Picking Out the Best MT Model: On the Methodology of Human Evaluation

Stepan Korotaev s.korotaev@effectiff.com
CTO, Effectiff LLC., Walnut Creek, 94596, USA

Andrey Ryabchikov a.ryabchikov@effectiff.com
Lead NLP Specialist, Effectiff LLC., Lauderdale by the Sea, 33308, USA

Abstract

Human evaluation remains a critical step in selecting the best MT model for a job. The common approach is to have a reviewer analyze a number of segments translated by the compared models, assigning those segments categories and also post-editing some of them when needed. In other words, a reviewer is asked to make numerous decisions regarding very similar, out-of-context translations. It can easily result in arbitrary choices. We propose a new methodology centered around real-life post-editing of a set of cohesive translated texts coming from *homogeneous* source documents. The homogeneity is established using a number of metrics on a preselected corpus. The key assumption is that two or more identical in length translated texts coming from different but homogeneous source documents should take approximately the same *effort* when edited by the same editor. Hence, if one text requires more effort, it is an indication of a relatively lower quality of machine translation used for this text. We proceed to show how this new methodology can be applied in practice and share results of an experiment carried out for the English > Russian language combination. We also discuss other possible applications of the methodology and directions of future research.

1. Introduction

Today, machine translation (MT) is available in a multitude of forms and shapes. The market is saturated, with dozens of providers competing for the supremacy in different language combinations and domains (Intento, 2021). From a practical standpoint, it means any party faced with a task of applying MT in their processes needs to select the best option among the available alternatives. There are two main tools for that:

- automatic metrics;
- human evaluation.

Metrics (like BLEU, hLEPOR, BERTScore, etc.), which are normally used for primary selection and narrowing down options, are beyond the scope of this paper. We will focus on the next step, human evaluation. It normally takes place after several models are picked out based on their higher automatic metric scores. Then, as part of the common methodology, a human reviewer is asked to review and, in some cases, post-edit a number of automatically sampled segments translated by different engines. The results of such evaluation are used to determine a winner. We present our critique of this approach in the next subsection.

1.1. Critique of the Common Methodology of Human Evaluation

The methodology outlined in this subsection is widely used in the translation industry, with slight modifications (incl. in Intento, 2021).

Each reviewer is asked to perform two kinds of work: to evaluate a number of segments by assigning them *categories* (like types of errors found in those segments) and to *post-edit* a different set of segments, which allows to calculate a distance (i.e., represent an amount of changes made as a number). We will call these two types of work *categorizing* and *post-editing*, respectively. A reviewer, consequently, must be both a *categorizer* and a *post-editor*. Typically, a reviewer will have to deal with the output of several engines that they will need to categorize and post-edit. For extra reliability, larger studies usually seek to engage several reviewers working in parallel with the same task, and then average the results. There are several problems, however, that hinder this process and, consequently, the trustworthiness of the human evaluation step as a whole.

Qualification requirements: Not every translator or editor can be a categorizer. It is a separate qualification requiring a certain personal disposition and a number of skills that are not very easy to come by.

Long preparation and training: Even if a researcher has enough categorizers at their disposal, they still need to be trained to make sure they understand the instructions, which can be quite extensive and complex. A researcher will also have to spend time on creating instructions or adapting them given the exact nature of the experiment.

Loss of focus during the post-editing stage: For the post-editing stage of the evaluation, reviewers are asked to post-edit various translated versions of the same source segment provided by all engines in the running. Then an amount of changes in each post-edited translation is calculated, which allows to rank engines by how much work each of them required. A reviewer's task is worded along the following lines: *amend each and every translation* to a state that you would call satisfactory but *don't try to replicate changes*—each translation should be *changed individually* based on its unique structure and possible shortcomings, *without taking into account changes made to other versions*. The problems caused by this approach and its expectations are obvious. Machine outputs can be quite similar, and it is almost impossible to 1) post-edit all of them as if each of them was unique—the *fatigue bias* on the part of a reviewer, and 2) change them all to a more or less equivalent degree—again, the fatigue bias caused by the repetitive process. As a result, different post-edited versions might end up being either very similar (i.e., changed based on a once found formula) or, conversely, amended inconsistently (some subjected to a deeper editing process, others left half-baked).

Lack of context during the post-editing stage: Not only are reviewers asked to work with several similar translations, those translations are usually also out-of-context and presented as a series of standalone sentences. It further hinders meaningful post-editing and contributes to the arbitrariness of the process.

Summing it up, the common methodology as outlined above requires too much time for preparation and training and might yield unreliable results due to a very likely fatigue bias on the part of reviewers.

1.2. Alternative Methodology of Human Evaluation

To overcome the limitations outlined in the previous subsection, we came up with a different approach based on the following concepts:

- No categorization: the methodology *only relies on the results of post-editing*.

- Instead of a set of out-of-context translations of the same source text, each reviewer *works with the cohesive, non-repetitive document* where the output of different engines is combined with human translation for benchmarking.
- Translation quality can be represented as an amount of *effort* required to edit a text to a desired state—hence we measure each editor's productivity across several metrics (*time spent*, *edit distance*, *percentage of segments changed*) and use these metrics to rank engines. The lesser the effort spent on an engine output, the higher the quality is deemed to be.
- No special requirements to the post-editors: they must be qualified enough to be able to work with a translated text; however, their style of editing and level of domain expertise are mostly unimportant as we are only interested in the *relative* effort—how much work is spent on each part of a text as compared to other parts. We are looking for a consistent correlation and pay little attention to the actual changes made to a text.

Below, we will describe the methodology in greater detail, present the results of its practical application, and discuss some of the interesting topics for future research.

2. Methodology

2.1. Key Assumptions and Process

The methodology is based on several key assumptions:

- Asking a reviewer to post-edit a cohesive, non-repetitive translation should produce better results compared to post-editing several similar, out-of-context translations of the same source text.
- Different but close in length (word count) and *homogeneous* texts (see below on how to determine homogeneity) take a reviewer approximately the same time to complete.
- It is possible to reliably determine if any two or more texts are homogeneous.
- If one of the engines' output consistently takes less effort to be post-edited across different homogeneous texts than the other's, it is proof that the first engine provides better quality for this language combination and domain.

Based on these assumptions, the following process can be set up:

1. Given the language combination and the domain that we are interested in, find several homogeneous texts (together called a *translation kit*).
2. Have different engines translate the whole translation kit.
3. Prepare a good human translation of the translation kit for the benchmarking purposes.
4. Shuffle machine and human translations to create *review kits*, which consist of the same parts as the translation kit, but with the condition that each of those parts is translated by a different engine (or a human).
5. Assign different post-editors to work with the review kits. Each post-editor works with one review kit.
6. Measure post-editors' productivity across all parts of the review kit: time spent, edit distance, percentage of changed segments.
7. Compare data measured for all post-editors to determine if there is a meaningful correlation between productivity metrics and the output of different engines.

2.2. Homogeneity

To establish the homogeneity of two or more texts (we will be calling them *documents*), we used the following principles:

- Documents should be of the same domain and genre.
- Documents should have similar complexity and/or readability scores based on selected metrics.
- Documents should be close in the density (number of occurrences) of specialized terminology.
- Documents should not have (or have very few) overlapping specialized terms—otherwise, the first part of a translation kit might require disproportionately more work to check terminology when it first occurs, with other parts benefitting from this work.

The practical application of these principles as regards our experiment is described below (see *Selection of Homogeneous Texts*).

2.3. Effort

Effort is calculated for each of the three measured metrics: time spent, edit distance, percentage of changed segments. In each case, we are *only interested in relative values*. One editor might feel more comfortable rewriting the text; another will only touch it in several places. For our methodology, it is not important. What is important, however, is how different parts of a review kit are stacked up against each other: which one has received more effort, regardless of whether large or small in absolute values, from a given reviewer?

2.4. Human Benchmark

Adding human translations to a mix for benchmarking purposes is a common approach. We did it as well but slightly modified this idea. Usually, human translations are taken from a "trustworthy" source like a large translation memory or other corpus. It is implicitly presumed that this translation must be, by definition, at least on par with MT, and most likely better. However, human translations in large corpora 1) are unpredictable in quality and 2) *can easily be not human at all*. The latter is especially valid and, in our view, largely overlooked in similar research. The use of MT in the industry is widespread, incl. by the translators copy-pasting MT for their own convenience without even telling anybody. It leads to a significant contamination of translation memories, presumed to only contain human translation, by machine output. On top of that, the real quality of any given human translation in a large corpus cannot be guaranteed. To solve these problems and create a reliable benchmark, we made sure to translate our translation kit by a trusted translator and then edit this text by an equally trusted and experienced editor. We also double-checked the final version of the translation for traces of MT. Though still subjective in nature, these measures helped us achieve a substantial level of confidence that our benchmark was reliable and high-quality.

2.5. Hypotheses

We formulated two hypotheses that we hoped to prove during our experiment.

Hypothesis 1 (H1): The average distribution of effort among documents will prove their homogeneity established based on our metrics. In other words, on average, all documents will require roughly the same amount of work.

Hypothesis 2 (H2): The human benchmark will be consistently shown to require less effort than any of the competing engines.

3. Structure of the Experiment

3.1. General Parameters

The experiment was set up with the following general parameters:

Language Combination: English into Russian.

Domain: Information Technology, Big Data & Machine Learning.

Genre: Popular Science (book).

Volume: Three parts of approximately two pages (500 words) each, total of six pages (1500 words) for each review kit. The volume was determined so it could be processed by a post-editor in one go, without getting too tired and thus losing speed.

Engines: Google Translate, Amazon Translate, Human (for benchmarking). For each engine, a stock version was used (no additional training had been performed).

Post-editors: Six post-editors, each working with a unique review kit.

3.2. Selection of Homogeneous Texts

This section contains a high-level overview of the procedure. For a more detailed description and code (Python scripts), see GitHub (2022). At the time of writing, it is being updated and expected to be finalized soon, with all relevant materials available for reference and download.

To find homogenous documents, we first looked for a corpus consisting of coherent sentences, written in more or less plain language and not overloaded with specialized terminology. The text had to be publicly available, not protected from use in our purposes (scientific research) and also not known to be published in the target language (Russian). We ended up with a monograph on big data (Richterich, 2018). Only the main text of the monograph was taken; other parts like the introduction, the table of contents, the reference aids and the bibliography were left out. The text was then cleaned using regular expressions to remove references to endnotes, endnotes themselves, bracketed references to literary sources, etc.

The resulting cleaned text was divided into paragraphs, and then consecutive paragraphs were combined into pieces of approximately 500 words. This way, each piece contained related paragraphs and was expected to be internally cohesive and providing enough context to a post-editor. In total, about 70 pieces were obtained for further processing.

The selected pieces then underwent tokenization (using the BlingFire library[1]) and segmentation (division into sentences). Pieces with an average sentence length less than six words were removed from the dataset.

Then the readability metrics and general textual statistical metrics were calculated for each piece. We used the following metrics as features for further clustering: Flesch Reading Ease, LIX, Dale-Chall Index, Characters Per Word, and Type Token Ratio. The first three metrics are based on the average number of words in a sentence and also include the average number of syllables in a word (Flesch Reading Ease), the proportion of long words (LIX), or the proportion of "difficult" words (Dale-Chall Index). In addition, metrics related to the number of characters in a word (Characters Per Word) and the proportion of different words (Type Token Ratio) were also used.

Metric values were then normalized using the min-max method and grouped into clusters using the DBSCAN algorithm. The distance between the points was calculated as a Euclidean metric, and the minimum number of pieces in the cluster was set as three.

[1] https://github.com/microsoft/BlingFire

The obtained clusters of homogeneous pieces were then checked for properties related to terminology based on the two stated principles: 1) homogeneous texts should not have a (significant) number of shared terms and 2) homogeneous text should have approximately the same density of terminology. For term extraction, we used a combination of seven methods based on the identification of individual frequent words, collocations, terms based on part-of-speech properties, and all suitable bigrams. For more details see GitHub (2022).

Finally, three pieces were taken from one cluster for which further subdivision into subclusters and sub-subclusters with respect to terminology did not lead to additional fragmentation for most term extraction methods that we used.

The numerical values of the readability and statistical metrics used for the selection are summarized in Table 1 (selected pieces were named *Doc I*, *Doc II*, and *Doc III*).

Texts	Flesh Reading Ease	LIX	Dale Chall Index	Char Per Word	Type Token Ratio	Words Per Sentence	Total Words	Total Sentences	Total Paragraphs
Doc I	33.08	58.42	12.00	5.43	0.52	23.22	534	23	5
Doc II	33.04	57.88	12.08	5.44	0.53	21.58	518	24	7
Doc III	32.35	58.09	12.10	5.40	0.51	22.44	561	25	7

Table 1. Readability and Statistical Metrics Used for the Selection of Homogeneous Texts.

As a final step, the selected documents were checked by a trusted human expert to make sure they looked similar in complexity to a human eye.

3.3. Preparation of Review Kits

Once a translation kit of three documents was formed, we proceeded to translate it using the engines we intended to compare (Amazon and Google). We also had the kit translated by a trusted linguist. The human translation was then edited and double checked to ensure quality. Linguists involved in the translation and editing at this stage did not participate in the other stages of the experiment.

As we wanted to study the results of the experiment for various potential correlations, we opted for a combinatorial approach in preparing the review kits. Having six post-editors as participants, we had prepared six unique review kits (all possible permutations without repetitions).[2]

Each of these kits consisted of the same documents in the same order (Doc I, Doc II, and Doc III), with each document translated by a different translator, machine or human. The idea behind this arrangement of review kits was to facilitate the detection of correlations between effort spent and any given engine or document. If it turned out that the correlation with effort was stronger for a particular document (e.g., Doc I always required more work than other parts, regardless of the engine), it would indicate that we did not do a good enough job finding homogeneous documents (see our hypothesis H1). However, if the correlation were to be stronger for a particular engine (e.g., Google consistently required more effort regardless of the document it was used for), it would offer evidence that this engine's output was poorer in quality than the competitor's.

3.4. CAT Environment and Instructions

Translation and post-editing were carried out in Memsource, a cloud-based CAT environment. It provides useful statistics that we needed to measure the effort, incl. editing time for each

[2] They were as follows (*A* stands for Amazon, *G* for Google, *H* for Human; *DocI-A* means that Doc I was translated by Amazon): {DocI-A, DocII-G, DocIII-H}, {DocI-A, DocII-H, DocIII-G}, {DocI-G, DocII-A, DocIII-H}, {DocI-G, DocII-H, DocIII-A}, {DocI-H, DocII-A, DocIII-G}, {DocI-H, DocII-G, DocIII-A}.

segment.[3] All post-editors were asked specifically to try to complete the job in one go, without distractions, to make time measurement more reliable. They were also warned that the job consisted of three documents, not directly related to each other. No other specific instructions were given. Our goal was to make this job as similar to any other as possible. The post-editors were not notified that the job included parts translated by different translators or engines or that MT was used at all. No glossaries or translation memories were included as part of the translation package.

3.5. Calculation of Effort

Effort was to be calculated for each of the three metrics (time spent, edit distance, percentage of segments changed). All document-level and editor-level values were averaged across segment-level scores. For a broader comparison, we also used aggregated (summed) or averaged values derived as an average of the three individual metric-level values. As we were only interested in relative values (i.e., a distribution of effort for every given post-editor across different parts of a review kit), in all cases, we standardized values as a *ratio to mean*.[4] Though this method is not scientifically strict, on a small dataset like ours it provides results very similar to a T-score standardization and has an added benefit of only producing positive values. In addition, effort values for different metrics standardized as ratio to mean are, in most cases, quite comparable in their absolute size and thus lend themselves well to aggregating and averaging. For time spent, we also additionally standardized values as words edited per minute (rather than per document) to account for variations in word count among the documents.

4. Results

For total results of the experiment, see GitHub (2022). Below we present the main findings, with several key visualizations.

4.1. Homogeneity (H1)

As stated above, we were interested to see if our methodology was actually capable of determining homogeneous texts, i.e., texts consistently requiring similar relative effort. The visualizations below show that the three selected texts proved to be close enough, if not perfectly similar.

The aggregate effort was obtained by summing average effort values for each of the metrics. Each metric-level effort, in turn, was averaged across all post-editors' efforts in the respective category.

[3] The time is measured between the moment a post-editor clicks into the translation field for a segment and the moment when they click into another segment. If a segment receives several sessions of post-editing (as is often the case), all sessions' times are summed.

[4] E.g., for changed segments, if Editor X changed 45% of segments in Doc I, 32% of segments in Doc II, and 15% of segments in Doc III, the effort would be calculated as 1.467391304 for Doc I (45/[(45+32+15)/3)]), 1.043478261 for Doc II (32/[(45+32+15)/3)]), and 0.489130435 for Doc III (15/[(45+32+15)/3)]).

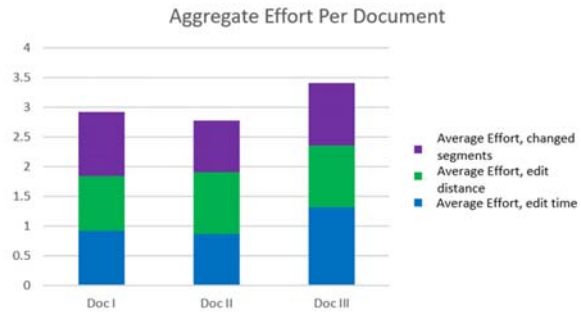

Figure 1. Aggregate Effort Across Documents.

As can be seen, Doc III required more effort than the other two, but it is not immediately clear how significant the margin is. Not only that but the difference is mostly connected to the time metric, which is inherently less reliable than the other two. In other metrics, Doc III was on par with the others as can be seen from the table below (used as the data source for Figure 1; selected are the largest values in each column):

	Time	Distance	Segments
Doc I	0.9143482	0.9199029	1.084156729
Doc II	0.8712726	1.038835	0.865417376
Doc III	1.3182554	1.0412621	1.050425894

Table 2. Metric-Level Efforts Averaged Across Documents (Ratios to Mean).

To make it more manageable, we averaged the metric-level efforts for each document and compared them using the confidence interval of 83% (recommended in Intento, 2021). Though somewhat arbitrary, this comparison shows that effort values are close enough:

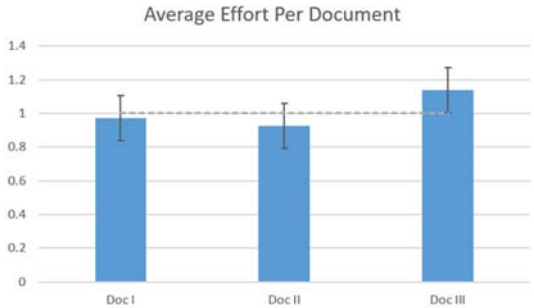

Figure 2. Average of Aggregate Effort Across Documents.

Based on the above, we cannot claim that H1 is proved. However, it cannot be rejected either, and, from a practical standpoint, the methodology seems to have yielded the results we had hoped for.

4.2. Human Benchmark (H2)

The human translation held its own against both engines and consistently required less effort from all post-editors.

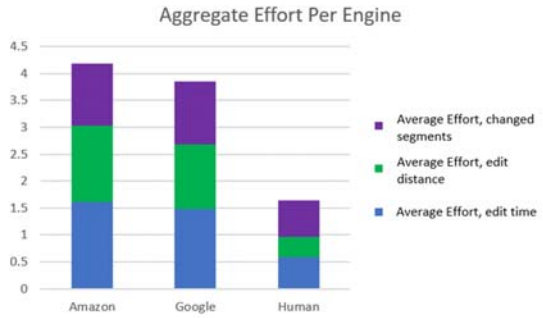

Figure 3. Aggregate Effort Across Engines.

Hence, H2 can be considered proved. It is important as it shows that the results, post-editor to post-editor, are not random, despite a significant variance in the absolute values. Some of the post-editors spent more time or made more changes on the whole than the others; however, all of them would consistently work less on a document translated by a human. It gives us more reason to believe that the relative distribution of effort between the engines was not random either and did reflect the quality.

4.3. Comparison of Engines

Amazon and Google, on the whole, performed very closely. It came as no surprise as major stock models, based on our experience, have become pretty similar in their output in recent years. However, in our case, we still could see consistent evidence in favor of Google. For practical purposes, it can be deemed enough to make a justifiable choice.

As shown in Figure 3 above, Amazon required more effort on aggregate. Below are the values for each metric:

	Time	Distance	Segments
Amazon	1.6108999	1.4126214	1.160136286
Google	1.4801239	1.2063107	1.162521295
Human	0.5869889	0.381068	0.677342419

Table 3. Metric-Level Efforts Averaged Across Engines (Ratios to Mean).

Drilling down to the post-editor-level, Figure 4 below shows how the aggregate effort varied across all participants of the experiment. Note that all values are relative. Long bars do not indicate that a given post-editor spent more absolute effort on a given document. It only shows that this particular document took this particular post-editor much more effort as compared to the other two documents in this post-editor's review kit.[5]

[5] In fact, the results are somewhat skewed in case of Editor 2 as she changed very little in *all* documents. The relative values turned out to be drastically different (and in favor of Amazon), but the absolute difference in effort behind this relative discrepancy was very small.

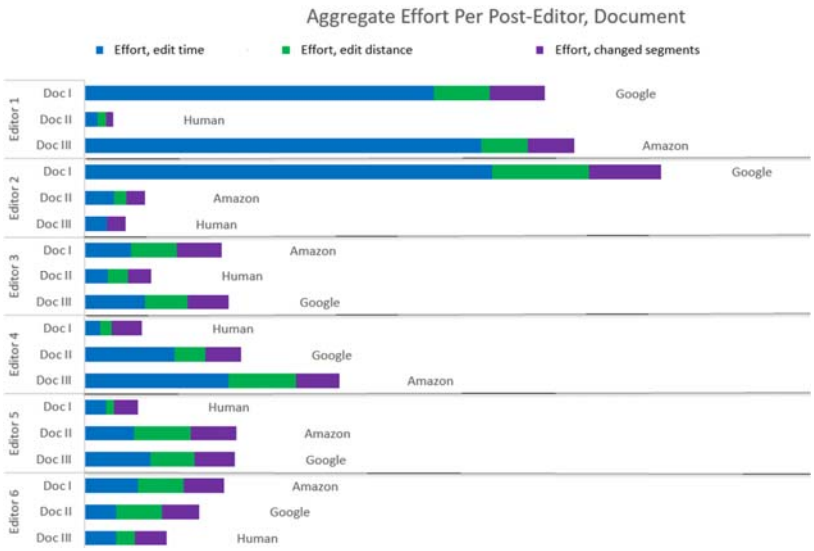

Figure 4. Aggregate Effort Across Post-Editors, Documents.

Another way to break down this data is to rank the engines based on their relative performance in each post-editor's set of documents. If a document required the least effort as compared to two others within a given post-editor's set of documents, we assigned the respective engine (or human) one point. Two points were given to the runner-up, and three points to the most effort-consuming engine. The greater the final score (summed across all post-editors), the poorer the performance.

Engine	Effort (from least to most, total for all post-editors and documents)			Score
	Least effort (1 point)	Middle effort (2)	Most effort (3)	
Amazon	0	2	4	16
Google	0	4	2	14
Human	6	0	0	6

Table 4. Aggregate Effort Ranking (Across All Post-Editors, Documents).

Again, the results are pretty close, yet Google scored slightly better.

5. Discussion

The main benefit of the proposed methodology lies in its relative simplicity and independence from unreliable human preferences and biases. Instead of creating a set of new requirements for the evaluation stage (which necessitates training and narrows the selection of candidates for the job), the methodology relies on parameters obtained through a typical editing process performed by regular (post-)editors. As was shown, the methodology provides interpretable, actionable results when applied to a real-life problem (selection of the best engine for a particular language combination and domain). The reliability of the results was corroborated by a consistent preference given to the benchmark (human) translation by all participants of the experiment.

In connection with the proposed methodology, we have also developed a separate methodology to establish homogeneity among different documents or parts of the same document. It can be used for various purposes, incl. outside the realm of MT (e.g., to quickly evaluate

complexity of any given corpus or its part). It remains to be seen, however, if this methodology is reliable enough to ensure accurate selection of homogeneous documents.

5.1. Limitations

The methodology as described in this paper relies on a combinatorial approach where each post-editor is given a unique review kit. It works well enough as long as we compare two engines (plus a human benchmark): we only need 3! = 6 review kits and, consequently, post-editors. Even for three engines, the number goes up significantly (4! = 24), which renders the procedure unpractical.

One possible solution would be to do away with combinatorics and create identical review kits (e.g., Doc I is always translated by Engine I, Doc II by Engine II, etc.). It will work if all documents are reliably homogeneous, i.e., any difference in effort could be traced to the MT quality and not to the general complexity of a document.

Another limitation is that time, which serves as one of the three main metrics, cannot be measured 100% reliably. During our experiment, we tried to take special precautions to make sure the time measurement was done in a right way; however, it is not always possible to ensure that. A solution could be to either exclude time from the set of metrics (and focus on edit distance and number of changed segments only) or reduce the weight of this metric.

5.2. Future Research

Below are several possible directions of future research. Our hope is that other researchers will join in exploring at least some of them.

As the methodology is perfectly suited for a two-engine setup, it can be used to test a custom, trained version of a model against a previous or stock version. Currently, this evaluation is often based on automatic metrics and/or subjective opinions.

To further test the methodology for establishing homogeneity, it will be interesting to see if the distribution of effort can be shown to be consistent for non-homogeneous documents as well. In other words, will documents with a higher complexity score (based on our methodology) actually require more effort, on a consistent basis? An experiment could be set up along the lines of the one described in this paper.

The apparatus used to calculate effort could be enhanced to make it stricter, incl. possible normalization of all values.

The methodology seems to be useful in verifying human parity claims (i.e., claims that a certain engine is capable of outputting translations of the same quality as provided by a good human translator). As of now, such claims can only be taken at face value. Our methodology offers a way to prove or debunk them.

It would also be intriguing to see how our results might stack up against those of a more traditional evaluation process (akin to Intento, 2021) for the same set of parameters (languages, domain, etc.).

6. References

GitHub. (2022). *GitHub*. Available at: https://github.com/Effectiff-Tech/homogeneity-scripts. Effectiff LLC.

Intento. (2021). *The State of Machine Translation 2021*. Intento, Inc.

Richterich, A. (2018). *The Big Data Agenda: Data Ethics and Critical Data Studies*. University of Westminster Press, London.