À tradire
ISSN : 2968-3912

3 | 2024
Comment continuer à former aux professions langagières à l'heure de l'intelligence artificielle ?

Reasoned Translation: Putting Neural Machine Translation and Generative Artificial Intelligence Systems to the "Delisle Test"

Intelligence artificielle générative contre traduction automatique neuronale : l'épreuve de la traduction raisonnée « à la Delisle »
Poniendo a prueba la traducción automática neuronal y la inteligencia artificial generativa: ¿pueden razonar como lo enseña Delisle?

Julián Zapata

https://lodelpreprod.univ-rennes2.fr/atradire/index.php?id=513

DOI : 10.56078/atradire.513

Electronic reference
Julián Zapata, « Reasoned Translation: Putting Neural Machine Translation and Generative Artificial Intelligence Systems to the "Delisle Test" », *À tradire* [Online], 3 | 2024, Online since 13 May 2026, connection on 22 September 2026. URL : https://lodelpreprod.univ-rennes2.fr/atradire/index.php?id=513

Copyright
Licence Creative Commons – Attribution 4.0 International – CC BY 4.0

Reasoned Translation: Putting Neural Machine Translation and Generative Artificial Intelligence Systems to the "Delisle Test"

Intelligence artificielle générative contre traduction automatique neuronale : l'épreuve de la traduction raisonnée « à la Delisle »
Poniendo a prueba la traducción automática neuronal y la inteligencia artificial generativa: ¿pueden razonar como lo enseña Delisle?

Julián Zapata

OUTLINE

Introduction
- On "reasoning", or performing discourse analysis in translation
- The reasoned translation test: experimenting with NMT and GenAI
  - Pilot experiments
  - Experimental design
    - Data gathering
    - Assessment of translated segments and marking of the results
- Main experiment data analysis
  - Data from chapters on syntactic difficulties (stage 1)
  - Data from chapters on syntactic difficulties (stage 2)
  - Data from chapters on lexical difficulties (stage 2)
  - Data from chapters on both syntactic and lexical difficulties (stage 2)
  - Instances where all systems passed or failed the test
- Should prompt engineering be integrated into translator training?

Conclusion

TEXT

Introduction

1 *La traduction raisonnée* or "Reasoned Translation" is the title translation scholar Jean Delisle gave to his celebrated pedagogy textbook, which has been used in English-French translator training in Canada and around the world for four decades. With such a title, the premise of the book is straightforward: a trained (human) translator will have the ability to reason; a good translation is one that was reasoned.

2 Delisle's textbook was first published in 1993, with subsequent editions every ten years (2003 and 2013) to which other trainers and researchers contributed as advisors, co-authors or solo authors of some chapters (*objectifs*, in French). Today, over ten years after the 3rd edition was published, it is worth revisiting the book in the current neural machine translation (NMT) and artificial intelligence (AI) era, putting the notion of "reasoned translation" into perspective.

3 In recent times, the field of natural language processing (NLP) has experienced notable progress, primarily attributed to the emergence of large language models (LLMs) such as generative pre-trained transformer (GPT) models, commonly referred to as generative AI (GenAI). These models are trained on extensive amounts of text data and can generate human-like language for various NLP tasks, including language translation (Brown *et al.*, 2020). LLMs have garnered significant attention due to their ability to generate text within seconds based on prompts by the user (e.g. "Write a short story in the style of Molière" or "Give me 10 reasons to attend the 2026 FIFA World Cup in North America"), in addition to their conversational capabilities. This success has resulted in a shift in the field of NLP, with researchers exploring ways to fine-tune these models for specific tasks such as translation (Hendy *et al.*, 2023) and translation assessment (Kocmi and Federmann, 2023).

4 That being said, can NMT and GenAI systems pass what we call the "Delisle test"? In other words, do these systems display any signs of the ability to reason, the way properly trained human translators do when translating from English into French? To explore these questions, we put NMT systems (MS Bing Translator, Google Translate and DeepL) and GenAI systems (ChatGPT, NotionAI and Gaby-T) to the test, using a selection of examples taken from 30 chapters from the 3rd edition of *La traduction raisonnée* (Delisle and Fiola, 2013a and 2013b), analyzing the results in light of the background information and explanations provided in the textbook. Can these systems produce "reasoned" translations? What might the implications of the answer to this question be for the future of translation pedagogy?

5 Table 1 below provides a snapshot of the corpus compiled and analyzed, and the tests performed for this study; more details will be provided in Sections 3 and 4.



Table 1. Total number of segments in the corpus analyzed per condition, experiment stage and section of the book, and grand totals.

Table 1. Total number of segments in the corpus analyzed per condition, experiment stage and section of the book, and grand totals.

6 The remainder of this paper is organized as follows: Section 2 will build upon what inspired Delisle's work and what he means by *traduction raisonnée*; Section 3 will present two pilot experiments and the methodology of our main experiment. Section 4 will focus on the results of the main experiment; and Section 5 will offer some insights into potential research avenues in translation pedagogy and beyond, before the concluding remarks in Section 6.

On "reasoning", or performing discourse analysis in translation

7 Delisle was first inspired by (and borrowed the notion of *reasoned translation* from) French linguist, author, and translator Jean Darbelnet, one of the most important contributors to the field of comparative stylistics (CS) (Delisle and Fiola, 2013a: 18). This field focuses on comparing the ways in which two texts in two different languages structure their information grammatically and rhetorically, conceptualizing comparative features so that, when applied to translation practice, translators are more adept to avoid the impulse of adhering too closely to the expressive form of the source text. Darbelnet applied these observations in a 1969 paper, titled precisely *La traduction raisonnée*, and filled with CS concepts and examples comparing English and French. He suggests that untrained translators are often "victims of the tyranny of form" (Darbelnet, 1969: 7, our translation). For his part, Delisle suggests that:

> Learning to translate is learning to approach a text in a "reasoned" way, to progressively discover all the things involved in the transfer of the meaning of a text from a language into another, a task more difficult than it may appear at first glance (Delisle, 2003: 17, our translation).

8 He also reminds the reader/learner that the textbook is not to be used as a recipe book, and that the example or model translated phrases, sentences or texts in French provided throughout the book are not *the* translations of the corresponding English phrases, sentences, or texts. They only go to show that "the solutions to every translation problem are multiple" and "always depend on context" (*ibid.*, our translation).

9 Delisle's earlier (doctoral) work, which later developed into *La traduction raisonnée*, proposed "discourse analysis as a method of translation" (Delisle, 1988) focusing on the complexity of the intellectual mechanisms involved in translation. Discourse analysis is crucial for understanding complex texts. It helps reveal the author's intent and assists translators in making informed decisions. By examining the discourse surrounding the text, translators can identify key themes, rhetorical strategies, and cultural references that are essential to understanding the original text and objectivize its translation. But what is *discourse*? Shi-xu presents this concept as:

> a set of diversified and competing constructions of meaning associated with particular groups of people [...] [It is] a construction of meaning—representing and acting upon reality—through linguistic means in concrete situations. It is thus a unity of both form and meaning. And it is not merely a form of talking or writing, but also a way of thinking (2005: 1).

10 In other words, discourse is a complex process that involves language and thought, as well as the social and cultural contexts in which communication takes place. Moreover, to define *discourse analysis*, Phillips and Hardy (2002: 4) explain first that discourses are "embodied and enacted in a variety of texts," which may take a variety of forms (written texts, spoken words, pictures, symbols, artifacts, etc.), and that texts are not meaningful in isolation: "[I]t is only through their interconnection with other texts, the different discourses on which they draw, and the nature of their production, dissemination, and consumption that they are made meaningful." Thus, *discourse analysis* is the practice of exploring:

> how texts are *made* meaningful through these processes and also how they contribute to the constitution of social reality by *making*

> meaning [...]. Discourse analysis is thus interested in ascertaining constructive effects of discourse through the structured and systematic study of texts (*ibid.*; emphasis in the original).

11 Translation studies (TS) scholars such as Delisle and others (e.g., Brisset, 2010; Munday and Zhang, 2017; Schäffner, 2004; Zhang *et al.*, 2015) have unsurprisingly been interested in discourse analysis for decades, since there is much the discipline can learn from the structured and systematic study of texts. Discourse analysis in TS has been revealing of contextual factors; linguistic features; pragmatics; intertextuality; ideology and power relations; and discourse communities; all of which helps to objectivize the translation process. By applying discourse analysis to the act of translation, translators can identify strategies for producing top-quality texts in the target language—an ability that Delisle seeks to develop among translators-in-training who carefully study *La traduction raisonnée*. In a review of the 2013 edition of the book, Kumbe concludes:

> [Jean Delisle] and his collaborators deal with questions relating to the reasoned approach of the translator, as well as linguistic and stylistic differences. Incorporating this knowledge into translator training programs will alert future translators to the pitfalls of translation and, if the learning does happen, translators are able to avoid unnecessary trial and error (2016: 736, our translation).

12 The goal of the reasoned-translation method is thus to cultivate a reflex in future translators that promotes a conscientious approach to translating, analyzing the discourse, and avoiding obvious shortcuts (syntactic or lexical calques, the choice of the first equivalent word provided by a bilingual dictionary, etc.; more on this in our experimental data below). Instead, trained translators will reason to get to the message and reformulate that message idiomatically in the target language —in this case French— and using proper terminology. Ultimately, they can contribute to preserving the structure and style typical of the language, and the richness of its vocabulary.

13 In short, *La traduction raisonnée* encourages trainees to reason when translating, i.e., to adopt an attentive mindset that resists shortcuts, ultimately leading to a more thoughtful and effective translation process (Delisle and Fiola, 2013a: 422). With this idea in mind, let us

now dive deeper into our experiments testing machines' ability to reason when translating from English into French, illustrated by several examples.

The reasoned translation test: experimenting with NMT and GenAI

14 In this section, we present two pilot experiments which informed this study, and the methodology of the main experiment, also described in this section. The results of the main experiment will follow in section 4.

Pilot experiments

15 During the initial phases of this study, two pilot experiments were conducted. For the first one (May 24, 2023), a quick test was run machine-translating a 230-word text using DeepL (an NMT system), and ChatGPT (a GenAI system). In the case of ChatGPT, a "zero-shot" prompt was used: "Translate the following text into French: [text in English]." Zero-shot prompts directly instruct the model in a straightforward manner, without any examples or demonstrations (Saravia, 2022). The text was taken from Delisle and Fiola (2013a)'s "Lexical Networks" chapter (*Objectif 73 - Réseaux lexicaux*), which opens with a quote by Valery Larbaud: "A single word, used by the author in two different passages, will not always be translatable by the same word in the two corresponding passages. Yet this seems contrary to logic" (2013a: 613, our translation). The chosen text for our first pilot exemplifies exactly that. The example and analysis were borrowed by Delisle and Fiola from linguist and translator Maurice Pergnier, who explains the notion of *semantic fields* and how the same word (in the case of this excerpt, the word *land*) repeated eight times in English, cannot be translated eight times by the same word in French. As a matter of fact, in the proposed translation, this word is translated in six different ways: by the most obvious equivalent *terre*(s), but also by *régime seigneurial*, *seigneurie*, *parcelles*, *domaine* and *fiefs* (while acknowledging that other equivalents could have been possible). Delisle and Fiola support that a translator would "commit a methodological error" (2013a: 616, our translation) if they limited the translation of this word to a single equivalent, or even to the equivalent(s)

provided by bilingual dictionaries (the limits of which are also discussed in the third chapter [*Objectif* 3] of the textbook). They conclude from their analysis in *Objectif* 73 that "language units" do not simply have "language value" but rather "discourse value". It is only through discourse analysis that, in this case, a precise word can be found in the French translation at every instance of the word *land* in English. While thoroughly analyzing our first pilot test is beyond the scope of this paper, it is worth noting that the systems failed the test, consistently translating the eight instances of *land* by "*terre*" (singular) or "*terres*" (plural) instead of using the variety of more precise terms that would be preferred in five of the eight cases. Both systems made the common methodological error highlighted by Delisle and Fiola: failing to provide a variety of more precise terms in translation.

16 The second pilot (June 7, 2023) involved drafting a "few-shot" prompt intended for GenAI systems. Few-shot prompts are characterized by the inclusion of examples and demonstrations (Saravia, 2022). The prompt consisted of providing the system with two examples taken from *La traduction raisonnée*: source segments in English and the French translations provided by Delisle and Fiola. More specifically, we took these examples from the chapter focusing exclusively on the difficulty of translating "available" (*Objectif* 30) which can be mistranslated as "*disponible*" in French if no analysis takes place during the translation process. In this test, we wanted to "have a conversation" in French with the GenAI system, asking it to explain, based on the two examples provided, what strategy the translator used, and why the latter avoided translating "available" as "*disponible*". After this prompt was tested, a second few-shot prompt was drafted using the same two examples, this time asking the system to look at the two examples provided and then translate a third segment which also contained the word *available* "by also avoiding '*disponible*'". Again, while the discussion of this second pilot is beyond the scope of this paper, it is worth noting that the replies provided by the system displayed signs of "reasoning" (or a good imitation of it) in a way a trained linguist/translator would. These initial observations led us to design and conduct a formal experiment, described in the following section.

Experimental design

17 In this subsection we will describe the methodology of the main experiment, which was conducted in two stages: first in June 2023 and again in December 2023.

Data gathering

18 In the first stage of the project (June 2023), we compiled a corpus consisting of English segments extracted from 15 of the chapters in the section of Delisle and Fiola (2013a)'s textbook dedicated to syntactic difficulties (*Objectifs* 48 to 62). Instead of using longer texts (like the text in the first pilot), we decided to use only short segments from different chapters of the book. (The segment-based approach and its limitations will be discussed in the conclusion.) Such segments appear either under the "*Exercices d'application*" or the "*Exemples de traduction*" sections. Before and/or after the latter, the authors provide explanations as to why it is necessary to reason when translating texts containing the difficulty in question (e.g., *available*), and give examples of translated versions, often offering more than one possible translation, or "variations" of the translation of such segments. When the example segments were taken from the "*Exercices d'application*" section, we used the French version provided in the "*Livre du maître*" version of the textbook (Delisle, Fiola, 2013b), which provides trainers with example translations in French of the various segments and texts in the students' textbook, for reference.

19 We randomly extracted three English segments per chapter, for a total of 45. Each segment was translated using the six systems selected, namely MS Bing Translator (Bing), Google Translate (GoogleT), DeepL, ChatGPT, NotionAI and Gaby-T; in other words, for each of the 45 English segments, we had six machine-generated French translations. Naturally, no prompt was used when translating with the three NMT systems. In the case of GenAI, a zero-shot prompt was used: "Translate the following text into French: [segment in English]".

20 For the second stage (December 2023), we repeated the operation for 15 more chapters of the book, in this case those found in the section

devoted to lexical difficulties (*Objectifs* 30 to 44). We compiled a total of 90 segments, each machine-translated six times: three times with no prompt, with NMT, and three times with a simple zero-shot prompt, with GenAI.

21 In addition, we engineered a total of 30 few-shot prompts, one for each of the 30 chapters studied. These advanced prompts were elaborated with the model from the second pilot. We had already extracted three English segments from each chapter. Now, the first two, and their corresponding French translation provided by Delisle and Fiola served as the two examples from which the system was to "learn". In addition, to provide more context to the AI system, for each few-shot prompt we summarized in one paragraph (of approximately 100–150 words), to the best of our knowledge and capacity, the contents of the chapter. This meant briefly explaining the translator's "reasoning" (as explained by Delisle and his collaborators themselves, or borrowed from others) when translating the two segments, offering supplementary explanations, and then ending the prompt by asking the system to translate a third sentence, while reminding it what (not) to do; in other words, what it takes for the translation to be considered "reasoned" (and pass the "Delisle test"). Figure 1 below presents, in English, the structure of the few-shot prompts that were provided to the GenAI systems in French:


Figure 1. Structure of the few-shot prompts

22 The segments translated with the few-shot prompts were also included in the corpus in both phases of the experiment: 15 from June 2023 for the chapters dealing with syntactic difficulties, the same 15 segments translated in December 2023, and 15 new segments for the chapters dealing with lexical difficulties. In total, we had 630 machine-translated segments to analyze in the second stage; in total, 945 segments were analyzed in both stages. Figure 2 provides an example of how the translations of each segment were compiled in the corpus in a Google Doc.



Figure 2. A cropped screenshot of the corpus document

23 With the corpus compiled, we proceeded to assess whether the systems passed the test when translating each segment, as explained below.

Assessment of translated segments and marking of the results

24 To mark the data, on Google Sheets, a simple Yes/No-question method was used to indicate whether each one of the systems tested passed the test for each one of the example segments translated. Figure 3 below displays an example of the marking. While analyzing the corpus data, this author considered a system to have "passed the Delisle test" if, when translating the given segment, it produced an apparently "reasoned" translation both by avoiding the syntactic or lexical calque that is common in untrained translators (and criticized by Delisle's method), and suggesting a solution that corresponds to what a trained, human translator would have suggested (an example is provided in the following paragraph).



Figure 3. A cropped screenshot of the marking spreadsheet

25 Using the example in Figure 2 above (segment #1 from *Objectif* [OBJ] 48) and the marking in Figure 3, we can observe in the row for example (EX) segment #1 that all three systems were marked with the number "1" (to facilitate the total count later) under "N" (No) because they "failed" the test: they all failed to reason the translation in a way that avoids using the *comparatif elliptique* (or incomplete comparison) in French; in the English language, the incomplete comparison is widely used and accepted. Specifically, the three systems translated "older", "wiser", and "more cautious" with expressions that include "*plus*" + the equivalent adjectives for "old", "wise", and "cautious" (*âgé*, *sage*, *prudent*, respectively, and by default in the masculine form). The reasoned-translation method teaches that, in French, the second element of the comparison needs to be made explicit, i.e., the answer to the question: "older, wiser and more cautious than who?" Delisle and Fiola (2013a: 420) provide the following reasoned translation in French for that part of the English

segment shown in Figure 2 that contains the difficulty in question: "*Aujourd'hui, maturité et sagesse aidant, je redouble de prudence [...].*" On the other hand, we also observe in Figure 3 that the three NMT systems passed the test and were marked under Y (Yes) in the case of the example segment #2. That is because all three systems produced translations that the reasoned translation method would have considered satisfactory: they avoided the incomplete comparison in French and used more idiomatic expressions, as shown in Figure 4 below. Had they translated "easier" by "*plus facile*" (without the second element of the comparison made explicit, i.e., the answer to the question: "easier than what?"), the systems would have failed the test again.



Figure 4. An example of a segment that passed the "Delisle test" for all three NMT systems.

Figure 4. An example of a segment that passed the "Delisle test" for all three NMT systems.

26 It is important to note that, in our experiment, "passing the test" does not necessarily mean that the translation is flawless. While we were focusing only on the word or syntactic structure that was dealt with in each of the chapters, we could observe, even in some of the segments that passed the test, other types of issues that would be criticized by the reasoned translation method or deemed unacceptable by an evaluator (for instance a *contresens*, or saying the exact opposite, e.g. "*l'emploi*" as the translation of "unemployment"). Such analyses are beyond the scope of this paper but remain possible for future work looking at the corpus from other angles.

Main experiment data analysis

27 In this section, we begin by providing "vertical analyses", or total counts and percentages, for each column from the marking in the two stages of the experiment. First, the focus is on the data from the chapters on syntactic difficulties collected in June 2023, and then the data for the same chapters collected in December 2023, and how it compares with the numbers from June. Then we focus solely on the data from the chapters on lexical difficulties collected in December and look at the totals from both syntactic and lexical difficulties.

Lastly, we offer insights based on "horizontal analyses" of the data (e.g., by looking at specific rows or groups of rows in the marked spreadsheets, what are the segments or chapters where all systems, or one type of system, consistently failed the test?).

Data from chapters on syntactic difficulties (stage 1)

28 Having compiled the corpus and marked the data in the first stage of the experiment, we soon noticed that NMT systems outperformed GenAI systems in the zero-shot-prompt condition, yet the success rates were not very high in any of the cases. In the case of NMT, the average success rate was 25.2%. We observed a tie between Google Translate (GoogleT) and DeepL with 28.9% scoring the highest, and the lowest performance for Bing with 17.8%, as shown in Table 2:



Table 2. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments from chapters on syntactic difficulties collected in stage 1 (n = 45).

Table 2. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments from chapters on syntactic difficulties collected in stage 1 (n = 45).

29 In the case of GenAI systems with the zero-shot prompt, the average success rate was 19.3%. The system with the highest performance was NotionAI with 24.4%, and the lowest score was noted for Gaby-T with 13.3%, as shown in Table 3:



Table 3. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments from chapters on syntactic difficulties collected in stage 1 (n = 45).

Table 3. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments from chapters on syntactic difficulties collected in stage 1 (n = 45).

30 Because the first two example segments from each objective were included in the few-shot prompts, comparisons between the various prompting conditions (no prompt, zero-shot prompt, few-shot prompt) are based on a set consisting of the remaining 15 segments,

i.e., the third segment from each of the objectives included. For this sub-set ("the comparable sub-set" hereafter), we observe a higher average success rate for NMT with 35.6%, a tie between Bing and DeepL with the lowest performance, and a "win" for GoogleT with 40%, as shown in Table 4:



Table 4. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 1 (n = 15).

Table 4. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 1 (n = 15).

31 The results for the GenAI systems were virtually the same as when considering all segments (table 3 above). NotionAI displayed a lower performance this time and a tie with ChatGPT with 20%, as shown in Table 5 below. The average success rate was naturally lower as well (17.8%).



Table 5. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 1 (n = 15).

Table 5. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 1 (n = 15).

32 That being said, a big boost in the numbers was quickly observed with the few-shot prompt: the average success rate observed was 51.1%, with NotionAI scoring the highest with 60%, Gaby-T going from worse (13.3% in Table 5 above) to second best with 53.3%, and ChatGPT scoring the lowest with 40% (still, it was 100% higher than in the zero-shot-prompt condition), as shown in Table 6:



Table 6. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a few-shot prompt. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 1 (n = 15).

Table 6. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a few-shot prompt. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 1 (n = 15).

Table 6. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a few-shot prompt. Segments in the comparable sub-set from chapters on syntactic diffi-

culties collected in stage 1 (n = 15).

33 Now, what happens when collecting data for the same segments using the same six systems and under the same conditions, six months later? Can we expect the results to remain unchanged even with the systems' training data evolving over this period? Let us now look at the results from December 2023 for the same 45 segments. We expect to observe a change in performance of these systems over a six-month period, which can likely be attributed to model updates and improvements, among other reasons. Indeed, the systems' models are continuously being refined, and updates to the underlying algorithms, training data, or architecture of these models may impact their performance over time.

Data from chapters on syntactic difficulties (stage 2)

34 In comparing tables 7 and 8 below to tables 2 and 3 respectively, it can be observed that the results *are* different, with one exception, and the averages are slightly higher. In the case of NMT, the average performance was 29.6%, and in the case of GenAI it was 21.5%. The performance of DeepL was the same, while that of Bing increased by 62.36% to tie with DeepL. GoogleT improved slightly, as shown in Table 7:



Table 7. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments from chapters on syntactic difficulties collected in stage 2 (n = 45).

Table 7. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments from chapters on syntactic difficulties collected in stage 2 (n = 45).

35 In the case of the GenAI systems, the performance was virtually the same for ChatGPT (a slight increase) and NotionAI (a slight decrease), which tied at 22.2%, yet noticeably higher for GabyT, which increased from 13.3% (as observed in Table 3) to 20%, as shown in Table 8:



Table 8. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments from chapters on syntactic difficulties collected in stage 2 (n = 45).

Table 8. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments from chapters on syntactic difficulties collected in stage 2 (n = 45).

36 Remarkable differences were also noted in the case of the data for the comparable sub-set when compared with that from June (Tables 4 and 5 above). For NMT, the average performance was 40% in December. In addition, as shown in Table 9 below, while the performance was unchanged for GoogleT and DeepL (see Table 4), Bing went from worst to best, scoring 46.7% in December.



Table 9. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 2 (n = 15).

Table 9. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 2 (n = 15).

37 The difference for Bing was made by two of the segments that failed the test in June but passed it in December, as exemplified in Figure 5 below.



Figure 5. An example of a segment translated by MS Bing in June 2023, when it failed the test, then in December 2023, when it passed the test. The segment is taken from *Objectif* 52 ("While").

Figure 5. An example of a segment translated by MS Bing in June 2023, when it failed the test, then in December 2023, when it passed the test. The segment is taken from Objectif 52 ("While").

38 The average success rate for GenAI was also slightly higher in December with 20%, but differences were noticeable for all systems: ChatGPT scored 33.5% higher, and Gaby-T scored 50.38% higher than in June, while NotionAI scored 33.5% lower, as can be observed by comparing Table 5 above and Table 10 below:

 Table 10. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 2 (n = 15).

Table 10. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 2 (n = 15).

39 Remarkable differences were also evidenced in the case of the few-shot-prompt condition. As a reminder, we are comparing the output from the same systems tested for the same segments, under the same conditions, and using the same prompts, six months apart. Can we expect the results to remain unchanged over this period? It was observed that the average success rate decreased to 46.7% in December (from 51.1% in June). Individually, NotionAI went from best to worst (from 60% to 33.3%), Gaby-T remained second but decreased from 53.3% to 40%, and ChatGPT went from worst to best, increasing from 40% to 66.7%, as shown in Table 11.

 Table 11. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a few-shot prompt. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 2 (n = 15).

Table 11. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a few-shot prompt. Segments in the comparable sub-set from chapters on syntactic difficulties collected in stage 2 (n = 15).

40 Having observed the changing performance of the systems over time, let us now focus on analyzing the data from the 15 chapters dealing with lexical difficulties collected only in December 2023.

Data from chapters on lexical difficulties (stage 2)

41 Just like with syntactic difficulties, we soon noticed that NMT systems outperformed GenAI systems in the zero-shot-prompt condition. The success rates were not very high in any of the cases, yet higher than for syntactic difficulties. For NMT, the average success rate was 37%. GoogleT displayed the lowest success with

31.1%, and DeepL the highest with 44.4%. Bing was placed second best with 36.6%, as shown in Table 12:



Table 12. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments from chapters dealing with lexical difficulties collected in stage 2 (n = 45).

Table 12. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments from chapters dealing with lexical difficulties collected in stage 2 (n = 45).

42 In the case of GenAI systems, the average success rate was 23%. The system with the highest performance was NotionAI with 28.9%, and the lowest score was noted for ChatGPT this time, with 17.8%, as shown in Table 13:



Table 13. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments from chapters on lexical difficulties collected in stage 2 (n = 45).

Table 13. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments from chapters on lexical difficulties collected in stage 2 (n = 45).

43 When considering the comparable sub-set, we observe a higher average success rate for NMT with 48.9%, and a tie between Bing and DeepL with the highest performance (53.3%); GoogleT's performance was the lowest with 40%, as shown in Table 14. Figure 6 shows an example of a case where both Bing and DeepL passed the test, and GoogleT failed.



Table 14. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments in the comparable sub-set from chapters on lexical difficulties collected in stage 2 (n = 15).

Table 14. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. Segments in the comparable sub-set from chapters on lexical difficulties collected in stage 2 (n = 15).



Figure 6. An example of a segment translated by MS Bing, GoogleT and DeepL in December 2023. The segment is taken from Objectif 30 ("Available").

Figure 6. An example of a segment translated by MS Bing, GoogleT and DeepL in December 2023. The segment is taken from *Objectif* 30 ("Available").

44 The performance was lower for GenAI systems with the simple zero-shot prompt, with NotionAI and Gaby-T tying at 33%, and ChatGPT displaying the worst performance with 20%, as shown in Table 15 below. The average success rate was naturally lower than for NMT as well (28.9%).



Table 15. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. Segments in the comparable sub-set from chapters on lexical difficulties collected in stage 2 (n = 15).

45 But again, a considerable boost in the numbers was observed in the condition with the few-shot prompt: the average success rate observed was 73.3% (it was 51.1% in June and 46.7% in December for syntactic difficulties), with NotionAI scoring the lowest this time (66.7%). ChatGPT placed second with 73.3%. Gaby-T displayed the highest percentage in the data: 80%, as shown in Table 16:



Table 16. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a few-shot prompt. Segments in the comparable sub-set from chapters on lexical difficulties collected in stage 2 (n = 15).

46 Overall, it can be observed in this analysis that both system types provide better results for segments in chapters on lexical than on syntactic difficulties, since the total counts and percentages are slightly (or sometimes much) higher in the case of lexical difficulties, particularly in the case of the condition with the few-shot prompt.

47 To finalize this vertical analysis, let us look at the grand totals from December 2023, that is, combining the data from all 30 chapters.

Data from chapters on both syntactic and lexical difficulties (stage 2)

48 Considering the previous analysis, it is not surprising that NMT performed better overall than GenAI with zero-shot prompts. The average success rate for the NMT systems was 33.3%, this time based on all 90 segments. GoogleT displayed the lowest success with 31.1%, and DeepL the highest with 36.7%. Bing placed second best with 32.2%, as shown in Table 17:



Table 17. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. All segments from chapters on both syntactic and lexical difficulties collected in stage 2 (n = 90).

Table 17. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. All segments from chapters on both syntactic and lexical difficulties collected in stage 2 (n = 90).

49 In the case of GenAI systems with a zero-shot prompt, the average success rate was 22.2% overall. The system with the highest performance was NotionAI with 25.6%, and the lowest score was noted for ChatGPT this time, with 20% (virtually a tie with Gaby-T, which scored 21.1%), as shown in Table 18:



Table 18. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. All segments from chapters on both syntactic and lexical difficulties collected in stage 2 (n = 90).

Table 18. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a zero-shot prompt. All segments from chapters on both syntactic and lexical difficulties collected in stage 2 (n = 90).

50 When considering the comparable sub-set alone, we observed a higher average success rate for NMT with 44.4%. Bing displayed the top performance with 50%, DeepL came second with 43.3%, and GoogleT was third on the podium with 40%, as shown in Table 19:



Table 19. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. All segments in the comparable sub-set from chapters on both syntactic and lexical difficulties collected in stage 2 (n = 30).

Table 19. Total counts and percentages of "pass" (Y) or "fail" (N) per NMT system. All segments in the comparable sub-set from chapters on both syntactic and lexical difficulties collected in stage 2 (n = 30).

51 Overall, and when looking only at the comparable sub-set, GenAI systems with a zero-shot prompt performed less well than NMT, with NotionAI and ChatGPT tying at 23.3%, and ChatGPT displaying the highest score with 26.7%, as shown in Table 20 below. The average success rate was naturally lower than for NMT as well (24.4%).



Table 20. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with zero-shot prompt. All segments in the comparable sub-set from chapters on both syntactic and lexical difficulties collected in stage 2 (n = 30).

Table 20. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with zero-shot prompt. All segments in the comparable sub-set from chapters on both syntactic and lexical difficulties collected in stage 2 (n = 30).

52 In contrast, as could be expected based on the previous separate analyses, the numbers were much higher overall in the condition with the few-shot prompts. This time, considering only the comparable sub-set across all 30 chapters, the average success rate was 60%, with NotionAI scoring the lowest (50%). Gaby-T placed second with 60%. Finally, ChatGPT displayed the highest percentage: 70%, as shown in Table 21:



Table 21. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a few-shot prompt. All segments in the comparable sub-set from chapters on both syntactic and lexical difficulties collected in stage 2 (n = 30).

Table 21. Total counts and percentages of "pass" (Y) or "fail" (N) per GenAI system with a few-shot prompt. All segments in the comparable sub-set from chapters on both syntactic and lexical difficulties collected in stage 2 (n = 30).

53 Based on the data, it can be concluded that while both system types demonstrate overall better performance with segments involving lexical challenges compared to syntactic ones, notable variations emerge when looking at the grand totals. For instance, in certain cases, a specific NMT system may occasionally equal (or eventually outperform) a specific GenAI system in the few-shot-prompt condition (for example, Bing scoring 50% [Table 19] and NotionAI with a few-shot prompt also scoring 50% [Table 21]). However, this trend is not consistent across all cases, suggesting that specific system characteristics may lend themselves better to particular types of difficulties. These patterns will be explored in more depth in the following discussion.

Instances where all systems passed or failed the test

54 When marking the data, we also observed cases where *all* systems would fail or pass the test for *all* segments analyzed from specific chapters. We indicated this in the spreadsheets with all the markups and then proceeded to create separate tables for the total counts of instances of all systems failing or passing, per chapter (rows) and per type of system and condition (columns).

55 In Table 22 below, for instance, which displays data on the syntactic difficulties, we can observe that: 1) the different system types have more instances of "all fail" than "all pass", in both stages of the experiment, 2) NMT systems perform better than GenAI with a zero-shot prompt, and 3) there was only one instance of a system that passed the test for all three segments translated (namely, *Objectif* 50, which deals with the difficulty of translating "on… basis"). We also noted that for *Objectifs* 55 ("*Disjonctions exclusives*"), 57 ("*Structures résultatives*") and 61 ("*Voix passive*"), specifically, all systems failed the test at both stages of the experiment. (This observation for specific chapters is not seen in the total-count tables 21 through 24 but are highlighted in our data.)



Table 22. Total counts of chapters on syntactic difficulties where all systems passed (Y) or failed (N) the test for all three segments translated in both stages.

Table 22. Total counts of chapters on syntactic difficulties where all systems passed (Y) or failed (N) the test for all three segments translated in both stages.

56 In the case of the chapters on lexical difficulties, which were only dealt with in stage 2 (see Table 23), we observed again that GenAI systems have more instances of "all fail" than NMT. In this case, only one instance of "all pass" per system type was noted, namely *Objectif* 33 ("Corporate") for NMT, and *Objectif* 37 ("Issue, to issue") for GenAI. Lastly, there were two specific chapters where all systems failed: 31 ("Challenge, challenging, to challenge") and 41 ("Problem"). For the latter, for instance, all six systems translated as "*problème*" the word "problem" in the three selected segments, e.g., segment #1: "It is true that I have a problem with hearing. I have had it since I was a baby" translated by ChatGPT as "*Il est vrai que j'ai un problème d'audition. Je l'ai depuis que je suis bébé*," which is precisely the "methodological error" (Delisle, Fiola, 2013a: 616) that the reasoned translation method deplores – not to mention other mistranslations (e.g., "*depuis que je suis bébé*") which were beyond the scope of this study. In the textbook, the following reasoned translation is suggested in French: "*Il est vrai que je souffre de déficience auditive, et cela depuis ma tendre enfance.*"

 Table 23. Total counts of chapters on lexical difficulties where all systems passed (Y) or failed (N) the test for all three segments translated in stage 2.

Table 23. Total counts of chapters on lexical difficulties where all systems passed (Y) or failed (N) the test for all three segments translated in stage 2.

57 To include the few-shot-prompt condition in this analysis as well, we proceeded with the total count, but instead of considering all three segments per chapter, only the comparable sub-set. Once again, we observed in the case of syntactic difficulties that NMT systems performed better than GenAI in the zero-shot-prompt condition. However, as shown in Table 24 below, GenAI performed much better with few-shot prompts: not only were there fewer instances of "all fail", but there were also more instances of "all pass". In this spreadsheet, we also noted "all-fail" cases for *Objectifs* 48 ("*Comparatifs elliptiques*"), 55 ("*Disjonctions exclusives*") and 60 ("*Participes présents, gérondifs et rapports logiques*"), and one "all-pass" case: *Objectif* 58 ("*Verbes de progression, verbes d'aboutissement*").



Table 24. Total counts of chapters on syntactic difficulties where all systems passed (Y) or failed (N) the test only for the comparable sub-set translated in both stages.

Table 24. Total counts of chapters on syntactic difficulties where all systems passed (Y) or failed (N) the test only for the comparable sub-set translated in both stages.

58 Lastly, in the case of the lexical difficulties, and looking at the comparable sub-set only for the three conditions, the trend was repeated: there were more instances of all failing for GenAI with zero-shot prompts than for NMT (which points to the fact that a prompt as simple as "Translate the following text into [language]" is not enough), but a much higher success rate for GenAI with the few-shot prompt, with only two instances of "all fail" compared to eight instances of "all pass" (out of 15), as shown in Table 25 below. Contrastively, in this case, we noted that there was no instance of a specific chapter in which all systems failed across conditions, and three in which all passed: 33 ("Corporate"), 36 ("To involve") and 37 ("Issue, to issue").



Table 25: Total counts of chapters on lexical difficulties where all systems passed (Y) or failed (N) the test only for the comparable sub-set translated in stage 2.

Table 25: Total counts of chapters on lexical difficulties where all systems passed (Y) or failed (N) the test only for the comparable sub-set translated in stage 2.

59 The insights gleaned from our experiments and the comprehensive quantitative data analysis presented in the previous sections led us to reflect on the potential of what we may call "advanced prompt engineering" as a promising area of research in translation technology and pedagogy. This inquiry will be addressed briefly in the next section before offering concluding thoughts on this study.

Should prompt engineering be integrated into translator training?

60 The use of prompts to perform natural-language communication tasks has become a subject of scholarly inquiry, including in transla-

tion (Castilho *et al.*, 2023; Yamada, 2023). In part, our study endeavoured to examine the impact of prompts on machines' ability to "reason" in translation; in other words, if machines can be "told", like human students, what the dos and don'ts are in translation and explained why, with examples. We sought to elucidate whether the provision of few-shot prompts led to improvements in translation quality (yet, only when translating short sentences or isolated segments, removed from their context; more on this in the conclusion).

61 As discussed above, from the pilot explorations to the main experiment, a notable enhancement in the overall performance of GenAI systems was evidenced when the systems were provided with few-shot prompts, signifying the potential of such human-expert interventions to augment AI-driven translation capabilities. Our Delisle-inspired prompts imparted examples, contextual cues and instructional guidance to AI models, thereby facilitating the production of translations that exhibit humanlike reasoning compared to unprompted counterparts. In contrast, we must acknowledge the complexity of prompt engineering, particularly concerning the pursuit of the discourse-analysis objectives espoused by Delisle and other TS scholars. While our few-shot prompts yielded improvements in translation quality, the process of crafting such prompts proved to be labour-intensive and time-consuming. Furthermore, despite the enhanced performance demonstrated by few-shot-prompt-engineered translations in our main experiment, they did not seem to fully adhere to the principles of the reasoned-translation method as delineated by Delisle, his collaborators, and others. In other words, the fact that a specific segment passed the test when translated using the few-shot prompt did not mean that it was flawless, and there may have been other issues that are also dealt with in the textbook, in other chapters included or not in this study; issues that a properly trained translator – with the reflex to reason – would have identified and fixed – or avoided in the first place. An example is provided in Figure 7 below, which illustrates a segment from *Objectif* 42, which deals with the difficulty of translating the lexical unit "system", All NMT tools failed the test because they all translated "system" by "*système*", as the GenAI also did with a simple zero-shot prompt. In contrast, with the advanced few-shot prompt, only

ChatGPT passed the test in that it processed the information about why a reasoned translation is needed when translating "system", avoiding "*système*". Yet, the translated segment contains other signs of "unreasoned" translation, such as failing to provide a more precise word or term (something that is extensively discussed in the textbook, particularly in *Objectif* 29 "*Mot juste*", which was out of the scope of this study) in lieu of "*dispositifs de stockage*", and also translating "problem" as "*problème*", a difficulty dealt with in *Objectif* 41, included in this study and exemplified above. The reasoned translation proposed by Delisle and Fiola (2013b: 177) considers all that: "*Dans les silos autres que ceux de la ferme, les insectes sont la source de graves ennuis.*"



Figure 7. An example of a segment translated by ChatGPT, NotionAI and Gaby-T under the few-shot-prompt condition, in December 2023. The segment is taken from Objectif 42 ("System").

Figure 7. An example of a segment translated by ChatGPT, NotionAI and Gaby-T under the few-shot-prompt condition, in December 2023. The segment is taken from *Objectif* 42 ("System").

62 The corpus compiled in this study remains a fertile ground for further analyses, for example, with the participation of several human evaluators such as translator trainers in an effort to objectivize the analysis and, together, revisit the notion of "reasoned translation" in the AI era. We also believe that the methodology of this study is replicable for future research to augment the corpus and perform much deeper analyses to draw stronger conclusions in this vein. A follow-up study could also consider collecting the data again, using the same segments, the same prompts and the same systems, but using the latest versions of the latter. Our study used the freely available ChatGPT 3.5 in both June 2023 and December 2023. Since then, a paid subscription and more advanced versions of the models have been introduced. As of this writing, the latest version of ChatGPT is "o1 pro mode", available to Pro subscribers: how would its performance differ from the discontinued 3.5? The models running the subscription-based NotionAI and the freely available Gaby-T may have also evolved since 2023 and may yield different results in a new study.

63 The question of whether to prompt or not to prompt (Yang and Nabity-Grover, 2024) deserves attention in TS. While prompts hold promise in enhancing the performance of GenAI systems in translation tasks, particularly when equipped with comprehensive guidance akin to advanced "few-shot prompts" based on discourse analysis theory, their efficacy remains to be investigated. It is also worth keeping in mind that while NMT is specifically designed for translation, GenAI is only used incidentally as a translation tool. In what way does the systems' primary purpose impact their potential trajectory as useful tools for translation? In other words, does the fact that MS Bing, GoogleT and DeepL were developed purposely *to translate* make them more useful than GenAI systems for language professionals such as translators and post-editors? As AI "copilots" are used for more and more applications, can we envision a future application where the human translator creates prompts for NMT engines so that these already powerful translation-specific tools produce reasoned translations?

64 For the time being, given the still-evident limitations of prompt-engineered translations in replicating the analytical and interpretive capacities of properly trained human translators, as observed in this study, this author believes that it may be prudent to prioritize continued investment in the training of human translators—and the training of trainers. Future research may also build upon the development of pedagogical materials such as *La traduction raisonnée* and similar resources in different formats (e-books, digital-learning resources) that cater to different language combinations and reflect real-world translation and intercultural communication situations.

Conclusion

65 The exploration of the "Reasoned Translation" approach in the context of contemporary AI-based translation systems presents intriguing insights into the evolving natural language processing landscape. Jean Delisle's seminal work, *La traduction raisonnée*, advocates for a translation approach rooted in reasoning—an attribute traditionally associated with human translators. As we assess the capabilities of NMT and GenAI models through the lens of Delisle's theory, we confront fundamental questions about the nature

of translation and the role of human intelligence in linguistic tasks (e.g., the intelligence that is required to perform discourse analysis).

66 In an interview with the Swiss *Neue Zürcher Zeitung* newspaper, French AI expert François Chollet argues that "investments in generative AI are based on false promises [...] and [that] the money being thrown their way could be put to better use" (cited in Fulterer, 2024). He explains that:

> Humans learn from data, absolutely. But not like LLMs [...]. Humans have a memory, they store information, but they also do much more with their brains. In particular, they can adapt to new situations, they can make sense of things they've never seen before. That's intelligence. The world is complex and ever changing, and full of novelty. Which is why you need general intelligence to operate in this world, as a human. Meanwhile, LLMs don't have intelligence. Instead, they have memory. They have stored online data and can call up facts and patterns. But they don't understand things that are different from what they have learned. [An LLM] has memorized hundreds of thousands or millions of times more things than you, and yet you can handle more new situations—because you can adapt on the fly. LLMs have way more memory, but only a fraction of your capabilities. (*ibid.*)

67 Our investigation into the translation capabilities of the LLMs that run the GenAI systems tested highlights both the promise and the limitations of AI-driven translation technologies, as well as the need for more research on translation-task-specific "advanced prompt engineering", and how such tools and topics ought to be included in translator-training programs. While NMT and GenAI systems continue to demonstrate remarkable progress in handling translation tasks, they often fail in replicating the depth of reasoning characteristic of trained human translators. The integration of AI into translation practice necessitates a deeper examination of how humanlike reasoning can be instilled within these systems, moving beyond surface-level language generation towards a more comprehensive understanding of semantic and cultural nuances. As we navigate this intersection of tradition and innovation, the concept of "reasoned translation", which in the case of English-to-French translation needed a doctoral-thesis-inspired textbook of no fewer than

716 pages and 75 chapters, serves as a guiding principle for shaping the future of translator training and translation technologies, and advancing our understanding of computational systems' linguistic competence.

68 While our study examined comparative performance of NMT and GenAI systems in translating individual sentences, it is essential to acknowledge the inherent limitation of evaluating machine-generated translations solely at the sentence level (Castilho, 2020). Our findings underscore the imperative for future investigations to expand the scope of analysis beyond isolated segments to provide a more comprehensive understanding of these systems' ability to reason "*à la Delisle*" when translating entire texts, as well as their limitations in real-world translation scenarios. As a matter of fact, traditionally, in machine-translation research, automatic measures of translation quality have focused on sentence-level rather than document-level analyses. However, recent efforts have been made to address this limitation. Castilho *et al.* (2023), for instance, tested different NMT systems and ChatGPT using a test suite to determine if providing solutions to context-related issues could enhance the systems' ability to deliver good-quality translations. The researchers conducted a small-scale manual analysis to assess the accuracy and fluency of translations and offered valuable insights into the choices made by these systems. They state that:

> although LLMs are the new hype in the AI world, further investigation on their translation capabilities is necessary. Future work should focus on more context-related issues [...] and also expanding the context span to verify whether that can have an effect on the translation output. (*ibid.*)

69 Our main experiment sheds light on machines' performance when translating isolated English segments into French. However, to truly grasp these systems' potential and limitations in real-world translation scenarios, future work should expand the scope of analysis to encompass entire texts (as suggested with our first pilot experiment presented in this paper) and include human participants: from students and professionals performing reasoned-translation tasks, to human evaluators assessing the quality of the translations performed by students, professionals, NMT and GenAI. It should also be mindful

of the translator experience, that is, the "translator's perceptions of and responses to the use or anticipated use of a product, system or service" (Zapata, 2016: 16). Lastly, future research efforts should also be considerate of the environmental impact of developing and using AI tools and, when integrated into translator training curricula, also promote critical thinking among students regarding the ethical and responsible use of AI-based technologies, promoting "life cycle thinking in the service of a transition to a sustainable society" (CIRAIG, 2022). In the words of Moorkens (2023):

> A concern is that hype and too little concern about ethics and sustainability will lead to the use of AI tools in inappropriate circumstances. Literacy about how they work and the data that drives them will be increasingly important.

70 This broader perspective will enable a more critical and comprehensive understanding of how technologies function and how they can *truly* serve as tools in practical contexts. By adopting such a holistic approach, researchers and developers can advance the deployment of tools with greater efficacy. And as these technologies continue to evolve, the need for skilled human translators will remain crucial, as it has been for thousands of years. University-level translator training will continue to be essential in the foreseeable future, as well as any other form of continued professional development training that encourages trainees to adopt an attentive mindset that resists shortcuts, ultimately leading to a more thoughtful and effective translation process, and to preserving the idiomatic character of the target language. In a world with more than 7,000 languages spoken and used, and increasingly easier access to information and communication technologies for all, further investing in the training of augmented human translators (O'Brien, 2023) is vital to ensuring global understanding and cooperation.

71 Acknowledgements

72 I would like to thank the reviewers for their thorough feedback and suggestions. I am also grateful for the generous support from three colleagues at various stages of this study, from ideation to validation of the methodology (e.g., the format and content of the "few-shot prompts" in French), to the interpretation and presentation of the

results: Sheila Castilho, Gabrielle Garneau and Elizabeth Marshman. I also acknowledge the timely collaboration of research assistants Tatiana Cruz Sánchez (data collection from NMT systems) and Lara Hanna (preparation of the final draft), and the financial support from the Office of the Dean of Arts, Toronto Metropolitan University. To all, *merci beaucoup*!

BIBLIOGRAPHY

Brisset Annie, 2010, "Cultural perspectives on translation", *International Social Science Journal*, Vol. 61, No. 199, p. 69-81.

Brown Tom B., Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared D., Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, Herbert-Voss Ariel, Krueger Gretchen, Henighan Tom, Child Rewon, Ramesh Aditya, Ziegler Daniel, Wu Jeffrey, Winter Clemens, Hesse Chris, Chen Mark, Sigler Eric, Litwin Mateusz, Gray Scott, Chess Benjamin, Clark Jack, Berner Christopher, McCandlish Sam, Radford Alec, Sutskever Ilya, Amodei Dario, 2020, "Language Models are Few-Shot Learners", in Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan and Hsuan-Tien Lin (eds.), *Advances in neural information processing systems (NeurIPS 2020)*, Vol. 33, p. 1877-1901.

Castilho Sheila, 2020, "On the Same Page? Comparing Inter-Annotator Agreement in Sentence and Document Level Human Machine Translation Evaluation", in Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-Jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri (Eds.), *Proceedings of the 5th Conference on Machine Translation*, Association for Computational Linguistics, p. 1150-1159.

Castilho Sheila, Mallon Clodagh Quinn, Meister Rahel, Yue Shengya, 2023, "Do Online Machine Translation Systems Care for Context? What About a GPT Model?", in Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, Helena Moniz (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Tampere (Finland), European Association for Machine Translation, p. 393-417.

CIRAIG, 2022, "About", [https://ciraig.org/index.php/about/], viewed on 7 October 2024.

Darbelnet Jean, 1969, "La traduction raisonnée", *Meta : Journal des traducteurs*, Vol. 14, No. 3, p. 135-140.

Delisle Jean, 1988, *Translation: An Interpretive Approach*, Ottawa, University of Ottawa Press.

Delisle Jean, 2003, *La traduction raisonnée. Manuel d'initiation à la traduction professionnelle de l'anglais vers le français*, 2nd edition, Ottawa, University of Ottawa Press.

Delisle Jean and Fiola Marco, 2013a, *La traduction raisonnée. Manuel d'initiation à la traduction professionnelle de l'anglais vers le français*, 3rd edition, Ottawa, University of Ottawa Press.

Delisle Jean and Fiola Marco, 2013b, *La traduction raisonnée. Manuel d'initiation à la traduction professionnelle de l'anglais vers le français. Livre du maître*, 3rd edition, Ottawa, University of Ottawa Press.

Fulterer Ruth, 2024, *Google researcher on AI hype: "Investments are a thousand times too high"*, [https://www.nzz.ch/english/google-researcher-says-ai-hype-is-skewing-investment-ld.1825122], viewed on 8 October 2024.

Hendy Amr, Abdelrehim Mohamed, Sharaf Amr, Raunak Vikas, Gabr Mohamed, Matsushita Hitokazu, Kim Young Jin, Afify Mohamed and Awadalla Hany Hassan, 2023, "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation", arXiv preprint, [https://doi.org/10.48550/arXiv.2302.09210].

Kocmi Tom and Federmann Christian, 2023, "Large Language Models Are State-of-the-Art Evaluators of Translation Quality", in Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, Helena Moniz (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Tampere (Finland), European Association for Machine Translation, p. 193-203.

Kumbe Kornebari B. 2016, "Delisle, Jean (2013) : La traduction raisonnée : manuel d'initiation à la traduction professionnelle de l'anglais vers le français. 3e édition. Ottawa, Les Presses de l'Université d'Ottawa, 716 p.", *Meta : Journal des traducteurs*, Vol. 61, No. 3, p. 734-737.

Moorkens Joss, 2023, "Generative tools using large language models and translation", in *Marie Curie Alumni Association Newsletter*, 35, June, p. 16-18.

Munday Jeremy and Zhang Meifang, 2017, *Discourse analysis in translation studies*, 1st ed., Amsterdam, John Benjamins Publishing Company.

O'Brien Sharon, 2023, "Human-centered augmented translation: against antagonistic dualisms", *Perspectives: Studies in Translation Theory and Practice*. [https://doi.org/10.1080/0907676X.2023.2247423], viewed on 10 November 2024.

Phillips Nelson and Hardy Cynthia, 2002, *Discourse Analysis*, Thousand Oaks, SAGE Publications, Inc.

Saravia Elvis, 2022, "Prompt Engineering Guide", December,

[https://www.promptingguide.ai/], viewed on 7 January 2025.

Schäffner Christina, 2004, "Political Discourse Analysis from the Point of View of Translation Studies", *Journal of Language and Politics*, Vol. 3, No. 1, p. 117-150.

Shi-xu, 2005, A *Cultural Approach to Discourse*, London, Palgrave Macmillan.

Yamada Masaru, 2023, "Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability", in Masaru Yamada, Felix do Carmo (Eds.), *Proceedings of Machine Translation Summit* XIX, *Vol.* 2, Asia-Pacific Association for Machine Translation, p. 195-204.

Yang Ning and Nabity-Grover Teagan, 2024, "To Prompt or Not to Prompt: The Impacts of Generative AI on Programming Education as Perceived by Educators", *AMCIS 2024 Proceedings*, 8, [https://aisel.aisnet.org/amcis2024/is_education/is_ed viewed on 11 October 2024.

Zapata Julián, 2016, *Translators in the Loop: Observing and Analyzing the Translator Experience with Multimodal Interfaces for Interactive Translation Dictation Environment Design*, PhD thesis, University of Ottawa, 243 p.

Zhang Meifang, Pan Hanting, Chen Xi and Luo Tian, 2015, "Mapping Discourse Analysis in Translation Studies via Bibliometrics: A Survey of Journal Publications", *Perspectives*, Vol. 23, No. 2, p. 223-239.

ABSTRACTS

English

This paper reports on a 2-stage experiment aimed at assessing machines' ability to "reason" in translation The experiment was built around Jean Delisle's pedagogy textbook *La traduction raisonnée* which has been used in English-French translator training for more than 40 years, in Canada and around the world. We put Bing Translator, Google Translate and DeepL – neural machine translation systems – as well as ChatGPT, NotionAI and Gaby-T – generative AI systems – to the test, using a selection of examples taken from the textbook and analyzing the results considering the background information and explanations provided, more specifically in the 30 chapters from where the examples were taken. In this paper, we first explore what inspired Delisle's work and what he means by "reasoned translation". Then, we focus on presenting the methodology and the results of our two pilot main experiments. Lastly, we offer some insights into potential future research avenues in translator training and beyond.

Français

La traduction raisonnée est le titre que le traductologue Jean Delisle a donné à son célèbre manuel de pédagogie, utilisé depuis 40 ans dans la

formation des traducteurs anglais-français au Canada et dans le monde entier. Ce titre résume bien la prémisse du manuel : un traducteur (humain) bien formé a la capacité de raisonner ; une bonne traduction en est une raisonnée. Paru en 1993, le manuel a été réédité tous les dix ans; pour les deuxième et troisième éditions, Delisle a recruté d'autres formateurs et chercheurs en tant que conseillers, co-auteurs ou auteurs uniques de certains objectifs (des chapitres portant sur une difficulté ou un sujet particulier). Aujourd'hui, plus de dix ans après la publication de la troisième et dernière édition, le manuel mérite une relecture à l'ère de la traduction automatique neuronale (TAN) et de l'intelligence artificielle (IA) afin de mettre en perspective la notion de « traduction raisonnée ».

Ces dernières années, le domaine du traitement automatique des langues connaît des progrès fulgurants, principalement attribuables à l'émergence de grands modèles de langage tels que les modèles de transformateurs génératifs pré-entraînés, communément appelés « IA générative ». Ces modèles, entraînés sur de vastes quantités de données textuelles, suscitent un certain engouement en raison de leur capacité à générer du texte en quelques secondes sur la base de requêtes de l'utilisateur, en plus de leurs capacités conversationnelles. Cette percée a entraîné une nouvelle vague de recherches en linguistique informatique et dans des domaines connexes, notamment pour explorer les moyens d'affiner ces modèles dans le cadre de tâches spécifiques comme la traduction et l'évaluation des traductions.

Cela dit, les programmes de TAN et d'IA générative peuvent-ils passer ce que nous appelons le « test de Delisle » ? En d'autres termes, montrent-ils des signes d'une capacité à raisonner, comme le font d'ailleurs les traducteurs humains adéquatement formés lorsqu'ils traduisent de l'anglais vers le français ? Pour le découvrir, nous avons mis à l'épreuve des programmes de TAN (MS Bing Translator, Google Translate et DeepL) et d'IA générative (ChatGPT, NotionAI et Gaby-T) en utilisant une sélection d'exemples tirés de 30 objectifs de la troisième édition de La traduction raisonnée et en analysant les résultats à la lumière des explications fournies dans le manuel. Ces outils produisent-ils des traductions « raisonnées » ? En quoi cette capacité ou incapacité se répercutera-t-elle sur la pédagogie de la traduction à l'avenir ?

Dans son manuel, Delisle suggère qu'apprendre à traduire, c'est apprendre à aborder un texte de manière « raisonnée », c'est-à-dire à découvrir progressivement tous les éléments qui interviennent dans le transfert interlinguistique. Il rappelle également à l'apprenant que le manuel ne doit pas être utilisé comme un livre de recettes. Les exemples ou les modèles de phrases ou de textes traduits de l'anglais au français fournis tout au long de l'ouvrage ne sont pas les seules traductions acceptables : ils servent plutôt à montrer que les solutions à tout problème de traduction sont multiples et dépendent toujours du contexte.

Les travaux de doctorat de Delisle, qui ont mené à la première édition de La traduction raisonnée, prônaient l'« analyse du discours comme méthode de traduction » ; ils mettaient l'accent sur la complexité des mécanismes intellectuels impliqués en traduction. L'analyse du discours appliquée à la

traduction est particulièrement utile lorsqu'il s'agit de textes complexes qui nécessitent une compréhension nuancée des intentions de l'auteur original. En examinant le discours qui entoure le texte, les traducteurs peuvent repérer les stratégies rhétoriques, les références culturelles et les grands thèmes essentiels à l'interprétation du texte original et à l'objectivation de la traduction. Pour définir l'analyse du discours, Phillips et Hardy expliquent que les discours sont « matérialisés et mis en œuvre dans une variété de textes » (2002 : 4, notre traduction) aux formes diverses (langue écrite ou parlée, images, symboles, artefacts, etc.). Ils ajoutent que les textes ne revêtent aucun sens considérés isolément : le sens émerge plutôt de leur interrelation avec d'autres textes, des différents discours sur lesquels ils reposent et de la nature de leur production, de leur diffusion et de leur consommation. D'après ces auteurs, l'analyse du discours consiste à explorer la manière dont ces processus font émerger le sens des textes et comment cette production de sens contribue à construire la réalité sociale. L'analyse du discours s'intéresse donc aux effets constructifs du discours à travers l'étude structurée et systématique des textes (ibid.).

Sans surprise, l'analyse du discours est pratiquée depuis des décennies par des traductologues comme Delisle et d'autres (par exemple, Brisset, 2010 ; Munday et Zhang, 2017 ; Schäffner, 2004 ; Zhang et al., 2015), car la traductologie ne peut que s'enrichir d'une étude structurée et systématique des textes. L'analyse du discours a mis en évidence les facteurs contextuels, les caractéristiques linguistiques, la pragmatique, l'intertextualité, l'idéologie et les relations de pouvoir, ainsi que les communautés discursives comme autant de pistes à suivre pour objectiver le processus de traduction. En appliquant l'analyse du discours à l'acte de traduire, les langagiers peuvent cerner des stratégies pour produire des textes de qualité dans la langue d'arrivée — une capacité que Delisle cherche à inculquer aux traducteurs en formation qui étudient attentivement La traduction raisonnée. Cette méthode de traduction raisonnée vise donc à cultiver chez les traducteurs en herbe un réflexe menant à une approche consciencieuse de la traduction qui privilégie l'analyse du discours et met en garde contre les raccourcis (calques syntaxiques ou lexicaux, choix du premier mot équivalent fourni par un dictionnaire bilingue, etc.). Les traducteurs formés raisonneront pour comprendre le message et le reformuler de manière idiomatique dans la langue d'arrivée (en l'occurrence, le français) tout en utilisant la terminologie appropriée. À terme, ils contribueront à préserver la structure et le style propres à la langue, ainsi que la richesse de son vocabulaire. En bref, La traduction raisonnée encourage les apprenants à raisonner lorsqu'ils traduisent, c'est-à-dire à adopter un état d'esprit attentif qui résiste aux raccourcis et qui mène à des textes réfléchis (Delisle et Fiola, 2013a : 422).

Durant la phase initiale de cette étude, nous avons mené deux expériences pilotes. Pour la première, nous avons testé deux outils de différents types, DeepL et ChatGPT, en leur faisant traduire un texte de 230 mots. Pour ChatGPT, nous avons utilisé la requête « Traduisez le texte suivant en français : [texte en anglais] ». Le texte a été tiré de l'Objectif 73 — Réseaux lexicaux de la troisième édition du manuel (Delisle et Fiola, 2013a). Les auteurs

ont emprunté l'exemple et l'analyse de cet objectif au linguiste et traducteur Maurice Pergnier, qui explique la notion de champs sémantiques et pourquoi un même mot (soit land dans le cas analysé) répété huit fois en anglais ne peut pas être traduit huit fois par le même mot en français. Delisle et Fiola soutiennent que le traducteur commettrait une erreur méthodologique s'il limitait la traduction de ce mot à un seul équivalent, voire aux équivalents fournis par les dictionnaires bilingues. Ils concluent que les unités linguistiques n'ont pas simplement une valeur de langue, mais plutôt une valeur de discours. Ce n'est que grâce à l'analyse du discours qu'on arrive, dans ce cas, à un mot précis dans la traduction française pour chaque occurrence du mot land. Les deux systèmes ont échoué au « test de Delisle », car ils ont traduit systématiquement les huit occurrences de land par « terre » ou « terres » au lieu d'utiliser d'autres termes plus précis qui conviendraient mieux dans cinq des huit occurrences.

La deuxième expérience pilote consistait à rédiger une requête destinée aux programmes d'IA générative. La requête donnait à l'outil des exemples tirés de l'Objectif 30, consacré à la difficulté de traduire le mot available : deux segments en anglais et les exemples de traductions françaises correspondantes du manuel. Dans le cadre de cette expérience, nous voulions converser en français avec l'outil et lui demander d'expliquer, d'après les deux exemples fournis, la stratégie employée par le traducteur et la raison de son choix. Après ce test, nous avons formulé une deuxième requête en utilisant les deux mêmes exemples, mais en demandant cette fois à l'outil de tenir compte des deux exemples fournis pour traduire un troisième segment qui contenait la même difficulté d'ordre lexical, available. Les réponses de l'outil présentaient des signes indicateurs d'un « raisonnement » (ou une bonne imitation de celui-ci) comme le travail d'un traducteur adéquatement formé. Ces premières observations nous ont amenés à la conception d'une expérience formelle.

L'expérience principale a été menée en deux étapes, en juin et en décembre 2023. À la première étape, nous avons compilé un corpus composé de segments en anglais extraits de 15 objectifs de la section du manuel consacrée aux difficultés d'ordre syntaxique. Nous avons utilisé des textes plus courts que dans la première expérience pilote. Nous avons extrait au hasard trois segments anglais par objectif, pour un total de 45. Chaque segment a été traduit à l'aide des six outils sélectionnés. Pour chacun des 45 segments en anglais, nous disposions donc de six traductions en français générées de façon automatique; trois produites par la TAN et trois par l'IA générative. Bien entendu, aucune requête n'a été utilisée lors de la traduction à l'aide des outils de TAN. Dans le cas des outils d'IA générative, la requête était simple : « Traduisez le texte suivant en français : [segment en anglais] ». Pour la deuxième étape de l'expérience, nous avons répété l'opération avec 15 autres objectifs du manuel, cette fois tirés de la section consacrée aux difficultés d'ordre lexical. Ainsi, nous avons compilé un total de 90 segments, tous traduits six fois.

Ensuite, nous avons formulé 30 « super-requêtes », une par objectif étudié et inclus dans l'expérience, en suivant le modèle de la deuxième expérience

pilote. Nous avions déjà extrait trois segments en anglais de chaque objectif. Les deux premiers, ainsi que leur traduction en français fournie dans le manuel, ont servi d'exemples à partir desquels le programme devait « apprendre ». En outre, pour fournir un contexte étoffé à l'IA, nous avons résumé le contenu de l'objectif dans un paragraphe de 100 à 150 mots pour chaque super-requête. Il s'agissait de résumer le « raisonnement » du traducteur, tel qu'expliqué dans le manuel, lors de la traduction des deux segments, de fournir des explications supplémentaires, puis de terminer la requête en demandant au système de traduire une troisième phrase, tout en lui rappelant quoi faire ou ne pas faire, soit la condition à respecter pour produire une traduction « raisonnée ». Dans le corpus, nous avons également inclus les segments traduits à l'aide des « super-requêtes » au cours des deux phases de l'expérience. Au total, nous avions 585 segments à analyser à la deuxième phase. Une fois le corpus compilé, nous sommes passés à l'évaluation des outils de TAN et d'IA générative pour déterminer s'ils avaient réussi le « test de Delisle » en traduisant chaque segment.

Lors de l'analyse des données, nous avons indiqué « réussi » lorsque la traduction d'un segment avait l'apparence d'une traduction « raisonnée », c'est-à-dire qu'elle évitait le calque syntaxique ou lexical fréquent chez les traducteurs sans formation adéquate (et critiqué par la méthode de Delisle) avec une proposition semblable à celle d'un traducteur humain bien formé. Soulignons que, dans notre expérience, le fait de « réussir le test » ne signifiait pas nécessairement que la traduction était parfaite. Notre analyse ne portait que sur le mot ou la structure syntaxique traitée dans chacun des objectifs, mais nous avons relevé, y compris dans certains segments ayant passé le test, d'autres types de tournures critiquables selon la méthode de la traduction raisonnée ou qu'un évaluateur humain aurait rejetées. Dans l'ensemble, nous avons observé que les deux types d'outils à l'essai (la TAN et l'IA générative) fournissaient de meilleurs résultats pour les segments tirés des objectifs sur les difficultés d'ordre lexical que pour ceux des difficultés d'ordre syntaxique, en particulier dans les cas d'essai avec la super-requête. Cependant, d'importantes variations émergeaient d'une analyse des totaux généraux : par exemple, dans certains cas, la traduction de tous les segments d'un même objectif réussissait ou non le test selon le type d'outil.

L'analyse des données expérimentales nous a menés à une réflexion sur le potentiel de ce que nous proposons d'appeler la « rédactique avancée » (en anglais, advanced prompt engineering), sujet de recherche prometteur en traductique, en pédagogie de la traduction et d'autres domaines. Cette forme de rédactique (prompt engineering) se pencherait sur l'utilisation de requêtes avancées (en faisant appel à des few-shot prompts, c'est-à-dire des requêtes assorties d'exemples et de démonstrations) pour effectuer des tâches langagières, y compris traduire. Notre étude visait en partie à examiner l'incidence des requêtes sur la capacité des machines à « raisonner » en traduction. Nous voulions déterminer s'il est possible, comme on le fait d'ailleurs pour les apprenants humains, de « dire » aux machines quoi faire ou ne pas faire en traduction et de leur expliquer pourquoi à l'aide d'exemples. Des phases exploratoires à l'expérience principale,

nous avons constaté une amélioration notable de la performance globale des programmes d'IA générative lorsqu'ils recevaient les super-requêtes, ce qui indique le potentiel de l'expertise humaine appliquée à l'augmentation des capacités de traduction pilotées par l'IA. Ces super-requêtes ont fourni aux modèles d'IA quelques exemples, des indices contextuels et des instructions pour faciliter la production de traductions qui présentaient un raisonnement comparable à celui de traducteurs humains. En revanche, nous devons reconnaître la complexité inhérente à la rédactique, notamment en ce qui concerne la poursuite des objectifs de l'analyse du discours prônés par Delisle et d'autres traductologues. Bien que nos requêtes avancées aient permis d'améliorer la performance des outils à l'essai, leur processus d'élaboration s'est avéré laborieux et chronophage. Malgré certaines améliorations, les traductions générées à l'aide de requêtes avancées dans notre étude ne semblaient pas adhérer pleinement aux principes de la méthode de la traduction raisonnée promue par Delisle, ses collaborateurs et d'autres chercheurs. En effet, même si un segment donné « réussissait le test » grâce à la requête avancée, il contenait parfois d'autres types d'erreurs traitées dans le manuel : des fautes de langue ou de transfert qui auraient été repérées, corrigées, voire complètement évitées par un traducteur adéquatement formé, c'est-à-dire ayant acquis le réflexe de raisonner en traduisant.

L'intérêt de recourir aux requêtes avancées mérite d'autres études en traductologie. Si les requêtes semblent prometteuses pour améliorer la performance des outils d'IA générative en traduction, leur efficacité n'a pas été évaluée. Pour l'instant, étant donné les limites manifestes des traductions générées à partir de requêtes dans le but de reproduire les capacités d'analyse et d'interprétation de traducteurs humains bien formés, nous estimons qu'il demeure préférable d'investir en priorité dans la formation des traducteurs humains — et dans celle de leurs formateurs. La recherche pourrait également favoriser la conception de ressources pédagogiques semblables à La traduction raisonnée offertes sur différents supports (livres électroniques, plateformes d'apprentissage numériques, etc.) et adaptées à différentes combinaisons de langues et aux situations réelles de traduction et de communication interculturelle.

En conclusion, l'exploration de l'approche de la « traduction raisonnée » dans le contexte des outils de traduction contemporains basés sur l'IA offre de fascinantes perspectives dans le paysage évolutif du traitement automatique des langues. En pédagogie de la traduction, l'œuvre pionnière de Jean Delisle préconise une méthode fondée sur le raisonnement, attribut traditionnellement associé aux traducteurs humains. L'évaluation des capacités des modèles de TAN et d'IA générative sous l'angle de la théorie de Delisle nous confronte à des questions fondamentales sur la nature de la traduction et sur le rôle de l'intelligence humaine dans l'accomplissement de tâches linguistiques (par exemple, l'intelligence requise pour effectuer l'analyse du discours). Notre étude sur les capacités de traduction de divers programmes de TAN et d'IA générative met en évidence les promesses et les limites des technologies de traduction pilotées par l'IA, ainsi que la nécessité de poursuivre les recherches sur la « rédactique avancée ». Elle signale

également l'intérêt d'inclure ces technologies et les sujets connexes dans les programmes de formation des traducteurs. Bien que l'IA continue de faire des progrès remarquables, elle ne parvient souvent pas à reproduire la profondeur de raisonnement caractéristique des traducteurs humains ayant reçu une formation adéquate. Tout effort d'intégration de l'IA dans la pratique de la traduction nécessite un examen approfondi de la manière dont on peut enseigner le raisonnement humain à ces outils : comment peuvent-ils dépasser une production linguistique superficielle pour arriver à une compréhension élargie des nuances sémantiques et culturelles? À l'intersection entre tradition et innovation, le concept de « traduction raisonnée » sert de principe directeur pour façonner l'avenir de la pédagogie et des technologies de la traduction, ainsi que pour faire progresser notre compréhension des compétences linguistiques des machines. Malgré le rôle accru des technologies en traduction, il existe toujours une demande pour des traducteurs humains. Dans un avenir proche, la formation universitaire des traducteurs restera essentielle, de même que les autres formes de perfectionnement professionnel continu incitant les apprenants à adopter un état d'esprit attentif qui résiste aux raccourcis, comportement qui mène à un processus de traduction réfléchi et efficace, ainsi qu'à la préservation du caractère idiomatique de la langue d'arrivée.

Español

Este artículo presenta un experimento en dos fases diseñado para evaluar la capacidad de «razonamiento» de las máquinas en el contexto de la traducción. El experimento se basó en el libro *La traduction raisonnée* de Jean Delisle, que se ha utilizado en la enseñanza de la traducción inglés-francés durante más de 40 años, en Canadá y otras partes del mundo. Pusimos a prueba los sistemas de traducción automática neuronal *Bing Translator*, *Google Translate* y *DeepL*, así como los sistemas de IA generativa *ChatGPT*, *NotionAI* y *Gaby-T*, usando segmentos extraídos de 30 de los capítulos del texto y analizando los resultados según las lecciones impartidas en los mismos. Primero analizamos en qué se inspira el libro de Delisle y qué se entiende por «traducción razonada». Luego, presentamos la metodología y los resultados de los experimentos piloto y principal. Por último, sugerimos algunas pistas de investigación, sobre todo en el campo de la didáctica de la traducción.

INDEX

Mots-clés

intelligence artificielle générative, traduction automatique neuronale, grands modèles de langues, traduction raisonnée, pédagogie de la traduction, analyse du discours, rédactique

Keywords
generative artificial intelligence, neural machine translation, large language models, reasoned translation, translator training, discourse analysis, prompt engineering

Palabras claves
inteligencia artificial generativa, traducción automática neuronal, grandes modelos de lenguaje, traducción razonada, didáctica de la traducción, análisis del discurso, ingeniería de instrucciones

AUTHOR

Julián Zapata
julian.zapata[at]torontomu.ca
Dr. Julián Zapata is an Assistant Professor of Translation Studies at the Department of Languages, Literatures and Cultures, Toronto Metropolitan University (Canada). He is also a certified translator and an entrepreneur. He completed his PhD in Translation Studies (2016) at the University of Ottawa, where he also taught for nearly a decade. His research interests include human-computer and human-information interaction; translation processes; translation technologies; speech technologies; mobile and cloud computing; multimodal interaction; ergonomics in translation; and translation pedagogy. His work has been funded, notably, by the Social Sciences and Humanities Research Council of Canada, the *Fonds de recherche du Québec – société et culture*, and Entreprise Ireland.