WMF Design Strategy **Mar 2022**

Machine translation post-editing process

Annotated bibliography

Prepared by Eli Asikin-Garmager

[Overview](#) [Bibliography](#) [Summaries](#)

Overview and themes

By reducing the burdens of information gathering and translation tasks, machine translation-aided tools increase both the speed and ease with which content can be created.

One such translation tool, Content Translation, has been used to create over 1 million Wikipedia articles, reducing the content and knowledge gaps between larger and smaller wikis.

This annotated bibliography is part of the Language Team's Machine Translation, Human Editors project, which aims to understand how editors work with machine translation outputs to publish new Wikipedia articles.

Machine translation post-editing, such as that used in Content Translation, has grown increasingly common.

The goals of this annotated bibliography are threefold. First, to briefly survey existing literature on the topic of the machine translation post-editing process, whereby human editors receive and improve initial machine translation outputs to arrive at a translated text. Secondly, to understand the relevance of past studies for the current Language Team study. Finally, to generate key takeaways and questions for Content Translation stemming from this prior work.

Before jumping directly into the summaries, let's take a brief look at 5 themes that emerge in the literature surveyed in this document.

Theme 1
Searching for ways to reduce human effort needed to post-edit machine translations.

Much of the literature on machine translation post-editing focuses on understanding human effort with the goal of finding ways of reducing it. The Nunes Vieira (2020) book chapter highlights this in their broad survey of past work. One specific example of this work is Daems et al. (2017), whose article is focused on identifying machine translation errors with the greatest impact on post-editing effort. Examples include how errors of coherence affect post-edit durations.

The Nunes Vieira chapter also highlights some findings about post-editing that have relevance for anyone designing a post-editing system. For example, they cite work showing that for informative texts, the highest post-editing level is redundant compared to a more moderate level when it comes to the impact on the reader's experience.

Possibly surprising is the role that monolinguals can potentially play in machine translation post-editing, thereby further reducing demands on multilinguals who traditionally carry out this work. One study surveyed shows that for monolingual participants who are domain specialists, over 90% of their post-edits are ‘completely correct’.

Finally, some limitations highlighted by Nunes Vieira’s survey include a relative dearth of rigorous cross-linguistic comparison. In addition, most of the post-editing output used for analysis is generated in experimental conditions, with very little reliance on real-world data and behavior patterns.

Theme 2

There are three types of post-editing effort: temporal, technical, and cognitive.

A number of authors, including Daems et al. (2017) reference three types of post-editing effort. Daems et al. discusses these effort types as relevant for a *process* analysis of post-editing (vs. *product* analysis, which is more focused on the nature of edits made). Temporal, the easiest to measure, is simply related to time needed to post edit. Technical, which is harder to measure, involves efforts required to implement the desired changes to the machine translation outputs. Finally, cognitive, which relates to general mental processes and cognitive load required for post-editing, is often measured via eye fixation data, assuming that longer fixations are the result of higher cognitive load.

Theme 3

We can use machine translation post-edit analysis to improve estimation of machine translation quality.

Assessments of machine translation quality are often performed through fully automated measures, such as BLEU, or via subjective human judgements. Snover et al. (2006) essentially tries to combine automation with human annotation to produce results at least as accurate as automated measures such as BLEU, and argues this approach is less noisy when compared to subjective human judgements alone. Aziz and Specia (2012) similarly try to develop a means for assessing machine translation quality via post-edits, and in addition also develop a tool for post-editing.

Theme 4

Machine translation post-edits are frequently lexical in nature, but also differ according to translator experience.

Blain et al. (2011) shows that at least for some language pairs and contexts, the vast majority of post-edits are lexical in nature. However, one should keep in mind that, as with many of these studies, results are often generated from studies of limited, or even one, language pair under tightly controlled experimental conditions. At least based on the survey of literature performed for this document, it’s again worth underscoring the need for more cross-linguistic data to understand to what degree reported trends are language-pair-specific.

Daems et al. (2017) also underscores that post-edit differences may not only be found across language pairs, but also post-editor types. For example, they report that the students in their study focused more on grammatical and lexical issues, whereas professional translators post-edits were more influenced by considerations of coherence and more discourse-level factors.

Theme 5

We can understand machine translation post-edits through typological frameworks. Are all post-edits necessary? ...probably not.

A few prior studies have attempted more qualitative analyses of post-edits made to machine translation outputs. For example, Blain et al. (2011) presents a post-editing typology of edits. Although outside of their focus, given that different machine translation errors affect post-editing effort differently, one might ask if such typologies should be overlaid with a weighting system. This is also relevant as we consider the impact of different error types on the reader's experience, although there appears to be less research in this area.

Finally, Koponen and Salmi (2017) similarly perform an analysis focused on understanding the types of post-edits made, finding they're comprised mostly of form changes and insertions. This provides some corroborating evidence to Blain et al.'s (2011) findings that the majority of post-edits may involve noun phrase modifications. Koponen and Salmi, however, go one step further, asking to what degree all post edits are correct and necessary. They confirm in their work that most post-edits are correct, but as many as 34% of those they analyzed may not be necessary to the extent they're needed to resolve grammatical or semantic errors.

How the remainder of this document is organized

Having briefly introduced these 5 themes, the remainder of this document is organized as follows: Immediately following this overview, an inventory of the sources surveyed is provided. These sources include journal articles, conference proceedings, and one book chapter.

After the list of sources, a concise summary is provided for each source. Along with each summary are key takeaways, questions, and considerations for Content Translation and the Language Team's ongoing study. Finally, there's also an overview of the current Content Translation quality system, especially the machine translation limits system which drives this system (see *Content Translation limits system mechanics*).

Bibliography

Below is the list of works surveyed in this document. For each work below, this document includes both a brief summary, as well as key takeaways and questions in the context of Content Translation.

Aziz, W., & Specia, L. (2012). **PET: A tool for post-editing and assessing machine translation.** *Proceedings of the 16th EAMT Conference*.

Blain, F., Senellart, J., Schwenk, H., Plitt, M., & Roturier, J. (2011). **Qualitative analysis of post-editing for high quality machine translation.** *Proceedings of the 13th Machine Translation Summit*.

Daems, J., Vandepitte, S., Hartsuiker, R. J., & Macken, L. (2017). **Identifying the machine translation error types with the greatest impact on post-editing effort.** *Frontiers in Psychology*, *8*(1282).

Koponen, M., & Salmi, L. (2017). **Post-editing quality: Analysing the correctness and necessity of post-editor corrections.** *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 137–148.

Nunes Vieira, L. (2020). **Post-editing of machine translation.** In *The Routledge Handbook of Translation and Technology* (pp. 319–335). Routledge.

Santhoshtr. (2022). **Content Translation machine translation abuse calculation.** https://github.com/wikimedia/mediawiki-extensions-ContentTranslation/blob/master/doc/MTAbuseCalculation.md

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

Wikimedia Foundation. (2022). **Help:Content translation/translating/translation quality.** https://www.mediawiki.org/wiki/Help:Content_translation/Translating/Translation_quality

PET: A tool for post-editing and assessing machine translation (Aziz & Specia, 2012)

Aziz and Specia describe a standalone machine translation (MT) post-editing (PE) tool (PET) that has two purposes. First, it facilitates the PE of outputs from any MT system so they reach publishable quality. Secondly, it collects sentence-level information from the PE process, such as time and keystroke statistics. A few other similar PE tools are noted, including SDL Trados, Wordfast, and Deja Vu X2. Translog is also mentioned, a tool that logs detailed information about operations performed during PE, which can be used for measuring translation quality and diagnosing translation problems. It even allows playback as though the PE process was a video.

Examples of the tool's interface design are provided on page 3984, which is customizable along a few parameters, including display and assessment options. Editors are asked to provide manual feedback about the PE process as they proceed through translations. Editors proceed sentence-by-sentence, but the tool is not apparently optimized for mobile devices based on images and its description. The authors provide examples of a number of different uses of the tool:

- Compare PE of different MT systems
- Compare manual and MT-aided translation in time and effort
- Quality estimation of MT systems
- Compare experiences and data across various language pairs

Key points, discussion questions, and opportunities for Content Translation:

1. A number of tools (listed above with links) may be relevant for considering as part of an updated competitive analysis for CX.
2. Use of Translog with consenting CX users could be an interesting approach for future studies examining the role of editors in the CX post-editing process and getting a detailed understanding of their PE process.
3. While statistics about the PE process may be relevant for CX design/development, what about feedback for editors? The Language Team could consider various

metrics about the PE process and to what degree these may be relevant and/or helpful for CX users. As one example, overall translation duration could be used to show editor efficiency improvement over time. Another example might include an assessment and presentation of what number of minimal post-edits are most strongly correlated with articles that go on to survive and be further edited and improved.

Qualitative analysis of post-editing for high quality machine translation (Blain et al., 2011)

Machine translation post-editing (PE) differs from traditional translation review because of the nature of the errors to correct. Past approaches of measuring PE effort have used time, keystrokes, and even eye movement measures. Like others, the authors of this article are interested in how we may reduce the effort of PE.

Blain et al. proposes that PE activity can be modeled by a set of rules, resulting in decomposition and qualitative analysis of PE results. To do so, they propose extracting minimal and logical edits that map closely to post-editing intents. For example, multiple mechanical moves may be the result of a single post-editor intent. Take, for example, the French modification of 'le bord est affiche' to 'la bordure est affichée'; in this example of a machine translation post-edit, there are three words substituted by three different words (mechanical changes), but the editor intent was to correct a single word (the head and modifier in this case).

The article provides an overview of a post-editing action (PEA) typology, the purpose being to define the minimal logical edits relevant to post-editing. A brief/summarized view of this typology is shown below:

- **Noun phrase (NP)** - lexical changes
    - Determiner choice
    - Noun meaning choice
    - Noun stylistic change (e.g., synonym)
    - Noun number change
    - Case change
    - Adjective choice to fit noun
    - Multiword change
    - Noun phrase structure change (but sense preserved)
- **Verb phrase (VP)** - related to grammatical changes
    - Verb agreement
    - Verb phrase structure change
    - Verb meaning choice
    - Verb stylistic change (e.g., synonym)
- **Preposition change**
- **Co-reference change** (e.g., definite to possessive determiner)
- **Reorder** (e.g., adjective or adverb ordering)
- **PE error** - post editor made a mistake in review
- **Misc.** - cannot be classified

Having established this typology, the authors discuss how they've attempted to automate analyses of PEs. Later in the article, a caveat is offered, which is that this approach is specific to when machine translation quality is high and post-editors do ‘light editing’. The protocol for this process begins with linguistic annotations and constituent tags, which are generated for the initial translation (machine translation output) and post-edited version (using SYSTRAM syntactic analyzer). Next, both sentences from the MT output and post-edited version are aligned in order to identify all changes made during the PE process. Then, PEAs are identified through pattern matching. Finally, a subset of the data is compared with a manual analysis by humans. Interestingly, the authors note that around 90% of the changes they analyzed involved an NP change (more specific breakdown on page 6).

Key points, discussion questions, and opportunities for Content Translation:

This article raises a number of ideas for how we might approach our analysis of post-edits found in CX publications. It also highlights some areas that may be unique to CX and our analysis, which we need to consider further.

1. The typology used by these authors could serve as a good starting point for our analysis of post-edits with CX publications. However, we may want to consider how it’s not a system that counts mechanical changes, and so for the purposes of evaluating the current limits system, we might have to consider mechanical edit count as well. Also, this typology likely will need some adjustments based on both the languages we’re analyzing and the specific nature of Wikipedia articles. Consider, for example, the removal/addition of content and possible dialect shift/change. Also, is ‘reorder’ too broad of a category?
2. Given that the authors note that 90% of PEA were NP-related, this sets up a clear prediction for our study, but a more nuanced view of NP change type would likely be helpful, and may vary across languages and language pair machine translation quality.
3. If most NP changes are terminological, the authors note that this information could in theory be fed back into a system with the goal of improving it. Unlike the authors’ study, in the case of CX we should also consider PE continuity issues, such as terminological consistency across sections of the same article, or even across articles related to the same topic. Although we may not be able to access editor intent and cross-article comparison, such changes may manifest via semantic NP modifications, and particularly through synonym substitution.

4. To what degree are PEAs fed back into MT engines in the context of CX? We've heard from event organizers about the toll of making repeated terminological changes, raising the need for translation memory. While we should observe for repeated terminological changes within articles in our study, we may not be able to evaluate across-article issues given that articles were randomly sampled.
5. Because the article was not focused on the reader experience of post-edited translations, the authors didn't discuss the topic of weighted typologies. For example, whether we consider pure mechanical changes or changes via the typology presented in this article, we may hypothesize that they don't all have an equal impact on the final reading experience of post-edited machine translations. We may wish to explore prior work examining which post-edits have the greatest impact on the final reading experience. Even rudimentary weighting could be used by the current CX limits system in refining its evaluation/estimation of quality.

Identifying the machine translation error types with the greatest impact on post-editing effort (Daems et al., 2017)

Ideally translation tools could predict in which cases it's more effortful to post-edit (PE) a machine translation (MT) or start a translation from scratch. Letting humans make this decision costs time and effort. This study confirms that MT quality affects PE effort indicators (with some exceptions of specific indicators), but provides a more nuanced view. For example, Daems et al. concludes that if PE speed is the main goal, then a language service provider should focus on measurements of coherence as these impact duration the most. In other words, MT error types affect different measures variably. And, in order to correctly estimate PE effort, more fine-grained MT quality analyses are required.

The article begins with a concise summary of prior work in the area of PE effort analysis, and distinguishes measures of PE effort via *product* analysis vs. PE effort via *process* analysis. Via product analysis refers to how the MT output can be compared to reference translations to evaluate MT performance. Limitations of this approach include that all edits are assumed to require roughly equal effort, which is suspect. Measuring PE via process analysis usually includes reference to one of three types of effort: temporal (easiest to measure; time), technical (harder to measure; efforts required to implement changes), and cognitive (mental processes and cognitive load, usually measured via eye fixation data assuming that longer fixations are a result of higher cognitive load). These three types of PE effort are related; for example, temporal effort is determined by a combination of technical and cognitive effort. Quite relevant for CX, for which the base of users is quite diverse, the article looks at the impact of translation experience on PE effort.

The authors offer answers to four hypotheses examined in their study:

1. Hypothesis 1: PE effort indicators are influenced by MT quality. Answer: Overall, yes, but not all measures are impacted equally.
2. Hypothesis 2: Product effort indicators are influenced by other MT error types than process effort indicators. Answer: Yes, because product effort indicators measure different things than process indicators.
3. Hypothesis 3: There is overlap in the error types that influence the various process effort indicators. Answer: No, not all process effort indicators are influenced by the

same MT error types. For example, duration is influenced most by coherence, whereas fixation duration is influenced by other meaning shifts

4. Hypothesis 4: Effort indicators of student translators and professional translators respond to different error types in different ways. Answer: Yes, but patterns were not as strong as expected. Nonetheless, students were more influenced by grammatical and lexical issues, whereas professional translators were more heavily influenced by coherence.

Key points, discussion questions, and opportunities for Content Translation:

1. As CX is used by editors of a wide range of language pairs, this article raises the question of whether it's fully appropriate to always lead with a MT output, particularly in cases in which MT quality may be quite poor. Although it may not be in scope of the Language Team to develop a system of predicting effort and automating the choice for users, it may be appropriate to remember editor preferences in use (and non-use) of MT outputs, adapting presentation to these preferences.
2. Although automated effort analysis is a large lift, there's value in better understanding what types of MT output errors are most costly to Wikipedia editors. Such information could be strategically used for product decisions, such as decisions about which feature development/improvements should be prioritized to overall reduce user effort.
3. The fact that professional translators may focus more on discourse-level MT errors, raises the question of how in our study we will measure discourse-level edits. For example, a lexical change may be made in response to discourse considerations, but it may be difficult to determine intention behind edit types since we're only working with the actual output changes and don't have access to the editors who produced the edits. Nonetheless, this is a topic we should monitor when developing the instrument we'll use for analysis.
4. Given the authors' finding that professional and student translators focus PE efforts more/less on different aspects of MT errors, should we be more focused on collaboration in future development of CX? And, for the team's current study, if feasible, we should certainly take a look at the data broken down by some proxy for

editor experience although we don’t have indicators for level of professional translation experience. Experience more broadly could be a predictor of overall edit types, which gives some view into the overall published translation. And, if the priority were to better support CX newcomers, then with this information we could prioritize product improvements supporting certain types of edits.

Post-editing quality: Analyzing the correctness and necessity of post-editor corrections (Koponen & Salmi, 2017)

This article reports a pilot study analyzing edits made by university students in a post-editing (PE) English to Finnish translation task using machine translation (MT) outputs. While many past studies have examined editor effort, Koponen and Salmi focus more on analyzing the nature, correctness, and necessity of edits. Many studies assume post-edits are correct, but Almeida (2013), cited in this article, shows that post-editors failed to make essential changes in 11-15% of analyzed cases and introduced new errors in 5% of cases. Moreover, this same author showed that 'preferential' changes (those in which the unchanged sentence would have been grammatically correct without a change) accounted for between 16-25% of cases analyzed (mostly lexical changes). Similarly, Kopenen and Salmi (2017) demonstrate that while most edits performed are correct, a significant number of them (34%) are unnecessary. Overall, better understanding the nature of edits made by post-editors has implications for practice and training.

Before presenting the current study, these authors quickly review some past work on PE effort, citing Temnikova's (2010) classification for machine translation (MT) errors, ranked in terms of presumed cognitive effort. It's noted that incorrect word forms are easier to correct, while word order and reordering edits involve greater effort. No specific reference is made to discourse or larger contextual edits.

For this study, 16 translation students post-edited a short MT text. The participants were native Finnish speakers studying translation, and the MT text was produced based on an English source text. They used an edit distance metric (HTER), which compares the MT and PE versions of a sentence and computes the minimum number of word-level changes divided by the number of words in the PE version. In addition, each word was manually annotated with one of the following categories:

- Unedited; no change
- Form changed; different morphological form
- Word changed; different lemma
- Deleted; word removed
- Inserted; word added
- Order; position of word changed
- Mixed; combination of 2 or more changes

Furthermore, each of these changes were annotated according to correctness and necessity. Correctness was defined as accurate in terms of the source language meaning and proper grammar and spelling of the target language. Necessity was defined as a change needed to make the target language sentence comprehensible, either grammatically or semantically.

A summary or results show that the most common edits were form changes and insertions (10% of all changes, each). Meanwhile order changes accounted for 2.9% of changes and ‘multiple changes’ accounted for 2.3% overall. As for correctness and necessity, all unedited words were deemed correct, but there were 36 cases (3% of all unedited words) in which a necessary correction hadn’t been made. Overall, 91% of PEs were correct, but only 61% necessary. This means that 38% of all edits made were deemed unnecessary.

Key points, discussion questions, and opportunities for Content Translation:

This article is particularly relevant to the current Language Team’s study as it similarly provides a manual analysis of PEs, and raises the question of whether or not we should also consider *correctness* and *necessity* as topics to evaluate.

1. While we may or may not want to track PE errors, with rates as high as 15-25% we probably want to evaluate unnecessary and/or preferential (optional/stylistic) edits.
2. This article reported that most unnecessary PEs were related to word order and personal pronoun deletion, generally unnecessary for the language pair (English - Finnish). This suggests that language pair is a determining factor for unnecessary changes. If tracking unnecessary edits, we should perform a breakdown to compare these edits across languages.
3. Given that Temnikova (2010) argues that word order and reordering edits involve a greater effort, it’s relevant to note that if this is the case, it means that the more difficult PEs are not accounted for in the current CX limits system (given it doesn’t treat a reordering as a PE). Similarly to the point in #2, we should evaluate any correlations between general edit patterns and language pair as this would be a first step in helping us determine if accounting for reordering changes is necessary for the limits system.

4. It is likely we will have to come up with a system for annotating ‘mixed’ edits. However, we should try to do so in a way that doesn’t lump them all together so we can describe the most common combinations.

Post editing of machine translation (Nunes Vieira, 2020)

Nunes Vieira's book chapter provides an overview of how post-editing of machine translation has evolved as a practice and service. The role of the human translator relative to the machine translation (MT) outputs has changed over time. Whereas early on the paradigm was one in which human editors assisted the machine, with modern CAT tools, machines are now aiding the human editors. And, as MT outputs continue to improve, so will the role of humans. For example, with very high quality MT outputs, human translators could eventually reach a point of simply providing terminological checks and content sign-off, a possible reversal back to a situation in which human editors aid and provide checks for machines. One risk that may increasingly appear as MT quality improves is that humans may miss MT errors in regards to content, particularly in cases in which the source text is not readily and easily available for cross-checking.

The chapter overviews a number of previous research studies, noting a general interest in the topics of quality and post-editing effort. For studies of effort, there are three dimensions, including: (1) cognitive - editing decisions, (2) technical - implementation of edits, and (3) temporal - time required by any activity. One line of work tries to understand how post-editing effort can be predicted to a degree by the source-text genre and complexity, both known to impact MT output quality. In the stream of work investigating what features of source text predict post-editing effort required, noun quality and sentence length appear to be possible predictors (e.g., not surprisingly, longer sentences require more cognitive effort).

In addition to studies of post-editing effort, this chapter covers the impact of editing on the reader experience. For example, one study tested four levels/degrees of MT output editing and found for informative texts, the highest post-editing level was redundant. It generally did not improve the end-user's perception of content compared to a 'moderate' (less extensive) level of post-editing. As for which types of post-edits are most impactful for the end-user, the author notes the Translation Automation User Society's (TAUS) guidelines for 'good enough' post-editing, which focus on semantics and comprehension. Syntax, style, grammar, and formatting only appear in guidelines when the goal is working towards human translation quality.

This chapter is not the only source on this topic to raise the idea of monolingual post-editing, an idea that holds promise, 'where the MT output is edited by domain specialists,' for example. In a 2014 study, Schwartz is cited as reporting that, 'over 90% of sentences post-edited monolingually by a domain expert were found to be completely correct'. Conversely, other studies highlight cases in which bilingual post-editing is superior - namely in terms of 'adequacy,' or the extent to which the source-text meaning is conveyed; presumably because the bilingual has the ability to reference and understand the source text more comprehensively.

Key points, discussion questions, and opportunities for Content Translation:

1. MT quality varies significantly by language pair, and is improving at different rates for various languages. This means that the typical tasks of CX users may vary substantially according to language and language pair. Design may wish to understand how the experience of these editors vary based on MT output quality and evaluate to what degree workflows may vary and vary in how they're supported relative to MT quality and the resulting workflow of editors.
2. If a system for predicting the level of effort required for the post-editing of articles/sections could be repurposed or recreated, it could be used to tailor suggestions based on a consideration of predicted effort and the editor's experience with CX, scaffolding the experience for newcomers.
3. As a baseline and competitive comparison of CX with modern CAT tools, the Language Team may wish to pursue a competitive analysis of specialized CAT tools used among professional translators. Although not all CX users are professional translators, this audit could highlight possible gaps and opportunities for features such as translation memories and other terminological resources.
4. On the topic of monolingual MT post-editing, given that research shows that domain experts have proven success publishing quality translations, we should ask in the context of Wikipedia how we might find a proxy for domain expert (e.g., previous edit histories), and how translation suggestions could be tailored for post-editing by monolinguals. Another option is to consider how the role of monolingual (or 'weak/passive multilinguals') may have in the role of content

translation. What tasks could they productively complete in the context of content growth via translation?

Content Translation limits system mechanics (Santhoshtr, 2022; Wikimedia Foundation, 2022)

This section provides a brief overview of the current machine translation limits system created by the Language Team to encourage high quality translations by avoiding too much unmodified machine translation, which may degrade the quality of a CX-published article.

As machine translation (MT) quality varies substantially by language pair and MT engine, to aid article quality, the Content Translation (CX) tool imposes limits on how much unaltered machine translation can be present in the published article. In its current form, it's a measure of how many words have been added, removed, or modified when comparing the initial MT output that an editor receives and the draft for publication (post-edited form).

This system works based on calculations that are made based on plain text. Of the source and target set, the biggest and smallest set is identified, and then the intersection of sets is found; this 'unmodifiedtokens' set (what doesn't differ when comparing the MT output and draft for publication) provides a basis for understanding what percentage of the initial MT output has been modified. Spelling and case adjustment counts as an instance of an edit, but reordering is not taken into account - that is, it doesn't signal an edit to the MT output. Finally, consecutive whitespaces are treated as a single whitespace, and for CJK languages, tokenization happens at the character level.

Measurements are made at both the paragraph and whole article level, with different limits applied at each. By default, for article level limits, an editor cannot publish a translation with 99% or more of unmodified contents. For the paragraph level limits, a paragraph is flagged as problematic if it contains more than 85% of the initial machine translation unaltered; editors must modify 15% or more of the initial MT output. However, if the editor reviews and marks paragraph-level alerts as resolved, the 85% limit is increased to 95%. Also, publication is blocked if there are 50 or more problematic paragraphs (insufficient post-edits) ; those with 10-49 problematic paragraphs are tracked for review.[1]

[1] For more information refer to Machine translation abuse calculation and Content Translation quality.

This limits system can be adjusted on a per wiki basis with feedback from the community, and has been modified on a number of occasions. Also, not all article content is included in the limits system review. For example, very short section titles, citations, references, images, tables, section headings, infoboxes, lists, math formulas, definition lists, and poems are all excluded from the check.

Discussion/follow-up questions for the Language Team:

1. Which wikis have requested modifications to the default limits:
   a. Wikis requesting stricter limits:
   b. Wikis requesting less strict limits:
   c. Other information about rationale provided by communities:
2. Are there any additional features of this limit system that have been tried before? If so, which and why were they discontinued?
3. What limitations or possible changes to the system are team members interested in learning about more?
4. What other general questions are top-of-mind for team members about the limits system and how it’s currently working?

A study of translation edit rate with targeted human annotation (Snover et al., 2006)

Overall, Snover et al. is interested in developing a way of involving humans in the assessment of machine translation (MT) quality in a way that's less expensive and noisy than subjective judgements, but at least as accurate as common automated measures, such as BLEU. The authors present work that accomplishes this, except they note it's still costly in terms of human time. More specifically, they introduce two measurements: Translation Edit Rate (TER) and Human-targeted translation edit rate (HTER), and show that HTER yields higher correlations with human judgements than BLEU.

Translation Edit Rate (TER) is calculated by taking the number of post-edits (PE) to a machine translation (MT) output and dividing by the average number of reference words. 'Reference' is used to refer to the unedited MT output. It factors in various edit types including: insertion, deletion, substitution of single words, and shifts of word sequences. Similar to how other authors treat punctuation, such tokens are treated as normal words, and capitalization changes are counted as edits.

Human-targeted translation edit rate (HTER) builds on TER, compensating for how TER ignores notions of semantic equivalence that may be present in MT output to edited transformations. Basically HTER uses human annotations to make TER a more accurate measure of translation quality; a 'human-in-the-loop' evaluation.

In the end, both of these measures are shown to be good predictors of human judgements of MT quality, but HTER is less subject to the noise of human judgements (as it's more rooted in actual human transformations, and post-editing, of MT outputs). While superior to subjective human judgements, HTER is expensive, requiring 3-7 minutes per sentence for a human to annotate.

Key points, discussion questions, and opportunities for Content Translation:

1. TER and HTER are in many ways similar to the current abuse calculation used in the CX limits system, with the exception that the abuse calculation doesn't involve

human annotators and TER treats reorderings as a count of an edit, whereas the abuse calculation disregards reorderings.

2. As this article is really focused on developing an approach for evaluating MT output quality, it doesn't present data that allows us to understand trends for language pairs as measured by TER. However, it does support an approach for analyzing post-edits that factors in modifications of semantic similarity; for example synonym substitution as a semantic transformation.
3. Should the Language Team wish to modify the current MT abuse calculation, the measures presented in this article may be helpful, especially in helping determine the details for how any changes in the post-edited linearization patterns (compared to MT output patterns)