Translators' Perceptions and Edit Traces: Quality of MT as a Tool in Canadian Parliamentary Translation

Jeniffer Leal-Wyss Jeniffer.Leal@nrc-cnrc.gc.ca
Gabriel Bernier-Colborne Gabriel.Bernier-Colborne@nrc-cnrc.gc.ca
Delaney Lothian Delaney.Lothian@nrc-cnrc.gc.ca
Michel Simard Michel.Simard@nrc-cnrc.gc.ca
Rebecca Knowles Rebecca.Knowles@nrc-cnrc.gc.ca
Digital Technologies Research Centre, National Research Council of Canada, Ottawa, ON, Canada

Abstract

The Parliament of Canada's translation workflow includes access to a specialized neural machine translation (NMT) system. This study analyzes post-editing (PE) activity to identify the types of edits translators make when interacting with the NMT system, as well as the frequency, nature, and severity of errors encountered. We compare translations produced with and without the use of this NMT system to evaluate potential differences in edit patterns. To complement this analysis, we draw on insights from a user study. Our findings explore how translators' perceptions align with observed PE patterns and how their feedback can inform strategies to better understand, and possibly mitigate, some of the errors observed.

1 Introduction

At the Parliament of Canada, the transcripts of the debates in the House of Commons (and the Senate) are translated into both official languages, French and English. These translation services are provided by the Translation Bureau, the federal institution responsible for delivering translation services to the Government of Canada. As part of their workflow, translators for the House of Commons have the option to use a custom machine translation (MT) system called Hawkeye as a translation aid. Hawkeye is an NMT system integrated into their work environment based on the Sockeye toolkit (Hieber et al., 2022), trained on parliamentary text, and updated multiple times a year.

A previous study examining logs of user interactions with Hawkeye (Simard et al., 2026) in the House of Commons debates and committees texts showed an increase in MT usage, from 16% in May 2023 to close to 60% in December 2025. In that same work, a comparison of the number of changes made by revisers (members of the parliamentary translation team who revise the translations before publication) to texts produced by translators with or without the help of Hawkeye showed that the texts produced with Hawkeye required a statistically significantly smaller number of reviser edits than texts produced without. While that work quantifies some differences in translations with and without this MT system, it doesn't provide us with information about the types of changes being made to the MT output or to the translators' reactions to MT errors. This motivates us to explore the impact of MT on translation and revision patterns in greater detail.

In this paper, we present a study that examines post-edited text samples, where we automatically detect spans that were modified during post-editing (PE) and revision, and then perform an analysis of error types, severity, and frequency using the MQM framework (Lommel et al., 2013). Additionally, we compare edits made during revision of texts translated with and without Hawkeye. Our analy-

sis shows that translations produced with the help of Hawkeye not only require fewer revisions, they also exhibit slightly lower error severity than those translated without Hawkeye. We complement this data with insights from a user study[1] involving parliamentary translators through questionnaires and interviews, highlighting their perceptions of MT quality. Through the analysis of text samples and users' perceptions, we gained valuable insights into the nature, severity and frequency of Hawkeye's errors, and its utility in this translation workflow. Our study supports that complementing data analysis with user feedback can contribute to a better understanding of translation quality.

2 Background and Related Work

In this study, we perform an assessment of MT output in a PE workflow. Thus, related research in translation studies and natural language processing provides valuable insights to inform this work.

The concept of quality in translation has often been debated, as it is inherently tied to expectations. However, research in translation quality evaluation (TQE) reminds us that quality should be defined in relation to the intended purpose of the translation or MT system being assessed (Bowker, 2020). In line with this perspective, we emphasize that the MT system under evaluation is designed to serve as an aid for translators, and will therefore be assessed according to that purpose.

Different methodologies and procedures have been used for TQE of MT, including human evaluation, which typically includes rating, ranking and annotation, as well as automatic methods, which are often based on lexical and/or semantic similarity to a reference translation. Human TQE is typically considered the gold standard, and it can be grouped into three main categories: manual scoring, semi-automatic, and task-based (Liu et al., 2024).

Semi-automatic evaluations are based on annotations from which a score is computed. For that purpose, a text is first annotated or post-edited, and then an edit distance algorithm can be used to identify edited spans in MT output (Snover et al., 2006). In cases where post-edited data are accessible, as in this study, this approach allows for analysis of corrections to MT output. This step can be complemented by various evaluation frameworks that typically involve categorizing errors and assigning scores based on their number, type, and severity. One widely used framework is MQM (Lommel et al., 2013), which offers a fine-grained approach to TQE, although other valuable frameworks also exist (O'Brien, 2012; Graham et al., 2013; Kocmi et al., 2024; Magdy et al., 2026, i.a.). The use of a framework such as MQM can be valuable to understand error frequency, type, and severity, and to assess the general quality of a system, as shown by various studies (Ortiz-Boix and Matamala, 2017; Freitag et al., 2021; Park and Padó, 2024; Li et al., 2025). As MQM has a mostly standardized error taxonomy and is widely used for human evaluation, it can enable comparison of results across studies.

Other studies focusing on perceptions of MT in translation environments have employed more qualitative data collection methods. These perception-based methods are different from the ones described above in that they do not assess specific translation samples, but rather rely on subjective assessment, using instruments such as questionnaires, focus groups or interviews (Vieira and Alonso, 2020; Cadwell et al., 2018) or a combination of quantitative and qualitative data collection methods (Moorkens et al., 2018; Yang et al., 2023; Heinisch and Lušicky, 2019). Such research methods are helpful for understanding the usability of MT output for system users, capturing examples that might not be found in the samples selected for post-hoc evaluation by researchers, and assessing perceptions of MT quality as a computer-assisted translation (CAT) tool.

Closely related to the work described here are prior publications on MT and the Canadian Parliament, specifically the quantitative analysis of MT usage by translators over time in Simard et al. (2026) and the analysis of translators' reasons for adopting (or not adopting) the Hawkeye MT systems as part of their workflows described in Leal-Wyss et al. (2026). The latter draws on the the same pool of user study data that we analyze here, though it focuses on aspects of translator experience rather than questions of MT quality.

3 Methodology and Data

This paper uses two main approaches to assess quality and utility of MT: annotation and evaluation of two sets of translation samples (Sec. 3.1) and analy-

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

sis of user perceptions of system quality (Sec. 3.2).

We were provided access to a large collection of Parliament translation log files containing the source text, the Hawkeye MT output, the initial translation (which may or may not have been based on Hawkeye output), the revised translation, a flag indicating whether the translator invoked Hawkeye to produce the initial draft, and other metadata (for more details, see Simard et al., 2026). We sampled data for annotation (Sec. 3.1.2) from this collection.

3.1 Annotation and Evaluation

We refer to the two subsets of our annotation and evaluation data as "revision data" (REV for short) and "post-editing and revision data" (PE+REV). For REV, we compare the translation drafts produced by translators to the final revised version; we split this into two groups: translation drafts produced with the use of Hawkeye or without the use of Hawkeye. For PE+REV, we compare the original Hawkeye MT output to the final revised version of the translation; this encompasses two steps: PE and revision together (hence the name).

3.1.1 Sample Annotation Schema

Two members of our team (fluent in English and French, with formal training in translation studies and experience in translation) performed blind MQM annotation on texts from both the REV and PE+REV datasets; these annotations were revised by the other annotator and disagreements were resolved through discussion. The PE+REV samples are a subset of the REV samples.

We automatically compared the first target text (Hawkeye MT in the case of PE+REV, translator's draft in the case of REV) to the final revised version, highlighting and numbering the edited spans to make them clearly visible to the annotators. We then used the MQM framework (more specifically, MQM-Core[2]) for annotation, which organizes translation issues into top-level categories, such as *Accuracy* and *Terminology*, each of which includes subcategories. This is a subset of the full MQM schema comprising eight high-level error categories and the most common subtypes. It also includes a scoring model that rates errors by severity. In our sample analysis, we applied this framework to TQE of texts produced with and without the use of Hawkeye, then calculated the frequency of error types and severity. Each highlighted error was assigned a top-level error type, subtype, and a severity level. In cases where multiple automatically-detected edits would be better interpreted as a single error, the annotators could cluster them into a single error. We do not compute global error scores weighted by severity, but rather use this analysis to understand edits made to MT output and to compare texts that used and didn't use Hawkeye as a translation aid to assess which sample requires more edits and which sample shows more severe errors.

The annotation schema selected is not without challenges. One limitation is the absence of an explicit category for *preferential* changes. In such cases, we applied the severity level *Neutral* to indicate that these are considered stylistic preferences, and we classified them under *Style – Unidiomatic*. In many instances, this classification was chosen not because the initial draft was unidiomatic but because the revised version offered a more idiomatic rendering. We highlight that all errors classified as *Style – Unidiomatic* that have a *Neutral* severity are preferential – for those that were considered required changes, the *Minor* severity was used. We also found that certain errors could reasonably fit multiple categories, so the annotators discussed ambiguous errors and established their own guidelines to resolve these.

It is important to note that these are post-hoc analyses of edits; we do not have access to the actual tracked changes, which is why we use an edit distance algorithm to identify edited spans. It is possible that a translator has in fact deleted the MT output and then typed their own text from scratch (we do not have a way to automatically detect this), and we cannot assume that the non-Hawkeye subset of REV is MT-free, as translators may have post-edited the output of other (non-Hawkeye) MT systems.

3.1.2 Dataset for Annotation

From our collection of data, we sampled "chunks" (roughly paragraphs, which are further subdivided into "segments", which are usually a single sentence, but some contain two or three) for which Hawkeye MT was used at translation time, uniformly at random; these are annotated in both the PE+REV and REV settings. We also sampled an

[2]As described in Lommel et al. (2024) and https://themqm.org/the-mqm-typology/

equal number of chunks for which Hawkeye MT was not used; these are annotated in only the REV setting. This produced a dataset containing 1393 segments. Statistics on this dataset are shown in Table 1. See Appendix A for details on data sampling; we focused only on House of Commons debates data for this annotation.

	PE+REV	REV
# chunks	114	230
# segments	784	1393
# words	14,737	26,197

Table 1: Numbers of chunks, segments, and words (i.e. space-separated tokens) in the sampled dataset.

3.2 User Study

The user study employed a mixed-methods explanatory sequential design (Creswell and Creswell, 2017). First, a questionnaire was used to collect data on users' experience with Hawkeye.[3] The findings from this stage informed the design of semi-structured interviews in the second stage, which were intended to expand upon the results of the questionnaire.

All participants for both stages were recruited from a pool of 66 parliamentary translators from the Translation Bureau. Both the questionnaire and interview protocol included consent forms, and participation was strictly voluntary. Participants were recruited via an email written by the research team and distributed by the Translation Bureau.

3.2.1 Questionnaire

The questionnaire was designed to gauge respondents' experience with the Hawkeye MT tool. It was written in English and translated into French, and was deployed online using the Voxco survey software platform.

The questionnaire includes a matrix question regarding error types and frequencies, where participants were provided with MQM high-level categories along with a short explanation of each; a second matrix question regarding error severities and frequencies; a question regarding overall assessments of quality improvements over time; as well as a question about the error types that users would most like to have eliminated. For the matrix question, participants were presented with the eight top-level MQM-Core categories and asked to indicate the frequency of each type of error. We chose not to include the subcategories, as these would have been difficult to assess based solely on impressions and memory without reference to specific translation samples. This design aimed to capture general perceptions of the frequency of error types in the MT output. We intentionally kept error types and frequency distinct from severity, using two separate matrix questions, since integrated designs are more appropriate when evaluating specific translation samples. Treating them separately allowed us to assess these dimensions as two complementary but independent aspects of MT performance.

We sent 66 questionnaire invitations and received 39 responses (total response rate: 59%). Of these, 4 responses were excluded after determining that participants were confused about the tool under evaluation. In this paper, we present only the section of the questionnaire intended for participants who had used Hawkeye (a subset of the original 35 valid responses, which is n=26), since they are the only participants who can assess the system's quality. We analyzed the quantitative data using descriptive statistics and used thematic analysis (Naeem et al., 2023) for the open-ended questions.

3.2.2 Interviews

The second stage of this user study used semi-structured interviews, which asked participants about their perception of the Hawkeye MT system and other CAT tools. Invitation e-mails were sent to participants who had volunteered through an optional form at the end of the questionnaire. Interviews were held remotely in French or English, led by a primary interviewer with support from a note-taker. They lasted up to 45 minutes and were all recorded. In total, 10 participants were interviewed. Thematic analysis used an inductive coding process, in which themes were identified and collected from participant open answer data.

4 Analysis of Edits

Overall error counts and frequencies in both the REV and PE+REV data are shown in Table 2. An error in this context is a manually annotated error (with MQM category and severity label) which

[3]For the full text of the questionnaire, see the appendices of Leal-Wyss et al. (2026).

	REV-ALL		REV-NO-HMT		REV-WITH-HMT		PE+REV	
	EN-FR	FR-EN	EN-FR	FR-EN	EN-FR	FR-EN	EN-FR	FR-EN
# errors	379	381	206	194	173	187	631	641
# segments	668	725	282	336	386	389	398	386
Errors/seg.	0.567	0.526	0.730	0.577	0.448	0.481	1.585	1.661

Table 2: Error frequencies. "HMT" means Hawkeye Machine Translation. An error (manually annotated) can correspond to one or more of the automatically-detected edit spans (merged by annotators when multiple automatically-detected errors spans could best be understood as a single error).

can correspond to one or more of the automatically-detected edit spans. The frequency distribution of error counts per segment is shown in Appendix C. While PE+REV counts are shown in the same table, they should not be directly compared to the REV counts, as the former covers the full set of changes from raw MT output to revision, while the latter covers only changes from the first translator's draft to revision.

4.1 Edits when Hawkeye is Used

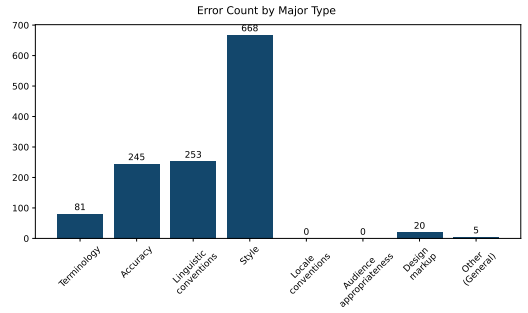

Figure 1: Frequency distribution of error categories in the PE+REV data

In this section, we present our results related to edits made to the Hawkeye MT output in PE+REV with a focus on two aspects: error type and severity. As shown in Fig. 1 and Fig. 2, the four most frequent error categories identified in PE+REV are *style*, *linguistic conventions*, *accuracy* and *terminology*, while the two most frequent error severity categories identified are *Minor* and *Neutral* errors. Note that we obtained similar distributions for the REV data (see Fig. 8 in Appendix D and Fig. 9 in Appendix E).

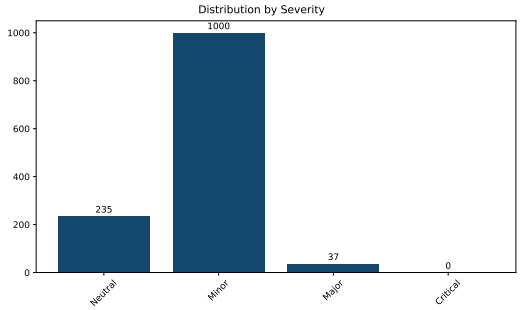

Figure 2: Frequency distribution of error severities in the PE+REV data

In Figure 3, we present a Sankey diagram showing the main categories and subcategories of errors in the PE+REV data along with their corresponding severities (the analagous diagram for the REV data is shown in Fig. 12, Appendix G).

A closer examination of style errors reveals that many of these corrections relate to unidiomatic or awkward phrasing, as well as non-compliance with organizational style. The latter occur when the MT output does not adhere to institutional guidelines, for example, using regular spaces instead of non-breaking spaces, regular hyphens instead of non-breaking hyphens, or quotation marks that differ from those recommended in internal style guidelines, among others.

Accuracy errors most often involve mistranslations or omissions. In the case of mistranslations, we observed recurring issues with false friends and overly literal renderings. Although these issues may appear quite different from a style error that results in unidiomatic phrasing, these behaviours are not entirely unrelated in terms of their explainability, as they are both caused by a translation that is too close to the source text. Moreover, in the context of MT, it may be preferable for a system to produce a trans-

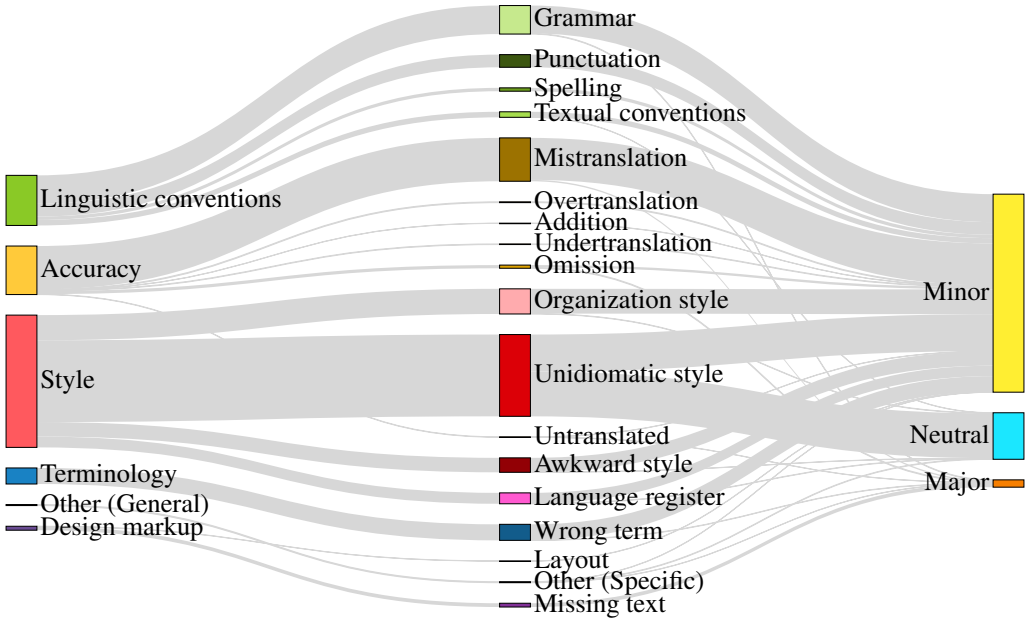

Figure 3: Sankey plot of Error Categories, Subcategories, and Severities in the PE+REV data.

lation that remains closer to the source text than one that introduces hallucinations.

Errors related to terminology corresponded in all instances to the use of an incorrect term (e.g. "bills" being translated as "*projets de loi*" instead of "*factures*"). Terminology issues were not necessarily tied to parliamentary language, and involved a variety of possible explanations, including the training data and literal translations (false friends).

For linguistic convention issues, we found that these most often involved grammar, textual conventions (e.g. cohesion, coherence, repetition), spelling, and punctuation issues. In many instances, errors occurred in the use of a tense, gender, or number within a given verb form, which could be attributed to the MT system's limited context scope (individual segments). We often noticed cases of wrong gender, especially when translating into French, (for example, "*le député*" instead of "*la députée*"). These patterns suggest that these errors are in many cases related to the fact that the system operates at the sentence level and therefore lacks access to contextual information beyond individual sentences, or world knowledge.

Our analysis of PE+REV samples did not find any issues related to locale conventions (such as number, currency, measurement, date or time format), or issues related to audience appropriateness (which can include culture-specific references or offensive language). We suspect this to be related to the parliamentary data used for training.

As noted above, most errors were found to be *minor* or *neutral* in severity, with only a few identified as *major* and none as *critical*. This distribution reflects the desired behaviour for an MT system intended for PE, where corrections involve non-critical issues that, if left unedited, would have minimal impact on the target audience. Thus, the MT output quality is deemed as adequate for PE in a human-centred process where the MT tool serves as an aid for translators.

Regarding the mapping between edited spans and discrete translation errors, the distribution of the number of spans per error (shown in Appendix F) indicates that the vast majority of errors map to a single edited span, but some errors map to 2 or 3 spans. In rare cases (likely, preferential stylistic edits that we merged together), the number of edits is even higher.

4.2 Impact of Hawkeye MT on Revision

We observe that REV data (revisions of translator drafts) for which Hawkeye was used in the initial draft contained fewer errors than segments for which it was not, as shown in Table 2.

REV samples show 0.567 errors per segment EN-FR and 0.526 FR-EN, on average. Of these, texts translated using Hawkeye exhibited 0.448 errors per segment in EN-FR and 0.481 in FR-EN, while texts translated without Hawkeye (REV-NO-HMT) showed a higher number of errors per segment in EN-FR (0.730) than in FR-EN (0.577). This aligns with the results of the initial log file analysis that text translated with the help of Hawkeye required less revision than text translated without it. The difference in errors/segment between REV-NO-HMT and REV-WITH-HMT is statistically significant only in the EN-FR direction.[4]

The distribution of error severity did not show notable differences between texts translated with Hawkeye and texts translated without Hawkeye. In both cases, a majority of errors were found to be either *minor* or *neutral*, with very few *major* errors (see Fig. 9 in Appendix E). We observe, however, that the only critical error found in our sample comes from a text that did not use Hawkeye, corresponding to a sentence that was entirely omitted, and that the number of major errors is higher in texts that didn't use Hawkeye, suggesting that its use can contribute to quality improvements.

In terms of error types, the texts translated with and without Hawkeye followed similar trends, though with some variation in distribution. The most common error category for translations produced with Hawkeye was style errors, followed by issues related to linguistic conventions, accuracy, terminology, and design/markup. Similarly, translations produced without Hawkeye also showed a predominance of style errors, followed by linguistic conventions, accuracy, terminology, and design/markup.

5 User Perceptions of MT Quality

Our questionnaire aimed to identify the types of errors translators perceived to be present in the MT output. To do so, the 26 participants who had used Hawkeye were presented with a matrix question based on MQM error categories, with frequency options. Based on the responses, summarized in Fig. 4, the most frequent MT errors are perceived to be related to *Accuracy* with 23 and *Terminology* with 21 responses indicating that these are perceived to occur sometimes or often, followed by *Style* errors, with 14 respondents reporting them sometimes or often. *Linguistic conventions* errors are perceived to occur less frequently, with 9 responses noting that these appear sometimes or often, followed by *audience appropriateness* with 4 participants reporting them as sometimes or often, while *design and markup* errors are reported to appear sometimes or often by 5 participants, and *locale conventions* errors are reported to be less frequent with 3 participants noting these appear sometimes or often.

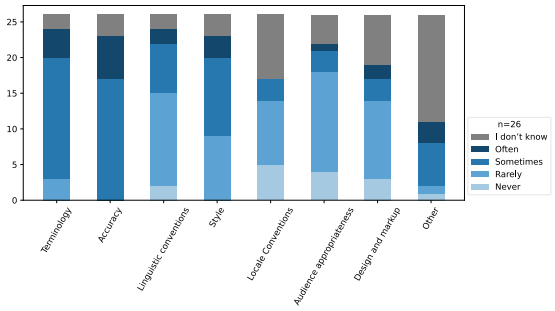

Figure 4: Question: *When using the Hawkeye translation tool, what types of errors do you usually observe, and how frequently do you observe them?*

We also asked these 26 participants a matrix question about error severity to determine how often *critical, major, minor,* and *neutral* errors were perceived to be found in the NMT output. Responses are summarized in Fig. 5. The most frequently reported severity was *minor errors* with 22 participants reporting these appeared sometimes or often, followed by *neutral errors* with 21 reporting them sometimes or often. Conversely, *major errors* were reported to be less frequent, with 13 participants reporting they appear sometimes and no participants reporting they appear often, and *critical errors* were the least frequent with 2 participants reporting them sometimes and no participants reporting them often. This suggests that less severe errors are perceived to occur more frequently, whereas more severe errors are relatively rare.

[4]Mann-Whitney U test (since the distribution is not normal) with $p < 0.01$.

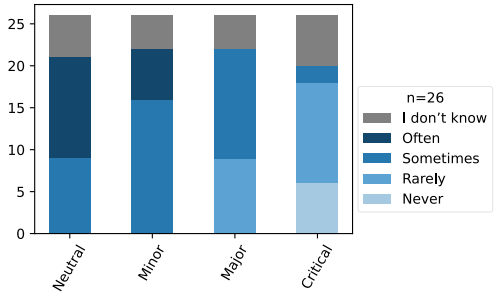

Figure 5: Question: *Based on the above defined error severities, please indicate the frequency of error severity occurrence in the Hawkeye translation output.*

We also asked these 26 participants about the types of errors they would most like to see eliminated from the MT output (multiple responses allowed). The three most frequent answers were accuracy (23 responses), terminology (17), and linguistic conventions (7). They also had the option to select "Other" and provide comments in an open text box. Here, translators noted that they often modify the MT output because it is too literal or too close to the source text. Others highlighted issues such as inconsistent terminology, some hallucinations involving proper names, reduced lexical variation, and stylistic problems specific to the House of Commons.

During the interviews, we explored quality in greater depth. We asked participants (n=10) about the types of errors they observed in the MT system, and how these compared to human errors (*Consider the types of errors that you observe in the Hawkeye machine translation output. How do they compare against errors you might have observed in human translation?*). Participants mentioned issues commonly associated with MT systems, including challenges with informal or everyday language, idiomatic expressions, figurative language, literal formulations, long and complex paragraphs, co-reference (difficulties identifying correct pronouns), proper names, omissions, hallucinations, additions, as well as misguided attempts to disambiguate vague discourse. Many interviewees also noted that the system's output was generally accurate when handling specialized parliamentary language (6/10 participants) but tended to perform less effectively with more general language—a difference likely related to the data used to train the system. We illustrate this issue with an example taken from the interviews:

"I remember a specific example: someone was thanking his wife for taking care of the cat litter on her own, and Hawkeye rendered it as something like a 'trash can' for the cats or something along those lines, even though it's really not a specialized or difficult term to translate." (English translation, Interview Participant 4)[5]

This set of challenges for MT have been previously discussed in the literature (Naveen and Trojovský, 2024) and some are known limitations of NMT systems. While some of these issues may be difficult to fully resolve, others may be easier to address. For instance, co-reference and consistency issues might be possible to improve on through the use of larger context windows or with external information sources. Overall, even considering the observations about MT output quality, it appears that translators perceived the quality to be acceptable in a PE setting, as most errors are reported to be either minor or neutral, and only a minority are perceived to be critical or major.

6 Comparison Between Perceptions and What Translators Change

While we have found differences between actual post-edits observed in our sample and perceived MT errors reported by translators, we have also found convergence in other areas. For example, in terms of error severity, most errors in our sample analysis were found to be *minor* or *neutral*, with few *major* errors, which matches participants' perceptions as reported in our user study.

Regarding perception of error types reported in our user study, we observed slight differences compared to the annotated samples, but overall trends remained similar. In our sample, the most common post-edits fell within the categories of *style*, *linguistic conventions*, *accuracy*, and *terminology*. This finding is noteworthy, as these categories were also perceived as frequently found in the MT system and as those which translators would most like to see eliminated. However, the distribution shows slight variations. One source of these differences could be that user study participants are describing all of their experiences with Hawkeye (both debates and com-

[5]Original French versions of quotations are provided in Appendix B.

mittees), while our annotations only focus on debates data; in particular, it may be the case that some committees use more challenging technical terminology. These differences could also be explained by participants' varying levels of familiarity with the MQM framework, despite the questionnaire including a brief explanation of its categories, or because some MQM categories may be particularly open to variable interpretations. Other aspects, such as how memorable errors are, or how challenging they are to revise, may also play a role in this discrepancy and in the similarity between the distribution of the errors perceived as frequent and those that translators would most like to see improved.

Errors related to *linguistic conventions* were the second most common errors in our annotations, while participants perceived them as less common than terminology, accuracy, or style errors. The most common of these errors observed were related mainly to grammar, including changes in tense, gender or number agreement, pronouns, and prepositions. Although these errors were not perceived as particularly frequent in the questionnaire responses, feedback collected during the interviews did identify them as common issues in Hawkeye's output. This finding aligns with our sample analysis, as illustrated by one participant: "*Certainly, one of the most common errors that I pick up in Hawkeye is... and you know, I translate into English, I don't think this is as big a problem going into French, they have their own set of problems, but... Using 'he' or 'she' instead of 'it'. That happens a lot in Hawkeye.*" (English original quotation, Interview participant 2).

Additionally, *terminology* errors were perceived to be among the most common error types, while our annotation sample found them to be less common. This may reflect the fact that certain issues, such as the incorrect translation of a term, could have been interpreted as mistranslations, which fall under the category of *accuracy*. For example, in our sample, the translation of "food" as "nourriture" was corrected to "denrées alimentaires." This example shows a case that could cause confusion because "food" may not necessarily be considered a term in most contexts, although it was in this specific context. This example highlights a limitation of this framework, as well as of others: certain nuances may rely on personal interpretation.

The most common errors found in our samples were *style*-related. Although translators also perceived *style* issues as frequent, they did not consider them the most common error type in Hawkeye's output. Certain client-required revisions, categorized as *organizational style* in our analysis, may have been classified under different categories by translators. For example, this could occur when corrections are made to comply with capitalization rules that are specific to the House of Commons. While capitalization issues would normally be considered *linguistic convention* errors in many cases, certain capitalization corrections would be acceptable in other contexts. Because the correction is made to ensure compliance with the client's style guide, the error is more appropriately categorized as a *Style – Organizational Style* error.

Regarding *accuracy* errors, translators identified these as one of the most frequent categories in Hawkeye's output. They noted frequent mistranslations of idiomatic expressions and overly literal translations, as well as occasional omissions and hallucinations, particularly when the MT attempted to disambiguate vague discourse. This is consistent with the findings from our sample analysis, in which *mistranslation* was the most common subcategory of *accuracy* issues.

Additionally, the types of errors that translators reported as infrequent, such as *audience appropriateness*, which can include cultural-specific references or offensive language, and *locale conventions*, which can include number, currency, and measurement formats, among others, were not identified in our sample analysis. This is likely because the translators worked with greater volumes of texts, where such issues may arise occasionally, but they did not appear in our relatively small sample.

An interesting finding from our analysis is that many of the edits made by translators are closely tied to the text type. Parliamentary translators often work with transcripts of speeches, which are naturally vague, ambiguous, or loosely structured due to the characteristics of spoken language. As a result, translators frequently have to restructure sentences or adapt them for the target audience, and add information or cues that are implicit in the context of the speech but not explicitly stated by the speaker, helping to clarify meaning for the reader, or otherwise adapt the text to the target audience. Currently, most MT systems are not capable of performing this

kind of contextual enrichment, making human intervention necessary. A comment from one of the interviewees supports this observation: "*To be honest, sometimes there are things that just can't be translated—or that translate extremely poorly—and at that point, it's no longer a matter of translation; it's outright adaptation, so… and I don't think that's what we're asking from the tool for now*." (English translation, Interview Participant 3)

We would like to highlight that the feedback provided by translators during our user study complements the data collected through our sample analysis. The insights gathered offer valuable examples and explanations about parliamentary translation that are essential for understanding Hawkeye's quality and could help guide future work with the system. Moreover, the feedback sheds light on recurring issues and potential areas for improvement.

7 Limitations and Future Work

A key limitation of this study concerns the sample evaluation. Although our research team made use of all available user reference materials and engaged annotators with expertise in translation, the fact that the annotators were not the actual users introduces the possibility of judgment discrepancies. Users may categorize certain errors differently, drawing on their own knowledge and reference materials that were not accessible to our team. Despite this limitation, we consider the evaluation framework to be valuable, as it provides a meaningful overview of the types of errors identified in the MT system under study. Further, collecting multiple, independent annotations would allow us to measure agreement and identify the most confusable pairs of error categories, and possibly revise our annotation schema.

While MQM analysis proved valuable for comparing perceived errors to actual PE, we acknowledge that this framework has limitations in this context. Identifying error types and assessing their severity provide meaningful insights into a system's overall performance and behaviour; however, to achieve a deeper understanding of the factors driving specific system behaviours, an additional dimension—error explainability—is needed. Going forward, we could perform a targeted analysis of error samples that considers error type, explanation of causes, and potential resolution. For example, if a terminology error is traced back to training data, an appropriate remedy could be to augment the training corpus with additional relevant data.

For convenience, we have treated a sequence of MT systems deployed overtime as a single system. Since the user study was conducted only once, it does not allow for a fine-grained analysis of changes in MT output over time. However, responses to a question on the evolution of system performance showed that a plurality of Hawkeye users perceived an improvement in MT output quality. Further information on the system's evolution can be found in Simard et al. (2026), which examines the evolution of system adoption and compares the number of revisions made to Hawkeye and non-Hawkeye translations over time. We also note that the main change in the updated systems during the period covered by this study was the use of increasing amounts of training data.

Finally, we note that in previous work (Simard et al., 2026), we found that most translators used Hawkeye either rarely or almost all the time. As a result, when comparing translations done with and without the help of Hawkeye, we are in fact comparing translations done by two mostly distinct groups of translators. This means that within our current study, we cannot fully disentangle differences based on the use of Hawkeye from individual translator differences.

8 Conclusions

Our study provides evidence that Hawkeye is a useful tool for translators working for the Canadian Parliament, as its use leads to a reduced number of edits by revisers compared to translations produced without the Hawkeye system, and to less severe errors. We note that our comparisons are observational; however, the insights gained from this study could help improve the system and, in turn, have a direct impact on translators' working conditions.

This study also provided valuable insight into how frequently certain types of errors occur and which of them may be more easily addressed. Such understanding would not have been possible through sample analysis alone; it was enriched by the feedback and perspectives of translators who regularly interact with the system. For this reason, we believe that close collaboration between developers and users is essential to ensure a process of continuous improvement of NMT systems.

Acknowledgements

We would like to thank our study participants, the translators working for the parliamentary translation team, as well as the Canadian government's Translation Bureau, without whom this work would not have been possible. We also thank Kerry Butt for his assistance with the deployment of our questionnaire. Lastly, we would like to thank our colleagues Marc Tessier and Jackie Lo for feedback on this work.

References

Bowker, L. (2020). Fit-for-purpose translation. In O'Hagan, M., editor, *The Routledge Handbook of Translation and Technology*, volume 1, pages 453–468. Routledge, 1st edition.

Cadwell, P., O'Brien, S., and Teixeira, C. (2018). Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, pages 1–21.

Creswell, J. W. and Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In Pareja-Lora, A., Liakata, M., and Dipper, S., editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Heinisch, B. and Lušicky, V. (2019). User expectations towards machine translation: A case study. In Forcada, M., Way, A., Tinsley, J., Shterionov, D., Rico, C., and Gaspari, F., editors, *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 42–48, Dublin, Ireland. European Association for Machine Translation.

Hieber, F., Denkowski, C. D., Canada, E., August, D., Québec City, Canada, August 31 - September 2, 2026. Volume 1: Research Track
C. D., Niu, X., Hoang, C., Tran, K., Hsu, B., Nadejde, M., Lakew, S., Mathur, P., Currey, A., and Federico, M. (2022). Sockeye 3: Fast neural machine translation with pytorch.

Kocmi, T., Zouhar, V., Avramidis, E., Grundkiewicz, R., Karpinska, M., Popović, M., Sachan, M., and Shmatova, M. (2024). Error span annotation: A balanced approach for human evaluation of machine translation. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Leal-Wyss, J., Lothian, D., Bernier-Colborne, G., Knowles, R., and Simard, M. (2026). "all those in favour will please say yea"': Understanding the factors behind machine translation adoption at the Canadian Parliament. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026)*, pages 593–618, Tilburg, The Netherlands. European Association for Machine Translation.

Li, Y., Suzuki, J., Morishita, M., Abe, K., and Inui, K. (2025). MQM-chat: Multidimensional quality metrics for chat translation. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3283–3299, Abu Dhabi, UAE. Association for Computational Linguistics.

Liu, T., Lo, C.-k., Marshman, E., and Knowles, R. (2024). Evaluation briefs: Drawing on translation studies for human evaluation of MT. In Knowles, R., Eriguchi, A., and Goel, S., editors, *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 190–208, Chicago, USA. Association for Machine Translation in the Americas.

Lommel, A., Gladkoff, S., Melby, A., Wright, S. E., Strandvik, I., Gasova, K., Vaasa, A., Benzo, A., Marazzato Sparano, R., Foresi, M., Innis, J., Han, L., and Nenadic, G. (2024). The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In Martindale, M., Campbell, J., Savenkov, K., and Goel, S., editors, *Proceedings of the 16th Conference of the Association for Machine*
pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.

Lommel, A. R., Burchardt, A., and Uszkoreit, H. (2013). Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Magdy, S. M., Alwajih, F., Mekki, A. E., El-Sayed, W., and Abdul-Mageed, M. (2026). Lqm: Linguistically motivated multidimensional quality metrics for machine translation.

Moorkens, J., Toral, A., Castilho, S., and Way, A. (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7:240–262.

Naeem, M., Ozuem, W., Howell, K., and Ranfagni, S. (2023). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. *International Journal of Qualitative Methods*, 22:1–18.

Naveen, P. and Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience*, 27(10):110878.

O'Brien, S. (2012). Towards a dynamic quality evaluation model for translation. *The Journal of Specialised Translation*, pages 55–77.

Ortiz-Boix, C. and Matamala, A. (2017). Assessing the quality of post-edited wildlife documentaries. *Perspectives*, 25:1–23.

Park, D. and Padó, S. (2024). Multi-dimensional machine translation evaluation: Model evaluation and resource for Korean. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11723–11744, Torino, Italia. ELRA and ICCL.

Simard, M., Leal-Wyss, J., Bernier-Colborne, G., and Knowles, R. a. (2026). A longitudinal study of the adoption of specialized mt systems in Canadian Parliamentary translation. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026)*, pages 133–141, Tilburg, The Netherlands. European Association for Machine Translation.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Vieira, L. N. and Alonso, E. (2020). Translating perceptions and managing expectations: an analysis of management and production perspectives on machine translation. *Perspectives: Studies in Translatology*, 28(2):163–184.

Yang, Y., Liu, R., Qian, X., and Ni, J. (2023). Performance and perception: machine translation post-editing in chinese-english news translation by novice translators. *Humanities & Social Sciences Communications*, 10(1):Article 798.

A Details on Data Sampling

We sampled data for annotation separately from two pools: data from 2023-2024, which ranges from 2023-05-15 to 2024-12-17 and data from 2025, which ranges from 2025-05-27 to 2025-06-20. While the full 2025 data available to us covered a wider span of time (from 2025-04-30 to 2025-12-12), there was very little data in April (explaining why our random sample does not cover that month) and we intentionally set a cutoff date of 2025-07-01 to make sure that we were only looking at data that translators had worked with prior to the start of our questionnaire. Gaps in the data can be accounted for by the fact that we focused only on translation for the debates of the House of Commons, which are translated almost exclusively while the House is sitting.

We sampled "chunks" for which Hawkeye MT was used at translation time from each of the datasets, uniformly at random. The size of the sample was constrained by the time it takes to complete the manual annotations; our annotated REV data comprises approximately 0.10% of the 77091 chunks in the 2025 data and 0.07% of the 209767 chunks in the 2023-2024 data, biasing our annotations slightly toward the data closest temporally to the questionnaire. Statistics for the two date ranges are shown in Table 3.

	PE+REV		REV	
	Chunks	Segs.	Chunks	Segs.
2023-2024	74	583	150	1024
2025	40	201	80	369
Total	114	784	230	1393

Table 3: Numbers of chunks and segments in the sampled dataset.

We intended to have equal numbers of REV-WITH-HMT and PE+REV segments (Table 2), but they are slightly different. These differences are due to the fact that (1) we accidentally missed one of the 115 chunks when producing the PE+REV dataset, as shown in Table 1; and (2) there were slight differences in the segmentation of chunks into segments, between the REV and PE+REV data.

B Original Interview Quotations and Notes on User Study

Here we provide the original French transcriptions of the two translated quotations. Translation was performed by a member of the research team with a background in French–English translation.

"Je me souviens un exemple précis, il y a quelqu'un qui remercie son épouse de s'occuper de la litière des chats toute seule, puis Hawkeye m'avait sorti comme une 'boîte à déchets' pour les chats ou quelque chose comme ça, alors que ce n'est vraiment pas un terme spécialisé, compliqué à traduire". (Original French Quotation, Participant 4)

"En même temps, si on est franc, parfois, il y a des trucs qui ne se traduisent pas ou qui se traduisent extrêmement mal, puis là on n'est plus dans la traduction, on est carrément dans l'adaptation, donc… et je ne pense pas que c'est ce qu'on demande à l'outil pour l'instant". (Original French Quotation, Participant 3)

For the full text of the questionnaire and the initial interview questions, see Appendix B of Leal-Wyss et al. (2026).

C Distribution of Errors/Segment

The frequency distribution of error counts per segment in the PE+REV data is shown in Fig. 6.

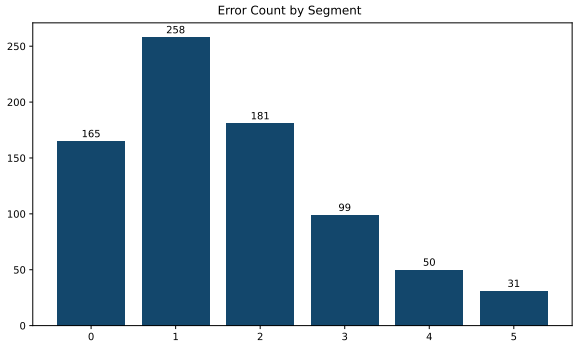

Figure 6: Frequency distribution of error counts per segment in the PE+REV data.

The frequency distribution of error counts per segment in the REV data is shown in Fig. 7.

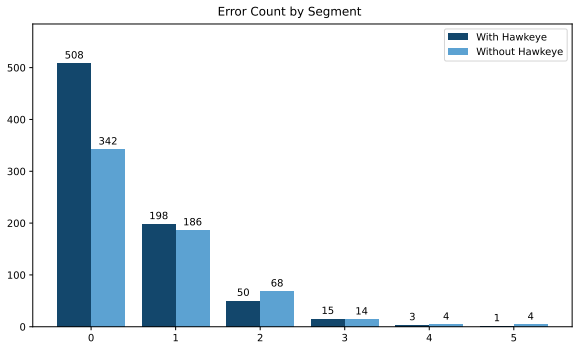

Figure 7: Frequency distribution of error counts per segment in the REV data.

D Distribution of Error Categories

The frequency distribution of error categories in the REV data is shown in Fig. 8.

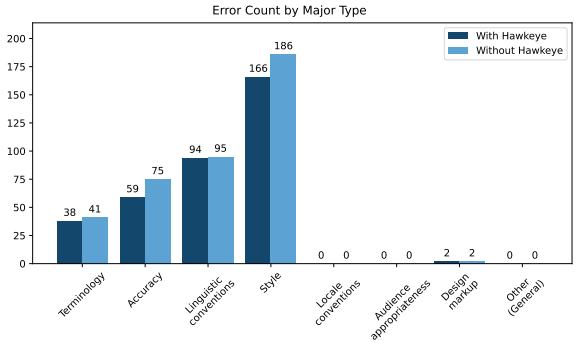

Figure 8: Frequency distribution of error categories in the REV data.

E Distribution of Error Severities

The frequency distribution of error severities in the REV data is shown in Fig. 9.

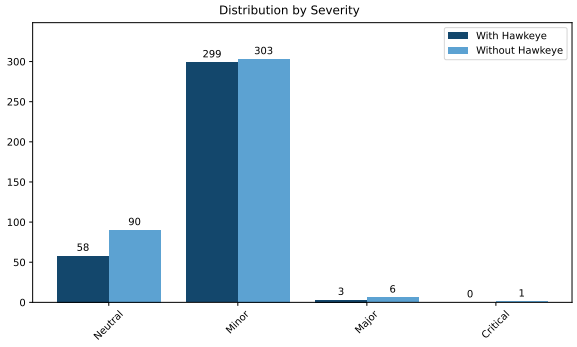

Figure 9: Frequency distribution of error severities in the REV data.

F Distribution of Spans/Error

The frequency distribution of span count per error in the PE+REV data is shown in Fig. 10.

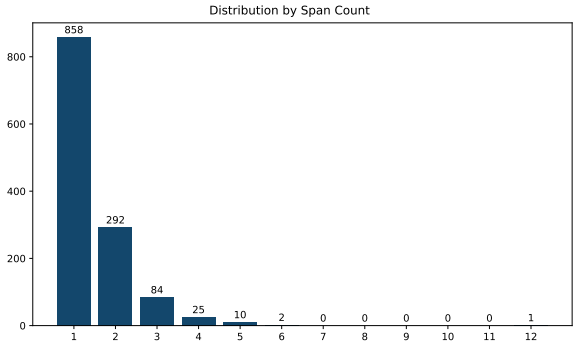

Figure 10: Frequency distribution of span count per error in the PE+REV data.

The frequency distribution of span count per error in the REV data is shown in Fig. 11.

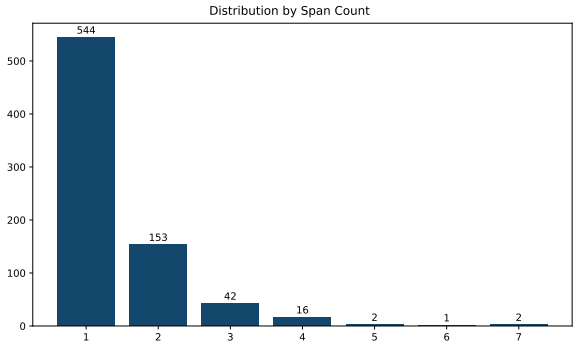

Figure 11: Frequency distribution of span count per error in the REV data.

G Relationship Between Error Categories, Subcategories, and Severities

The relationship between error categories, subcategories, and severities is shown in Fig. 12 for the REV data (the analogous figure for PE+REV was shown in Fig. 3).

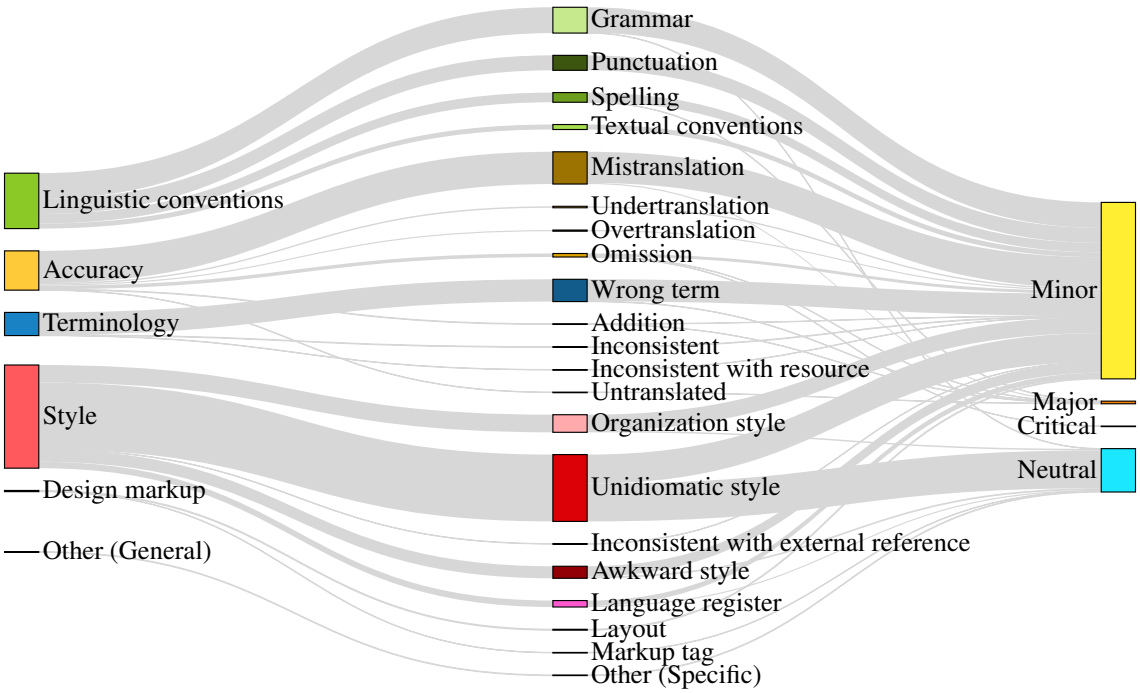

Figure 12: Error Categories, Subcategories, and Severities in the REV data.