Watching the process unfold: using API-based interaction data to understand student use of AI in translation tasks

JUDITH RAIGAL-ARAN and NUNE AYVAZYAN

Universitat Rovira i Virgili, Spain

This article describes a methodology to automatically create translation memories for subtitling, using translated books adapted into films and recognizing extra-linguistic markers to differentiate character interventions from narration. This methodology includes the automatic identification, extraction and alignment of the dialogues. The aligned bi-texts served as translation memories in the subtitling of the adapted films. Results show an overall 95% extraction rate for English dialogues and 85% for Spanish dialogues. Alignment showed an accuracy of 90%. Results for the translation memory performance showed that hits between 70% and 100% matches accounted for 15% of the corpus. The results reinforce the claim that dialogues in books can be used as reference material for the translation of subtitles.

Keywords: audiovisual translation, translation memories, subtitling, natural language processing

Introduction

Anthony Pym has made a sustained and influential contribution to both translator education and the pedagogical use of translation as a means of learning languages. Over the years, his work has helped shape how translation is taught, not only as a professional skill but also as a valuable communicative practice within language learning (Pym et al. 2013; Pym and Ayvazyan 2017). Across his work on translator training, Pym has consistently argued that translation skills should be complemented with wider translator competence, including risk management, critical reflection and trust (Pym 2015; 2025). His research on translation and language learning has further highlighted how translation tasks can be used to develop language awareness and intercultural competence, challenging rigid binary separations between “language

teaching" and "translation teaching" (Pym et al. 2013) and calling for empirically grounded accounts of classroom practice (Pym 2011). Pym has also underscored the need to integrate emerging and evolving technologies into pedagogy in ways that are both exploratory and critically informed (Ayvazyan et al. 2024). The rapid diffusion of large language models (LLMs) in translation studies extends these concerns into a new technological moment. Pym's recent work with Yu Hao on generative AI (2025) explicitly frames such tools as potential means of augmenting language skills while practicing translation.

The integration of machine translation and, more recently, LLMs into both translator training and language learning has renewed longstanding interrogations about questions such as what competences should be cultivated, how technology mediates learning and what role risk, trust and critical reflection play in the training of translators, questions that were already central in the EMT 2022 competence framework.

These questions have become increasingly urgent, yet empirical research into how students actually use AI during translation tasks remains limited. Much of the existing evidence on generative AI use in translator training is based on self-reported data such as surveys, interviews or post-hoc reflections (*cf.* Almahasses et al. 2024; Belhassen and Hamda 2025; Bouyzourn and Birch 2025) which, while valuable, cannot capture the granular, process-oriented dimension of AI-mediated practice: how students formulate queries, revise prompts, selectively accept or reject suggestions and negotiate with the system across iterative exchanges. This gap between students' accounts of their AI use and the actual texture of that use constitutes a significant methodological challenge for the field.

This study responds to this challenge by reporting on an experiment designed to move beyond self-report studies towards the systematic analysis of interaction logs. By situating our study within the concerns that have shaped Pym's research trajectory—technology, risk and trust applied to translator training and language learning—we seek to contribute to an empirically grounded understanding of how generative AI is being incorporated, resisted and renegotiated in translator and language training classrooms.

Literature review

Process-oriented research on AI-assisted translation tools such as Google Translate and DeepL has historically employed methodologies such as eye tracking, keylogging to capture fine-grained data on user behavior, as well as think-aloud protocols (Muñoz and Rojo 2025). These approaches make it possible to examine attention patterns, revision sequences and decision-making processes as they unfold in real time. Eye-tracking studies reveal how

users allocate visual attention across source texts, machine output and external resources, keylogging provides detailed records of text production and editing, while think-aloud protocols offer access to participants' verbalized reasoning. Together, these methodologies try to shed light on the cognitive processes underlying translation decisions, while these remain largely inaccessible and can only be inferred indirectly through such observable data.

The analysis of student prompting strategies represents a new avenue for process-oriented research; as such, relatively few studies have examined actual interactions through prompts. Zhang et al. (2025) studied student prompting strategies for five translation-related tasks: understanding, transfer, documentation, revision and analysis. The data revealed that the students had used generative AI mostly for transfer and revision. The students also used the tool to obtain explanations of domain-specific terms, proper names and background information. With regard to prompting behavior, approximately half of the students produced a single prompt, whereas the remaining participants engaged in multi-turn interactions with the system. Further, most prompts were expressed in the imperative but also included requests and suggestions. Similarly, Su et al. (2026) found that students tend to adopt direct prompting strategies when interacting with general-purpose generative AI tools, often requesting complete answers or executing straightforward commands rather than engaging in more exploratory or reflective forms of interaction.

Some studies have tried to examine trust in generative AI through surveys. Bouyzourn and Birch (2025) surveyed 115 university students and found generally high levels of trust in ChatGPT. Frequent use increased trust, while greater technical understanding reduced it. Translation was considered moderately reliable and only 8.7% of participants expressed negative views of the tool. Almahasees et al. (2024) surveyed 102 English students and found generally positive perceptions of ChatGPT across several dimensions. Students reported high trust in its translations and believed they preserved source meaning. Many participants also expressed trust in the system's ability to maintain the confidentiality of their translated work. Belhassen and Hamda (2025) surveyed 150 students and found high but conditional trust in AI-generated translations. Most participants considered outputs accurate (83.3%) and expressed trust in AI tools (66.6%). However, a third noted limitations in handling complex tasks requiring creativity, cultural awareness, or deeper understanding.

To our knowledge, no study has yet examined prompt formulation behavior among translation students working specifically with legal texts. This chapter seeks to address this gap.

Methodology

The data reported in this article were collected during a single in-class session in November 2025 as part of *Direct Translation I*, an elective course taken by third- and fourth-year students enrolled in a bachelor's degree in English. Although 19 students were present on the day of the experiment, the analysis includes only the 16 participants who carried out the task successfully (completed both pre and post questionnaires). The study received clearance from the Ethics Committee for Research on People, Society and the Environment at Universitat Rovira i Virgili, with the approval code CEIPSA-2024-PR-0012.

The course is one of two translation courses in the degree, alongside the mandatory third-year course *Introduction to English Translation*. Students are generally expected to complete the mandatory course where foundational translation skills are taught *before* enrolling in the elective. So, the students had some prior training in translation. The elective course *Direct Translation I* incorporates components of specialized translation, including technical, academic, economic and legal domains. In the course, students are required to translate from English (their L2/L3) into their preferred L1, Spanish or Catalan. Legal translation is generally regarded by students as particularly demanding due to its reliance on system-specific, dense and often non-transparent terminology and the potential consequences of interpretative errors. For this reason, a legal text was selected for the experiment, as it was expected to elicit a greater number and variety of student queries related to meaning negotiation, term selection and reformulation strategies when interacting with the GPT API.

The legal text provided was a standard force majeure clause in English (189 words), part of an agreement. It contained several translation problems primarily related to terminology (e.g. "act of God"), expressions that are inconsistently interpreted across legal systems (e.g. "best efforts") as well as references to applicable law (e.g. "Cal. Com. Code § 2-615"). The full text can be consulted in Section 4, Results.

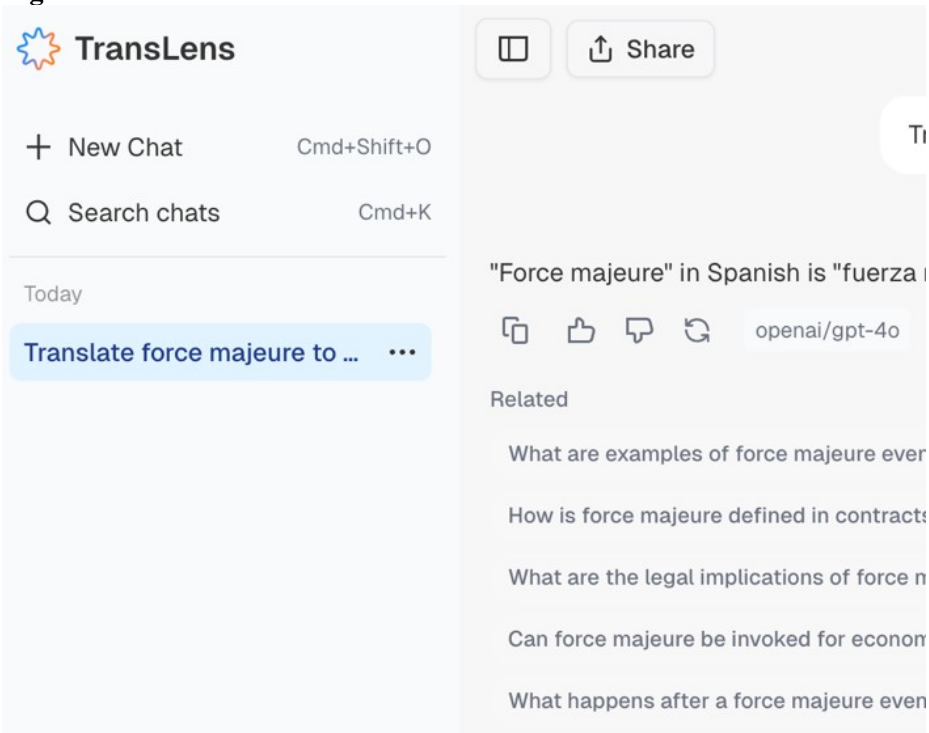
The session was designed as a timed individual translation activity which lasted 70 minutes, structured around four sequential steps that were explained to the students prior to the experiment. The four steps consisted of a pre-task questionnaire, translation, a post-task questionnaire and a final class discussion of the experiment. The study aimed to answer the following research questions:

- Q1. What interaction and revision patterns can be observed in students' use of generative AI?
- Q2. How do students formulate prompts when using generative AI for legal translation? and
- Q3. Does students' trust in AI-generated translations increase after the task?

The students began by completing a pre-task questionnaire designed to capture their prior experience with AI translation tools and their expectations regarding the task ahead. They were then asked to translate a 189-word legal text from English into either Spanish or Catalan, both working languages of our students. To contextualize the translation task, students were provided with a translation brief: a company based in Tarragona was preparing to sign a service agreement with a Californian company and required a translation of the document in order to assess several key clauses before signing. This framing was intended to situate the task within a recognizable professional scenario and to provide a communicative purpose for the translation.

For the translation step, the students were encouraged, though not required, to use GPT-4o as part of their workflow. Access to the tool was provided through a dedicated account linked to a paid API subscription set up specifically for this activity. Each student was assigned an individual anonymized username and password, a configuration that served a dual purpose: it allowed students to access a version of the tool with greater capacity than the free tier and it enabled the researchers to retrieve the complete interaction logs associated with each account for subsequent analysis. A screenshot of the tool interface can be seen in Figure 1.

Figure 1: Screenshot of the tool interface.



The tool, designed to resemble the interface of ChatGPT (or any chat-based assistant), was specifically developed for this activity and enabled the systematic recording of all interactions between students and GPT-4o. These interactions were subsequently made available for retrieval in the form of an Excel file. The dataset includes columns for anonymized user identifiers (created expressly for the activity), conversation id, message id, timestamps corresponding to each message (day and hour), message authorship (user or assistant) and the message content itself.

The students were informed that the session data would be recorded and used for research purposes and were instructed not to include any personal information in their queries. They were also free to consult any other resources they considered relevant, meaning that AI use was positioned as one option among several rather than as a mandatory tool. Screen activity was not recorded; instead, students self-reported their use of additional resources in the post-task questionnaire.

Upon completing their translation, the students responded to a post-task questionnaire. This second questionnaire aimed at gathering post-task reflections on their use of the tool, the difficulties encountered and their assessment of the translation they had produced. The class then finished with a 20-minute discussion where the students discussed the activity.

This design allowed to triangulate four different data sources: the pre- and post-task questionnaires, the final translations submitted by students, the full logs of student-AI interactions captured through the API and the class discussion. In this article, we report on the data we obtained from the student-AI interactions through the API but also address key questions on trust in the pre- and post-task questionnaires. Due to the space constraints of the present article, the remaining data will be presented in a separate, related article.

Results

This section presents the results obtained from 16 students. Within the dataset generated by the GPT API interface, the participants were labeled as "users." For consistency reasons, the participants are therefore identified as "users" followed by their numerical identifier as assigned by the GPT API, while the term "students" is employed in the narrative when referring to the cohort more generally.

Interaction counts, time-on-tool and time-on-task

Interaction data was extracted from the conversation logs. We computed each student's number of user-initiated messages (i.e., prompts sent to GPT), total messages exchanged, session start time (first prompt) and end time (last GPT response).

Combining student-initiated prompts and GPT responses, the total number of messages exchanged per student ranged from 2 to 24 (M = 8.12 exchanges), with a grand total of 132 messages across the full sample (see Table 1). The distribution was markedly asymmetric and polarized into three clusters. Eight students (50%) exchanged four messages or fewer, equivalent to one or two conversational turns. This suggests a largely transactional use of the tool. A middle cluster of four students (user12, user13, user14, user15) exchanged six messages each, consistent with a brief but slightly more iterative engagement. At the upper extreme, two students engaged in 16 interaction turns (user11 and user16) and two students (user2 and user4) accounted for 48 of the 132 total messages (36.36% of the entire sample's conversational output), with 24 messages each.

Table 1: Interaction counts, time-on-tool and time-on-task

Student	Interactions (n)	GPT session duration	Total task time
user1	6	11m 25s	43m 41s
user2	**24**	21 m 41s	51m 12s
user3	6	11m 10s	36m 6s
user4	**24**	10m 53s	40m 16s
user5	6	2m 7s	32m 53s
user6	2	0m 7s	36m 29s
user7	6	4m 40s	35m 3s
user8	2	0m 10s	39m 13s
user9	2	0m 8s	35m 38s
user10	4	**23m 13s**	(no pre-task)
user11	16	7m 20s	38m 56s
user12	4	**22m 44s**	33m 47s
user13	4	0m 11s	32m 30s
user14	4	2m 42s	49m 2s
user15	4	2m 58s	33m 52s
user16	16	8m 53s	21m 9s
Total	132	-	-
Mean	8.12	8m 9s	37m 34s

The total interaction time of each participant with the tool was calculated using the timestamp of the participant's first input and the timestamp of the

assistant's last response. This ranges from as little as seven seconds (user6) to over twenty-three minutes (user10), with a mean of 8 minutes and 9 seconds.

Total task duration, which is the time spent on the pre-task questionnaire, the translation task itself (including GPT interaction) and the post-task questionnaire, measured from the start of the pre-task questionnaire to the completion of the post-task questionnaire, ranged from 21 to 51 minutes (M = 37 minutes and 34 seconds). All the participants completed all the stages of the experiment, except for user10, who did not complete the pre-task questionnaire.

What is interesting here is that the time spent on the GPT API does not seem to reflect the number of messages: user2 spends 21m 41s on the tool to interchange 24 messages, while user4 only needs 10m 53s for the same number of interactions. On the other hand, user10 and user12 use roughly the same amount of time on the tool as user2, but with only four interactions each. Thus, time spent on the tool does not necessarily indicate the amount of interaction. Further, the quality of the final text cannot be inferred from interaction volume alone. Interaction metrics must be interpreted in relation to qualitative aspects of use, such as the complexity of prompts and the extent of revision, together with the final text quality.

The time logs also show that several participants (e.g., user6, user8, user9) interact with the system for only a few seconds, yet their total task time remains comparable to that of other students. This suggests that a substantial portion of the translation process takes place outside the GPT interface, which may include reading, revising, or evaluating the generated output.

Prompt patterns and query strategies

Half of the students (n = 8) wrote their prompts in Spanish, their L1 and the target language of the translation, while the other half chose English, the source language of the text (see Table 2). This split suggests that students did not treat prompt language as a principled choice aligned with the translational direction of the task.

In terms of syntactic form, imperative (e.g., "Translate this text into spanish" [sic], as expressed by user7) constructions dominated (n = 9, 56.3%), while only two students framed their initial query as a question (e.g., "Can you translate [...]", as expressed by user11). The very limited use of interrogative forms further reinforces this tendency, as few students framed the interaction as a request for assistance or dialogue. Instead, prompting was largely reduced to issuing instructions, which may constrain opportunities for iterative refinement or exploratory engagement with the system. The register was markedly direct: only one student greeted the system ("hey chat", user1) and only one used "please" (user3), which indicates that most participants approached the tool as a utility rather than a conversational agent.

All 16 students who used GPT requested a translation of the full text at some point during the session. Most included the source text in their initial prompt (n = 15, 93.75%) and approximately two thirds specified the target language (n = 10, 62.5%) or provided a translation brief (n = 9, 56.25%), which was actually the commission given to them in the activity instructions. Very few mentioned the source language explicitly (n = 2, 12.5%) and only five students added instructions beyond a simple directive to translate. For example: "Traduce el siguiente texto jurídico" [translate this legal text] (user11), "Traduce este fragmento de un contrato, en el que se especifican las causas de fuerza mayor y como [sic] esto afecta a dicho contrato, del inglés al español" [Translate this excerpt from a contract, which specifies force majeure circumstances and how they affect the contract, from English into Spanish] (user14).

Table 2: Initial prompt characteristics

Variable	Category	n	%
Language of prompts			
	Spanish (L1)	8	50%
	English (L2)	8	50%
Prompt form			
	Imperative	9	56.3%
	Question	2	12.5%
	Teacher-provided default instructions (unmodified)	5	31.3%
Initial prompt content			
	Requested translation of the full text	15	93.75%
	Included the source text	15	93.75%
	Included the translation brief	9	56.25%
	Specified target language	10	62.5%
	Specified source language	2	12.5%
	Added other instructions	5	31.25%

To provide a clearer illustration of a fully developed prompt, the following example shows how one student formulated their request (in italics) by asking the tool to translate the text and by including both the source text and the brief outlined in the instructions:

> Please help me with the Spanish translation of the following text. Take into account possible specific terms and keep the same tone and language: "6.1 In case of earthquake, typhoon, flood, fire, war or unavoidable force majeure events (including but not limited to act of God, strike, riot, act of war or outbreaks of infectious diseases), and

> thereby causing direct impact on the performance of this Agreement or this Agreement cannot be performed according to the terms agreed, the party who encounters the aforesaid force majeure event shall immediately inform the other party, provide written report on details of force majeure within 5 business days after the occurrence of force majeure event, and submit valid supporting documents. Based on the event's degree of impact on agreement performance, all parties may decide whether or not to exempt from performing the obligations of this Agreement, or delay the performance of this Agreement. The performance affected by such force majeure event may be delayed or excused under Cal. Com. Code § 2-615. Neither party may propose a claim for compensation for the loss caused by force majeure. In addition, the parties shall use best efforts to mitigate the impact of the force majeure event and to resume performance of this Agreement as soon as reasonably possible. Pay special attention to these guidelines: Descripción del encargo. Una empresa con sede en Tarragona tiene previsto firmar un contrato de intercambio de datos y servicios con una empresa californiana. Antes de la firma, la empresa de Tarragona ha identificado dudas significativas sobre varias cláusulas clave, como la responsabilidad, las condiciones de resolución de disputas, la protección de datos personales, la fuerza mayor y la ley aplicable. Para tomar una decisión informada y minimizar riesgos, la empresa te solicita una traducción del documento.

The student's prompt demonstrates a structured strategy that combines task definition, contextualization and guidance. The request includes both the full source text and the translation brief, which situates the task within a professional context and may support more appropriate output. The reference to tone and terminology indicates awareness of domain-specific challenges. However, the instructions remain somewhat general, as no concrete constraints or priorities are specified other than the reference to the brief.

Reformulation behavior (see Table 3) was less common but qualitatively significant. Three students asked GPT to revise the full translation (user2, user7 and user11), though only user11 did so specifying that it was a legal text. Three others requested reformulation at the level of individual terms or fragments (user4, user7 and user11) and four students introduced targeted changes to punctuation or word choice. Perhaps most noteworthy is the behavior of three students who explicitly challenged GPT's translation decisions. They questioned, for instance, the capitalization of "Agreement" or the choice of a specific equivalent (translation of "acts of God" as "actos de Dios"). One of the students (user4) asked GPT to implement each change directly in the chat, effectively using it as a text editor and delegating the writing process itself, thereby turning GPT into an active co-editor.

Table 3: Reformulation behavior

Types of revision requests	n	%

Requested reformulation of full translation	3	18.8%
Reformulation included specific instructions	1	6.3%
Requested reformulation of a term or fragment	3	18.8%
Requested specific changes (punctuation, terminology, etc.)	4	25.0%
Challenged GPT's translation decisions	3	18.8%
Introduced changes from an external source	2	12.5%

When it comes to terminology-specific queries (see Table 4), seven students did at least one terminology-related query (user12 and user14 did one, user10 and user 13 did two, user4 did three, user16 did 4 and user11 did six). Across these seven prompts, the users engaged GPT as a combined translation, revision and editing tool, asking for legal terminology equivalents (e.g., "performance", "party" and references like "Cal. Com. Code § 2-615"), questioning translation choices and consistency (especially capitalization and term selection) and issuing direct instructions to modify the text (such as removing commas, replacing words, or inserting translated segments).

Table 4: Terminology queries

Presence of query	n	%
At least one terminology query	7	43.8%
No terminology queries	9	56.3%

A comparison with Table 1 shows that, unsurprisingly, most students with more interactions also produced a greater number of queries: user4 with a total of 24 interactions and user11 and user16 with 16 interactions each. Interestingly, user2, who also engaged in 24 interactions and spent roughly double the time spent by user4 on the GPT API, had no terminology queries. These findings suggest that while a higher number of interactions tend to correlate with a greater number of queries, this relationship is not straightforward. The case of user2 demonstrates that extensive interaction and longer time spent on the tool do not necessarily translate into deeper or more focused engagement with specific translation challenges such as terminology.

Acceptance and modification of AI-generated output

Table 5 presents the number of modifications (if any) per student. Of the 16 students, nine (56.25%) made no changes to the text provided by GPT and submitted the AI-generated text without any modifications (100% similarity). The seven remaining students did edit the text (between 7 and 46 changes). It is noteworthy that during the post-task discussion, one student reported consulting GPT but ultimately relying on a machine translation tool to complete the translation.

The students who revised the text made such changes as lexical substitutions (selecting different Spanish equivalents for English terms), structural reformulations and, in some cases, additions of words not present in the GPT output. The student with the highest modification rate (user13, 29.1%) only had a single query but consistently reformulated terminology and restructured several clauses, suggesting an active and critical post-editing process that unfolded outside the GPT interface. At the other extreme, five students submitted only a single query to the tool and made no changes.

Table 5: Modification of AI-generated output

User	Words in GPT output	Words in student text	Changed words	% changes	Interactions (n)	GPT session duration
User1	235	238	**45**	**19.1**	6	11m 25s
User2	231	231	**17**	**7.4**	**24**	21 m 41s
User3	234	234	0	0	6	11m 10s
User4	242	242	0	0	**24**	10m 53s
User5	193	239	**46**	**23.8**	6	2m 7s
User6	241	241	0	0	2	0m 7s
User7	219	219	0	0	6	4m 40s
User8	229	229	0	0	2	0m 10s
User9	227	227	0	0	2	0m 8s
User10	240	239	**32**	**13.3**	4	**23m 13s**
User11	238	238	0	0	**16**	7m 20s
User12	238	237	7	2.9	4	**22m 44s**
User13	220	236	**64**	**29.1**	4	0m 11s
User14	241	247	26	10.8	4	2m 42s
User15	241	241	0	0	4	2m 58s
User16	240	240	0	0	**16**	8m 53s

As can be seen in Table 5, there is no consistent relationship between interaction metrics and revision behavior. For example, user2 engaged in 24 interactions and spent over 21 minutes on the tool but introduced only limited changes (7.4%), while user13 made the highest number of changes (29.1%) with only four interactions and almost no recorded time on the GPT interface. Similarly, several users with multiple interactions (e.g., users 4, 11 and 16) made no changes at all. In the case of these users, their interactions dealt with local adjustments to punctuation, capitalization ("Acuerdo"), specific terms (e.g. "días laborables" vs. "días hábiles") and the explicit reference to the *"Cal. Com. Code § 2-615."* User14 similarly concentrated on micro-level choices such as the translation of "party" as "partes" and the phrasing of "as

soon as reasonably possible", without prompting the system for substantially different versions of the clause and not changing the initial output.

This indicates that neither interaction frequency nor time spent on the tool reliably predicts the degree of post-editing. The data point to a potential tendency toward acceptance of GPT output, which, in light of the nature of the text, is particularly worrying: the prevalence of zero-change submissions suggests that for many students, GPT functions less as a draft to be critically revised and more as a final product to be adopted with minimal intervention.

Translation of "act of God"

Students were asked to share their translations of specific translation problems in the post-task questionnaire. One of them was "act of God." Their translations fall into three categories. The vast majority of students (13, 81.25%) opted for a literal calque, rendering the term as "acto(s) de Dios" (though with minor variation in number and capitalization: eight used the standard lowercase plural "actos de Dios", two capitalized it as "Actos de Dios", one used the singular "acto de Dios" and one wrote "actos de dios", all lowercase). Only two students (12.5%) chose the conventional Spanish legal equivalent "fuerza mayor": user2 rendered it as "casos de fuerza mayor" and user3 as "eventos de fuerza mayor", while a single student (6.25%), user7, produced a descriptive paraphrase ("fenómenos naturales inevitables" [unavoidable natural phenomena]).

Looking back at the data, the three students who diverged from the literal calque show quite different interaction profiles. User3 ("eventos de fuerza mayor") did ask specific terminology questions. This student explicitly directed GPT to *change* "actos de Dios" to "eventos de fuerza mayor" mid-session. This means the non-literal translation was not the result of GPT's initiative or of a genuine exploratory question—the student already knew (or had searched it somewhere else) the functional equivalent and simply instructed the tool to apply it. User7 ("fenómenos naturales inevitables") did not ask specific terminology questions, but asked GPT to *look deeper into the sentence "including but not limited to act of God"*—a meta-level reformulation request rather than a direct terminological query. This produced a naturalized, descriptive phrase, but not a legally grounded one, which suggests that open-ended prompts of this kind lead GPT towards paraphrase rather than towards established legal equivalence.

Self-assessment of legal translation competence, trust and risk awareness

The pre-task questionnaire indicated that the students had low confidence in their legal translation competence, with a mean score of 1.94 (n = 16) on a 1–5 scale. This is consistent with the profile of our undergraduate students with limited prior experience in legal translation. After the task, the mean score

modestly rose to 2.13 (n = 16). Only three students (18.75%) reported an increase in self-confidence and thirteen students (81.25%) remained at the same score. No student reported a decrease in self-confidence following the activity.

Similarly, the students reported a mean confidence in GPT's legal translation competence of 2.125 (n = 16). After completing the task, mean confidence in GPT rose to 2.75 (n= 16). The distribution shifted markedly toward the midpoint: ten students (63%) assigned a score of 3 ("medium confidence"). Three students scored 1, one scored 2, one scored 4 and one student (user15) assigned the maximum score of 5.

The written justifications that students provided for their post-task trust scores—both in relation to their own capacity and to ChatGPT—allowed us to identify three patterns.

The most repeated pattern was the difficulty in evaluating the quality of the output. Various students recognized that they did not have enough domain knowledge to know if GPT's translation was correct or not. For example, user4 wrote that "I think it has done an acceptable job, but I don't know to what extent it is a good translation or not" and user9 expressed something similar: "I still lack the knowledge to judge how it did it." These students were conscious of their limitations, but this consciousness did not help them to act in a different way.

A second pattern was related to the impact of AI use in the learning process. User11 reflected that the fact of delegating the translation to the tool had impeded real learning: "I did not actually work on the translation—it was done by the artificial intelligence. Therefore, I have learned absolutely nothing that could serve me in a future legal translation task." This is the unique case in which a student connected the use of AI directly to a loss in their own formation as a translator. In line with this reported no trust in GPT for translating legal texts and stated that, although ChatGPT had been useful for completing the exercise, they still had no confidence that an AI system could translate any text correctly or consistently without human intervention.

The third pattern was what we could call conditional trust, which is not put into practice. Some students said that the ChatGPT output was acceptable as long as it was revised by someone with knowledge, but these same students submitted the text without making any modifications. User15, for instance, wrote that the translation "could perfectly be used for a legal text, as long as there is post-translation editing", despite having changed zero words. This contradiction between what the student says and what the student does is especially relevant from a risk perspective, because it shows that having the correct procedural knowledge is not sufficient if the student lacks the competence to apply it.

Discussion

The findings reported in Section 4 (Results) can be understood on two levels. On the surface, they show some behavioral patterns in how undergraduate students of English used a generative AI tool during a timed legal translation task. On a deeper level, they relate directly to concerns that have influenced Pym's work on translator training, such as competence in a broader sense, the management of risk and uncertainty and how trust is adjusted in professional and technological contexts. The following discussion looks at each of these themes in relation to the research questions and considers what the methodological approach used in this study—based on API interaction logs instead of self-report—allows us to see in a new way.

Q1 asked what interaction and revision patterns can be observed in students' use of generative AI. The interaction data resist any linear interpretation that connects the number of interactions with the degree of critical engagement. The distribution of messages—polarized between a majority with two to four exchanges and two students with twenty-four each—might, at first glance, suggest that the more active users were also the more reflective ones. However, user2, who produced 24 messages and spent over 21 minutes on the tool, introduced only limited changes in the final text (7.4% of words modified), whereas user13 made the highest proportion of changes in the sample (29.1%) with just four interactions and almost no recorded time on the GPT interface. Similarly, several users with multiple interactions (e.g. users 4, 11 and 16) submitted the GPT translation without any modification. Thus, interaction frequency does not reliably indicate critical engagement, as both extensive and minimal use of the tool can result in either substantial revision or complete reliance on GPT output.

With regard to Q2 on how students formulate prompts when using generative AI for legal translation, our students formulated prompts in varied but generally simple ways when using generative AI for legal translation. Most rely on direct instructions, often requesting a full translation with minimal specification. A smaller group incorporates contextual information, such as the translation brief, tone, or target audience, resulting in more structured prompts. Some students engaged in iterative prompting, refining outputs through follow-up queries, including terminology questions or requests for reformulation. However, many prompts remained underspecified, with limited attention to legal nuances. Overall, prompting behavior ranged from instrumental, one-off requests to more exploratory and interactive strategies and reflected differing levels of engagement and competence. This echoes recent research showing that student translators' prompting practices tend to be intuitive and uninformed and that explicit AI literacy training is needed to foster more strategic use of GenAI tools in translation tasks (Zhang et al. 2025). It shows that students do not naturally use generative AI in a

strategically informed way, which might highlight a gap between access to the tool and prompting competence.

Regarding Q3, whether students' trust in AI-generated translations increased after the task, the pre- and post-task questionnaires showed that trust both in their own legal translation competence and generative AI competence rose after the task. This can be interpreted as a result of hands-on interaction with the tool, which may have fostered a sense of familiarity and perceived control, regardless of very little student-GPT interaction.

Building on interaction and text modification data patterns identified in Section 4, we identify four user profiles. The largest group consists of instrumental users, who submit a single request for full-text translation, receive the output and deliver it without further interaction or modification. In these cases, the tool is treated as a black-box translation machine and AI output is accepted as a final solution rather than a draft. A second group, referred to as exploratory users, engages in four to six exchanges, including occasional terminology queries or partial reformulations. However, this increased interaction does not result in substantial post-editing, as their engagement remains at the level of consultation rather than intervention. A third, less frequent profile is that of active post-editors, who interact iteratively with the system, question specific decisions and produce translations that differ significantly from the original output. Finally, one participant represents an autonomous user who uses the tool extensively as a parallel resource rather than a drafting aid, producing an independent translation. This profile aligns most closely with the pedagogical goals of critical AI use. These profiles should not be seen as fixed categories, but as context-dependent responses influenced by task difficulty and domain familiarity.

An observation that falls outside the main scope of this study but deserves attention is that no student selected Catalan as the target language, despite it being explicitly offered as an option. The reasons behind this unanimous choice cannot be determined from the available data, but several plausible explanations emerge. These include contextual factors, such as the instructions being provided in Spanish, individual language preferences and, more significantly, a potential lack of trust in GPT's ability to produce high-quality translations into a less represented language.

Perhaps the most significant finding emerges from the comparison of two data sources that, when considered independently, would suggest very different interpretations. The post-task questionnaires indicate that students are aware of the limitations of the tool, as several participants explicitly stated that the text produced by GPT would require expert revision before it could be used in a professional legal context. The interaction data, however, reveal that 50% of the participants submitted the translation provided by the system without any modification. The discrepancy between these two sets of data cannot be explained as an isolated mismatch. It constitutes a systematic

pattern across the sample and perhaps would not have been observable in a study based solely on self-reported data. The "act of God" example shows this competence–trust asymmetry clearly: most students accepted GPT's translation (acto(s) de Dios or a paraphrase) and did not see that it was a serious legal error, even though they had said that such clauses should be checked by an expert. In studies on human-machine interaction, this phenomenon has been described as automation bias, understood as the tendency to follow the outputs of an automated system even when there are reasons to question them (Romeo and Conti 2025). In the context of this study, automation bias might be present not as blind trust in the system, since the students do not necessarily consider GPT to be infallible. Rather, it might take the form of epistemic surrender. Faced with the inability to verify whether the output is correct, the student implicitly treats it as acceptable. Zhang and Doherty (2025) argue that novice students may lack the competence to detect errors in AI translations, which can lead to over-reliance on these tools and hinder the development of critical thinking and creativity. However, students are likely to use generative AI for translation regardless. The key issue is therefore not whether to allow its use, but how to frame it pedagogically. Raising students' awareness of their own limitations, as well as those of the tool, can foster more critical engagement and support the development of translator competence. This possibility opens up avenues for future research, for example, in an experiment consisting of various sessions, where students are allowed to see their results and act upon them.

The data reveals a paradox: far from generating additional caution, the specialized nature of the text appears to have had the opposite effect. Students did not trust GPT less because of the difficulty of the legal text, but rather more. Several participants explicitly acknowledged in the open-ended questions that they were aware that an error in a text of this nature could have serious legal consequences and that the result should be reviewed by an expert. And yet, 50% submitted the tool's translation without any modification. This combination of simultaneous awareness of risk and inaction cannot be explained as negligence; it is the result of a specific competence asymmetry. As several students expressed, they did not feel capable of assessing whether GPT's translation was correct or not. The specialized nature of the text, instead of increasing scrutiny, increased reliance on the tool precisely because it exposed the limits of the students' own knowledge. In other words, the less you know, the more you trust, in this case, the machine (Tully et al. 2025). Trust does not stem from the belief that it performs well, but from the lack of any means to determine whether it is wrong.

Conclusions

A defining feature of our study lies in its reliance on behavioral rather than self-reported data. Instead of focusing on what students claim to have done, the analysis is grounded in what they actually did during their interaction with the tool. Access to interaction logs captured through the GPT API makes it possible to observe the translation process as it unfolds in real time, rather than reconstructing it retrospectively through questionnaires or interviews. This methodological orientation situates the study within the tradition of process-oriented research in translation and enables a more fine-grained understanding of student behavior, alongside studies with eye tracking, keylogging and think-aloud protocols.

The present study is subject to several limitations. First, it does not include a quality assessment of the final translations produced by the students. Such an analysis would have made it possible to determine to what extent the GPT API suggestions were incorporated into the final texts and whether their use had a measurable impact on translation quality. This limitation will be addressed in a separate article, in which the students' final translations will be analyzed in relation to their interaction data. A further limitation concerns the absence of full screen recordings. Such data would have made it possible to identify whether students relied on additional tools during the experiment, thereby offering a more comprehensive view of their workflow and enabling a more accurate interpretation of their interaction with the GPT system. Other limitations are the students' language and educational level. Taken together, these limitations point to the need for further research on interaction data in order to develop a more comprehensive account of how students engage with generative AI in translation tasks.

Taken together, the experiment shows that interaction logs uncover a clear gap between what students say about AI and what they actually do in legal translation tasks: many acknowledge the risks but still submit unedited GPT output. This suggests that GenAI-based tasks must be designed explicitly as learning opportunities that require students to scrutinize and revise AI translations, rather than simply use them as shortcuts. Further work with interaction data in other contexts will be needed to refine this account of how translation students learn to work with generative AI.

Declaration of AI use

In preparing this manuscript, the authors used Claude to reformulate and improve selected expressions. All AI-assisted content was subsequently reviewed, edited and verified by the authors, who take full responsibility for the accuracy and integrity of the final text.

References

Almahasees, Zakaryia, Hussein Abu-Rayash, Samer Naser Olimat, and Dana Mahadin. 2024. "An analytical study on present perceptions of using ChatGPT in language translation". *Language Value* 17(2): 1-23. Universitat Jaume I ePress: Castelló, Spain. http://www.languagevalue.uji.es

Ayvazyan, Nune, Yu Hao and Anthony Pym. 2024. "Things to do in the translation class when technologies change. The case of AI text generation". In Yuhong Peng, Huihui Huang, Defeng Li (eds.) *New Advances in Translation Technology*. Springer.

Belhassen, Saleh, and Achwak Hamda. 2025. "Translation Students' Reliance on and Trust in Artificial Intelligence for Successful Translation Projects: Opportunities, Challenges, and Implications". Arab World English *Journal for Translation & Literary Studies* 9 (2): 106-119. http://dx.doi.org/10.24093/awejtls/vol9no2.7

Bouyzourn, Kadija, and Alexandra Birch. 2025. "What Shapes User Trust in ChatGPT? A Mixed-Methods Study of User Attributes, Trust Dimensions, Task Context, and Societal Perceptions among University Students". *arXiv preprint*, https://doi.org/10.48550/arXiv.2507.05046

EMT. 2022. European Master's in Translation Competence Framework. https://comission.europa.eu/system/files/2022-11/emt_competence_fwk_2022_en.pdf. Visited March 2026.

Muñoz Martín, Ricardo and Rojo López, Ana María. 2025. *Research methods in cognitive translation and interpreting studies*. Amsterdam: John Benjamins. https://doi.org/10.1075/rmal.10.intro

Pym, Anthony. 2011. "Training translators". In Kirsten Malmkjær and Kevin Windle (Eds.) *The Oxford Handbook of Translation Studies*, pp. 475-489. Oxford: Oxford University Press.

Pym, Anthony, Kirsten Malmkjaer, and Mar Gutiérrez. 2013. *Translation and Language Learning*. Luxembourg: European Commission.

Pym, Anthony. 2015. "Translating as risk management". *Journal of Pragmatics*, 85, pp 67-80, ISSN 0378-2166. https://doi.org/10.1016/j.pragma.2015.06.010.

Pym, Anthony, and Nune Ayvazyan. 2017. "Linguistics, translation and interpreting in foreign-language teaching contexts". *The Routledge Handbook of Translation Studies and Linguistic*s. 393-407. 10.4324/9781315692845

Pym, Anthony. 2025. *Risk Management in Translation.* Elements in Translation and Interpreting. Cambridge University Press.

Pym, Anthony, and Yu Hao. 2025. *How to Augment Language Skills. Generative AI and Machine Translation in Language Learning and Translator Training*. Routledge.

Romeo, Giuseppe and Conti, Daniela. 2025. “Exploring automation bias in human–AI collaboration: a review and implications for explainable AI”. *AI & Soc* 41, 259–278. https://doi.org/10.1007/s00146-025-02422-7

Su, Hanyu, Huilin Zhang, and Shihui Feng. 2026. “Comparing the Impact of Pedagogy-Informed Custom and General-Purpose GAI Chatbots on Students’ Science Problem-Solving Processes and Performance Using Heterogeneous Interaction Network Analysis”. https://doi.org/10.48550/arXiv.2604.03022

Tully, Stephanie M.; Longoni, Chiara, and Appel, Gil. 2025. “Lower Artificial Intelligence Literacy Predicts Greater AI Receptivity”. *Journal of Marketing*, 89 (5). https://doi.org/10.1177/00222429251314491.

Zhang, Jia, and Stephen Doherty. 2025. “Investigating novice translation students’ AI literacy in translation education”. *The Interpreter and Translator Trainer*, 19(3–4), pp. 234–253. https://doi.org/10.1080/1750399X.2025.2541478

Zhang, Jia, Xiaoyu Zhao, and Stephen Doherty. 2025. “Prompt engineering in translation: How do student translators leverage GenAI tools for translation tasks”. *Proceedings of Machine Translation Summit XX* 1, pp. 420–431, Geneva, Switzerland. European Association for Machine Translation.