quality because the generic grammar rules are not applied.

Lastly, the study showed that style guides do not always have an LP. Certain clients have a style guide per Target Language (TL), independent of the Source. To accommodate both, two templates were developed, one for LP and one for TL.

3.1.2 Model Instructing

The initial stage of the research was to automate the process, and the best way was through a Python script. Due to token restrictions, TowerLLM is unable to perform the extraction and application. As such, we proceeded to its experimentation with GPT API, more specifically GPT 4o. For future iterations, we intend to test Gemini on this task.

First, two input files were loaded: the client's style guide and the template, defining the output schema, whose sections are initialised as empty arrays in a dictionary structure. Second, the style guide is processed in chunks, depending on its size, and each chunk was submitted to the API with a prompt to extract and return information. The data was extracted from all chunks, one at a time, and rewritten as a clear and structured prompt, aggregated by category into the structured dictionary, ensuring that multiple relevant rules per section are retained. Lastly, a second prompt was applied, incorporating the template and the aggregated data, and directing the model to organise the extracted information and properly add it to the template, according to its rules. For any empty sections, it should add generic rules or discard. The completed template is written to a configurable output file.

3.2 Data

For the development stage, and as testing data, three different clients of different domains were used: legal, with banking data; marketing, with product description; and media, with promotional text. Each had style guides that vary in length, format, and structure. All the necessary materials — templates, scripts, and prompts — for the pipeline were tested using these clients, until satisfactory results were achieved across the board.

Following the validation of the methodology, a testing framework was established with a different client from the previous three as the primary case-study. This client was selected as it had both the resources and the interest in implementing this study. The previous three clients did not have the necessary data availability or the data was too simple to evaluate the style guide adherence.

The client's data encompassed multiple brands targeting distinct demographic segments, including adult and children's markets, with product categories spanning various domains. Translation tasks were centred on detailed product descriptions. For this purpose, we selected 28 products, each with four types of descriptions: title; consumer description; shopper description; and product description. Each product, with the four descriptions, had around 20 to 30 sentences.

This research was conducted in seven languages: Japanese (ja-JP), es-ES, Danish (da-DK), Korean (ko-KR), Polish (pl-PL), pt-PT, and Standard German (de-DE). Two of them were selected as they had two style guides each, PL and DE.

The style guides were similar to each other. They were *.docx* files, with similar structures, around 10 pages long, and with a list of rules for grammar, format, and tone of voice. The secondary files for PL and DE comprised a detailed list of formatting rules for specific scenarios, and they were longer and more dense. The processed style guides had around 1500 tokens.

The style guides were processed and post-edited by the author, needing minor annotations. For the translation step, two Tower models were tested in instruction-following, to determine the best when

applying the pipeline to a translation workflow. The models were TransPerfect's proprietary models, TowerZen 9B, the most recent model, and Tower+ 72B (Rei et al., 2025), tested in the three-tier approach assessed in Section 2. For unbiased answers, the models were labelled as A, corresponding to TowerZen 9B, and B, Tower+ 72B.

The annotations were performed by a research linguist or freelancer, with high proficiency in the TL, and prior experience with similar tasks. Detailed guidelines were prepared, covering task descriptions, rules, evaluation criteria, and key considerations. To ensure clarity, an open communication channel was maintained between the author and the annotators, and a dedicated step for comments was incorporated. Both models were given a score of 1 (worst) to 5 (best) for translation quality.

It is important to note that the quality scores were separate from the style guide application, one translation could receive a score of 3, and yet perfectly follow the style guide. To further distinguish between these aspects, two additional questions were included: "Best Style Guide adherence" and "Best quality". For each question, the annotator would select "A", "B", "Equally Bad" if both models failed, or "Equally Good" if both models presented good quality.

4 Results

This section presents the empirical findings from the two experimental phases: the automated information extraction from client style guides and the evaluation of translation quality achieved through the application of these extracted style guides.

4.1 Information Extraction and Application

Overall, this process had quite positive results, requiring post-edition, but to a short extent. The issue was in the system's tendency to omit more specific rules while maintaining accuracy in the information it did extract. On average, four rules were eliminated for being overly general or unnecessary, and 12 rules were missing and added in post-edition. In fact, no incorrect information was generated, suggesting that the extraction methodology prioritises precision. However, in opposition, the system would add information that an LLM would already know. While not incorrect, it was generic or irrelevant, making it unnecessary.

The extraction quality varied according to the specificity and clarity of the source material. In cases where style guides contained ambiguous or unclear rules, the system occasionally produced information that diverged from the intended specifications. Generic style conventions, such as punctuation and formatting rules, were well-captured, whereas domain-specific or client-particular requirements that opposed the general rules, frequently required manual edition, as seen by the higher quantity for addition as opposed to deletion.

Despite these limitations, the extraction framework produced outputs that were well-structured and accurately represented the original style guides provided by the client, indicating that the methodology serves as a viable foundation for semi-automated style guide generation, provided that human oversight is maintained for validation and completion of specialised requirements.

The performance of the script varied depending on the format of the style guide. The *.txt* files and *.docx* / *.doc* consistently yielded the most reliable results. When rules were missing, it did not appear to be due to formatting constraints. *PDF* files, with more complex *PDF* layouts — multiple columns, graphics, tables, and distinct text boxes — were more challenging. A substantial portion of the content was not captured. This suggests that while it generalises well across most common file formats, documents with highly complex visual layouts may benefit from a pre-processing step prior to extraction, in order to ensure a more comprehensive capture of stylistic content.

4.2 Translation and Instruction-Following

	Style Guide Adherence				Translation Quality		
	Equally Good	Equally Bad	A	B	A mean	B mean	Δ (B-A)
ES	4	13	4	7	3.28	3.6	+0.32
DE	1	4	14	9	2.78	2.96	+0.18
KO	0	7	0	21	2.57	3.14	+0.57
PT	4	8	5	11	3.36	3.18	-0.18
PL	11	0	9	8	3.8	3.78	-0.02
DA	22	4	0	2	2.21	2.21	0
JA	18	0	7	3	2.6	2.75	+0.15
	30.6%	18.4%	19.9%	31.1%	2.9	3.1	

Table 1: Style guide adherence and translation quality results

The evaluation results in Table 1 indicate a marginal difference between the two models, with A (TowerZen 9B) achieving a mean translation quality score of 2.9 and B (Tower+ 72B) a mean of 3.1. For four of the seven languages, Model B received a higher score, with the most significant difference in ko-KR (Δ = +0.57) and es-ES

(Δ = +0.32), although the gap between the two remains modest.

Across languages and models, the adherence produced a number of positive results. However, several recurrent patterns indicate areas for improvement. The most consistent shared challenge was trademarks (words tagged with ® or ™), and product names. Both models translated terms that ought to have been kept. Measurement and number formatting were flagged across many languages as well. Both models comprised non-localised units alongside the localised ones rather than replacing them, and did not modify the numbers, not localising thousand separators.

In es-ES, Model B was selected by the annotator nearly twice as often as Model A, (7 vs. 4), although a notably high number of "Equally Bad" in style guide adherence (13) indicates that neither model consistently satisfied the requirements. Literal translations occasionally occurred, e.g. "great value" translated to "expensive," and "walkways" translated as "pasarelas," a term more appropriate for a bridge or a fashion runway than the setting indicated in the source. However, most of the annotated errors were precisely the no conformity to rules, as the language itself was quite positive, and neither of the rules not followed were critical errors.

For pl-PL, it also stands out as a language without "Equally Bad", suggesting that both models produced outputs that the annotator deemed acceptable. Both pl-PL and pt-PT had a preference in translation quality for Model A or was rated in pair with B (Δ = -0.02 and Δ = -0.18, respectively). The de-DE accompanies this trend, but with a larger distinction between the models, with A demonstrating a higher overall adherence. However, both models had register issues, occasionally defaulting to formal structures when an informal tone was required.

For da-DK, it shows identical mean scores for both models in translation quality (2.21), with A having a preference for adherence, showing the uncertainty of the annotator for this language. It had the most individual errors, with major mistranslations in Model A ("gingerbread man" translated as "pejsmand") and cross-language interference in Model B due to Norwegian spellings. In contrast, ja-JP was a balanced language, with both models assessed similarly in 64% of situations, presenting mostly compliance difficulties largely limited to trademark and measurement norms that appeared consistently across assignments. However, these languages recorded the highest number of "Equally Good" in adherence (22 and 18, respectively), indicating satisfaction with both models.

The models performed well more often than not. Taking into account the responses in which both models were rated good (30.6%), Model A showed positive results in approximately 50% of cases (19.9% + 30.6%), while Model B reached around 62% (31.1% + 30.6%). In contrast, responses rated as "Equally Bad" represented 18.4% of the cases, indicating that clear failure was the least common outcome.

Across LPs, the two models show closely marched quality scores, with differences of less than 0.32 points in all cases. PL, pt-PT, and es-ES came out as the strongest performers, with both models achieving mean scores above 3 on a 5-point scale. DA consistently presents the lowest quality scores, which may reflect the relative scarcity of training data.

The co-occurrence of high quality and high "Equally Good" in pl-PL and pt-PT suggests that the style guides did not negatively impact translation quality, maintaining a high baseline of acceptability.

The results also show substantial cross-lingual variation in both adherence and quality, which constitutes the most salient finding of this evaluation. This variation likely reflects differences in the complexity and specificity of each language style guide, the typological distance between source and TL, and the relative volume of language-specific fine-tuning data available to each model. These findings motivate future work on the formulation of language-aware style guides and model selection strategies tailored to individual TL.

5 Conclusion and Future work

This research presents an exploratory evaluation of style guide information extraction through GPT 4o, and application as system prompts to MT translation across seven LPs, comparing two Tower models under a qualitative analysis annotation.

Building on the three-tier approach proposed in (Graça, 2025) and discussed in 2 — which itself yielded positive results — the present work introduces a fourth step that both extends and reinforces these findings. Together, they offer converging evidence that style guide prompting is a viable option

to improve translation quality. Model B demonstrated a consistent, if moderate, advantage in adherence across most languages examined. That said, the cross-lingual variation observed in the results serves as an important reminder that its effectiveness is shaped by language-specific factors, including the availability of training data, the structural complexity of the TL, and the degree of specificity of the style guide itself.

The main limitation of this study was the use of a single annotator per language, a constraint imposed by time and budget considerations. While this allowed us to conduct a broad, cross-lingual evaluation, it means that inter-annotator agreement was not assessed and that bias may influence the results. Future work will incorporate multiple annotators per language, and apply standard inter-rater measures to ensure better evaluations. Additionally, the evaluation was conducted in one text domain (product description), which limits the generalisability of the findings. Style guide compliance and model behaviour may differ across domains or other languages not tested here.

This work was a first step toward a systematic understanding on how to best optimise client style guides through a more automatic approach, extracting its information through a script that proved itself successful, and of how style guide can be operationalised as system prompts in LLM-based pipelines.

Due to the results of this work, this study is now part of a larger ongoing research that examines style guides across multiple client domains and text types. Subsequent evaluations will involve a pilot that will begin with multiple clients from different industries, using diverse style guides and source material. These studies will allow the assessment of whether the observed patterns generalise beyond the description of the product or not, and to assess the liability of this application in scale to the translation pipeline.

Together, these efforts aim to provide a more comprehensive picture of the conditions under which style guide system prompts are effective for production translation workflows.

6 Acknowledgments

This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI); Fundação para a Ciência e a Tecnologia (FCT), through Portuguese national funds under projects UID/50021/2025 (DOI:https://doi.org/10.54499/UID/50021/2025) and UID/PRR/50021/2025 (DOI: https://doi.org/10.54499/UID/PRR/50021/2025) and by CLUL, UID/214/2025 (https://doi.org/10.54499/UID/00214/2025)

References

American Translators Association. 2019. The whys and hows of translation style guides. A case study. *The Savvy Newcomer*. Accessed February 10, 2026.

Aranbure, Yaiza, Silvia Arenas, Javier Augusto, Ana Azuaje, Naiara Ortega, and Sonia Solana. 2018. The use of style guides in technical and scientific translations.

Ghosal, Pamela. 2024. Why you should work with a translation style guide. *Phrase*. Accessed February 10, 2026.

Graça, Leonor. 2025. Contextual parameters in domain-specific translation: A three-tier model. Master's thesis, School of Arts and Humanities, University of Lisbon, Lisbon, Portugal. `http://hdl.handle.net/10400.5/118388`.

Heo, Juyeon, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley Ren, Udhay Nallasamy, Andy Miller, and Jaya Narain. 2025. Do llms "know" internally when they follow instructions?

International Organization for Standardization. 2015. ISO 17100:2015 Translation services — Requirements for translation services. Technical report, ISO. Accessed February 18, 2026.

Jang, Joel, Seonghyeon Ye, and Minjoon Seo. 2022. Can large language models truly follow your instructions? In *NeurIPS ML Safety Workshop*.

Jaroslawicz, Daniel, Brendan Whiting, Parth Shah, and Karime Maamari. 2025. How many instructions can llms follow at once? `https://arxiv.org/abs/2507.11538`.

Lingualinx. 2024. The crucial role of style guides in language translation. *News and Updates*. Accessed February 10, 2026.

Rei, Ricardo, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms.

The Business Research Company. 2026. Ai in lnaguage translation market report 2026. Technical report, The Business Research Company.

Young, Richard J., Brandon Gillins, and Alice M. Matthews. 2025. When models can't follow: Testing instruction adherence across 256 llms.