A Longitudinal Study of the Adoption of Specialized MT Systems in Canadian Parliamentary Translation

Michel Simard Jeniffer Leal-Wyss Gabriel Bernier-Colborne Rebecca Knowles
National Research Council Canada
{Michel.Simard, Jeniffer.Leal,
Gabriel.Bernier-Colborne, Rebecca.Knowles}@nrc-cnrc.gc.ca

Abstract

Since 2023, translators for the Parliament of Canada have had the option to use neural machine translation (NMT) technology provided by the National Research Council of Canada (NRC) to support their work in translating parliamentary publications between French and English. We present our analysis of an anonymized dataset of translators' interactions with our Hawkeye MT systems, collected since their introduction and covering a period of 2.5 years. This data provides a unique perspective on how translators interact with the systems, how their use evolved over time and how it impacts the nature of their translations.

1 Introduction

As machine translation is becoming a standard fixture in professional translators' work environments, we see a growing literature on translator's changing work habits (do Carmo and Moorkens, 2022; O'Brien, 2024), attitudes towards technology (Nunes Vieira and Alonso, 2019; Guerberof Arenas, 2025), motivations for adoption and non-adoption (Cadwell et al., 2018; Zaretskaya, 2015), etc. Such studies can be extremely rich in details and serve a useful function in understanding the impact of such a transformation of translation work environments on the professional practices and lives of translators. In most cases, however, these studies can only offer a snapshot of this rapidly-evolving landscape at a given point in time. We rarely have the opportunity to witness and document the transformation process itself, as it is happening. This paper is about one such opportunity.

Since 2023, the National Research Council of Canada (NRC) has been providing custom MT technology to the Government of Canada's Translation Bureau (TB), to assist the linguists of its parliamentary service in translating documents published by the Parliament of Canada, most notably the Hansard, i.e. the transcripts of the House of Commons debates, in both of Canada's official languages, French and English. The Hawkeye systems are neural machine translations (NMT) systems, created from Canadian parliamentary texts. Parliamentary translators have access to a number of MT tools, however Hawkeye is the only one that is currently directly integrated into Prism, the Parliament of Canada's publication workflow management application, which is the translators' main work tool (O'Brien, 2002). No translation memory tools are integrated into Prism. Unlike many work environments where postediting MT is mandatory, use of Hawkeye MT is optional: translators are actively encouraged to use it but translations are not automatically pre-populated with MT and translators who want to use it have several user interface options for inserting Hawkeye MT output into their editing window.

Since the official deployment of the systems in May 2023, user interactions with Prism and Hawkeye have been systematically recorded. This anonymized data offers us a unique view on how translators interact with the systems and how their use of Hawkeye MT evolved over time. Because the data include the text of the translations at different stages—initial MT, translator's draft and final revised version—we also have a view of how MT impacts the nature and quality of translators'

work (cf. shorter-term studies like Macken et al. (2020)). In the following pages, we present our analysis of over two and a half years of these user interaction logs.[1]

We first analyze how the use of Hawkeye MT globally evolved over time: we observe that, from 16% of texts when Hawkeye was first deployed in May 2023, the use of the system climbed to 60% in December 2025. We then look at individual (anonymous) translators' work habits with MT to see whether patterns of use emerge: we find that while users are free to decide for each paragraph whether or not to use Hawkeye, in practice, they either use it all the time or not at all. Finally, we examine how the use of Hawkeye affects the translations themselves: we see that when translators chose to insert Hawkeye MT into their translation window, they produce translations that are much more similar to the initial MT than those who don't. This suggests that Hawkeye indeed provides a useful initial draft for those translators.

2 Hawkeye and Prism

The Hawkeye systems are neural machine translations (NMT) systems, trained exclusively with Canadian parliamentary bilingual data. They are based on the Sockeye NMT toolkit (Hieber et al., 2022). Hawkeye systems are created and maintained for four different domains: House of Commons debates, House of Commons committees, Senate debates and Senate committees, but in this work we only focus on House of Commons systems. Distinct systems are produced for English-French and French-English translation. All systems for a given language direction are based on a common system, trained from scratch on all available parliamentary data. Each domain-specific system is then produced by fine-tuning this base system using only the domain-specific data. Table 1 summarizes the data used to produce the latest updated systems, in December 2025. The systems are trained using all parliamentary texts in the proprietary "House of Commons" XML format; as a result, systems handle structured XML input, producing translations with valid XML markup. All systems are updated three times a year, by recreating the systems from scratch, using all data available at that time. More details can be found in Knowles et al. (2024) and Knowles et al. (2023).

		Words	
Domain	**Segments**	**English**	**French**
HoC Debates	7.7 M	135.3 M	148.6 M
HoC Committees	16.6 M	227.1 M	245.7 M
Sen. Debates	0.6 M	11.1 M	12.5 M
Sen. Committees	1.9 M	29.6 M	32.4 M
Total	**26.8 M**	**403.1 M**	**439.3 M**

Table 1: Training data for Hawkeye systems (Dec. 2025).

Hawkeye systems are integrated into Prism (O'Brien, 2002), the workflow application that is the primary means of producing the Hansard and other publications of the Canadian parliament. Within this environment, linguists produce and revise translations using a dedicated three-paned user interface (UI) window: the left-hand pane is used to select from a list of documents that have been assigned to the translator or reviser, the middle pane displays the source-language text to be translated, and the right-hand pane is used to edit the target-language version. The source and target texts within each document are divided into text *chunks*: paragraph-like units that most often correspond to interventions from a single speaker. Both versions also contain structural XML tags, which in the UI are displayed as grey labels that the user can manipulate in various ways.

Translators are free to decide whether or not they use Hawkeye MT: the translation editing pane is not pre-populated with machine translation. Translators who want to use Hawkeye need to explicitly insert MT into their target translation, by activating various UI components. This operation can be performed for all chunks at the beginning of a new document, or individually for each chunk.

3 The Prism-Hawkeye Logs

Systematic recording of data on the use of Hawkeye MT within Prism began shortly after the first systems were deployed, in May 2023. In this work, we examine a dataset of 286,858 records that were collected over a period of 31 months, between 15 May 2023 and 12 December 2025.[2] In what follows, we refer to this dataset informally as the "logs". Table 2 presents general dataset statistics.

Each record in the logs corresponds to a text

[1]The described study design was reviewed and approved by the National Research Council of Canada's Research Ethics Board (NRC-REB #2024-22).

[2]This number of records is computed after filtering out texts with no translation: those that may have been incomplete at the moment of collection or which were translated outside of the Prism interface.

HoC Domain	EN-FR rec.	EN-FR words	FR-EN rec.	FR-EN words	Total rec.	Total words
Debates	104 k	10.9 M	37 k	3.8 M	141 k	14.7 M
Committees	95 k	6.7 M	51 k	3.6 M	146 k	10.3 M
Total	199 k	17.7 M	88 k	7.3 M	287 k	25.0 M

Table 2: Number of records and words in dataset.

chunk, i.e. a segment of text about the size of a paragraph (87.2 words per chunk on average). Each record contains the source-language version of that text (we refer to this field as *SrcText*), along with two or three target-language versions: 1) *MTText*: the machine translated version produced by Hawkeye (regardless of whether or not it was used by the translator), 2) *TrText*: the translation produced by a parliamentary translator, and 3) *RevText*: optionally, a revised translation, as produced by a reviser.

As mentioned above, Hawkeye systems are trained using data from both houses of the Canadian Parliament: the House of Commons and the Senate. However, only House of Commons systems were available to parliamentary translators during the period covered by our analysis. Therefore, in this study, we consider only the two House of Commons domains: *House of Commons Debates* and *House of Commons Committees*, in both translation directions.

We note that parliamentary translators are organized in two teams: the Committees team and the Debates team. Both teams translate documents from the House of Commons and the Senate, have both English and French translators (with some doing both directions), and also have revisers (most of whom also perform translation work). Translations of Debates are required to be systematically revised and to be ready for publication the next morning. Because House of Commons debates normally take place during the day (the House typically sits between 10am and 7pm), the Debates team usually works the evening shift. Translations of committees are not subject to the same tight deadlines and translators of that team normally work the day shift. However, they are often asked to also translate Debates on days with high volumes.

Regarding language, 69.3% of texts in the logs were translated from English into French and 30.7% from French into English. This is representative of the distribution of language interventions in Parliament. The source language of each chunk is also recorded in the logs.

A number of records have an empty *RevText*, i.e. the field containing the final version of the translation, after revisions: 138,557 (48.3%). Most of these are chunks that were ultimately not revised; in these cases, the translator's version *TrText* is the final version.[3] As noted above, translations of Debates are all revised; most Committees' translations in the logs are not. We made this information explicit by adding a Boolean field *Revised* to each record, whose value is based on whether *RevText* is instantiated or not.

Each record also contains anonymized identifiers of the translator and reviser who handled the translation, as well as various timestamps.

Finally, one particularly relevant piece of information is contained in the *HMTinserted* field, a Boolean value indicating whether the translator inserted the Hawkeye MT output into their translation window using dedicated UI components.

4 Analysis

4.1 Use of Hawkeye MT

As mentioned above, the *HMTinserted* field is a Boolean value indicating whether the translator inserted the Hawkeye MT output into their translation. We use that information to analyze the use of Hawkeye MT (sometimes abbreviated HMT below) by translators. Of course, just because a translator had HMT inserted into their translation for a given segment of text does not necessarily mean that they actually based their translation on it; they may have inserted it initially and then later chose to delete it. Conversely, the fact that a translator did not insert HMT does not mean that they did not use any MT at all: parliamentary translators have access to other MT tools, such as DeepL Pro[4] or GCtranslate,[5] which they sometimes use rather than Hawkeye. However, because these MT tools are not integrated into Prism, translators need to manually insert the MT into their translation, either by cutting and pasting from a different window or by typing. We currently have no way of detecting these events. Therefore, in what follows,

[3] A small number of records with empty *RevText* may be ones for which the revision was not yet completed at the moment of data collection.

[4] https://www.deepl.com/en/pro

[5] https://www.canada.ca/en/public-services-procurement/news/2025/09/gctranslate-using-artificial-intelligence-to-build-a-more-agile-modern-and-bilingual-public-service.html

when we use phrases such as "translated with the help of Hawkeye MT", we merely mean that the *HMTinserted* flag was set to *True*. Conversely, when that flag is set to *False*, we cannot assume that the translation was produced without the help of any MT; we simply do not know whether some MT was used or not.

Globally, 46.9% of all text chunks translated in Prism during the period covered by the logs were translated with the use of Hawkeye. Figure 1 shows global HMT use depending on translation direction and the origin of the texts. We observe that HMT was used more for translation into French than for translation into English, and more for the translation of Committee hearings than for Debates.

Figure 2 shows the evolution of Hawkeye use month-by-month between May 2023 and December 2025. Overall, from 16% in May 2023, HMT use gradually increased to over 50% in the fall of 2024 and appears to hover between 50 and 60% since then. Figure 2 also shows this evolution as a function of text origin (middle plot) and language (lower plot). We note that, while HMT was used more for Committees than for Debates overall, since the summer of 2025, HMT usage is comparable for both domains. This suggests that while translators of the Debates team were slower to adopt HMT, their usage later increased, to a point where they now use it equally often.

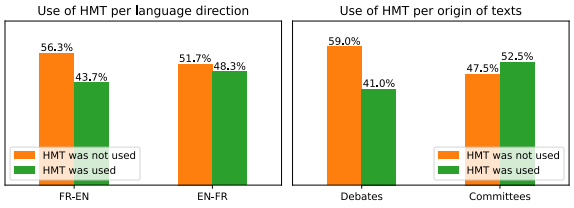

Figure 1: Hawkeye MT use as a function of translation direction (left) and origin of text (right).

4.2 Translators

During the period covered by the logs, a total of 145 linguists (as identified by distinct anonymized IDs) either translated or revised House of Commons texts. Table 3 gives a more detailed breakdown by task, language, and origin of texts. It is worth noting that most revisers also worked as translators during the covered period. Many translators worked on both debates and committees: this is true of 56 French and 25 English translators.

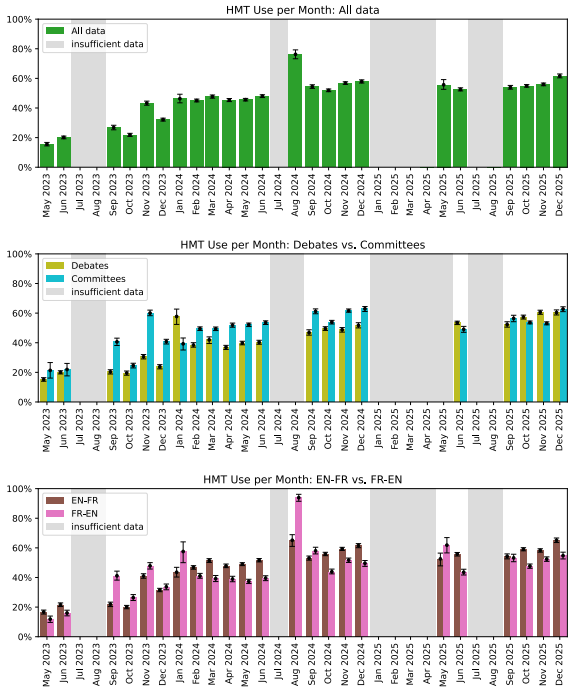

Figure 2: Monthly percentage of Hawkeye MT Use: global (top), compared per text origin (middle) and per language (bottom). Only months during which 20,000 words or more were translated for all conditions are reported; confidence intervals computed by bootstrap resampling (1000 iterations, $p = 0.01$). The gaps correspond to periods during which Parliament was not sitting, most notably the prorogation of January 2025 and the summer pauses.

	Translators		Revisers	
Language	**Debates**	**Comm.**	**Debates**	**Comm.**
EN-FR	75	88	21	11
FR-EN	33	41	11	6

Table 3: Numbers of linguists involved in each parliamentary task.

Finally, 12 translators did both English and French translation, 7 of whom also worked as revisers.

We saw in Section 4.1 that translators have used Hawkeye to help translate approximately 47% of all text between May 2023 and December 2025. But is it that 47% of all translators used HMT all of the time, or that all translators used HMT 47% of the time? To find out, we computed the proportion of text for which each translator used HMT, during different periods.

Figure 3 shows the distribution of translators based on their use of HMT, and how this distribution evolved over time: each shade of blue denotes a different time period. Within these histograms, each column shows the percentage of translators whose use of HMT fell within a given range during the given period. Looking at the lightest shade

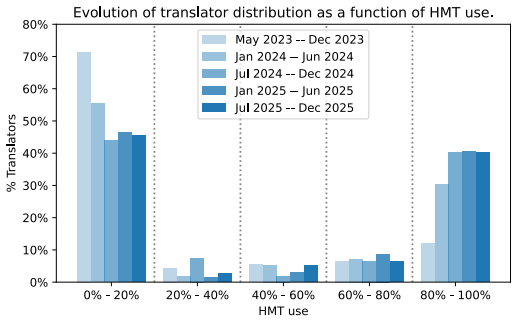

Figure 3: Evolution over time of the distribution of translators based on the proportion of documents for which they used Hawkeye MT.

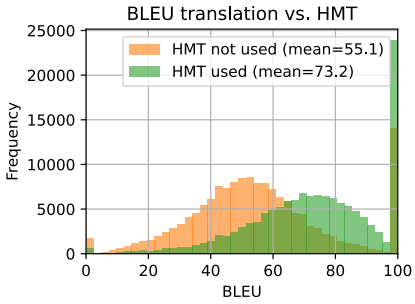

Figure 4: Distribution of textual similarities (BLEU metric) between machine translation output and draft translations, whether Hawkeye MT was used (green) or not (orange).

columns, we see that, between May and December 2023, the vast majority (more than 70%) of translators used HMT for less than 20% of the texts that they translated; for simplicity, we refer to this group as *non-users* of Hawkeye MT. During the same period, about 12% of translators used HMT for more than 80% of the texts that they translated; we call this group *regular users* of Hawkeye. In between these two extremes we find what we call *occasional users*: translators who used Hawkeye for anywhere between 20% and 80% of all their translations: between May and December 2023, these amounted to about 15% of all translators.

As time progresses (darker shades), the proportion of non-users of HMT diminishes, while the proportion of regular users increases. Both groups seem to stabilize around July 2024: non-users at about 45% and regular users at 40%. All this time, the size of the group of occasional users remains relatively stable. We performed a brief manual examination of how individual anonymous translators' use of Hawkeye changed month-to-month. In some cases, the translators who appear (in Figure 3) to be occasional users would be more accurately described as users who are in the process of migrating from one group (non-users or regular users) to the other. That is, they are increasing (or decreasing) their use of Hawkeye over the course of the log file collection time span. This suggests that the vast majority of parliamentary translators are either non-users of Hawkeye (meaning they never or very rarely use it) or regular users (meaning they use it systematically).

4.3 MT Usage and Impact on Revision

We finally examined whether translators actually make use of the Hawkeye machine translation output. This was done by measuring the textual similarity between the MT output (field *MTText*) and the translator's draft (field *TrText*); we then compared the distributions of these similarities under the two conditions: *Hawkeye MT was used* and *Hawkeye MT was not used*. The result of this comparison, using BLEU as similarity metric, is shown in Figure 4.[6] For cases where the *HMTinserted* flag is *True*, the translator's draft is much more similar to the MT output (average similarity is 73.2) than for cases where that same flag takes the value *False* (55.1). This indicates that when translators import MT into the editing pane, they do effectively tend to base their translation on that proposed MT.

A more important question perhaps is whether translations produced with the help of Hawkeye MT differ from those that do not use Hawkeye MT in terms of their quality. As a proxy, we measure the amount of revision edits that were performed by revisers on translators' drafts.[7] For this, we compute the Translation Edit Rate (TER) between the translator's draft (field *TrText*) and the final reviser's version (field *RevText*).[8] This metric can be interpreted as the percentage of words of the text chunk that were edited—inserted, deleted, modified or moved—by the reviser (higher values indicate more revision). We then compare the distributions of these rates under the two conditions: *Hawkeye MT was used* and *Hawkeye MT was not used*. The results are shown in Figure 5.

The dotted lines in this figure show the global

[6]Note that, in this context, BLEU is *not* used to measure MT quality, but merely textual similarity between the MT and the translator's draft – see Appendix A.

[7]This analysis is limited to chunks that were actually revised, i.e. which had non-empty *RevText* and were thus marked with the Boolean field *Revised* set to true.

[8]Note that, in this context, TER is *not* used to measure machine translation quality directly, but rather to measure the amount of revision work performed by revisers – see Appendix A.

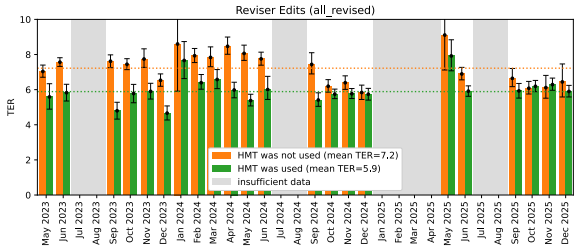

Figure 5: Monthly comparison of the average amount of revision per chunk (measured in TER) performed by revisers on translator drafts, depending on whether these were done with or without the help of Hawkeye MT. Only months during which at least 500 chunks were translated are reported; confidence intervals computed by bootstrap resampling (1000 iterations, $p = 0.05$).

average TER for translations done with and without the help of HMT. Globally, we observe lower TER values for translations produced with the help of Hawkeye MT than for those produced without: the average number of revision edits for chunks produced without Hawkeye MT is 7.22; for those produced with the help of Hawkeye MT, it is 5.89, yielding a statistically significant difference of 1.33 ± 0.16 between the two conditions (estimated through 1000 iterations of bootstrap resampling with $p = 0.01$). However, while this tendency is quite strong in 2023 and 2024, it seems to diminish substantially in 2025, to a point where most differences between the two conditions are not statistically significant. It is difficult to determine whether these variations over time carry any significance. At this point, we simply observe that, based on the amount of edits performed by revisers, translations produced with the help of Hawkeye do not appear to be of lesser quality than those produced without.

Note that this assessment reflects the amount of revisions in terms of surface-level edits, not the severity of errors that revisers observe in draft translations. Also worth noting: as far as we know, apart from the name of the translator, Prism does not provide revisers with information that would allow them to determine whether a given translation was produced with or without the help of Hawkeye.

5 Conclusions

We presented our analysis of user interactions of the translators and revisers of the Translation Bureau's parliamentary service with the Hawkeye MT components of the Prism parliamentary publication workflow management software, based on user interaction data collected between May 2023 and December 2025. Overall, we found that during the covered period, translators used Hawkeye machine translation for 47% of all the texts they translated using the Prism environment. This use increased steadily in 2023 and 2024, from 16% in May 2023 to approximately 55% in the fall of 2024, plateauing at about that level throughout 2025. During that year, Hawkeye was used more frequently for translation into French (57.8%) than into English (50.0%), and more frequently for the translation of Debates (56.4%) than for committee evidence (53.9%).

We observed that most translators either rarely use Hawkeye MT (*non-users*) or almost always use it (*regular users*). In the period between July and December 2025 (last six months of logged period), these two groups represented 44% and 41% of all translators, leaving only about 15% of translators qualifying as *occasional users* of Hawkeye MT.

In terms of quality, translations produced with the help of Hawkeye required fewer revisions on average than those produced without, although the difference between the two conditions seems to fluctuate over time. Again, we note that this analysis of quality was based solely on the amount of edits performed by revisers (in terms of the number of words and punctuation marks revised), not on the gravity of the errors they encountered.

Collection of user interaction data in Prism is ongoing, and we hope to repeat this analysis periodically and follow the evolution of the observed trends and better our understanding of how Hawkeye MT is used by parliamentary translators. We note in passing that the current data collection procedure does not allow us to reliably measure translator and reviser productivity. If measuring this is critical, then additional instrumentation of the Prism translation user interface and/or voluntary user disclosure would be required. Another thing that the current data collection procedure does not allow us to detect is situations where translators (or revisers) use other translation tools, such as DeepL Pro or GCtranslate, to produce translations. Again, this would require different instrumentation.

Finally, and as noted above, there seem to exist differences in the quality of the translations produced with and without the help of Hawkeye MT. This warrants a more in-depth analysis, in which we would examine the nature of the revisions performed on each kind of translation.

Acknowledgments

We wish to thank the parliamentary translation team and all our partners at the Canadian government's Translation Bureau, without whom this work would not have been possible.

References

Cadwell, Patrick, Sharon O'Brien, and Carlos Teixeira. 2018. Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, 26(3):301–321.

do Carmo, Félix and Joss Moorkens. 2022. Translation's new high-tech clothes. In Huertas-Barros, Elsa, Gary Massey, and David Katan, editors, *The human translator in the 2020s*, pages 11–26. Routledge.

Guerberof Arenas, Ana. 2025. Perspectives on machine translation, post-editing, and automation. In Walker, Callum and Joseph Lambert, editors, *The Routledge handbook of the translation industry*, pages 186–204. Routledge.

Hieber, Felix, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with pytorch.

Knowles, Rebecca, Samuel Larkin, Marc Tessier, and Michel Simard. 2023. Terminology in neural machine translation: A case study of the Canadian Hansard. In Nurminen, Mary, 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, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 481–488, Tampere, Finland, June. European Association for Machine Translation.

Knowles, Rebecca, Samuel Larkin, Michel Simard, Marc A Tessier, Gabriel Bernier-Colborne, Cyril Goutte, and Chi-kiu Lo. 2024. Some tradeoffs in continual learning for parliamentary neural machine translation systems. In Knowles, Rebecca, Akiko Eriguchi, and Shivali Goel, editors, *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–118, Chicago, USA, September. Association for Machine Translation in the Americas.

Macken, Lieve, Daniel Prou, and Arda Tezcan. 2020. Quantifying the effect of machine translation in a high-quality human translation production process. *Informatics*, 7(2).

Nunes Vieira, Lucas and Elisa Alonso. 2019. Translating perceptions and managing expectations: An analysis of management and production perspectives on machine translation. *Perspectives*, 28:1–22.

O'Brien, Audrey. 2002. Prism: The house of commons integrated technology project. *Canadian Parliamentary Review*, 25(2):20–22.

O'Brien, Sharon. 2024. Human-centered augmented translation: against antagonistic dualisms. *Perspectives*, 32(3):391–406.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Zaretskaya, Anna. 2015. The use of machine translation among professional translators. In *Proceedings of the EXPERT Scientific and Technological Workshop*, pages 1–12.

A Measures of Textual Similarity

In different parts of this study, we find the need to quantify the differences between different translations of the same text. We base these comparisons using *textual similarity metrics*: functions that produce values that can be interpreted as the degree of similarity between two strings of text.

In practice, the metrics that we use in this study are standard machine translation evaluation metrics: BLEU, ChrF and TER. Here, however, we use these MT evaluation metrics strictly as textual similarity metrics, and *not* to directly measure MT (or postedited text) quality against a reference. For example, in Section 4.3, we use BLEU to compare the similarity between Hawkeye outputs (hypotheses) and the translators' drafts (references) under two different conditions; in that same section, we use TER to measure the similarity between translators' drafts (hypotheses) and final, revised translations of the same texts (references). In both of these examples, these metrics' values should not be interpreted as denoting machine translation quality. In the case of BLEU, we are examining how much of the original MT appears in the translator draft. In the case of TER, we are examining

how many revisions were made to the translator's draft, conditioned on whether Hawkeye MT was selected at translation time or not (without knowing directly whether it was used or not).

We use the `sacrebleu`[9] implementation of BLEU, ChrF and TER machine translation evaluation metrics Post (2018), with signatures provided in Table 4.

[9] `https://github.com/mjpost/sacrebleu`

BLEU	`nrefs:1|case:mixed|eff:yes|tok:13a|smooth:add-k[1.00]|version:2.4.3`
ChrF	`nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.3`
TER	`nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.4.3`

Table 4: Sacrebleu signatures for metrics used in this paper. Because we average similarities between relatively short segments of text, we need to compute BLEU scores with smoothing: we use *add-1* smoothing.