Machine Translation Post-editing: Assessing Effort, Text and Bias

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering *by Research*

by

Arafat Ahsan
2019701030
arafat.a@research.iiit.ac.in

International Institute of Information Technology
(Deemed to be University)
Hyderabad - 500 032, INDIA
March 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Machine Translation Post-editing: Assessing Effort, Text and Bias" by Arafat Ahsan, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Advisor: Prof. Dipti Misra Sharma

To my parents

Acknowledgments

First and foremost I thank my supervisor, Professor Dipti Misra Sharma for her continued supervision and guidance as I wrestled with this thesis in fits and starts. Her gentle prodding and suggestions have only made me cover this ground more thoroughly. I have greatly enjoyed our discussions over many aspects of this present work and other tasks that we worked upon over the last three years. Her marvelling at a certain linguistic phenomenon, or unraveling the syntax of a turn of phrase in the data and examples that we constantly came across, is surely a sign of someone with an ear finely tuned to the workings of language. I do hope that some of it has also rubbed off on me.

I would also like to acknowledge my fellow lab mates, Vandan Mujadia and Pruthwik Mishra, who acted not only as a sounding board for the many ideas and results in this present work, but were also involved, directly or indirectly, in some of the experiments conducted as part of these explorations. They have been most generous with their time and most earnest in their share of the tasks that we worked upon together.

I thank Samar Husain for his inputs on the design of the behavioral experiment that constituted a crucial part of this thesis. His suggestions and pointers to relevant literature in this regard were most helpful.

I thank Nazia Akhtar for putting up with my many idiosyncrasies, past and present. And many more in future too, I pray.

Finally, I am indebted to my parents for always being the shore whenever I was at sea.

ہم سے کچھ آگے زمانے میں ہوا کیا کیا کچھ
تو بھی ہم غافلوں نے آ کے کیا کیا کیا کچھ

ham se kuchh aage zamāne meñ huā kyā kyā kuchh
to bhī ham ġhāfiloñ ne aa ke kiyā kyā kyā kuchh

Not long before us much happened in this world, and yet some more.
Still we came heedless, perpetrated much, and yet some more.
– Mir Taqi Mir[1]

[1]The English translation is Faruqi's from Mir and Faruqi (2022).

Abstract

Machine Translation aided workflows have become increasingly prevalent across industry and within large institutional translation efforts. We evaluate the machine translation post-editing process end-to-end on three aspects: the usefulness of the process itself, as to whether it leads to reduction in temporal, technical and cognitive effort compared to unaided translation; the linguistic nature of the products of this process, in terms of the quality of translations generated measured against unaided human translations; and the suitability of utilizing these outputs towards a standard machine translation evaluation task. The language direction we study is English-Hindi, a mid to high resource language pair.

We first conduct a behavioral experiment utilizing professional translators. In our analysis, employing mixed-effects modeling techniques, we find that for this language pair, post-editing reduces translation time by as much as 63%, consequently increasing productivity; it reduces technical effort measured as keystrokes logged by 59%; and reduces cognitive burden measured as reduction in number of pauses by 63%.

We then investigate the nature of the translations generated during post-editing by comparing them against raw MT outputs and unaided human translations on linguistic indicators we define and implement. We find significant differences between the three corpora in terms of lexical richness, normalization and interference by the source language. We also detect and confirm the presence of translation universals in our data.

We then go on to test the suitability of translated data thus created for a machine translation evaluation task. We detect engine-reference bias towards the engine utilized during post-editing, inflated engine scores across the board on post-edited references, and find the utilization of multiple references to be the most prudent choice when conducting meta-evaluation tasks. We run these evaluations across multiple string-based and pre-trained metrics and issue general recommendations on constructing evaluation test sets and metrics to score against.

We also contribute a set of post-editing guidelines with examples culled from a real-world translation task for translators and post-editors working with the English-Hindi pair.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The quest for Machine Translation is almost as old as the quest for general Artificial Intelligence (Nilsson, 2009). Having seen off its own share of seasonal *winters* over the years[1], Machine Translation (MT) technology today has slowly but surely permeated into many translation workflows, especially over the last decade (Gaspari et al., 2015). A number of studies have reported on the advantages of integrating MT engines into translation workflows, either on their own, or in conjunction with other Computer Aided Translation (CAT) tools, such as Translation Memories (TM). Some of the reasons given by translators themselves in favor of MT adoption in their workflows are: "for speed or productivity gains; because of the perceived good quality of the MT output; for inspiration, to kick-start the translation process, or for new ideas; to reduce typing or clicking" (Cadwell et al., 2016). However, this has not been without some continued resistance from other translators. Some of the top reasons for non-adoption are cited as: "because of perceived poor quality of MT output; because of MT's negative influence on a translator's abilities" (Cadwell et al., 2016).

These impressions about the usefulness or impact of MT in translation environments have often been obtained through large scale surveys or focus group activities (Gaspari et al., 2015; Cadwell et al., 2016). And these perceptions, especially those about speed and productivity gains, have been empirically validated by many recent studies (De Sousa et al., 2011; Plitt and Masselot, 2010; Green et al., 2013). But, more often than not, these studies have tended to focus on language pairs that have contained high-resource languages. Thus bench-marking the usefulness of MT for translations into, or, from mid-to-low resource languages remains an understudied area, notwithstanding some recent work for languages like Manipuri, Mizo, and Hindi in the Indian context (Meetei et al., 2020). However, many more controlled studies, sit-

[1] https://en.wikipedia.org/wiki/ALPAC

uated within real-world translation environments are needed, not just to establish productivity baselines but also to gain insights into translator behavior in general and gauge the quality levels of the text produced in MT-aided workflows.

Translation as a distinct cognitive task has been studied and theorized about for many years, especially in the area of Translation Studies (Newmark, 1981; Toury, 2012). Lately, there has been a re-orientation towards the use of corpus based approaches to compare text produced during the process of translation with those originally created in the target language, and presence of certain *translation universals* have been postulated (Baker, 1993). Some of these universals have been dubbed as *translationese* (Gellerstam, 1986) and later efforts towards automatically discriminating between original and translated texts have uncovered their underlying lexical features (Baroni and Bernardini, 2006) .

Into this mix, we must now also introduce post-editing (PE) ("revision of rough machine translation by a competent human translator" (Krings, 2001)) as a cognitive activity distinct from unaided human translation (HT). We may therefore enquire, that just like human-mediated text (HT), do *translationese* exist for machine generated text (MT), and machine-human mediated text (PE), as well? And, if yes, then how do they differ from the source and amongst themselves?

Finally, what uses translated text (generated from human, machine, or machine-human mediation) are eventually put to matters too. The misleading conclusions drawn, when using translated text as source sentences during MT evaluation, have previously been flagged (Zhang and Toral, 2019; Toral et al., 2018a). The contention being, that translated source segments displaying translationese features are often *simpler* than data originally created in the source language. With translation and post-editing being so prevalent today in any large translation effort, the provenance of data being used, especially for evaluation, becomes very important. Thus one may ask: do references, if originally post-edited, and then used in MT evaluations, show bias towards the engines used to produce them? And what metrics or class of metrics might be most prone to such biases? Some of these are urgent questions awaiting further exploration.

Thus, we see that Machine Translation, Post-Editing, and the Text generated from these systems and activities that may later be utilized for purposes such as evaluation, are cyclically linked, raising many important questions. This thesis is an effort towards tackling a few of these questions in some detail by putting the English-Hindi translation direction under the lens for the first time.

1.2 Research Objectives

With the above discussion in mind, we divide our investigations into three broad categories.

1.2.1 Quantifying Post-Editing Effort

Post-editing effort may refer to the time spent, the characters typed, and the cognitive load incurred in editing machine translation outputs. We propose to discuss post-editing productivity and effort estimation literature and follow it up with our own behavioral study for our language pair of interest, English-Hindi. We seek to answer the following: whether we have reached a point in MT quality for this pair where we can recommend post-editing of texts, rather than translation from scratch? Does improved MT quality always imply lesser PE effort? On what dimensions and how best can post-editing effort be reliably measured? Which MT quality metrics correlate best with post-editing effort dimensions?

Formally, we pose the following research questions (RQ):

- RQ1: Is post-editing effort as measured on temporal, technical and cognitive dimensions *lesser* in the PE condition than the HT condition for the English-Hindi direction?
- RQ2: Is the quality of post-edited segments *equal to* translated segments as ascertained by human raters?
- RQ3: Do automatic MT evaluation metrics *correlate* with PE effort indicators, both measured at the segment level?

1.2.2 Quantifying the Nature of PE Texts

While a majority of previous studies have shown MT post-editing to be a more efficient activity than translating from scratch, not many have addressed the issue of the linguistic quality and features of translations thus produced in any detail. Also, while researchers in the field of Translation Studies have previously hypothesized about the presence of *translationese* in text produced under a translation setting, and how they might differ from text produced in monolingual settings, it is only recently that some attention has been paid to this phenomenon in the context of machine translation aided workflows. In proposed experiments addressing these concerns, borrowing and building on recent work applying these ideas to raw MT outputs and post-edited texts, we scrutinize if these hypotheses stand true for our corpora and language pair.

The research questions we pose are as follows:

- RQ1: Do post-edited text and human-translated text exhibit different characteristics?
- RQ2: Do text created by different translators exhibit different post-editese and translationese behavior?
- RQ3: Does raw MT track closely with post-edited or human-translated text?

1.2.3 Quantifying the Impact of PE Texts on MT Evaluation

Natural Language Processing (NLP) in general, and MT in particular, constantly reuse text produced under various settings for training and evaluation of newer models, as well as for various bench-marking tasks. But what if we find text created under varying conditions to exhibit different properties, what impact might they have on a downstream activity like MT evaluation? Do they carry biases which might adversely affect the outcome and reliability of these downstream tasks? These are some of the questions we address in experiments around gauging the impact of these target text. More formally, we study the following:

- RQ1: Are post-edited (PE) references biased towards MT engines they were created from compared to their unaided human translated (HT) references?
- RQ2: Do the number of references used affect MT evaluation outcomes?
- RQ3: Which current MT evaluation metrics fare best under varying reference conditions?

1.3 Thesis Contributions

In studying the three Research Objectives outlined earlier and the questions brought forth in each, we make the following contributions:

1. We design and conduct the first behavioral post-editing effort estimation study for English-Hindi under controlled conditions utilizing professional translators.
2. We bring out the differences in machine-generated, post-edited, human-translated and original monolingual text using both established as well as novel metrics that we propose.
3. We show the impact of these differences when text generated in these settings are used for MT evaluation, and make recommendations towards creating benchmark evaluation data.
4. We contribute a unique dataset of translations from English into Hindi produced under two conditions by 10 professional translators.
5. We contribute post-editing guidelines with English-Hindi examples that may be utilized for post-editing activities in future.

1.4 Thesis Outline

Having introduced our motivation and research objectives in this chapter (Chapter 1), in Chapter 2 we present the design, results, and findings from our post-editing behavioral experiment. In Chapter 3, we analyze the text produced in the behavioral experiment conducted

earlier and contrast the nature of these text using various technical metrics. In Chapter 4, we utilize these text in the task of MT quality evaluation and scrutinize them for inherent biases. Finally, Chapter 5 concludes this thesis summarizing the results and laying out promising future research directions related to this topic.

Chapter 2

Assessing Post-editing Effort

2.1 Introduction

Translation workflows that are based on post-editing of Machine Translation output are being increasingly adopted in the industry (Gaspari et al., 2015). Gains that accrue from a post-editing based workflow, measured over multiple post-editing effort indicators, have been reported to be considerably significant by a number of previous studies over multiple language combinations (Plitt and Masselot, 2010; C. M. de Sousa et al., 2011; Green et al., 2013). But in order to extend post-editing beyond its current silos, it is imperative to put new and less-studied language pairs under the lens to make a case for its wider adoption via empirically backed evidence.[1]

Post-editing effort is often quantified across three different dimensions, each focusing in turn on a different aspect of post-editing behaviour (Krings, 2001). The dimensions studied are the following: *Temporal* — understood as the time taken to complete a translation task, often reported per segment or word; *Technical* — estimate of the physical labour of the translation activity, measured in terms of keystrokes logged or edit operations performed; and *Cognitive* — an indirect estimate of the extent of cognitive processes underlying the translation task, inferred from keylogging pause or eye-tracking data, as it is not possible to observe these directly (Moorkens et al., 2015).

If it can be shown that post-editing machine translation output is temporally efficient, technically less laborious, and cognitively less demanding, then it can be recommended as the default workflow for large translation jobs. But this first calls for a comparison between machine translation based post-editing behaviour (henceforth PE) and unaided human translation

[1] Gaspari et al. (2015)'s survey reveals a heavy skew towards English and other European language combinations.

from scratch (henceforth HT). Thus, the research questions (restated from Chapter 1) that we pose are the following :

- RQ1: Is post-editing effort as measured on temporal, technical and cognitive dimensions *lesser* in the PE condition than the HT condition for the English-Hindi direction?
- RQ2: Is the quality of post-edited segments *equal to* translated segments as ascertained by human raters?
- RQ3: Do automatic MT evaluation metrics *correlate* with PE effort indicators, when both are measured at the segment level?

Most of this chapter will focus on answering the first question in some detail. We are equally interested in the other two as well, but will only be presenting some initial results from a first attempt at tackling them in this chapter, with more to follow in chapters 3 and 4.

The rest of this chapter is organized as follows: Section 2.2 discusses some past studies including those that have previously studied the English-Hindi post-editing direction. In Section 2.3 we detail our experimental setup. Section 2.4 presents our results and analysis and in Section 2.5 we draw our conclusions and sketch the outlines of our future work pertaining to these outcomes.

2.2 Related Work

We now take a more detailed look at some of the past efforts towards contrasting the two settings. Plitt and Masselot (2010) compared HT and PE when translating from English into 4 European languages (French, Italian, German, and Spanish) and reported an overall productivity gain of 74% which converted into time savings of 43%. They also observed a 70% reduction in keyboard time and 31% in pause time for the PE setting. C. M. de Sousa et al. (2011) also report PE to be 40% faster than HT in the English-Portuguese direction when translating movie subtitles.

Other studies (Läubli et al., 2013) have reported more modest gains, with estimated time savings of 15–20% when translating between a European language pair (German-French) within a *realistic* translation environment.[2] Garcia (2011) also finds only marginal productivity gains when studying the English-Chinese pair and, additionally, reports an impact of directionality when source and target languages are switched.

All of these earlier experiments, however, were based on the output of Phrase based Statistical Machine Translation (PBSMT) systems. With Neural Machine Translation (NMT) and its

[2]Experimental settings for these studies may deploy specialized interfaces for accurate measurements or make use of environments already familiar to professional translators.

subsequent iterations being the current state of the art and outperforming PBSMT (Bahdanau et al., 2014; Vaswani et al., 2017; Castilho et al., 2018), this shift in the technology paradigm from PBSMT to NMT must then be addressed in post-editing studies as well.

Läubli et al. (2019) conduct such a study, this time utilizing the output of an NMT system to compare PE with HT in the German-French and German-Italian translation directions.[3] They report significant overall productivity gains, but with marked differences between the pairs: 59.74% for the former and only 9.26% for the latter. Another interesting comparison of HT, PBSMT, and NMT post-editing settings performed on a literary text (chapter from a novel) reports an increase in productivity by 36% for the NMT based setting over HT (Toral et al., 2018b).

We also notice in previous studies, that throughputs vary considerably depending on the language pair under the lens (Green et al., 2013; Läubli et al., 2019). We now discuss some earlier efforts that have included an Indian language in their experiment.

Shah et al. (2015) conducted an experiment where students post-edited parts of a specialized English language textbook on *bioelectromagnetism* into 7 languages, 3 of them being Indian languages including Hindi. They reported an increase in post-edit time by a factor of 3–5 when the target language was an Indian language. They put this down to greater terminological distance between English and Indian languages compared to other languages in their experiment. They do not study and compare against the HT condition, or report on technical or cognitive indicators.

Carl et al. (2016) also report results on Hindi (amongst 6 languages) comparing the HT and PE conditions. Their English-Hindi results are based on an existing multilingual translation database that contains experimental data of translators' activities in both conditions. They find in favour of the PE condition across all languages when measured on temporal indicators, but report translating into Hindi to be the slowest amongst the 6. They do not quantify average throughput gain or time savings.

Meetei et al. (2020) compare PE behaviour when translating from English into 3 Indian languages (Manipuri, Mizo and Hindi). They conduct light post-editing and report Hindi to be the fastest to post-edit amongst the three languages.[4] They ascribe it to the availability of relatively mature MT systems in the English-Hindi direction compared to Mizo and Manipuri, which are low-resource languages. They use student volunteers as translators and do not investigate cognitive indicators.

Ahmad et al. (2018) present an industry perspective and claim a 2–3 fold increase in productivity for users using their tools in combination with MT. However, they base this on longitudinal tracking of their users.

[3]The HT condition is aided by a domain specific translation memory (TM).

[4]It is also often referred to as *good enough* translations and is lower than publishable quality translations.

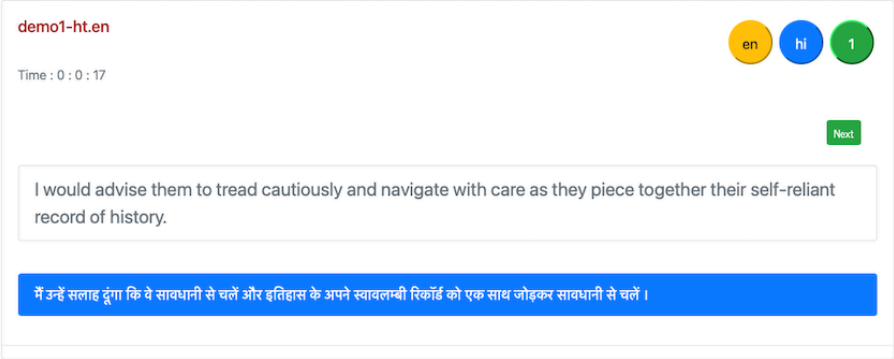

Figure 2.1: Web-based Translator workbench used in the study.

While all these past studies have certainly helped in providing insights into post-editing behaviour in Indian languages in general and in Hindi in particular, we sense a need for a more in-depth look along all PE effort indicators within at least one Indian-language setting. With our current work, we seek to address this gap.

We see our contribution differentiated from previous related work as follows: *(i)* we present in-depth results from all three primary indicators of PE effort (temporal, technical and cognitive) for the English-Hindi direction; *(ii)* we account for per translator and per item variation with the use of mixed-effects models; *(iii)* we utilize professional translators in order to accurately gauge the impact of contrasting conditions; *(iv)* we conduct a human quality-rating exercise comparing target text produced in both conditions; *(v)* we present correlations of automatic MT metrics with PE effort indicators.

2.3 Experimental Setup

We conducted this experiment under a 2 (translation conditions) $\times$ 200 (source segments) mixed design. All subjects saw both factor levels (HT and PE), but only one combination for each level, as having been exposed to a source segment in one condition would have affected their translation in the other. The study was conducted online over 5 consecutive days with 2 sessions per day. The sessions were not time bound. Subjects translated 2 files of 10 segments each in alternating conditions in each session.

All subjects participated in a warm-up translation exercise a week prior to the start of the actual task.[5] This was done to establish familiarity with the interface used in the study. We chose to adapt an existing beta version of a web-based translation workbench by adding extensive keystroke logging features along with some other minor tweaks.[6] The UI itself was

[5]They were introduced to a summarized version of the post-editing guidelines from Appendix B

[6]https://posteditme.in

kept clean and uncluttered, serving one segment at a time to the translators. This meant that while translators had previous context of the text under translation, they could not navigate ahead for context. A timer was displayed once a translator navigated to each new segment. This was meant to prompt the translator to focus on the activity at hand. Figure 2.1 shows the workbench interface as seen by translators under the PE condition.

We instructed the participants to aim for publishable translation quality, which entailed full post-editing as defined in Appendix B. They were free to conduct web searches and consult online or offline dictionaries, but were discouraged from spending too much time doing so.[7] It was deemed acceptable to transliterate any technical terms or terminology into Hindi if they could not find its translation even with the aid of resources available to them. However, they were strictly prohibited from consulting any online MT engines during the task. Subjects were encouraged to complete each task (consisting of 10 segments) in a single sitting without a break.

Previous studies, such as those discussed earlier, have noted the impact of a number of different variables (language pairs, MT paradigms, text domains, translation environments, translator competencies) on translation throughputs. This calls for not only careful experiment design, but also utilization of techniques that can help with the testing and inference of results. Green et al. (2013) utilized one of the first such designs for post-editing productivity studies and deployed mixed-effects models (Baayen et al., 2008) to account for inter-language, inter-subject, and inter-item variability.

Mixed-effects models are able to model this variability in two ways: *(i)* through random intercepts, that can account for the differences between translators seen in their differing throughputs (or differences between linguistic items due to the features inherent to them); *(ii)* and through a random slope that accounts for how different subjects may experience the change of condition differently. Accounting for these variabilities allows us to isolate the effect of condition, generalize our findings beyond our sample, and avoid the 'language-as-fixed-effect fallacy' (Clark, 1973).

In fitting our mixed-effects models we follow a methodology similar to the one described by Baayen et al. (2008) and followed by Green et al. (2013) and later Toral et al. (2018b). Maximal models were fit when possible (Barr et al., 2013); in case of convergence failure, a less complex model was fit by successively removing the random slopes of the by-subject and by-segment random effects component. Models thus obtained were compared via likelihood ratio tests. We also refit our final models after filtering data points with residuals deviating more than 2.5 standard deviations. This helps check for the influence of any atypical outliers (Baayen et al., 2008). We verify the residual plots for normality and homoscedasticity. We utilized the `lme4` package in R (Bates et al., 2015) for all mixed-effects models related analyses.

[7] This was done as technical terminology related difficulties have previously been noted for this language direction (Shah et al., 2015).

Day	S1-T1	S1-T2	S2-T1	S2-T2
Day 1	20.20 (A1)	26.30 (A1)	26.80 (A1)	21.30 (A2)
Day 2	20.90 (A2)	22.30 (A2)	24.90 (A2)	23.20 (A3)
Day 3	26.40 (A4)	24.60 (A4)	26.30 (A4)	22.40 (A4)
Day 4	22.40 (A4)	20.30 (A4)	18.00 (A5)	18.70 (A5)
Day 5	22.50 (A5)	16.00 (A5)	20.30 (A5)	23.00 (A5)

Table 2.1: Study Data. Average sentence lengths (in words) per session-task block as presented to translators. Also shown in parenthesis are the source articles used for each block.

2.3.1 Data

We assembled a corpus of recent English language news articles from two distinct online sources. The choice of news as a domain was motivated by observations of terminology-related difficulties in more specialized domains as reported by earlier studies (Shah et al., 2015). Each news article was segmented into sentences using the NLTK library and divided into *blocks* of 10 segments.[8] Only those blocks were used that fell within a $MEAN \pm SD$ threshold of the corpus mean (Table 2.1). We prioritized the continuity of a news article across blocks when making *block* selections in order to preserve the natural discourse of the text.[9] This methodology yielded a total of 200 unique source segments divided into 20 blocks of 10 segments each, spanning 5 different news articles: A1–A5. Conditions were counterbalanced to handle order effects.

2.3.2 Participants

The participants of our study are self-declared professional translators. We contacted a professional translation service provider to help assemble the pool.[10] A short questionnaire accompanied the registration form for the task. Of our participant pool of 10 subjects, 70% reported 2–5 years of experience translating in the English-Hindi direction, while 30% reported 0–2 years of experience. The same percentage breakdown was observed for a question related to previous post-editing experience. All subjects were paid the going market rates for the task regardless of the condition (PE, HT).

[8] https://www.nltk.org/api/nltk.tokenize.html

[9] In 2 cases out of 20 we had to skip the subsequent block, owing to short average sentence lengths of the blocks.

[10] http://www.ebhashasetu.com/

2.3.3 MT System

The English→Hindi MT engine used for the task is a transformer based neural machine translation system. This subword-based NMT system is trained on cleaned WAT 2021[11] English-Hindi training corpus using the Opennmt-py toolkit (Klein et al., 2020). The system also utilizes forward and backward translations on the IndicCorp monolingual corpus to obtain synthetic data for training.[12] It uses subwords as the basic translation unit with 20,000 merge operations on both source and target languages. The system obtained a BLEU score of *35.46* on cleaned WAT 2021 English-Hindi test data.

2.3.4 Pre-processing

Once we processed the activity logs for all 10 subjects across all 200 segments, they yielded 2000 unique observations. We found that 7 of these items (all from the HT condition) did not contain a final translation, so we discarded those. We think that in these cases, participants may have accidentally navigated to the next segment without having completed a translation. In the PE condition we found that one subject *P01* had not edited 68% of the MT outputs and had accepted them without modifications. This was almost 3 times the next highest proportion we detected across all other subjects. We decided to remove all data points generated by this subject. We were thus left with 1793 observations on which we base these results.

We calculated time per segment, source segment lengths (in words and characters), number of keystrokes (total, as well as those belonging to different categories: content, navigation and deletion), average pause duration, initial pause duration, and number of pauses. We also computed H-BLEU (Papineni et al., 2002), H-TER (Snover et al., 2006) and H-chrF (Popović, 2015) metrics on the post-edited segments.[13]

Having assembled this data, we set out to answer the research questions posed earlier in Section 2.1.

2.4 Results

2.4.1 Temporal Effort

We first present a view of temporal effort in terms of productivity measured as words per hour. We see productivity improvements in the PE condition across the board except for subject *P07*. Overall, this translates into a throughput increase from 359 words/hour to 979

[11] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/

[12] https://indicnlp.ai4bharat.org/corpora/

[13] H signifies that scores were computed using the reference generated in the PE condition by the same subject.

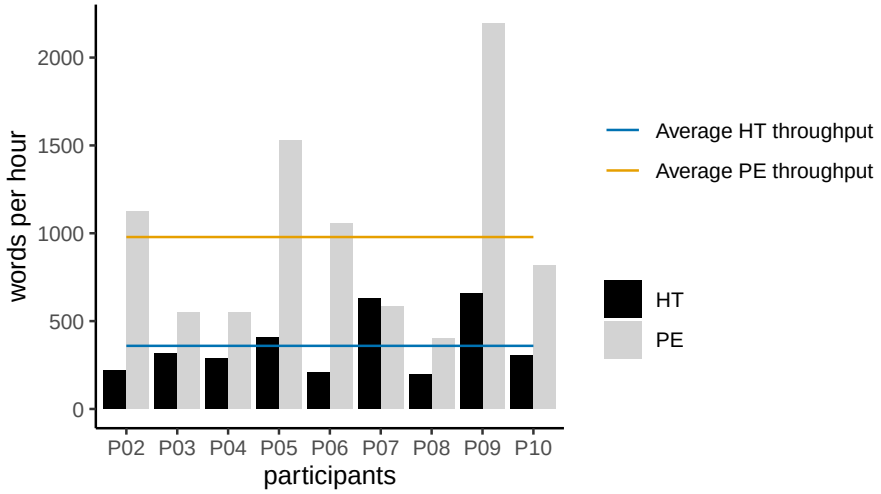

Figure 2.2: Individual translation throughputs in words per hour and average throughput in each contrasting condition.

words/hour. We thus observe an overall productivity gain of 172%, which amounts to 63% in time savings.

This is more than twice the 74% gain reported by (Plitt and Masselot, 2010) when studying European-language pairs and the 59.74% reported by (Läubli et al., 2019) recently. But we note that in the first case, the experiments were conducted on PBSMT outputs, and in the second, while NMT was used, the control condition was aided by a TM, thus pushing up the baseline throughputs. With this context in mind, the average productivity gain seen in our study does not appear to be unrealistic.

Figure 2.2 shows a comparison of individual throughputs in contrasting task conditions along with means aggregated for the two conditions. We also note a great variation in productivity gain amongst subjects ranging from -7% (P07) to 410% (P09).

Table 2.2 helps interpret this further. We contrast the number of unedited and edited MT proposals per subject and their individual productivity gain percentages. It follows that higher the acceptance of MT proposals without modifications by a subject, greater the gain in individual productivity. While this may point to high quality MT output, it also demands a closer scrutiny of the quality of translations generated in each condition. We address this in Section 2.4.4.

We now report the mixed-effects regression results. Plotting temporal data showed a right-skewed distribution, so we log transform all time data before proceeding further. As our goal is to predict translation time and establish the significance of conditions, we fit a linear mixed-effects regression model with two fixed-effect predictors (condition and segment length) and

Participant	Unedited (%)	Edited (%)	Productivity Gain (%)
P02	20	80	411
P03	3	97	73
P04	3	97	92
P05	18	82	275
P06	25	75	404
P07	0	100	-7
P08	2	98	105
P09	25	75	234
P10	16	84	169

Table 2.2: MT segments accepted without modifications and with modifications per subject along with individual productivity gain percentages.

Predictor	Temporal	Technical	Cognitive		
			number	average duration	initial duration
Segment length	$\uparrow^{***}$	$\uparrow^{***}$	$\uparrow^{***}$	$\uparrow^{***}$	$\uparrow^{***}$
Condition (PE vs. HT)	$\downarrow^{***}$	$\downarrow^{***}$	$\downarrow^{***}$	–	–

Significance levels: —($p > 0.1$), ($p < 0.1$), $^{*}(p < 0.05)$, $^{**}(p < 0.01)$, $^{***}(p < 0.001)$.

Direction: ($\uparrow$ $\downarrow$) arrows depict whether the predictor has a negative or positive correlation with the dependent variable.

Table 2.3: Significance levels of predictors in our final models across all modeled PE effort dimensions.

two random-effect predictors (subjects and segments), where on the subject predictor we also include a random-slope for task condition.

In the final model, we observe a significant main effect for both segment length as well as translation condition.[14] Temporal effort significantly increases with segment length, but decreases for the PE condition. Table 2.3 shows the significance levels and direction for each predictor in our final models across all PE effort dimensions that we study.

2.4.2 Technical Effort

We measure technical effort as the number of keystrokes used to generate the target text. We normalize it per source segment character. Figure 2.3 shows 1.33 keystrokes used per source

[14]We utilize the *lmerTest* package that extends results with p-values for models built with *lme4*.

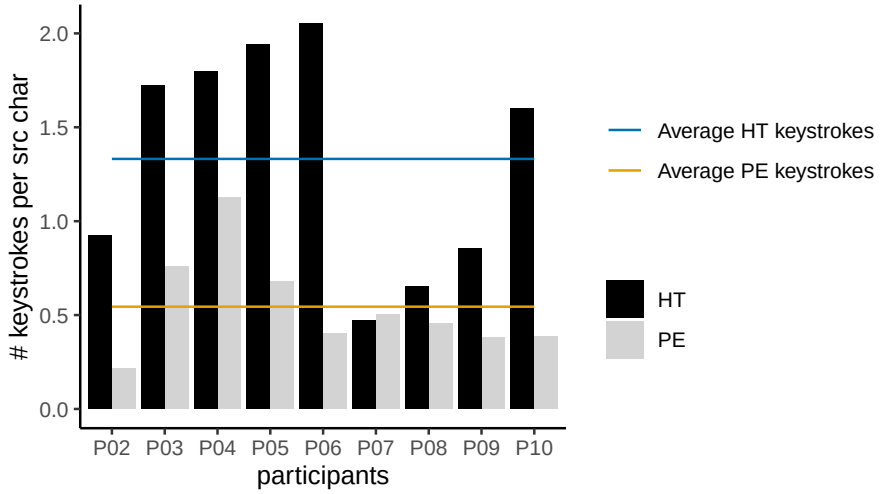

Figure 2.3: Technical effort estimated as number of keystrokes needed to generate target text per source character.

character in the HT condition and 0.54 keystrokes in the PE condition, amounting to an effort reduction of 59%. Contrast this with the 23% reduction reported by Toral et al. (2018b) when post-editing a literary text. Again, except for subject *P07* all participants show reduced effort in the PE condition.

We also classified each keystroke based on the type of the keystroke logged. We classify these into content, navigation, and deletion categories and report the percentage breakdown of the total into these categories in Table 2.4. We observe higher navigation and deletion operations in the PE condition (28% and 26%) than the HT condition (8% and 14%), while content keystrokes register a higher percentage in HT (77%) compared to PE (46%). Subject *P08* is an interesting case as they register the highest number of delete operations (they have high navigation numbers too) in either condition amongst all participants. This could point to frequent revisions made on the text.

As number of keystrokes is expressed as counts, we fit a Poisson generalized linear mixed-effects model to predict technical effort. We follow the same methodology as described in the previous section. We again find a significant main effect both for segment length as well as translation condition (Table 2.3), similar to what we saw for the temporal dimension earlier. Technical effort increases with increase in segment length, and decreases for the change in condition to PE.

Participant	HT (%)			PE (%)		
	Content	Navigation	Deletion	Content	Navigation	Deletion
*P*02	81	9	11	41	32	27
*P*03	94	1	5	58	8	34
*P*04	86	9	5	48	41	11
*P*05	85	5	10	56	30	14
*P*06	90	1	9	54	28	18
*P*07	66	11	23	56	12	32
*P*08	24	25	51	10	27	63
*P*09	78	12	10	35	52	13
*P*10	94	1	5	59	19	22
Mean± *SD*	77.55 ± 21.80	8.15 ± 7.89	14.30 ± 14.63	46.52 ± 15.91	27.69 ± 13.65	25.8 ± 16.1

Table 2.4: Types of keystrokes generated by subjects in each condition.

2.4.3 Cognitive Effort

Post-editing effort estimation studies based on eye-tracking data use *fixations* as a proxy to estimate cognitive load; the idea being that greater the number and duration of fixations, greater the cognitive load (O'Brien, 2011). In the absence of eye-tracking data, the use of *pauses* as a proxy for cognitive load is also well established (O'Brien, 2006). We report on three such cognitive indicators: number of pauses, pause duration, and initial pause duration. Findings related to the first two have been reported in previous post-editing literature (Green et al., 2013; Toral et al., 2018b).[15] The third (initial pause duration), we introduce in order to gauge differences in reaction times from when a subject first navigates to a new segment displayed in either condition to their first action on it.

We calculate the time difference between two subsequent key events and consider all observations above *1000ms* to be pauses following O'Brien (2006); Koehn (2009).

Figure 2.4 shows the differences in the frequency of pauses for each subject in the two conditions. We notice a reduction of 63% in the PE condition from 31 pauses per segment in HT to 12 pauses per segment in PE. This points to a much reduced cognitive load when post-editing.

However, a similar exercise on pause duration data reveals an increase of approximately 12% in the PE condition compared to the HT condition (Figure 2.5). Although, it is not significant, this is in line with findings reported previously comparing these two specific cognitive indicators (Green et al., 2013).

[15]There is also an indicator reported as pause ratio which we eschew in favour of initial pause time.

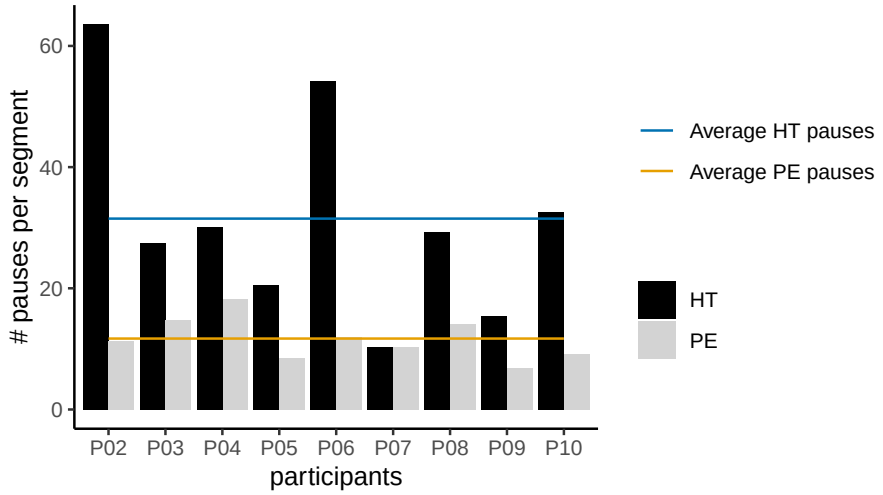

Figure 2.4: Cognitive effort estimated as average number of pauses per source segment.

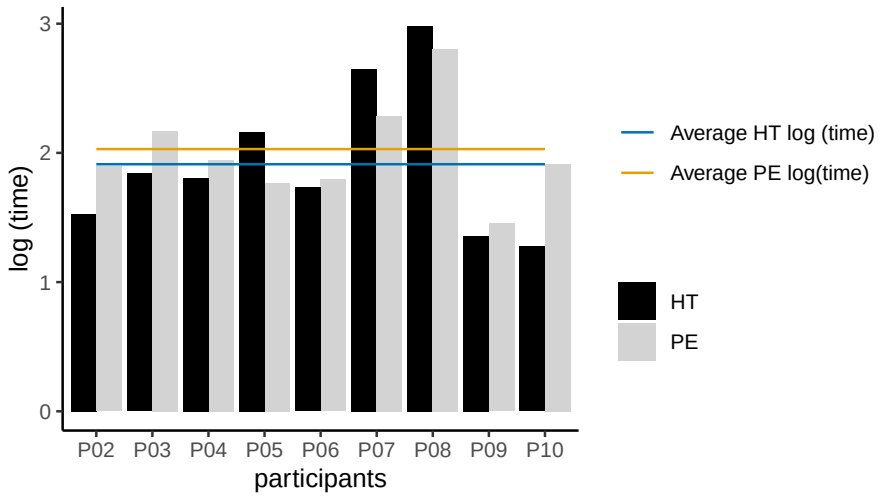

Figure 2.5: Cognitive effort estimated as average pause duration per source segment.

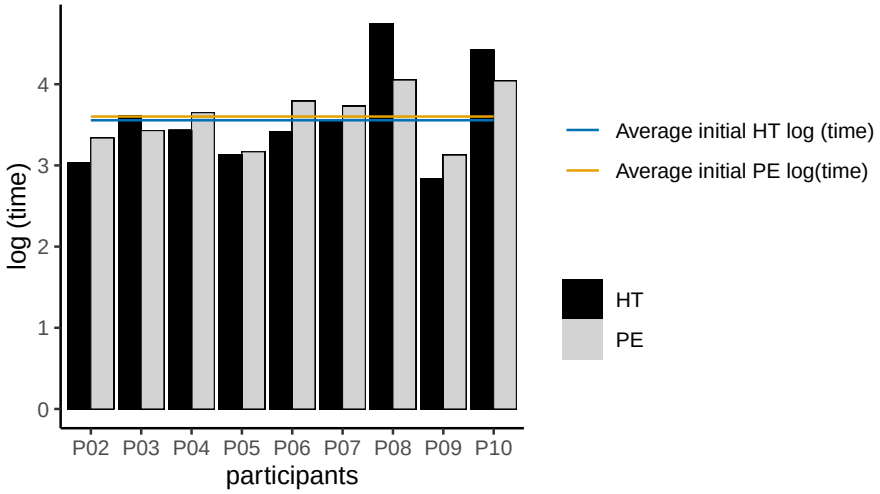

Figure 2.6: Cognitive effort estimated as average initial pause duration per source segment.

We finally compare initial pause duration between PE and HT. We expect this initial load to be higher for the PE condition given that there are two segments displayed to the subject in this condition: the source segment and the MT proposal, which have to be read and comprehended before starting the post-editing activity. This seems to hold, but not significantly, as we see only a small increase of about 5% for the PE condition (Figure 2.6). One explanation could be that in the PE condition, the MT proposal in spite of registering a higher cognitive load initially also later acts as a helpful prompt for the subject. An eye-tracking based experiment might prove useful in teasing apart these two opposite effects.

When comparing the means[16] for pause duration and initial pause duration we find pause duration (6.77s for HT and 7.71s for PE) to be considerably lower than initial pause duration (35.02s for HT and 36.67s for PE). The translator therefore, takes a longer initial pause to start formulating a response, but once they start the activity, they take considerably shorter pauses.

We go on to fit three mixed-effects models to validate these findings. A Poisson generalized mixed-effects model to estimate pause counts finds significant main effects for segment length and condition. Cognitive effort (measured as count of pauses) increases with segment length and decreases for the change in condition to PE.

The other two linear mixed-effects models fit to predict average pause duration, and initial pause duration find a significant effect only for segment length and not for condition (Table 2.3). This shows that while cognitive effort certainly increases with segment length, the change

[16] After transforming back from log scale. Also, note that pause duration does not include initial pause duration as a component, and is the duration of pauses after post-editing starts.

Evaluator	PE vs. HT		
	Win	Loss	Tie
$E1$	18	14	15
$E2$	27	15	5
$E3$	7	7	33
$E4$	14	11	22
$E5$	13	11	23

Table 2.5: Pairwise quality judgements on sampled target texts reported as win, loss, and ties for PE against HT.

in condition to PE, does not have a discernible effect on cognitive effort, when measured by the average and initial time duration indicators.

2.4.4 Quality Judgements

To evaluate whether the quality of texts created in the PE condition matched those created in the HT condition, we conducted a human judgement based pairwise ranking task Callison-Burch et al. (2011) on a small sample. We randomly sampled 3 target segments per condition from each subject. For each target text thus obtained, we paired it with another random sample after constraining on condition. We thus obtained 60 HT-PE pairs for evaluation. As we discovered issues (discussed earlier in Section 2.3.4) with subject *P01*'s data after the quality evaluation exercise had already been completed, we removed any pairs that had a segment translated by the subject. We report our results on this filtered set that consists of 47 pairs. Five evaluators were asked to judge each pair. Ties were allowed.

Table 2.5 shows the judgements from evaluators represented as win-loss statistics on the PE condition. We notice a high number of ties and a slight preference for the PE condition. However, the preference does not test to be significant on a sign test ignoring ties (p-value = 0.08). We conclude that translating in either condition produces similar quality target segments. However, we realise that the sample size was quite small compared to the number of possible combinations across all participants. We hope to conduct a more thorough review of quality in future.

2.4.5 Automatic Quality Metrics

Finally, we investigate the correlations of some popular automatic MT evaluation metrics with the post-editing effort indicators reported so far in this study. We generated metric scores

MT Metric	PE Indicator		
	temporal	technical	cognitive (# pauses)
$(H)BLEU$	r = −.56, ***	r = −.71, ***	r = −.48, ***
$(H)TER$	r = +.54, ***	r = +.70, ***	r = +.49, ***
$(H)chrF$	r = −.56, ***	r = −.73, ***	r = −.49, ***

Note: All coefficients for r(898). For TER lower is better hence the positive correlation. The other two cognitive indicators (average and initial pause duration) did not show any correlation with any of the metrics – coefficients were close to 0.

Table 2.6: Correlations of PE Effort indicators with automatic MT metrics.

by comparing the MT proposal against its post-edited reference. We calculate scores for H-(BLEU, TER, and chrF).

Table 2.6 shows moderate correlations for all 3 MT metrics on the temporal indicator, similar to what Tatsumi (2009) also reported for this indicator. Correlations then get stronger on the technical indicator and then fade for the cognitive indicator.

We believe this may be because cognitive effort is the only one out of the three PE dimensions we studied that is not directly observed (instead, inferred from pause frequency and pause duration data), whereas the technical and temporal indicators can be measured more directly. This is similar to findings previously reported by Moorkens et al. (2015) who note a similar correlation trend across the three PE effort dimensions. The technical effort indicator appears to be the one most strongly correlated with automatic metrics.

The other two cognitive indicators (average and initial pause duration), which did not test significant as per our mixed-effects models, also do not show any correlation with any of the MT metrics – coefficients obtained were close to 0. We omit reporting them in Table 2.6 due to space constraints.

2.5 Conclusion

We conducted a post-editing effort assessment study and presented detailed analysis of effort indicators along the temporal, technical and cognitive dimensions. We observed that in the temporal dimension, post-editing reduced translation time by 63%; in the technical dimension

it reduced number of key strokes by 59%; and in the cognitive dimension, it reduced the frequency of pauses by 63%. However, it increased average pause duration by 12% and average initial pause duration by 5%.

We then compared the quality of translations generated in each condition and found them to be similar.

And finally, we detected moderate to strong correlations for 3 automatic MT evaluation metrics across all PE effort indicators, with technical effort most strongly correlating with automatic MT metrics.

This study helped benchmark the gains accrued, when moving from a human translation only setup to a machine translation based setup for the English-Hindi pair. It uncovered insights on post-editing behavior when translating in HT and PE conditions and showed contrasting cognitive behaviors (depending on the indicator used) in the two conditions. There seem to be competing cognitive processes at work that call for a more detailed look in future.

The last two observations regarding human quality judgements, and MT metrics and their correlations also demand a closer look. We expect to undertake this as part of our future work. We also propose to extend this study by including a third condition in future, either as an additional MT engine to check if MT quality differences show up in PE effort indicators (Toral et al., 2018b), or by introducing the use of translation aids (TM) to gauge their impact in a similar manner (Läubli et al., 2013, 2019). In fact, another interesting experiment could entail the introduction of a previously translated segment as an MT proposal, which may help uncover not only behavioral insights about *reviewer* behavior, but also benchmark the Machine Translation quality upper bound. We would expect the technical, temporal and cognitive indicators to vary significantly in this case when this is undertaken as a *review* activity which may turn out to be a distinct cognitive task from post-editing and translation.

We also intend to study other language pairs, especially those within the multilingual Indian context.

Chapter 3

Measuring Post-edited Text

3.1 Introduction

Traduttore, traditore (Translator, traitor) goes the Italian proverb pointing to the impossible task of being loyal to both the source and target text simultaneously (Bar-Hillel, 1954). Inevitably, a translator strikes compromises in favor of one or the other. But how does this affect the text? And, how then, might the nature of translated texts be ascertained? Or more broadly, one may ask, what is the nature of *any* mediated text in translation, be it human or machine mediated? The area of Translation Studies has attempted a number of theories about the translation process and its artefacts over the years.

Newmark (1981, p. 7) supplies the following definition "Translation is a craft consisting in the attempt to replace a written message and/or statement in one language by the same message and/or statement in another language. Each exercise involves some kind of loss of meaning, due to a number of factors. It provokes a continuous tension, a dialectic, an argument based on claims of each language. The basic loss is on a continuum between over-translation (increased detail) and under-translation (increased generalization)". In his work, this definition is then brought into sharper relief by coining the terms: *Semantic translation*, where the translator attempts to produce the precise contextual meaning of the source language (SL) author; and *Communicative translation*, where the translator attempts to produce the same effect on the target language (TL) readers as was produced by the original on SL readers. The anchor though remains the source language, retaining its primacy in these earlier explorations.

Baker (1993) calls for a move away from this source-oriented notion of semantic equivalence towards a more corpus based approach. One of the primary efforts that they cite in this regard is the notion of *translation norms* proposed by Toury (2012). These norms are understood as regularities or choices frequently taken up by translators while translating within a given socio-cultural context. These patterns may be gleaned by scrutinizing corpora of source and

target text to uncover strategies that translators frequently resort to. As per Baker (1993), "The concept of norms tips the balance not only in favour of the target text (as opposed to the traditional obsession with the source text), but, more important, it assumes that the primary object of analysis in translation studies is not an individual translation but a coherent corpus of translated texts". Thus we see a re-orientation towards data or corpus oriented approaches in these later studies. Baker (1993) then goes on to list some features which typically occur in translated text rather than original monolingual text and refers to them as *translation universals*:

1. *Explication*: Understood as the tendency to make source information more explicit in target by adding more background information into the target.

2. *Simplification and Disambiguation*: Refers to efforts to resolve lexical ambiguity and to simplify difficult syntax.

3. *Normalization*: Refers to the preference for conventional 'grammaticality' such as rounding off unfinished sentences, grammaticizing ungrammatical utterances and omitting false starts and self-corrections to make the text more conventional.

4. *Reduced Repetition*: Points to the tendency to avoid source repetitions by either omitting or rewording them in target text.

5. *Exaggeration*: Refers to a tendency to exaggerate some features of the target language.

6. *Interference*: Understood as transference of some source features to target texts.

Many of the above, have also been referred to as *third code* (Frawley, 1984) or as *translationese* (Gellerstam, 1986).

Recent efforts have been made to find empirical evidence for some of these hypothesized universals by comparing translation corpora created in various domains (Pastor et al., 2008). Later studies have also sought to extend the search for these universals to text created with the mediation of machine translation engines and looked for similar markers dubbed as *post-editese*. We continue these investigations using established as well as new indicators/features and conduct a series of computational analyses in an attempt to answer the following questions:

- RQ1: Do post-edited text and human-translated text exhibit different characteristics?

- RQ2: Do text created by different translators exhibit different characteristics?

- RQ3: Does raw MT text track closely with post-edited or human-translated text?

3.2 Related Work

One of the earliest computational efforts towards detecting translationese looked for linguistic differences in translated and non-translated children's literature in Finnish (Puurtinen, 2003). It found high frequencies of non-finite constructions, lack of colloquial words, and specific uses of certain conjunctions as markers of translationese. A later effort sought to automatically discriminate between translated and non-translated monolingual corpora, and in the process detected the presence of translationese features within translated text (Baroni and Bernardini, 2006). They reported that the distribution of function words and morphosyntactic categories in general, and personal pronouns and adverbs in particular, were among the cues used by trained Support Vector Machine (SVM) models to perform the discrimination task. Koppel and Ordan (2011) extended this further by not only trying to classify translated and non-translated text, but also text created from different source languages into a common target. They reported that different source *interferences* could be successfully detected in their experiments and referred to them as different *dialects* of translationese.

Perhaps the first study that sought to compare *translationese* and *post-editese* (markers in post-edited text) by studying datasets created under each of those conditions was conducted by Toral (2019). They find PE to be simpler and more normalized and exhibit greater amount of interference from the source language than HT. Bizzoni et al. (2020) primarily focus on two structural metrics to compare translationese across human and machine translations from text and speech. They do not find structural differences displayed by human interpreting replicated in machine translation outputs. Bangalore et al. (2015) compare the role of syntactic variation in translation and post-editing analyzing behavioral data from an existing translators' database. They go on to manually annotate data for features such as, valency of the verb, voice and clause type and detect positive correlations between syntactic variation and translation production time.

Later work has sought to relate these variations and differences to the Machine Translation evaluation task. Popović (2020) looked at the differences amongst text generated by translators with different competencies, experiences and native languages, and called for careful selection of Machine Translation evaluation references given the variations seen by them. A study by Freitag et al. (2020) takes this further and demonstrates a stronger correlation between human judgements and automatic metrics, when the nature of references used is taken into account.

Our current work is in continuation of this later line of research. Our focus is on the three varieties of target text available to us, namely, MT, PE and HT. We are therefore interested in uncovering evidence for these *dialects* of the target: machine-translationese, post-editese and translationese. We also propose some new metrics to uncover these translation markers. We then go on to assess translator level variations for all our proposed metrics and finally, in a later chapter (Chapter 4), show the impact of our varied targets on a machine translation evaluation task.

Participants	PE-segments	HT-segments
P02	100.0	100.0
P03	100.0	100.0
P04	100.0	100.0
P05	100.0	98.0
P06	100.0	100.0
P07	100.0	100.0
P08	100.0	100.0
P09	100.0	100.0
P10	100.0	95.0
ALL	900.0	893.0

Table 3.1: Data Statistics

3.3 Experimental Setup

3.3.1 Data

Parallel Data

Our original dataset consisted of 200 unique English segments translated in either of the two conditions (PE or HT) by 10 professional translators. Thus for each English source segment, we had either a PE or an HT segment from each translator, originally yielding a total of 1000 PE and 1000 HT segments.

However, as stated earlier, upon closer inspection, we found issues with the translations generated by user P01. We therefore discarded all data points belonging to this user (200 segments). Further to this, we also came across a few data points (7 segments) where the translation field was empty (the user possibly had navigated ahead in the workbench without translating the segments), we discarded these too. We were thus left with a total of 1793 data points as seen in Table 3.1 on which we report all our experiments.

Monolingual Data

In order to compare some of the linguistic indicators calculated for parallel data against baseline values, we utilize monolingual annotated data from the Universal Dependencies Treebank

Language	Sentence	Token	Word	Lemma	PoS	Feat	Dep	LDep
English EWT	16622	254830	254830	17784	17	34	40	7
Hindi HDT	16647	351704	351704	15586	16	43	26	1

Table 3.2: Monolingual Corpora Statistics (Nivre et al., 2016)

(Nivre et al., 2016). We consider two monolingual corpora sources: UD English EWT[1] and UD Hindi HDTB[2](Bhat et al.; Palmer et al., 2009). The data statistics for both are summarized in Table 3.2.

The English and Hindi models utilized for part-of-speech and dependency annotation on data reported in Table 3.1 were also built from these sources (Qi et al., 2020). However, note that these are not comparable corpora and thus not utilized for any inferences regarding translationese or post-editese markers in our work. They are also quite small to draw any major statistical conclusions about the nature or properties of the two languages involved.

3.3.2 Linguistic Indicators as Metrics

Of the six translation universals hypothesized by Baker (1993), and following in part Toral (2019), we constrain our study to four[3]: *Explication (1)*, *Simplification (2), Normalization (3), and Interference (6)*. We eschew *Reduced Repetition (4)* and *Exaggeration* (5) at this time; the former, because owing to the smaller size and nature of our corpus, we are unlikely to find many instances of source language lexical or phrasal repetitions to test for against the target; and the latter, because detecting *Exaggeration* first calls for carefully defining them in terms of source and target linguistic properties (and unlike other translation universals, exaggeration is not as easily defined), which is currently beyond the scope of the current study.

We map each of the four Translation Universals that we propose to study with one or more Linguistic Indicators. These indicators are defined as linguistic features or properties of the text under analysis that are quantifiable and interpretable as a metric, and are meant to bring out key differences between the text. Table 3.3 lists this mapping.

We now go on and define each indicator in detail.

3.3.2.1 Lexical Richness

This metric is used to gauge the lexical variety present in the text and is typically calculated as the type-token ratio (TTR) of the text. Previous work in this area has found PE text

[1]https://universaldependencies.org/treebanks/en_ewt/index.html#udenglishewt

[2]https://universaldependencies.org/treebanks/hi_hdtb/index.html#udhindihdtb

[3]Toral (2019) did not consider Explication in their study

[4]We interpret the *Normalization* universal slightly differently and take it to be normalization by source

Translation Universal	Linguistic Indicator
(1) Simplification	(a) Lexical Richness (b) Lexical Density
(2) Explication	(c) Pronoun Density
(3) Normalization[4]	(d) Length Ratio
(4) Exaggeration	-
(5) Reduced Repetition	-
(6) Interference	(e) Part-of-speech Sequences (f) Dependency Minimization

Table 3.3: Mapping of Translation Universals with Linguistic Indicators.

to exhibit lower lexical variety than HT (Toral, 2019; Farrell, 2018). A comparatively lower lexical variety may point to the *Simplification* universal at work. However, TTR is known to be susceptible to variations in text length, we therefore, employ the Moving-Average Type-Token Ratio (MATTR) variant of this metric to control for the slightly differing lengths of our SRC, MT, PE and HT corpora (Covington and McFall, 2010). Note that lexical richness is also used as an indicator of morphological complexity when making language typology comparisons since Agglutinative languages tend to show some of the highest TTRs in cross-lingual analyses (Kettunen, 2014; Bentz et al., 2016). In our case, since the proposed comparisons are across three 'varieties' of the same target language corpus, chances of lexical richness scores being confounded with morphological complexity are controlled.

The MATTR formula used to measure Lexical Richness in our analysis is as follows:

$$MATTR_{Corpus} = \frac{1}{n}\sum_{i=1}^{n}\frac{num\_types(c_i)}{num\_tokens(c_i)} \tag{3.1}$$

where:

c = a 500 token slice of the Corpus

$i = i^{th}$ window of token size 500

n = number of moving windows, given by $(num\_tokens(Corpus) - 500)$

3.3.2.2 Lexical Density

Lexical density seeks to quantify the amount of information present in the text. This is done by calculating the ratio of content words (adverbs, adjectives, nouns and verbs) against the total number of words in a text.

Translated text have been found to be simpler than those created originally in the target language, exhibiting lower lexical density, and taken as evidence in support of the *Simplification* universal (Toral, 2019).

Conversely, we may also argue that given two monolingual comparable texts A and B, a higher lexical density score for A compared to B may also be indicative of the presence of higher *Explication* within the text.

Through the metric defined below we aim to test both hypothesis. We present a simplified version of the metric utilized. To control for slightly varying corpus lengths we adapted the below metric and implemented a moving average variant with a window size of 500 tokens.

$$lexical\ density = \frac{number\ of\ content\ words}{number\ of\ total\ words} \tag{3.2}$$

In order to compute lexical density, we annotate our text using the Stanza toolkit (Qi et al., 2020) employing Universal Dependency (UD) models trained on the English and Hindi UD treebanks[5]. The tagset (upos)[6] being universal, allows us to make comparisons across languages utilizing this annotated data.

3.3.2.3 Pronoun Density

We propose a new metric of pronoun density to detect evidence of *Explication*. Our hypothesis as per the *Explication* universal being that since translators tend to make background information more explicit in target text, one of the ways in which they accomplish this is by resolving ambiguous references by substituting them with their referent. As we define pronoun density to be the ratio of pronouns to nouns in the text, a reducing density trend may point to the dropping of pronouns and addition of more nouns in the target varieties.

$$pronoun\ density = \frac{number\ of\ pronouns}{number\ of\ nouns} \tag{3.3}$$

[5]https://stanfordnlp.github.io/stanza/available_models.html

[6]https://stanfordnlp.github.io/stanza/pos.html

3.3.2.4 Length Ratio

For a source text (ST) and its corresponding target text (TT) of any of the forms MT, PE, or HT, we compute the absolute difference in lengths (measured in characters) between ST and TT, and normalize it by the length of ST. This is done at the level of each sentential unit and a mean over the entire text is computed. The hypothesis being that target sentence lengths reflect *Normalization* driven by the source and track closely with source lengths. Therefore, as per our definition above, greater the *Normalization* universal at work, lower the length ratio.

$$length\ ratio = \frac{|length(ST)\ -\ length(TT)|}{length(ST)} \tag{3.4}$$

3.3.2.5 Part-of-speech Sequences

We move on to the *Interference* universal. Source and Target obviously differ in their syntactic structures. Typologically, the English-Hindi language pair is said to exhibit SVO and SOV word orders, respectively. How might we account for source interference if any?

One way to do this would be to estimate typical part-of-speech (POS) sequences in both source and target languages. We take these estimates to be the part-of-speech Language Models (LM) for source and target (LM_{source}, LM_{target}), respectively. Given two target samples, say A and B, if the POS sequences of A show more similarity with LM_{source} than the sequences of B measured against LM_{source}, we may conclude that A shows more source interference.

The proposed measure for this similarity test is based on *perplexity* (PP). We define the overall formula as follows:

$$PP\_diff = PP(T,\ LM_{source}) - PP(T,\ LM_{target}) \tag{3.5}$$

A high result for the above equation would indicate that a sample T is dissimilar to the source language and similar to the target. A lower result would indicate the opposite. Our hypothesis following (Toral, 2019) is that MT and PE text will show higher interference from source (will be similar to source) than HT text.

We build our part-of-speech language models on the universal (upos) part-of-speech sequences of the English and Hindi UD Treebank corpora introduced previously. We expect the corpus volumes to be sufficient given that our part-of-speech-vocabulary size is quite small (17 unique tags for English and 16 for Hindi – see Table 3.2). We utilize the SRILM toolkit and build our models with an ngram order of 5 and utilize Witten-Bell discounting (Stolcke, 2002).

3.3.2.6 Dependency Distance

Dependency distance, measured as the linear distance between two syntactically related words in a sentence, is said to be indicative of syntactic difficulty and memory burden. Various

large scale corpora analyses indicate that in order to reduce this memory burden, there might exist a universal preference for dependency distance minimization in human languages (Liu et al., 2017).

For example, for sentences in figures 3.1, 3.2, 3.3 and 3.4, we calculate the total dependency distance for each sentence as the sum of the linear distances between their inter-dependent nodes. As per (Futrell et al., 2015), for Sentences A and B either word order is deemed acceptable and the sentences are semantically equivalent. However, when comparing sentences C and D, again although semantically equivalent, English speakers find C to be more natural than D. This, as per the hypothesis, shows the preference for dependency minimization at work.

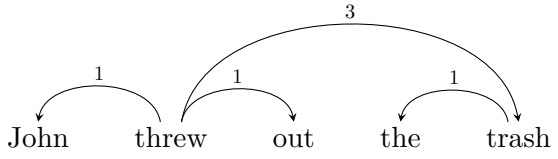

Figure 3.1: Sentence A. Total dependency length = 6. Example from Futrell et al. (2015)

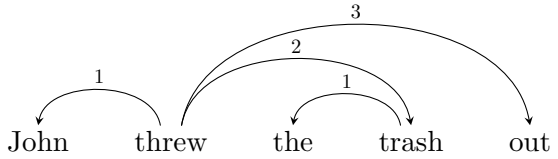

Figure 3.2: Sentence B. Total dependency length = 7. Example from Futrell et al. (2015)

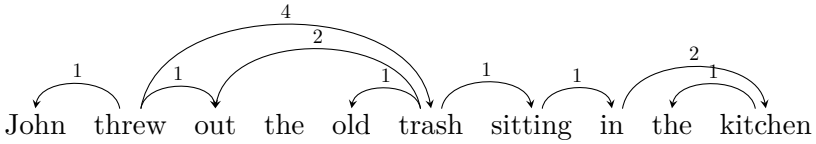

Figure 3.3: Sentence C. Total dependency length = 14. Example from Futrell et al. (2015)

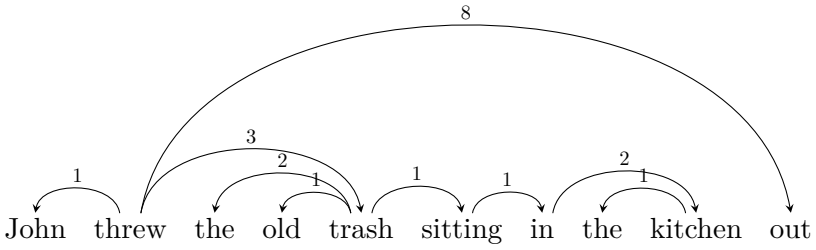

Figure 3.4: Sentence D. Total dependency length = 20. Example from Futrell et al. (2015)

It has also been argued that compared to artificially generated languages, human languages show shorter mean dependency distances (Futrell et al., 2015).

We thus combine the two insights gleaned from previous work in this direction and extend the hypothesis. We argue that as we go from MT to PE, and then on to HT corpora, we will see a reduction in dependency distance.

The proposed metric that we call mean dependency distance (MDD) adapted from (Liu et al., 2017) and applied to our target varieties is as follows:

$$MDD(sentence) = \frac{1}{n-1} \sum_{n=1}^{n} |DD_i| \tag{3.6}$$

where:

n = the number of words in the sentence

DD_i = dependency distance of the ith syntactic link of the sentence

To control for sentence length variations, we normalize the obtained dependency distance over the word length of each sentence and report the normalized corpus mean.

3.4 Results

3.4.1 Lexical Richness

Let us now look at the results for each indicator. Consider Table 3.4. The first three columns (*srcmt*, *srcpe*, *srcht*) show the MATTR scores calculated per participant and the means over all participants for the source text. The src prefix signifies that this text was the input used to generate the corresponding target text. Similarly, the next three columns (*mt*, *pe*, *ht*) show the MATTR scores for the target variants, raw MT, post-edited and human-translated, respectively. The last two columns (*mt_prcnt*, *pe_prcnt*) show the percentage difference for MT and PE as measured against the human (HT) baseline.

	Source			Target			% diff with HT	
	srcmt	srcpe	srcht	mt	pe	ht	mt_prcnt	pe_prcnt
P02	0.5362	0.5362	0.5479	0.5016	0.5132	0.5583	-10.16	-8.09
P03	0.5362	0.5362	0.5479	0.5016	0.5193	0.5369	-6.58	-3.28
P04	0.5479	0.5479	0.5362	0.5283	0.5338	0.5316	-0.63	0.41
P05	0.5479	0.5479	0.5363	0.5283	0.5218	0.5219	1.21	-0.04
P06	0.5479	0.5479	0.5362	0.5283	0.5343	0.5448	-3.03	-1.93
P07	0.5362	0.5362	0.5479	0.5016	0.5219	0.5220	-3.92	-0.02
P08	0.5362	0.5362	0.5479	0.5016	0.5141	0.5161	-2.81	-0.39
P09	0.5479	0.5479	0.5362	0.5283	0.5279	0.5119	3.19	3.11
P10	0.5362	0.5362	0.5494	0.5016	0.5098	0.5231	-4.12	-2.55
sd	0.0062	0.0062	0.0064	0.0141	0.0088	0.0149	4.00	3.13
MEAN	0.5414	0.5414	0.5429	*0.5134	0.5218	0.5296	-2.98	-1.42

Table 3.4: Lexical Richness: Comparisons between mt, pe, and ht text. The mt text tests to be significantly *less* richer than ht for an α level of 0.05 on a one-tailed t-test.

Analyzing the results, we find both MT and PE to be lexically degraded compared to HT by 2.98% and 1.42% points, respectively. However, only the comparison between MT and HT per participant means tests to be statistically significant. These results point to a limitation of the MT paradigm that appears to generate lexically poorer outputs compared to human translations. This has previously also been noted by Vanmassenhove et al. (2021). Comparing PE and HT, we do note a difference, but it appears that the post-editing process somewhat makes up for the lexical poverty introduced by MT. However, it does not catch up to HT. We therefore detect the *Simplification* translation universal at work in these text.

3.4.2 Lexical Density

Recall that we had introduced the Lexical Density metric in support of both the *Simplification* and *Explication* universals. Investigating Lexical Density scores, we note lower scores across MT and PE text when compared with HT. Interestingly, while we find PE to be 1.30 points less denser than HT, MT comes out to be only 0.47 points less denser. While the lower numbers for both fit the *Simplification* universal, they do not test to be statistically significant. In relative percentage difference terms our results are similar to those reported by Toral (2019). We do not detect any support for the *Explication* universal for this metric.

	Source			Target			% diff with HT	
	srcmt	srcpe	srcht	mt	pe	ht	mt_prcnt	pe_prcnt
P02	0.5016	0.5016	0.5097	0.4762	0.4756	0.4975	-4.27	-4.40
P03	0.5016	0.5016	0.5097	0.4762	0.4843	0.4960	-3.99	-2.35
P04	0.5097	0.5097	0.5016	0.4888	0.4829	0.4798	1.87	0.64
P05	0.5097	0.5097	0.5016	0.4888	0.4655	0.4597	6.33	1.25
P06	0.5097	0.5097	0.5016	0.4888	0.4906	0.4949	-1.24	-0.87
P07	0.5016	0.5016	0.5097	0.4762	0.4711	0.4798	-0.74	-1.82
P08	0.5016	0.5016	0.5097	0.4762	0.4747	0.4833	-1.47	-1.78
P09	0.5097	0.5097	0.5016	0.4888	0.4851	0.4789	2.08	1.31
P10	0.5016	0.5016	0.5072	0.4762	0.4717	0.4899	-2.79	-3.70
sd	0.0043	0.0043	0.0041	0.0066	0.0082	0.0119	3.39	2.07
MEAN	0.5052	0.5052	0.5058	0.4818	0.4779	0.4844	-0.47	-1.30

Table 3.5: Lexical Density: Comparisons between mt, pe, and ht text. The mt and pe text do not test to be significantly *less* denser than ht for an α level of 0.05 on a one-tailed t-test.

3.4.3 Pronoun Density

Recall our definition of *Explication* as the addition of more background information in the translated text. As apparent from Table 3.6, we see a pronounced effect for explication in terms of pronoun density score. The machine translated MT text comes out to be *less* denser in terms of the pronouns than the PE and HT text.

What might this indicate? Does machine translation have a tendency to drop pronouns in the source text, a frequent error it might be prone to (Hardmeier and Guillou, 2018)? Or are human translators in PE (to a lesser extent) and HT (to a greater extent) processes adding more background information, primarily made up of nouns? A closer analysis of pronoun and noun frequencies in each of these corpora reveal, that while MT does not add any additional pronouns, it does generate more nouns compared to the source text (Appendix A, Tables A.1, A.2). The human-mediated processes on the other hand (both PE and HT), add more pronouns and nouns in the generated target than the source. For now, we note a significant difference in pronoun densities between MT and HT text, but conclude that the lower pronoun density in case of MT is due to additional nouns added during generation. As to whether this 'extra information' in case of raw MT is meaningful or not remains to be seen.

	Source			Target			% diff with HT	
	srcmt	srcpe	srcht	mt	pe	ht	mt_prcnt	pe_prcnt
P02	0.1415	0.1415	0.1357	0.1358	0.1483	0.1522	-10.77	-2.60
P03	0.1415	0.1415	0.1357	0.1358	0.1489	0.1530	-11.19	-2.63
P04	0.1357	0.1357	0.1415	0.1251	0.1465	0.1784	-29.85	-17.88
P05	0.1357	0.1357	0.1410	0.1251	0.1732	0.1854	-32.50	-6.60
P06	0.1357	0.1357	0.1415	0.1251	0.1400	0.1769	-29.28	-20.86
P07	0.1415	0.1415	0.1357	0.1358	0.1686	0.1546	-12.13	9.07
P08	0.1415	0.1415	0.1357	0.1358	0.1652	0.1488	-8.74	11.02
P09	0.1357	0.1357	0.1415	0.1251	0.1249	0.1663	-24.75	-24.92
P10	0.1415	0.1415	0.1383	0.1358	0.1566	0.1451	-6.42	7.91
sd	0.0031	0.0031	0.0028	0.0056	0.0152	0.0148	10.46	13.46
MEAN	0.1390	0.1390	0.1385	*0.1311	0.1525	0.1623	-18.40	-5.28

Table 3.6: Pronoun Density: Comparisons between mt, pe, and ht text. The mt text tests significantly *less* denser than ht for an α level of 0.05 on a one-tailed t-test.

3.4.4 Length Ratio

We now move on the *Normalization* universal and its linguistic indicator, length ratio.

Table 3.7 shows the ratio (absolute difference between source and target length normalized by source length per sentence) to be highest for the *src-mt* comparison, before dropping for *src-pe* and increasing again for *src-ht*. Recall that a lower value indicates more normalization driven by the source length. This is a surprising finding as it appears that source length is a greater normalizing factor for PE and HT text than it is for raw MT.

This is also in contradiction to earlier reported results for this metric (Toral, 2019). However, there is a key difference between our study and the earlier work, where the MT engines used to calculate this metric were SMT and RBMT engines. In fact, Toral (2019) argued that this metric is necessarily an effect of those MT paradigms (SMT, RBMT) which are limited to producing outputs constrained by source length, thus they do not take the NMT paradigm into account in their work. Our findings, validate the intuition proferred in their study.

We therefore conclude that given an MT output (from a paradigm unconstrained by source length) and a Source, post-editors produce a text (PE) which shows greater normalization by source length, than translators who produce a text (HT) only given a Source.

	Source to Target(s) Ratios			% diff with HT	
	src-mt	src-pe	src-ht	mt_prcnt	pe_prcnt
P02	0.1218	0.0974	0.1058	15.05	-7.99
P03	0.1218	0.0980	0.1423	-14.44	-31.12
P04	0.1414	0.1330	0.1670	-15.30	-20.38
P05	0.1414	0.1236	0.1217	16.22	1.58
P06	0.1414	0.1193	0.1085	30.37	9.96
P07	0.1218	0.1129	0.1095	11.18	3.10
P08	0.1218	0.1031	0.1122	8.52	-8.11
P09	0.1414	0.1145	0.1358	4.12	-15.73
P10	0.1218	0.0957	0.1220	-0.17	-21.52
sd	0.0104	0.0131	0.0201	14.69	13.38
MEAN	0.1305	*0.1108	0.1250	6.17	-10.02

Table 3.7: Length Ratio: Comparisons for src-mt, src-pe, src-ht absolute differences normalized by src length. src-pe length ratio tests to be significantly *more* normalized than src-ht for an α level of 0.05 on a one-tailed t-test.

3.4.5 Part-of-speech Sequences

We now move on to the results of the *Interference* universal. The part-of-speech sequences metric seeks to gain insights on syntactic interference from the source language into the target.

In Table 3.8, we observe lowest perplexity difference from MT, followed by PE, and then HT, implying a reduction in source syntactic interference as we proceed from MT to PE and HT. Recall, that a low score implies similarity to the source language model (LM_{source}).

Our hypothesis that expected source interference in the MT and PE conditions appears to be validated. However, the differences do not test to be statistically significant. This again may point to the greater success of the NMT paradigm in achieving better reordering and hence improved target language fluency.

3.4.6 Dependency Distance

We finally, compare the dependency lengths of our target varieties. Based on Table 3.9, we can observe differences of 2.61% and 3.27% when MT and PE outputs are compared to HT. The hypothesis, that human languages tend to minimize dependency distances compared to artificially generated languages, appears to be valid when comparing MT with HT, and PE with HT. In both cases, the mean dependency distances are reduced.

	Source			Target			% diff with HT	
	srcmt	srcpe	srcht	mt	pe	ht	mt_prcnt	pe_prcnt
P02	-14.4920	-14.4920	-11.1863	8.4535	8.3461	7.9383	6.49	5.14
P03	-14.4920	-14.4920	-11.1863	8.4535	8.5507	8.5355	-0.96	0.18
P04	-11.1863	-11.1863	-14.4920	8.8167	8.4917	9.5492	-7.67	-11.07
P05	-11.1863	-11.1863	-14.3609	8.8167	9.8982	10.1204	-12.88	-2.19
P06	-11.1863	-11.1863	-14.4920	8.8167	8.9956	9.5138	-7.33	-5.45
P07	-14.4920	-14.4920	-11.1863	8.4535	8.9224	8.7024	-2.86	2.53
P08	-14.4920	-14.4920	-11.1863	8.4535	8.6301	8.3638	1.07	3.18
P09	-11.1863	-11.1863	-14.4920	8.8167	8.9109	9.1402	-3.54	-2.51
P10	-14.4920	-14.4920	-11.3606	8.4535	8.8790	8.4177	0.43	5.48
sd	1.7423	1.7423	1.7080	0.1914	0.4541	0.7034	5.71	5.41
MEAN	-13.0228	-13.0228	-12.6603	8.6149	8.8472	8.9201	-3.03	-0.52

Table 3.8: Part-of-speech Sequences show *Interference* from the source, the highest for MT, followed by PE.

We infer from these results, that greater the artificial scrambling (reordering) of a target text without constrains of memory burden, greater the dependency distance. The post-edited text also carries the effect of this artificial scrambling. It is only the HT process that appears to minimize this distance. However, some caution is advised as to the interpretability of these results which may be colored by the reported error rates of the parsers involed. Qi et al. (2020) report a *UAS* accuracy of **86.22** on the EWT treebank. They do not report Stanza accuracies for the HDT treebank model.

3.4.7 Translator Variation

We reported all our results per participant and as the overall means of all the participants that make up our data. Readers may have noticed the standard deviations of the participant scores as well. In Table 3.10, we extract and present the mean and standard deviations of the participants along each metric. We do this only for the PE and HT corpora, as they are the only text affected by human intervention, be it post-editing or translation.

We may infer from these results that the output produced by the participants appears to be fairly uniform. No large standard deviation values are noticeable. Our findings are different from those reported by Popović (2020), who found significant differences between text features (linguistic indicators) across various groups of translators, namely, crowd, professional, students, specialist etc. They also noticed differences when taking the translators native language into

	Source			Target			% diff with HT	
	srcmt	srcpe	srcht	mt	pe	ht	mt_prcnt	pe_prcnt
P02	3.1906	3.1906	3.1530	3.3695	3.3651	3.2663	3.16	3.03
P03	3.1906	3.1906	3.1530	3.3695	3.2745	3.1854	5.78	2.80
P04	3.1530	3.1530	3.1906	3.4317	3.5175	3.2552	5.42	8.06
P05	3.1530	3.1530	3.1963	3.4317	3.5025	3.4330	-0.04	2.03
P06	3.1530	3.1530	3.1906	3.4317	3.5290	3.1425	9.20	12.30
P07	3.1906	3.1906	3.1530	3.3695	3.3928	3.4247	-1.61	-0.93
P08	3.1906	3.1906	3.1530	3.3695	3.2482	3.3774	-0.24	-3.83
P09	3.1530	3.1530	3.1906	3.4317	3.5225	3.3747	1.69	4.38
P10	3.1906	3.1906	3.1685	3.3695	3.4206	3.3660	0.10	1.62
sd	0.0198	0.0198	0.0196	0.0328	0.1079	0.1052	3.57	4.72
MEAN	3.1739	3.1739	3.1720	*3.3971	*3.4192	3.3139	2.61	3.27

Table 3.9: Dependency Distance: Comparisons between mt, pe, and ht text. The mt and pe text tests show significantly *greater* mean dependency distances than ht for an α level of 0.05 on a one-tailed t-test.

account along with the translation direction and advised caution on the use of such data for tasks like MT evaluation.

In our case, the more uniform targets may be explained by the similar competencies of the participants in our study. Almost all shared the same native language they were translating into, and had some prior experience with post-editing tasks. Self-reported answers by the participants relating to HT and PE familiarity may be seen in Figure 3.5 with 30% reporting 0–2 years of experience in translating for this language direction and 70% reporting 2–5 years of experience. Similarly, for experience with post-editing machine translation output, 30% report some experience and 70% report a lot of experience.

3.4.8 Correlations with Productivity Metrics

Recall that in Chapter 2 we calculated post-editing effort indicators on the *Temporal, Technical* and *Cognitive* dimensions. How might our linguistic indicators correlate in these effort dimensions? Two of our linguistic indicators (*Length Ratio, Dependency Distance*) can be calculated at the sentence level, while the remaining are only applicable to a corpus. In Table 3.11 we present the correlations calculated separately for PE and HT sentences against the post-editing effort metrics.

Metric	pe_mean	pe_sd	ht_mean	ht_sd
Lexical Richness	0.5218	0.0088	0.5296	0.0149
Lexical Density	0.4779	0.0082	0.4844	0.0119
Pronoun Density	0.1525	0.0152	0.1623	0.0148
Length Ratio	0.1108	0.0131	0.1250	0.0201
POS Sequence	8.8472	0.4541	8.9201	0.7034
Dependency Distance	3.4192	0.1079	3.3139	0.1052

Table 3.10: Translators' mean and standard deviations on PE and HT text.

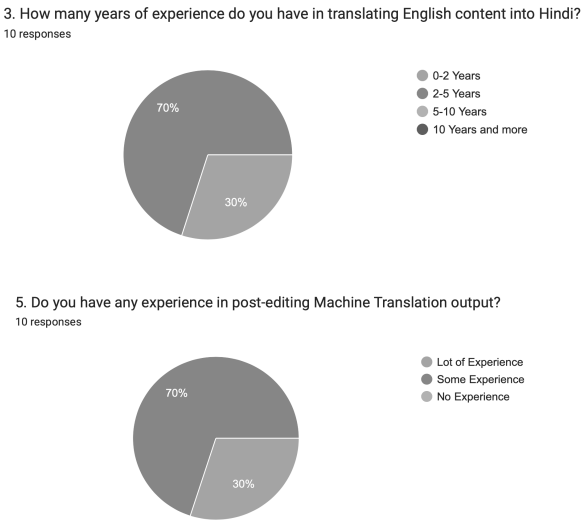

Figure 3.5: Self-reported experience by participants with HT and PE

	PE			HT		
Linguistic Indicator	temporal	technical	cognitive	temporal	technical	cognitive
Length Ratio	-0.12	0.09	-0.08	0.18	0.04	0.23
Dependency Distance	0.39	0.04	0.29	0.38	0.01	0.29

Table 3.11: Linguistic indicator correlations with Effort indicators. Pearson coefficient r reported for r(900) for PE and r(893) for HT.

Translation Universal	Linguistic Indicator	mt	pe	ht
Simplification	Lexical Richness	*0.5134	0.5218	0.5296
Simplification	Lexical Density	0.4818	0.4779	0.4844
Explication	Pronoun Density	*0.1311	0.1525	0.1623
Normalization	Length Ratio	0.1305	*0.1108	0.1250
Interference	POS Sequence	8.6149	8.8472	8.9201
Interference	Dependency Distance	*3.3971	*3.4192	3.3139

Table 3.12: Linguistic Indicator Results Summarized. * denotes tested statistically significant in comparison to HT.

Remember, that a low Length Ratio implied greater normalization by source which had earlier tested significant for the PE corpus. Here we do not find this correlated with any of the effort dimensions on post-edited sentences. Hence, we may infer that normalization by source is not a reliable indicator of post-editing effort, perhaps due to interference by the MT output. However, in the case of human-translations, weak correlations on the temporal and cognitive dimensions can be detected but in the opposite direction. For Dependency Distance, we detect a moderate positive correlation on the temporal scale across the board, implying that the time taken to create a translation in either of the conditions is related to the increasing dependency distance of the resulting target. These figures also point to an increasing cognitive overload incurred on more structurally complex sentences, regardless of whether they are translated in the PE or the HT condition. But we do not detect any discernible differences between the PE and HT slices of our corpus. These analyses appear to suggest that fine-grained target differences that we saw in this chapter are hard to infer solely based on post-editing effort metrics. Thus the quality and nature of translations produced must be validated independently.

3.5 Conclusion

In Table 3.12 we summarize our results. Through the six linguistic indicators, we have seen evidence of translation universals at work to varying degrees in the three text 'varieties' we analyzed. We conducted an experiment assessing target text created under three conditions: as raw machine translated output (MT), as post-edited text (PE), and as human translated text (HT). We defined linguistic indicators and mapped them to hypothesized translation universals. We reported the results per participant and text type for these indicators. Our results found evidence for the *Simplification*, *Explication*, *Normalization*, and *Interference* universals. While most differences tested to be statistically significant, others were merely indicative of trends. On the whole, we found post-edited text to be simpler, less diverse and more normalized than human translated text. We also found them to exhibit source interference. The primary reason for this we suspect is the priming of post-edited text by machine translation output as previously also noted by Carl and Schaeffer (2017). While this priming does increase productivity and reduces effort as we saw in the previous chapter (Chapter 2), our results in this chapter seem to indicate that it comes at the cost of greater source interference and lexically degraded target. The poorer machine translation text on the same parameters may be ascribed to the algorithmic bias of the MT paradigm itself (Vanmassenhove et al., 2021).

We also proposed and applied two novel linguistic indicators (Pronoun Density and Dependency Distance) to this task and found both to yield statistically significant results. Finally, we analysed the differences on these metrics amongst translators that participated in our study and noted very low deviations in the text produced by them.

To summarize, in terms of the research questions we floated in our Introduction (Section 3.1), we found MT, PE and HT text to exhibit different characteristics on most of our metrics; we did not find major differences in the text produced by different translators; and we found PE text to be more similar to MT than HT text in most cases.

A future extension of this work proposes to tackle the two translation universals (*Repetition* and *Exaggeration*) that we left out for want of data size. We also propose to take a closer look at some of the linguistic indicators like dependency distance by combining them with phrase (chunk) level syntax information. The translator text differences also require a deeper examination.

In this chapter (Chapter 3) we saw some marked differences between text created with the mediation of machine translation (PE) and those created unaided (HT). We also detected evidence of various translation universals at play amongst these text. We now look into utilizing the text produced for a machine translation (MT) evaluation task.

Chapter 4

Impact of Post-edited Text on MT Evaluation

4.1 Introduction

The MT evaluation landscape is quite vast if a translation workflow setting is assumed. In such a real-world scenario, not only do we need metrics to evaluate the quality of the MT output, but we must (ideally) also be able to quantify each of the following: the nature of the source itself; the quality of MT output in terms of various automatic and human evaluation metrics; the nature of the target produced in this process; and the effort needed to produce acceptable quality target text. Figure 4.1 provides us a glance of the possible metrics and methodologies that may be addressed. The highlighted sections are the ones that we have covered in some form during various sections of this study. In this chapter, we focus on the most crucial part, MT Quality metrics and evaluation methodologies.

Automatic MT evaluation metrics are used quite extensively in both the MT development process as well as in various MT evaluation shared tasks, since human evaluation, owing to high costs, lack of repeatability, subjectivity, and slowness of evaluating machine translation output remains infeasible at most times (Dorr et al.). But the use of automatic metrics has often been fraught with inconsistencies due to low correlations with human judgements (Callison-Burch et al., 2006).

Apart from the shortcomings of the metrics themselves, recent work has also focused on the nature of test sets too, uncovering inherent biases. Toral et al. (2018a) note the effect of directionality of the source, i.e. whether the source was original or created by translation and hence appeared simpler and more normalized, making it easier to translate for an MT engine.

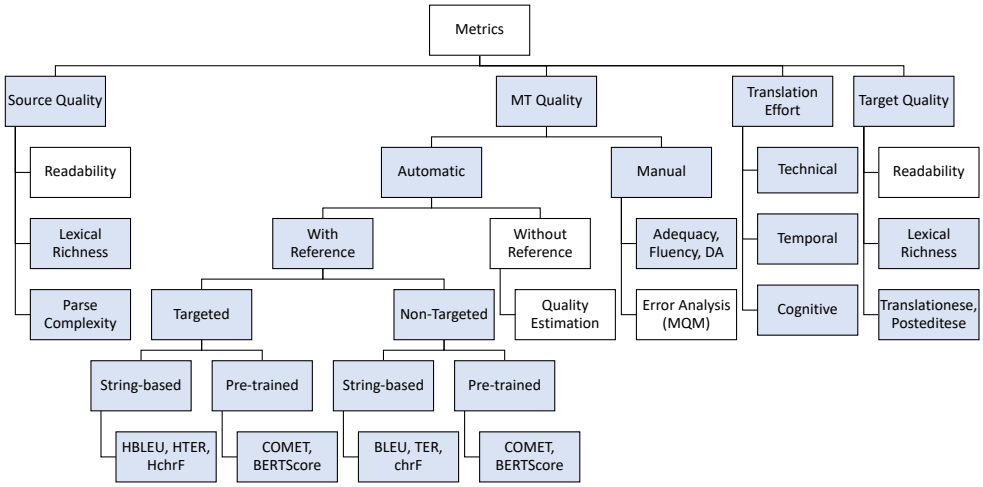

Figure 4.1: Metrics Landscape: Shows possible evaluation metrics that may offer insights in a typical MT workflow setting. The ones highlighted are the metrics we touched upon in various parts of this study.

Freitag et al. (2020) focus instead on the target side of evaluation sets and note the effect of paraphrased references on automatic evaluation.

Our study in this chapter, proposes to conduct experiments along similar lines. We seek to study the impact of references (target text) created under different conditions (PE vs. HT) on an MT evaluation task.

For these set of experiments, the research questions (RQ) we pose are the following:

- RQ1: Are post-edited (PE) references biased towards MT engines they were created from compared to their unaided human translated (HT) references?
- RQ2: Do the number of references used affect MT evaluation outcomes?
- RQ3: Which current MT evaluation metrics fare best under varying reference conditions?

4.2 Related Work

Recent work on reassessing MT evaluation processes has begun to note the importance of test set provenance in attaining fairer and less biased evaluation outcomes. These studies on provenance, as to how the source and target sides of evaluation test sets may be created and in which direction, have uncovered previously hidden biases in the MT evaluation task. Using translated Source (containing implicit translationese features) has been shown to lead to

misleading conclusions and even claims of human parity (Toral et al., 2018a). Zhang and Toral (2019); Graham et al. (2019) also note the effect of translationese on source sentences used as inputs and, crucially, find system rankings to be affected in meta-evaluation tasks.

The importance of using the right kind of references has also been studied. Recently, Kloudová et al. (2021) uncovered bias of post-edited references towards an online MT engine, that those references had been created from, leading to inflated scores for that particular engine in their evaluation task. Freitag et al. (2020) also note the prevalence of translationese in MT evaluation references and propose a unique paraphrasing task to make the references more *natural* and rid them of these biases. They also show improved correlations with multiple MT metrics including embedding-based ones when using such references.

In a later work, their focus turns to the methodologies of generating *gold* ratings, the human evaluation task itself. They propose Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014) based annotations, utilizing professional translators as the recommended rating method to adjudge system rankings (Freitag et al., 2021).

Our work in this study is along the lines of adjudging the impact of target (post-edited and translated) references on the MT evaluation task. We consolidate some of the previous work in this direction in the following ways: (i) we study the effect of both human translated and post-edited references on MT evaluation metrics; (ii) we study the effect of both single and multiple references on MT evaluation metrics; (iii) we compare and rank four different MT engine hypothesis for the same source on these varied references; (iv) we propose recommendations on creation of fairer references and the best MT metrics to utilize for MT evaluation tasks.

4.3 Experimental Setup

4.3.1 Data

The dataset for our study is the one used in Chapters 2 and 3 which consists of a total of 1793 data-points made up of 200 unique source segments. As described previously, 10 professional translators translated half the segments (100) in the PE condition using MT and the remaining (100) in the HT condition unaided. Post cleaning and filtering for issues, corpus statistics are reproduced below as Table 4.1.

It is important to get a sense of the distribution of HT and PE references in our dataset. Since each participant saw a source segment in either of the two conditions (HT or PE), we originally possessed 5 post-edited and 5 human translated segments per source. However, after discarding all of *P01*'s data due to issues previously mentioned, and a few other segments during cleaning and pre-processing, we were left with a variable number of references per source segment. Table 4.2 shows the HT references created from the same source by 4 different translators. Note the lack of an MT proposal in this case. Similarly, Table 4.3 lists the PE references created by

Participants	PE-segments	HT-segments
P02	100.0	100.0
P03	100.0	100.0
P04	100.0	100.0
P05	100.0	98.0
P06	100.0	100.0
P07	100.0	100.0
P08	100.0	100.0
P09	100.0	100.0
P10	100.0	95.0
ALL	900.0	893.0

Table 4.1: Data Statistics

5 different translators from the same MT engine (IIITH MT) proposal. The variations in references may be noted in both cases.

4.3.2 MT Systems

In addition to the in-house NMT system (IIITH MT) previously described in 2.3.3 we also generate MT outputs for system comparisons using the above source data from three publicly available MT systems, namely Google MT, Bing Translator and IndicTrans[1]. In our results we refer to these engines as IIITH, Google, Bing, and IndicTrans, respectively. However, note than none of the references in our study were created using the outputs of these later engines. All post-editing was conducted on the output of the in-house IIITH model only.

The training data and model architecture details of the public systems are unknown, but both (Google[2] and Bing[3]) seem to use some form of hybrid transformer-based models. The Google and Bing translations were obtained on *15 January 2022*.

IndicTrans is a multilingual Transformer model trained on the samanantar dataset[4]. We downloaded the model released on *05 June 2021*. It is trained on a multilingual Indic corpus

[1]https://indicnlp.ai4bharat.org/indic-trans/

[2]https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html

[3]https://www.microsoft.com/en-us/research/blog/microsoft-translator-now-translating-100-languages-and-counting/

[4]https://indicnlp.ai4bharat.org/samanantar/

Participant	Source	MT	HT
P04	Sundarlal Bahuguna: A Himalayan voice for ecological truth and social justice	-	सुन्दरलाल बहुगुणा: इकोलॉजी एवं समाजिक न्याय के महानायक
P05	Sundarlal Bahuguna: A Himalayan voice for ecological truth and social justice	-	सुंदरलाल बहुगुणा : हिमालय के पर्वतों से आने वाली पर्यावरणीय सत्य और सामाजिक न्याय की एक आवाज़
P06	Sundarlal Bahuguna: A Himalayan voice for ecological truth and social justice	-	सुन्दरलाल बहुगुणा: पारिस्थितिक सत्य और सामाजिक न्याय के लिए हिमालयीआवाज़
P09	Sundarlal Bahuguna: A Himalayan voice for ecological truth and social justice	-	सुन्दरलाल बहुगुणा : पारिस्थितिक सत्य और समाजिक न्याय के लिए हिमालय का पुरोधा

Table 4.2: HT References

	Source	MT	PE
P02	Sundarlal Bahuguna: A Himalayan voice for ecological truth and social justice	सुंदरलाल बहुगुणा : पारिस्थितिक सत्य और सामाजिक न्याय के लिए एक हिमालयी आवाज	सुंदरलाल बहुगुणा : पर्यावरनीय सत्य और सामाजिक न्याय के लिए एक आवाज-हिमालय से
P03	Sundarlal Bahuguna: A Himalayan voice for ecological truth and social justice	सुंदरलाल बहुगुणा : पारिस्थितिक सत्य और सामाजिक न्याय के लिए एक हिमालयी आवाज	सुंदरलाल बहुगुणा : पारिस्थितिक यथार्थ और सामाजिक न्याय की एक बुलंद आवाज थे।
P07	Sundarlal Bahuguna: A Himalayan voice for ecological truth and social justice	सुंदरलाल बहुगुणा : पारिस्थितिक सत्य और सामाजिक न्याय के लिए एक हिमालयी आवाज	सुंदरलाल बहुगुणा : पारिस्थितिक-संबंधी सत्य और सामाजिक न्याय के लिए एक हिमालयी स्वर
P08	Sundarlal Bahuguna: A Himalayan voice for ecological truth and social justice	सुंदरलाल बहुगुणा : पारिस्थितिक सत्य और सामाजिक न्याय के लिए एक हिमालयी आवाज	सुंदरलाल बहुगुणा : हिमालय से पारिस्थितिकीय सत्य और सामाजिक न्याय के लिए उठने वाली एक आवाज।
P010	Sundarlal Bahuguna: A Himalayan voice for ecological truth and social justice	सुंदरलाल बहुगुणा : पारिस्थितिक सत्य और सामाजिक न्याय के लिए एक हिमालयी आवाज	सुंदरलाल बहुगुणा : एक हिमालयी आवाज है पारिस्थितिक सत्य और सामाजिक न्याय के लिए।

Table 4.3: PE References

of $49.7M$ sentence pairs with a reported BLEU score of 38.6 on the WAT 2021 dataset. In comparison, our in-house system obtained a BLEU score of 35.46 on the same test data.

We also use a subsequent iteration of the IIITH engine that we call IIITH_v2 that was trained with some additional data. This engine was not used in the human evaluation ratings, thus makes an appearance only when we compare all the engines on MT evaluation metrics.

4.3.3 Pre-processing

References We process our original dataset to attach the following references to each participant-source combination, which is unique and sums up to a total of 1793 data points. This creates three distinct test sets only varying on the references used:

- slfref: denotes a single reference created by the translator themselves for that source, either in the ht or pe condition.
- perefs: denotes multiple pe references created by all other translators for that source, excluding the current translator in the pe condition.
- htrefs: denotes multiple ht references created by all other translators for that source, excluding the current translator in the ht condition.

Note that the number of the multiple references is variable owing to cleaning of problematic data points as discussed previously.

Views In addition to this, we calculate our scores separately on three *views* of the data:

- all: all source segments are considered regardless of the condition in which they were translated.
- pe: only those source segments that were translated in the pe condition are considered.
- ht: only those source segments that were translated in the ht condition are considered.

This segmentation should help us detect the effect of MT output bias (if any) with respect to PE and HT references, with the former, we hypothesize, leading to inflated scores.

4.3.4 Automatic Evaluation Metrics

We consider two classes of automatic metrics, string-based and pre-trained. The string-based metrics rely on the surface matching of lexical items in the MT hypothesis against a given reference(s), while the pre-trained metrics match these in either a monolingual or multilingual embedding space. They are thus said to address one of the vulnerabilities of string-based approaches, which is lack of soft-similarity or synonymous matches, owing to their surface

Metric	Type	Input Segments	Multi References	Variable References
Corpus BLEU	string-based	hypothesis, reference	Yes	Yes
Sentence BLEU	string-based	hypothesis, reference	Yes	Yes
TER	string-based	hypothesis, reference	Yes	Yes
CHRF	string-based	hypothesis, reference	Yes	Yes
BERTScore	pre-trained	hypothesis, reference	Yes	Yes
BLEURT[5]	pre-trained	hypothesis, reference	No	No
COMET[6]	pre-trained	source, hypothesis, reference	No	No

Table 4.4: Metric Features

nature. On the other hand, string-based metrics carry the advantage of general applicability across all languages since they do not rely on high quality pre-trained embeddings unlike the former. Table 4.4 compares the features of the metrics used in our experiments.

4.3.4.1 String-based Metrics

BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is a corpus level metric and has been the most widely used metric for MT evaluation. The BLEU score of a system output is calculated by counting the number of n-grams in the system output that occur in the set of reference translations. It is a precision-oriented metric since it measures how many of the n-grams of the system output are correct, rather than measuring whether the reference n-grams are fully reproduced in the system output. The following equation shows how the metric is computed.

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n log p_n) \tag{4.1}$$

where, p_n denotes the geometric average of the modified n-gram precisions, using n-grams up to length N and positive weights w_n summing to one. There is also a brevity penalty (BP) which penalizes outputs shorter than the references.

In our experiments, we use the SacreBLEU implementation of this metric (Post, 2018) and refer to it as cBLEU. Although, typically BLEU is calculated at the corpus level, we also use an implementation of sentence level BLEU in order to compare scores with all the other sentence level metrics in our study. We refer to this as BLEU and again use the SacreBLEu implementation.

[5]does not support multiple references natively

[6]does not support multiple references natively

TER (Translation Edit Rate) (Snover et al., 2006) is a sentence level metric that quantifies the edit operations needed on an MT hypothesis to turn it into a given reference. These operations are estimated at the word/token level and the edits include insertion, deletion and substitution of single words, as well as shifts of word sequences. This is formally defined as:

$$TER = \frac{\#\_of\_edits}{average\_\#\_of\_reference\_words} \tag{4.2}$$

TER in its Human-targeted form (HTER) has previously been shown to be useful in predicting the post-editing effort as well (Specia and Farzindar, 2010). The key difference between TER and HTER being that in the latter the reference utilized for comparison is created from the MT proposal itself. In case of multiple references the best score is returned.

In our experiments, we use the SacreBLEU implementation of this metric. It is the only metric in our experiments where a *lower* score implies a better MT output.

CHRF (character n-gram F-score) Popović (2015) is a sentence level metric that uses character n-grams as its countable units. The equation for calculating CHRF is given as:

$$CHRF\beta = (1 + \beta^2)\frac{CHRP \cdot CHRR}{\beta^2 \cdot CHRP + CHRR} \tag{4.3}$$

where CHRP and CHRR stand for character n-gram precision and recall arithmetically averaged over all n-grams, and β is a parameter which assigns β times more importance to recall than to precision. We utilize the recommended β value of 2 (Popović, 2016), and as with other string-based metrics we use the SacreBLEU implementation. CHRF has been shown to perform better for morphologically complex languages and indeed in Chapter 2, this was the metric with the best correlations in our human evaluation task.

4.3.4.2 Pre-trained Metrics

BERTScore Zhang et al. (2019) proposed BERTScore as an automatic evaluation metric for text generation. It computes a soft similarity score for each token in the candidate sentence with each token in the reference sentence based on contextual embeddings. On WMT evaluation sets it has been shown to better correlate with human judgements than popular string-based evaluation methods like BLEU.

Figure 4.2 illustrates the computation of RBERT - a recall based variant of the metric. The underlying contextual embeddings model is BERT (Devlin et al., 2019). The authors also claim that BERTScore is more robust to challenging examples compared to existing metrics.

In our experiments we report the F1 score variant for this metric as recommended by the authors.

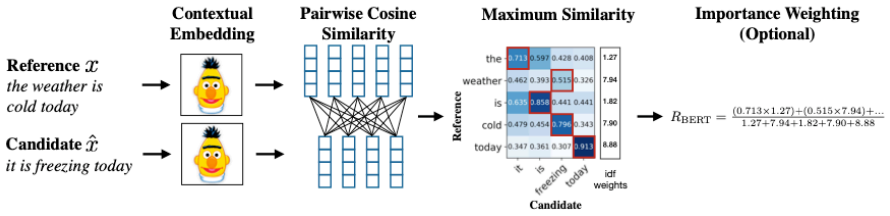

Figure 4.2: BERTScore: Shows the computation of Recall variant of the metric.

BLUERT Sellam et al. (2020) propose BLEURT, a learned evaluation metric based on BERT that is able to model human judgments with a few thousand training examples. In contrast to a BERT-based approach, without any learning from human judgement data, BLEURT claims to yield superior results even with fine-tuning on small amounts of training data. It also incorporates a pre-training scheme that uses millions of synthetic examples to help the model generalize. Owing to memory constraints we use the *BLEURT-20-D12* model for our evaluation task which is a distilled (12 layer) version of the large 32 layer model. However, note that BLEURT does not utilize any Hindi human judgement data in its fine-tuning phase.[7]

COMET Rei et al. (2020) propose COMET as a neural framework for training multilingual machine translation evaluation models. It is unique in that it incorporates the source as well in addition to the MT hypothesis and reference in scoring a given proposal in a multilingual embedding space. It is the only metric in our study to do so. Figure 4.3 shows the architecture of the estimator model (they also propose a ranking model) that we utilize. The entire model is trained by minimizing the Mean Squared Error (MSE) based on one of the possible human ratings (DA, HTER, MQM).

For our experiments, we utilize the DA based model.

4.3.5 Human Evaluation

We also conduct human evaluation for a randomly selected sample of 100 sentences from our dataset. We ran this evaluation task in terms of *Adequacy* and *Fluency* ratings for the four MT engine hypotheses in our experiments to establish correlations with automatic metric scores. We use a rating scale of 1—5 for both Adequacy and Fluency ratings.

Figure 4.4 shows a screenshot of an annotation item as visible to an evaluator. The first score is for Adequacy and the next one for Fluency for each MT proposal. We adapted a web-based workbench for our evaluation ratings (Girardi et al., 2014).

[7] https://github.com/google-research/bleurt

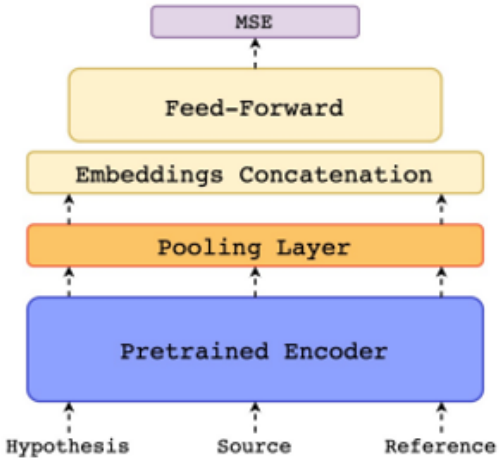

Figure 4.3: COMET: Shows the Estimator model architecture. The source, hypothesis and reference are independently encoded using a pretrained cross-lingual encoder.

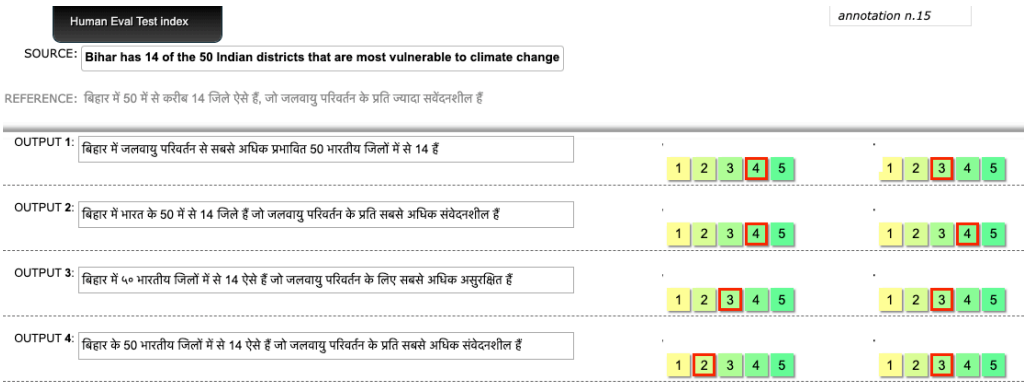

Figure 4.4: Human Evaluation: Screenshot of human evaluation ratings as conducted in an online workbench. The first score is for Adequacy and the next one for Fluency for each MT proposal.

4.4 Results

We calculate the automatic evaluation scores per engine and per metric against various sets of references and data views. For the sake of brevity we present detailed analysis only for the IIITH engine (originally used to generate the PE references). We then go on to show a summarized comparison with all other engines.

4.4.1 Automatic Evaluation Scores for IIITH MT

4.4.1.1 String-based Metrics

Corpus BLEU for IIITH MT Let us first look at the corpus BLEU scores, calculated per participant dataset, and then averaged across them. We calculate these scores for the IIITH engine across three reference sets and three data views. Note that SacreBLEu allows the use of variable number of references in case of a multi-reference setting[8].

In Table 4.5 the highest attained score on each data view (ALL, PE, HT) is marked in bold. As can be seen, it is the multiple PE references (*perefs*) in each view against which the highest BLEU scores are obtained. Unsurprisingly, the MT hypotheses when compared against single self (targeted) references don't do too well in the HT and ALL groupings, and fluctuate wildly. This is expected since these single references, especially in the HT view, carry no or minimal engine bias unlike the self references of the PE view. This clearly brings out the *engine-reference bias* when targeted post-edited references are used.

Another key insight is the magnitude of the effect of multiple HT references *htrefs*. We note that the average BLEU score rises from 22.98 to 41.99 when multiple human translated references are used. The jump in the quantum of scores is quite noticeable and somewhat expected, since the use of multiple references provides additional matches to the BLEU algorithm increasing its precision scores. However, would this go on to affect engine rankings as well? We address this point in a later section.

The last row of the table shows the $MEAN$ calculated over all participants. In all subsequent results we show these averaged scores only.

We now present the results of all the metrics for the IIITH engine in Table 4.6.

Sentence BLEU We have already analyzed corpus BLEU (cbleu) scores above, let us now look at the sentence BLEU (bleu) variant. In Table 4.6, we notice the same trend as with corpus BLEU with the highest scores obtained against multiple PE references in all the gorupings (ALL, PE, HT) and the lowest in case of a single *slfref* in the HT view. Note that in the HT groupings, *slfref* just refers to the human translation generated by the translator and since there was no MT hypothesis in this case, we do not expect it to carry engine bias unlike the *slfref* in the

[8]https://github.com/mjpost/sacrebleu#variable-number-of-references

	ALL (1793)			PE (900)			HT (893)		
	all-slfref	all-perefs	all-htrefs	pe-slfref	pe-perefs	pe-htrefs	ht-slfref	ht-perefs	ht-htrefs
P02	46.67	79.48	40.57	75.52	75.05	28.42	14.60	83.77	51.41
P03	37.76	84.39	40.29	51.91	85.01	28.42	23.14	83.77	51.03
P04	29.27	85.24	43.17	46.40	85.08	51.16	10.52	85.41	34.60
P05	38.69	84.63	42.44	58.82	83.98	51.16	15.57	85.31	33.03
P06	46.16	82.02	42.02	74.36	78.68	51.16	14.02	85.41	32.47
P07	45.20	84.23	36.73	48.94	84.69	28.42	41.23	83.77	44.49
P08	40.24	84.58	40.24	49.24	85.41	28.42	30.85	83.77	51.16
P09	43.60	84.58	40.16	64.89	83.77	51.16	20.91	85.41	28.28
P10	55.57	84.48	40.06	72.03	85.41	28.42	35.97	83.52	51.42
sd	7.31	1.83	1.86	11.75	3.64	11.99	10.77	0.88	9.77
MEAN	42.57	**83.74**	40.63	60.23	**83.01**	38.53	22.98	**84.46**	41.99

Table 4.5: Corpus BLEU for IIITH MT

	ALL (1793)			PE (900)			HT (893)		
	all-slfref	all-perefs	all-htrefs	pe-slfref	pe-perefs	pe-htrefs	ht-slfref	ht-perefs	ht-htrefs
cbleu	42.57	**83.74**	40.63	60.23	**83.01**	38.53	22.98	**84.46**	41.99
bleu	41.36	**87.56**	45.53	60.05	**86.31**	46.56	22.51	**88.82**	44.52
ter	49.81	**12.18**	52.86	30.60	**13.18**	52.34	69.18	**11.17**	53.36
chrf	62.64	**88.65**	61.39	75.32	**87.84**	61.45	49.86	**89.46**	61.35
brt	0.89	**0.97**	0.88	0.93	**0.96**	0.88	0.85	**0.97**	0.88
blr	0.66	**0.73**	0.56	**0.73**	**0.73**	0.55	0.59	**0.72**	0.57
cmt	0.71	**0.84**	0.49	**0.85**	0.84	0.46	0.56	**0.84**	0.52

Table 4.6: All metric scores for IIITH MT. Highlighted are the best scores for each *view* (ALL, PE, HT) of the data.

PE group. Again, we notice a jump in the magnitude of BLEU scores when multiple *htrefs* are used instead of a single reference.

TER (ter) being an edit distance metric, generates lower scores for hypothesis that require minimal editing. Therefore lower scores are better when evaluating against this metric. Again, in Table 4.6, we notice the same trend as in the other string based metrics seen so far. The best results are for the multiple PE references in all groupings, and the worst against a single *slfref* reference in the HT subset.

CHRF The chrf metric also exhibits the same trend in Table 4.6 with best results when hypothesis are evaluated against multiple PE references, and the worst against a single *slfref* reference in the HT view. Thus in all the string based metrics in our study we detect a bias for *targeted* post-edited references, those created from the same MT engine's proposal. We now move on to results for pre-trained metrics.

4.4.1.2 Pre-trained Metrics

BERTScore As discussed earlier, BERTScore (brt) is a contextual embedding-based metric which computes a soft similarity score for each token in the candidate sentence with each token in the reference sentence. We expect this metric to do as well on single references as on multiple references, since the soft similarity matches in the contextual embedding space would account for synonymous divergences in references, which is one of the weaknesses of string-based metrics.

We see this hypothesis bearing out to an extent when we look at the minimal differences in the quantum of scores when we compare single and multiple references. For example, in the HT grouping in Table 4.6 the mean BERTScore on single reference (*slfref*) stands at 0.85 while on multiple *htrefs* we obtain a score of 0.88.[9] Contrast this with the differences seen in similar comparisons on string-based metrics. Although, some caution may be in order in interpreting these results, since each metric is driven by a different scale and BERTScore is known to cluster towards the higher end of its scale. Nonetheless, the contrasts are striking but do require further statistical exploration.

BLEURT We now look at BLEURT (blr) and notice the trend changing for the first time. In case of the PE view, we find the scores to be equivalent on both single and multiple PE references 0.73. Remember, that BLEURT is a trained metric with some amount of fine-tuning on human judgement data. Also, quite important, BLEURT is computed with a single reference and cannot utilize either multiple or variable references out of the box. In order to compare against multiple reference scores given out by other metrics in our experiments, we take the

[9]Note that BERTScore calculates a score against each reference and returns the best score when multiple references are given.

maximum number of non-variable references for PE and HT groupings and report their mean scores.

COMET COMET (cmt) like BLEURT is a pre-trained embedding-based metric fine-tuned on human judgement data with two key differences: it incorporates the source segment in its calculation; and consequently, utilizes multilingual contextual embeddings. As with BLEURT before, COMET works against a single reference, therefore we report the mean scores in case of multiple variable references. On examining the results, in Table 4.6, what is immediately noticeable is that the trend shifts similar to BLEURT earlier, with the single reference *slfref* PE column attaining the highest score. Similarly, in case of the HT *slfref* and *htrefs* comparisons, we actually see multiple references hurting the COMET score. Does this imply, that for COMET, multiple references do not seem to help? The picture becomes clearer as we make MT engine comparisons in the next section.

4.4.1.3 Engine Comparison

We now bring in other MT hypotheses to evaluate against our varied source views and reference test sets. Recall that for engines other than the *IIITH* engine, the *slfref* single reference hereafter is NOT a *targeted* reference (was not created with the later MT engine hypothesis) and acts as an independent unbiased reference. We look at how different MT metrics evaluate the hypotheses from these MT engines. We present the results in a single table (Table 4.7) for ease of comparison with each metric block reporting results per engine and across all classes of references. The best scores are highlighted for each metric block.

When we look at the middle group of PE columns across all metrics, two trends are noticeable: that on a single reference *slfref* and multiple references *perefs* it is always the *IIITH* engine that comes out on top. The only metric to buck this trend is COMET, which ranks the *IndicTrans* engine higher. This result is not surprising, given that the single reference (being a targeted reference for IIITH MT) and the multiple PE references (primed by the same engine) are heavily biased towards the *IIITH* engine as we also saw earlier when we analyzed IIITH MT's scores in detail. Further, compare this situation with the rightmost group of HT columns in our table. Here, none of the references are likely to be biased towards a particular engine (including IIITH MT) since they were created in the unaided (without MT) condition. The trends we saw earlier, change quite dramatically in this grouping, with *Google* ranking as the best on all metrics except for BLEURT and COMET. Another important insight, comparing *ht-slfref* and *ht-htrefs*, both unbiased references, is that multiple references do not alter the engine rankings in any way except for one instance with the BLEURT score where the ranking is tied between Google and IndicTrans with a score of 0.60. Might we then infer, that single references suffice to reliably rank MT outputs?

Let us now address the results of the COMET metric for our engine hypothesis. We notice that COMET ranks the *IndicTrans* engine as the best, closely followed by *Google* across all data groupings. We also note that multiple references seem to hurt this metric, bringing down the scores achieved on single references whenever multiple references are used. While the ranking discrepancy will be resolved by looking at these scores in conjunction with human evaluation, we are puzzled by the negative influence of multiple references on COMET. Recall, that the metric itself is designed for single references.

4.4.2 Human Evaluation

Our Human Evaluation scores summarized in Table 4.8 bear out the COMET metric's rankings. We had previously seen how all the string based metrics and one of the pre-trained metrics (BERTScore) had almost unanimously adjudged Google's hypotheses to be the best, while COMET had ranked *IndicTrans* higher. Our human evaluation results correlate with COMET rankings.

How do we interpret this verdict? While, it may very well be the case that COMET, because it utilizes source segments in its scoring, is better able to make fine-grained distinctions, unlike say BLEURT which follows a similar training scheme but does not use source segments, the influence of references cannot be over-stated. We suspect COMET's use of source segments to be a moderating influence on biased references. We also suspect that even references carefully created in unaided human translation settings may inadvertently carry some online engine bias, particularly in favor of Google. Unless translators have been barred from utilizing any online translation aids (such as dictionaries, terminology and phrase look-ups and searches), given Google's prevalence in real-world translation settings, some bias for the engine may still make it into the translations (Fredholm, 2019). However, this needs more careful exploration in some detail.

4.4.2.1 Correlation with Human Evaluation

In the previous section, we saw COMET to be the best predictor amongst all metrics of MT output quality over four different engine outputs. We further run a correlation analysis against human evaluation ratings data to test the strength of this correlation.

In Tables 4.9 and 4.10, we can see the correlations of string-based and pre-trained metrics against the Adequacy and Fluency scores. Amongst the string-based metrics, we detect a moderate positive correlation for sentence bleu (bl) of 0.39 on Fluency ratings, and a moderate negative correlation of -0.37 for ter (tr) for Adequacy ratings. In both cases on multiple HT references.

Similarly, amongst pre-trained metrics, we detect a positive correlation against Fluency ratings with a value of 0.50 for the COMET metric both on a single reference and against

	ALL (1793)			PE (900)			HT (893)		
	all-slfref	all-perefs	all-htrefs	pe-slfref	pe-perefs	pe-htrefs	ht-slfref	ht-perefs	ht-htrefs
					Sentence BLEU				
bleu-Bing	29.55	52.62	44.61	36.20	52.08	45.24	22.84	53.14	44.00
bleu-Google	37.88	66.93	**61.13**	45.02	66.46	**61.34**	**30.71**	67.39	**60.98**
bleu-IIITH	**41.36**	**87.56**	45.53	**60.05**	**86.31**	46.56	22.51	**88.82**	44.52
bleu-IIITH_v2	30.65	55.04	47.01	37.65	54.87	47.70	23.60	55.20	46.33
bleu-IndicTrans	33.94	60.62	52.46	41.98	59.98	53.24	25.86	61.23	51.72
					Corpus BLEU				
cbleu-Bing	31.49	52.72	40.45	38.67	52.61	38.14	23.53	52.38	41.71
cbleu-Google	39.33	64.80	**55.45**	46.29	64.93	**51.47**	**31.58**	64.58	**57.26**
cbleu-IIITH	**42.57**	**83.74**	40.63	**60.23**	83.01	38.53	22.98	**84.46**	41.99
cbleu-IIITH_v2	32.16	54.24	42.48	39.30	54.65	40.29	24.36	53.84	43.97
cbleu-IndicTrans	35.27	58.92	47.02	42.93	59.13	44.25	26.85	58.67	48.83
					chrF				
chrf-Bing	56.30	70.38	61.87	62.13	70.13	61.83	50.44	70.63	61.93
chrf-Google	**62.69**	79.88	**71.15**	69.20	79.56	**70.78**	**56.16**	80.20	**71.56**
chrf-IIITH	62.64	**88.65**	61.39	**75.32**	**87.84**	61.45	49.86	**89.46**	61.35
chrf-IIITH_v2	57.16	72.61	62.87	63.44	72.39	62.77	50.85	72.83	62.99
chrf-IndicTrans	59.94	75.75	64.71	67.08	75.43	64.59	52.77	76.07	64.84
					TER				
ter-Bing	59.00	40.22	50.28	51.35	40.48	50.11	66.70	39.98	50.43
ter-Google	51.13	32.07	**36.66**	45.21	32.03	**36.97**	**57.07**	32.13	**36.29**
ter-IIITH	**49.81**	**12.18**	52.86	**30.60**	**13.18**	52.34	69.18	**11.17**	53.36
ter-IIITH_v2	56.41	38.58	46.05	49.87	38.57	46.08	62.98	38.61	46.00
ter-IndicTrans	54.37	36.41	43.95	47.63	36.61	44.00	61.15	36.22	43.87
					BERTScore				
brt-Bing	0.87	0.91	0.88	0.89	0.91	0.88	0.85	0.91	0.88
brt-Google	0.88	0.94	**0.91**	0.91	0.94	**0.91**	**0.87**	0.94	**0.91**
brt-IIITH	**0.89**	**0.97**	0.88	**0.93**	**0.96**	0.88	0.85	**0.97**	0.88
brt-IIITH_v2	0.87	0.92	0.88	0.89	0.92	0.89	0.85	0.92	0.89
brt-IndicTrans	0.88	0.93	0.89	0.90	0.93	0.89	0.85	0.93	0.89
					BLEURT				
blr-Bing	0.60	0.63	0.54	0.63	0.63	0.53	0.58	0.62	0.55
blr-Google	**0.66**	0.68	0.58	0.69	0.68	0.56	**0.63**	0.68	**0.60**
blr-IIITH	**0.66**	**0.73**	0.56	**0.73**	**0.73**	0.55	0.59	**0.72**	0.57
blr-IIITH_v2	0.62	0.64	0.55	0.65	0.64	0.53	0.59	0.64	0.56
blr-IndicTrans	0.65	0.68	**0.59**	0.69	0.68	**0.58**	0.62	0.68	**0.60**
					COMET				
cmt-Bing	0.60	0.66	0.47	0.67	0.66	0.43	0.54	0.65	0.49
cmt-Google	0.79	0.84	0.64	0.86	0.85	0.61	0.73	0.84	0.67
cmt-IIITH	0.71	0.84	0.49	0.85	0.84	0.46	0.56	0.84	0.52
cmt-IIITH_v2	0.64	0.70	0.50	0.71	0.70	0.47	0.58	0.70	0.53
cmt-IndicTrans	**0.81**	**0.87**	**0.67**	**0.87**	**0.87**	**0.65**	**0.74**	**0.86**	**0.70**

Table 4.7: MT Engine Comparisons. Best scores per column within each metric block are highlighted. IIITH_v2 is an iteration of the IIITH engine trained with additional data. It was not used for Human Evaluation.

Engine	Adequacy	Fluency
IIITH	3.91	3.44
Bing	3.67	3.30
Google	4.10	3.81
IndicTrans	**4.26**	**3.93**

Table 4.8: Human Evaluation Scores

	Fluency	Adequacy	bl_slfref	bl_perefs	bl_htrefs	tr_slfref	tr_perefs	tr_htrefs	ch_slfref	ch_perefs	ch_htrefs
Fluency	1.00	0.75	0.30	0.28	0.39	-0.26	-0.27	-0.37	0.32	0.33	0.32
Adequacy	0.75	1.00	0.25	0.27	0.36	-0.24	-0.31	-0.37	0.27	0.35	0.32
bl_slfref	0.30	0.25	1.00	0.29	0.28	-0.85	-0.23	-0.21	0.90	0.32	0.20
bl_perefs	0.28	0.27	0.29	1.00	0.17	-0.23	-0.88	-0.11	0.29	0.90	0.15
bl_htrefs	**0.39**	0.36	0.28	0.17	1.00	-0.21	-0.07	-0.88	0.33	0.27	0.89
tr_slfref	-0.26	-0.24	-0.85	-0.23	-0.21	1.00	0.23	0.19	-0.82	-0.28	-0.15
tr_perefs	-0.27	-0.31	-0.23	-0.88	-0.07	0.23	1.00	0.07	-0.23	-0.82	-0.07
tr_htrefs	-0.37	**-0.37**	-0.21	-0.11	-0.88	0.19	0.07	1.00	-0.25	-0.23	-0.86
ch_slfref	0.32	0.27	0.90	0.29	0.33	-0.82	-0.23	-0.25	1.00	0.40	0.30
ch_perefs	0.33	0.35	0.32	0.90	0.27	-0.28	-0.82	-0.23	0.40	1.00	0.29
ch_htrefs	0.32	0.32	0.20	0.15	0.89	-0.15	-0.07	-0.86	0.30	0.29	1.00

Table 4.9: String-based Metrics' Pearson coefficient r values for r(100). The highest value is highlighted in bold.

multiple PE references. Against the Adequacy ratings, we note a slightly stronger correlation 0.58, again for the COMET metric, against multiple PE references.

These scores further validate the engine rankings as measured by COMET. The picture on the use of multiple references is less clear. The correlation strengths definitely improve for string-based metrics, but in case of pre-trained metrics the improvements are marginal. Multiple references do not seem to negatively affect COMET correlations unlike what we noted earlier with the scores viewed independent of human ratings correlations. We conclude that it may be prudent to utilize multiple references even with pre-trained metrics.

4.5 Conclusion

We conducted a study on machine translation evaluation against single and multiple references generated in the post-editing and unaided conditions. We scored four MT engines against these references on both string-based and pre-trained automatic evaluation metrics currently prevalent in MT evaluation settings. We also independently conducted an *Adequacy* and *Fluency* based human evaluation task and correlated the automatic metrics' results with human evaluation findings.

	Fluency	Adequacy	bt_slfref	bt_perefs	bt_htrefs	ct_slfref	ct_perefs	ct_htrefs	br_slfref	br_perefs	br_htrefs
Fluency	1.00	0.75	0.35	0.32	0.39	0.50	0.50	0.49	0.45	0.45	0.37
Adequacy	0.75	1.00	0.35	0.41	0.42	0.54	0.58	0.57	0.40	0.50	0.44
bt_slfref	0.35	0.35	1.00	0.39	0.35	0.69	0.39	0.33	0.82	0.42	0.27
bt_perefs	0.32	0.41	0.39	1.00	0.25	0.43	0.62	0.42	0.43	0.76	0.38
bt_htrefs	0.39	0.42	0.35	0.25	1.00	0.40	0.37	0.59	0.35	0.37	0.60
ct_slfref	**0.50**	0.54	0.69	0.43	0.40	1.00	0.82	0.76	0.78	0.61	0.47
ct_perefs	**0.50**	**0.58**	0.39	0.62	0.37	0.82	1.00	0.85	0.53	0.81	0.55
ct_htrefs	0.49	0.57	0.33	0.42	0.59	0.76	0.85	1.00	0.47	0.64	0.74
br_slfref	0.45	0.40	0.82	0.43	0.35	0.78	0.53	0.47	1.00	0.62	0.48
br_perefs	0.45	0.50	0.42	0.76	0.37	0.61	0.81	0.64	0.62	1.00	0.66
br_htrefs	0.37	0.44	0.27	0.38	0.60	0.47	0.55	0.74	0.48	0.66	1.00

Table 4.10: Pre-trained Metrics' Pearson coefficient r values for r(100). The highest value is highlighted in bold.

To answer our originally posed research questions: we detected substantial engine bias in case of post-edited references; we found that it may still be prudent to utilize multiple references when adjudging engine rankings with automatic metrics; and we found COMET to be the metric that best correlates with human evaluation findings.

Based on these insights, we recommend the following:

- that MT evaluation sets should NOT be post-edited, but rather created in an unaided human translation condition.
- that multiple carefully curated high quality references be utilized for MT meta-evaluation tasks with string-based as well as pre-trained metrics.
- that a pre-trained metric that incorporates source segments in its scoring be used in conjunction with other popular string-based metrics.
- that human evaluation be conducted in conjunction with automatic evaluation as human evaluation still plays a crucial role in uncovering as yet undiscovered metric or test data biases.
- that one must also be alert to the use of online searches, bilingual dictionaries and term translations as another source of possible bias, especially in favor of online engines.

In terms of future work, we would like to expand on the human evaluation part of this study incorporating newer ranking or annotation strategies like MQM in addition to increasing the size of ratings data. A major bottleneck in wider adoption of pre-trained metrics remains the unavailability or quality of target language and multilingual embeddings for low and mid resource languages. It remains to be seen how metrics like BLEURT and COMET perform in these low resource settings. In such scenarios the challenge of automatic evaluation with fine-grained distinctions remains wide open.

Chapter 5

Conclusions

This thesis consisted of three independent but inter-related studies. We studied machine translation post-editing right from its usefulness to its applicability. First, as a process to establish whether post-editing helped or hindered the translation process; second, its artefacts, that is the text produced in this process, to establish their nature via various lexical and syntactic metrics; and third, its usage or applicability in an evaluation setting. Our experiments across these three sections and the unique datasets generated in this process offer valuable insights into the efficacy, nature and usage of the post-editing process and its outputs.

Assessing Post-editing Effort Is machine translation post-editing worth the effort (Koponen, 2016)? This question was investigated in some detail in this study for the English-Hindi language pair. A controlled behavioral study employing professional translators was conducted under two alternating conditions (PE vs. HT) in order to investigate the following questions:

- RQ1: Is post-editing effort as measured on temporal, technical and cognitive dimensions lesser in the PE condition than the HT condition for the English-Hindi direction?
- RQ2: Is the quality of post-edited segments equal to human-translated segments as ascertained by human raters?
- RQ3: Do automatic MT evaluation metrics correlate with PE effort indicators, when both are measured at the segment level?

For RQ1, on two of the effort dimensions, temporal and technical, we found post-editing productivity gains of up to 172% and technical effort reduction of 59%. In the cognitive dimension, employing three different measures, we saw reductions in frequency of pauses by 63%. However, we noticed an increase in average pause duration by 12% and average initial pause duration by 5%. For RQ2, we conducted a human evaluation ranking task and found both PE and HT segments to be of similar quality. For RQ3, we found moderate to strong correlations

for 3 automatic MT evaluation metrics across all PE effort indicators, with technical effort most strongly correlating with automatic MT metrics. We thus concluded post-editing to be *worth the effort* in improving productivity, producing translations equal in quality to human translations, and found the technical post-editing effort dimension to be the best predictor of MT quality as measured on automatic metrics.

The data points generated in these experiments were subsequently used in our later experiments.

Measuring Post-edited Texts In these set of experiments, we investigated the nature of the text produced during the post-editing process and contrasted them with human translated text. Both being forms of translated data, we looked for translationese, post-editese and machine-translationese markers in these text and tested them for evidence of hypothesized translation universals (Baker, 1993). We posed the following questions:

- RQ1: Do post-edited text and human-translated text exhibit different characteristics?
- RQ2: Do texts created by different translators exhibit different characteristics?
- RQ3: Does raw MT text track closely with post-edited or human-translated text?

For RQ1, we found PE and HT texts to exhibit significantly different characteristics on most of our metrics, with PE text appearing to be simpler, more normalized and showing greater source interference. We also noted the characteristics of raw MT output in comparison to PE and HT text and found them to be of degraded quality. For RQ2, we did not find major differences amongst the text produced by different translators. We ascribed this to a fairly uniform (in their competencies) pool of translators that we managed to assemble. For RQ3, as stated earlier, we found PE text to be more similar in nature to MT than HT text on most indicators. Given, these findings on the linguistic nature of these text, we advise caution and careful curation when employing them for any applied tasks, especially those centering on MT evaluation. To test their impact on a downstream applied task, we next conducted a study on MT evaluation using these data.

Impact of Post-edited Texts We conducted an MT evaluation study to observe the impact of text generated by post-editing and unaided human translation on MT engine rankings when used as references. We conducted evaluations for both widely used string-based metrics as well as the more recent pre-trained metrics. We posed the following questions:

- RQ1: Are post-edited (PE) references biased towards MT engines they were created from compared to their unaided human translated (HT) references?
- RQ2: Do the number of references used affect MT evaluation outcomes?

- RQ3: Which current MT evaluation metrics fare best under varying reference conditions?

For RQ1, we found evidence of bias in evaluation scores when using MT engine references created using the same engine during the post-editing process. We also saw some indication of a general bias towards a widely used online engine. For RQ2, we found that multiple references do not change engine rankings, but do lift up the quantum of metric scores across the board, except for one pre-trained metric (COMET), where using multiple references degraded the scores somewhat. However, we noted slightly better correlations even for this pre-trained metric when correlating with Human Evaluation ratings. We therefore think it prudent to utilize multiple references when available, especially for meta-evaluation tasks. For RQ3, we found the COMET metric to best correlate with human evaluation scores. In light of these findings, we issued some recommendations on the creation of test sets for MT meta-evaluation tasks.

Future Work We see multiple possibilities in extending this work in future. Further post-editing effort estimation studies are called for to assess the impact of language similarity on post-editing and translation productivity in general (Green et al., 2013), especially in the Indian language multilingual scenario. Of primary interest is a possible study on directionality in a post-editing setting, for example, how might translator behavior change when post-editing or translating within the same language family than across families? Cognitive effort as yet remains an understudied area, and use of eye trackers in a similar controlled setting may shed additional light on the impact of post-editing in this dimension. The experiments on translationese and post-editese markers looked mostly at lexical and a few syntactic features. Word order and dependency interactions as another marker of source interference may be a further direction worth investigating (Yadav et al., 2020). Similarly, studying syntactic entropy in translated texts (Bangalore et al., 2015), and its link to production times would further help link the first two experiments conducted in this current work. Under MT evaluation, it might be worth revisiting metrics that reward some of the linguistic characteristics we identified as lacking in machine translated outputs (Giménez and Màrquez, 2010).

Related Publications

In *Proceedings of the 18th International Conference on Natural Language Processing (ICON), 2021*

Ahsan, Arafat, Vandan Mujadia, and Dipti Misra Sharma."Assessing Post-editing Effort in the English-Hindi Direction." arXiv preprint arXiv:2112.09841 (2021).

Appendix A

Additional Results

A.1 Linguistic Indicators

	Source			Target			% diff with HT	
	srcmt	srcpe	srcht	mt	pe	ht	mt_prcnt	pe_prcnt
P02	110.0	110.0	110.0	120.0	130.0	140.0	-20.0	-0.0
P03	110.0	110.0	110.0	120.0	130.0	140.0	-20.0	-10.0
P04	110.0	110.0	110.0	110.0	140.0	180.0	-40.0	-20.0
P05	110.0	110.0	110.0	110.0	160.0	160.0	-30.0	0.0
P06	110.0	110.0	110.0	110.0	130.0	150.0	-30.0	-20.0
P07	110.0	110.0	110.0	120.0	160.0	150.0	-20.0	10.0
P08	110.0	110.0	110.0	120.0	160.0	140.0	-20.0	10.0
P09	110.0	110.0	110.0	110.0	120.0	150.0	-30.0	-20.0
P10	110.0	110.0	100.0	120.0	140.0	130.0	-10.0	10.0
sd	0.0	0.0	0.0	0.0	20.0	10.0	10.0	10.0
MEAN	110.0	110.0	110.0	110.0	140.0	150.0	-20.0	-0.0

Table A.1: Pronoun Frequency Counts

	Source			Target			% diff with HT	
	srcmt	srcpe	srcht	mt	pe	ht	mt_prcnt	pe_prcnt
P02	763.0	763.0	803.0	854.0	897.0	900.0	-5.0	-0.0
P03	763.0	763.0	803.0	854.0	893.0	948.0	-10.0	-6.0
P04	803.0	803.0	763.0	879.0	983.0	981.0	-10.0	0.0
P05	803.0	803.0	752.0	879.0	924.0	863.0	2.0	7.0
P06	803.0	803.0	763.0	879.0	914.0	859.0	2.0	6.0
P07	763.0	763.0	803.0	854.0	949.0	951.0	-10.0	-0.0
P08	763.0	763.0	803.0	854.0	938.0	954.0	-10.0	-2.0
P09	803.0	803.0	763.0	879.0	937.0	914.0	-4.0	3.0
P10	763.0	763.0	759.0	854.0	913.0	875.0	-2.0	4.0
sd	21.0	21.0	23.0	13.0	28.0	45.0	5.0	4.0
MEAN	781.0	781.0	779.0	865.0	928.0	916.0	-5.0	1.0

Table A.2: Noun Frequency Counts

A.2 Translator SD Comparison

Table A.3: Translator Standard Deviations for Automatic Metrics

	ALL (1793)			PE (900)			HT (893)		
	all-slfref	all-perefs	all-htrefs	pe-slfref	pe-perefs	pe-htrefs	ht-slfref	ht-perefs	ht-htrefs
					Sentence BLEU				
bleu-Bing	6.81	0.37	1.23	4.67	5.35	12.50	11.05	5.33	13.42
bleu-Google	12.32	1.29	2.00	7.14	10.94	22.40	18.82	11.25	23.32
bleu-IIITH	7.09	1.08	1.08	12.34	2.47	10.43	9.82	0.74	11.46
bleu-IndicTrans	7.77	0.46	1.18	4.77	5.98	15.65	12.62	5.88	16.42
					Corpus BLEU				
cbleu-Bing	7.47	0.94	2.12	5.36	4.44	13.69	12.17	5.02	11.69
cbleu-Google	12.95	2.10	3.98	7.58	8.74	24.15	20.51	9.43	21.66
cbleu-IIITH	7.31	1.83	1.86	11.75	3.64	11.99	10.77	0.88	9.77
cbleu-IndicTrans	8.10	0.91	2.46	4.87	4.76	16.62	13.93	5.52	14.12
					chrF				
chrf-Bing	6.82	0.44	0.89	4.69	3.13	8.69	10.32	2.79	9.22
chrf-Google	9.47	0.65	1.11	5.78	4.71	14.05	14.41	4.57	14.41
chrf-IIITH	6.23	0.90	0.73	7.55	1.83	7.68	9.51	0.07	8.24
chrf-IndicTrans	6.63	0.39	0.81	4.44	2.23	9.66	10.61	1.95	10.08
					COMET				
cmt-Bing	0.10	0.03	0.01	0.07	0.07	0.15	0.17	0.08	0.15
cmt-Google	0.12	0.02	0.02	0.09	0.10	0.11	0.16	0.11	0.11
cmt-IIITH	0.10	0.04	0.01	0.11	0.14	0.14	0.14	0.15	0.13
cmt-IndicTrans	0.09	0.02	0.02	0.06	0.07	0.12	0.14	0.08	0.12
					TER				
ter-Bing	7.89	0.43	1.27	4.73	4.36	10.30	12.83	4.31	10.89
ter-Google	12.92	1.02	1.87	7.92	9.44	17.97	18.88	9.57	18.38
ter-IIITH	7.24	1.25	1.15	12.55	2.50	6.31	10.86	0.10	7.28
ter-IndicTrans	9.07	0.38	1.24	4.77	5.32	12.75	14.51	5.17	13.25

A.3 Automatic and Human Evaluation Correlations

	Human		SBLEU			TER			CHRF			BERTScore			COMET			BLEURT		
	Fluency	Adequacy	bl_slfref	bl_perefs	bl_htrefs	tr_slfref	tr_perefs	tr_htrefs	ch_slfref	ch_perefs	ch_htrefs	bt_slfref	bt_perefs	bt_htrefs	ct_slfref	ct_perefs	ct_htrefs	br_slfref	br_perefs	br_htrefs
Fluency	1.00	0.75	0.30	0.28	0.39	-0.26	-0.27	-0.37	0.32	0.33	0.32	0.35	0.32	0.39	0.50	0.50	0.49	0.45	0.45	0.37
Adequacy	0.75	1.00	0.25	0.27	0.36	-0.24	-0.31	-0.37	0.27	0.35	0.32	0.35	0.41	0.42	0.54	0.58	0.57	0.40	0.50	0.44
bl_slfref	0.30	0.25	1.00	0.29	0.28	-0.85	-0.23	-0.21	0.90	0.32	0.20	0.84	0.31	0.22	0.56	0.30	0.18	0.69	0.30	0.10
bl_perefs	0.28	0.27	0.29	1.00	0.17	-0.23	-0.88	-0.11	0.29	0.90	0.15	0.25	0.86	0.14	0.28	0.52	0.29	0.30	0.66	0.29
bl_htrefs	0.39	0.36	0.28	0.17	1.00	-0.21	-0.07	-0.88	0.33	0.27	0.89	0.29	0.19	0.83	0.32	0.29	0.49	0.28	0.28	0.52
tr_slfref	-0.26	-0.24	-0.85	-0.23	-0.21	1.00	0.23	0.19	-0.82	-0.28	-0.15	-0.82	-0.28	-0.19	-0.56	-0.31	-0.18	-0.68	-0.32	-0.09
tr_perefs	-0.27	-0.31	-0.23	-0.88	-0.07	0.23	1.00	0.07	-0.23	-0.82	-0.07	-0.25	-0.84	-0.10	-0.32	-0.54	-0.33	-0.31	-0.66	-0.29
tr_htrefs	-0.37	-0.37	-0.21	-0.11	-0.88	0.19	0.07	1.00	-0.25	-0.23	-0.86	-0.27	-0.16	-0.87	-0.34	-0.33	-0.53	-0.26	-0.29	-0.50
ch_slfref	0.32	0.27	0.90	0.29	0.33	-0.82	-0.23	-0.25	1.00	0.40	0.30	0.89	0.35	0.29	0.64	0.37	0.28	0.78	0.40	0.20
ch_perefs	0.33	0.35	0.32	0.90	0.27	-0.28	-0.82	-0.23	0.40	1.00	0.29	0.36	0.91	0.28	0.40	0.59	0.40	0.43	0.74	0.39
ch_htrefs	0.32	0.32	0.20	0.15	0.89	-0.15	-0.07	-0.86	0.30	0.29	1.00	0.26	0.17	0.89	0.31	0.31	0.54	0.26	0.29	0.57
bt_slfref	0.35	0.35	0.84	0.25	0.29	-0.82	-0.25	-0.27	0.89	0.36	0.26	1.00	0.39	0.35	0.69	0.39	0.33	0.82	0.42	0.27
bt_perefs	0.32	0.41	0.31	0.86	0.19	-0.28	-0.84	-0.16	0.35	0.91	0.17	0.39	1.00	0.25	0.43	0.62	0.42	0.43	0.76	0.38
bt_htrefs	0.39	0.42	0.22	0.14	0.83	-0.19	-0.10	-0.87	0.29	0.28	0.89	0.35	0.25	1.00	0.40	0.37	0.59	0.35	0.37	0.60
ct_slfref	0.50	0.54	0.56	0.28	0.32	-0.56	-0.32	-0.34	0.64	0.40	0.31	0.69	0.43	0.40	1.00	0.82	0.76	0.78	0.61	0.47
ct_perefs	0.50	0.58	0.30	0.52	0.29	-0.31	-0.54	-0.33	0.37	0.59	0.31	0.39	0.62	0.37	0.82	1.00	0.85	0.53	0.81	0.55
ct_htrefs	0.49	0.57	0.18	0.29	0.49	-0.18	-0.33	-0.53	0.28	0.40	0.54	0.33	0.42	0.59	0.76	0.85	1.00	0.47	0.64	0.74
br_slfref	0.45	0.40	0.69	0.30	0.28	-0.68	-0.31	-0.26	0.78	0.43	0.26	0.82	0.43	0.35	0.78	0.53	0.47	1.00	0.62	0.48
br_perefs	0.45	0.50	0.30	0.66	0.28	-0.32	-0.66	-0.29	0.40	0.74	0.29	0.42	0.76	0.37	0.61	0.81	0.64	0.62	1.00	0.66
br_htrefs	0.37	0.44	0.10	0.29	0.52	-0.09	-0.29	-0.50	0.20	0.39	0.57	0.27	0.38	0.60	0.47	0.55	0.74	0.48	0.66	1.00

Table A.4: Automatic and Human Evaluation Correlations. Pearson coefficient (r) values for r(100).

Appendix B

Post-Editing Guidelines

B.1 Introduction

The technological advances in Machine Translation (MT) over the last two decades has led to its increasing adoption in translation workflows. This has been brought about not just by improved MT quality, but also because many end users have become increasingly familiar, with the sometimes flawed yet useful, MT output available via online generic MT engines like Google Translate and Microsoft's Bing Translator.

This acceptance of MT both in terms of quality and familiarity has led to it becoming a part of translation workflows across many domains and in many languages. Machine Translation is now counted as one of the most important tools in the Computer Aided Translation (CAT) ecosystem. It helps speed up the entire workflow, reducing turnaround time for translation projects, consequently, leading to greater volumes of translations being possible when projects are based on machine translation post-editing.

Post-Editing (PE) integral to MT aided workflows is the task of editing MT output for accuracy of meaning and target language fluency. It is often seen as an editor's task: of checking, making small changes, and signing off on the meaning and fluency of the MT proposal; but also at times a more creative translator's approach is required when MT has made a hash of meaning or syntax, or when it has failed to capture all the nuances present in the Source text. A post-editor is thus required to work across a spectrum of skills. In view of the above discussion we can see that successful MT adoption hinges on the following factors:

- Quality of MT output
- Translator post-editing skills

The first is measured through various human and automated metrics; while the second calls for learning of a new skill, which sometimes is the primary reason behind reluctance towards

adoption of MT technology on part of translators. In the absence of any training or guidelines for these actual end users of MT systems, MT adoption by small and medium translation projects has only worked in fits and starts. The real efficacy of an MT engine can perhaps only truly be measured by those skilled in the post-editing task.

Many have thus for long argued for designing courses for teaching post-editing skills (O'Brien 2002). While there are now a few such courses (Görög, 2014) and many guidelines documents (Hu and Cadwell, 2016) to familiarize translation practitioners with the post-editing process, there is nothing, as far as we are aware, that is specifically targeted towards post-editors and translators working in Indian Languages. This guidelines document is meant as a first step in that direction.

These guidelines intend to introduce translators and post-editors working in translating technical domain content from English into Indian Languages to the common issues that might arise during their task. The document presupposes the use of MT in the workflow and therefore is intended as a post-editing guidelines document. The document builds upon the categories and structure of the BOLT post-editing guidelines[1] (Translation) and comprises of three sections:

- **Capturing Source Meaning**: It deals with common issues that a post-editor may frequently encounter and provides recommendations on how to tackle them.
- **Avoiding Machine Translation Pitfalls**: This section lays out issues peculiar to Machine Translation outputs.
- **Minimizing Edits**: This section cautions post-editors against over-editing for style, spelling, or punctuation.

We have culled examples from ongoing real-world projects to demonstrate each point in the above sections where possible. At the moment, most, if not all examples are in Hindi, but we hope to collaboratively extend this pool of examples to as many Indian Languages as possible.

These guidelines do not claim to be exhaustive. We welcome both, more examples, as well as suggestions about additional sections from our readers. Finally, this currently is a draft version of what we hope will eventually become a living document. We hope to publish a 1.0 Version soon that will be available publicly as a technical document via IIIT-H's website.

B.2 Post-Editing Process

Post-Editing is the task of editing MT output for accuracy of meaning and target language fluency. We distinguish two types of post-editing activities that are common in practice:

[1] https://www.nist.gov/system/files/documents/itl/iad/mig/BOLT_P3_PostEditingGuidelinesV1_3_3.pdf

B.2.1 Partial Post-editing

As the name suggests, it is scaled back or light post-editing and has the following properties:

- used for content gisting or 'good enough' quality
- does not need to be of publishable quality
- might not be fluent
- no stylistic edits
- typically applied to social media, web forum and user-generated content

B.2.2 Full Post-editing

This type of post-edit is more common. Fluency is important here and it is differentiated from partial post-editing by the following properties:

- publishable quality is expected
- target fluency is important
- applied to all genres

These guidelines are aimed towards achieving Full Post-Editing.

B.3 General Post Editing Rules

B.3.1 Capturing Source Meaning

B.3.1.1 Acronyms

An Acronym can be substituted with its expanded form and vice versa. However, consider the following cases: If the Source has an acronym that has been presented in the same form by the MT engine, then there is no need to substitute it with a full form when post editing.

Source	It is Eighteenth module of AI that we are going to look at today.
MT	यह एआई का अठारहवाँ मॉड्यूल है जिसे हम आज देखने जा रहे हैं।
PE1	यह एआई का अठारहवाँ मॉड्यूल है, जिसे हम आज देखने जा रहे हैं।
PE2	यह आर्टिफिशियल इंटेलिजेंस का अठारहवाँ मॉड्यूल है, जिसे हम आज देखने जा रहे हैं।

Table B.1: Acronyms

In the above example both post-edits are acceptable but the first one is recommended and preferred since it utilizes the MT output and is true to the Source usage.

Alternatively, if the Source has an acronym that has been presented in an expanded form by the MT engine, there is no need to substitute it with the shortened form when post editing, as long as the expanded form is correct. We consider the same source example discussed above but now with a different MT output. Both post-edits are acceptable here, but the second one has minimal edits.

Source	It is Eighteenth module of AI that we are going to look at today.
MT	यह आर्टिफिशियल इंटेलिजेंस का अठारहवाँ मॉड्यूल है जिसे हम आज देखने जा रहे हैं।
PE1	यह एआई का अठारहवाँ मॉड्यूल है, जिसे हम आज देखने जा रहे हैं।
PE2	यह आर्टिफिशियल इंटेलिजेंस का अठारहवाँ मॉड्यूल है, जिसे हम आज देखने जा रहे हैं।

Table B.2: Acronyms

The guiding principle therefore is to be faithful to the Source as much as possible, but to also consider minimizing the number of edits performed on the MT output. You must strive to utilize as much of the MT output as possible without compromising on accuracy and fluency of the Target translation.

Let us now consider a slightly different case using the example below. Should we punctuate acronyms if they are not punctuated in source?

Consider the example below:

MT seems to have fully translated the noun phrase 'Food Safety and Standard Act'. If the translation proposed by MT is prevalent and acceptable in conventional usage in the target language, then we must use it. If not use the term or phrase as-is with transliteration. You will find more information about handling domain terms and phrases in a later section.

Now to the Acronym that derives from the above term: if the Acronym which is based on the letters of the Source is in wide and conventional usage, then it can be used in the transliterated form, like ए आई for AI in the example above.

It is recommended to punctuate Acronyms with space instead of the period symbol or follow the convention prevalent in your target language. In this example, PE2 becomes the preferred segment.

Source	We are going to see the objectives of the Food Safety and Standard Act (FSSA).
MT	हम खाद्य सुरक्षा और मानक अधिनियम (FSSA) के उद्देश्यों को देखने जा रहे हैं।
PE1	हम फूड सेफ्टी और मानक अधिनियम (एफ.एस. एस. ए.) के उद्देश्य जानेंगे।
PE2	हम खाद्य सुरक्षा और मानक अधिनियम (एफ एस एस ए) के उद्देश्यों को देखने जा रहे हैं।

Table B.3: Acronyms

B.3.1.2 Synonyms

Synonyms are acceptable as long as they convey the same source meaning.

In the example below, two post-editors seem to have picked two different words to convey the meaning of the source phrase 'food habits'. Both, being synonymous are acceptable. There will always be stylistic preferences - we talk about this more in a later section - but remember that we are editing foremost for meaning and fluency.

Source	It can be easily imagined that research touches every area be it law, be it marketing, technology, be it even food habits.
MT	यह आसानी से कल्पना की जा सकती है कि अनुसंधान हर क्षेत्र को छूता है, यह कानून हो, विपणन हो, तकनीक हो, यहां तक कि खाद्य आदतें भी हों।
PE1	इसकी आसानी से कल्पना की जा सकती है कि रिसर्च हर क्षेत्र को छूता है, चाहे यह कानून हो, चाहे वह विपणन हो, तकनीक हो, चाहे वह भोजन की आदतें हों।
PE2	इसकी आसानी से कल्पना की जा सकती है कि रिसर्च हर क्षेत्र को छूता है, चाहे यह कानून हो, चाहे वह विपणन हो, तकनीक हो, चाहे वह खान पान की आदतें हों।

Table B.4: Synonyms

But, sometimes, one can have too many Synonyms to choose from as in our next example. Consider 'barrier, बाधा, अवरोध.

If 'communication' and 'psychological barrier' are marked as domain terms and need to be carried over in transliteration into the target language then PE1 is a good edit; else MT also produces a decent output in the example given below and with a partial edit be turned into PE2.

Source	Our next barrier to communication is a psychological barrier.
MT	संचार के लिए हमारी अगली बाधा एक मनोवैज्ञानिक बाधा है।
PE1	कम्यूनिकेशन के लिए हमारा अगला अवरोध साइकलॉजिकल बैरियर है।
PE2	कम्यूनिकेशन के लिए हमारी अगली बाधा एक मनोवैज्ञानिक बाधा है।

Table B.5: Synonyms

B.3.1.3 Symbols

Symbols such as the percent symbol (%), the currency symbols ($), the degree symbol (°) should not be edited as far as possible. The guidance here is again the same as in the Acronym

section. Go with what is conventionally acceptable in the target language and try to utilize the MT proposal as-is if that is acceptable and conventional.

In the example below, the post-editor could have avoided changing डिग्री सेल्सियस back to °C since this expanded transliterated usage is common in Hindi.

Source	The first stage of fermentation takes place at 35°C temperature, and at this stage, the acidity of the milk is 0.8% lactic acid.
MT	किण्वन का पहला चरण 35 डिग्री सेल्सियस तापमान पर होता है, और इस स्तर पर, दूध की अम्लता 0.8% लैक्टिक एसिड होती है।
PE1	फरमेंटेशन का पहला चरण 35°C तापमान पर होता है, और इस लेवल पर, दूध की एसिडिटी 0.8% लैक्टिक एसिड होती है।

Table B.6: Symbols

These kinds of judgement calls also depend on other factors as well. For example, if these symbols occur as part of a mathematical equation, then we would use them without transliterating or in their expanded form.

B.3.1.4 Emoticons

Emoticons such as ☺ may appear in some types of source text and should not be removed. They are more common in social media and web forum texts and their usage and prevalence also depend on the social relationship between the interlocutors.

Note that it could also be the case that MT might not recognize the emoji and you might find it missing. In that case you may need to insert it back during the PE phase.

Source	I just found out that I passed my exams. ☺
MT	मुझे अभी पता चला है कि मैंने अपनी परीक्षा पास कर ली है।
PE1	मुझे अभी पता चला है कि मैंने अपनी परीक्षा पास कर ली है।☺

Table B.7: Emoticons

B.3.1.5 Numbers

It is acceptable to interchange between numerical (101) and word forms (one hundred and one). But given a text to post-edit, the post-editor must strive to be consistent in their edits. So if you see the MT proposals to be regularly using one form over the other, then try to be consistent with that usage in your edits.

Again, a reminder that other factors are also at play. You must be mindful of other points like: how is the chosen usage affecting fluency? is it part of an equation? etc.

In the example below, the post-editor has chosen to convert the word form of the cited year to its numerical form, we consider it to be an acceptable edit, but by the principle of minimizing edits we could have utilized the MT proposed word form as well. Of course, there was a lot more in this MT proposal that needed to be corrected.

Source	Blended learning according to T Barker in two thousand and five means a learning program where more than one mode of delivery is used.
MT	टी बार्कर के अनुसार दो हजार और पांच में मिश्रित सीखने का अर्थ है एक सीखने का कार्यक्रम जहां एक से अधिक विधाओं का उपयोग किया जाता है।
PE1	टी. बार्कर, 2005 के अनुसार, "ब्लेंडेड लर्निंग एक ऐसा लर्निंग प्रोग्राग होता है जिसमें एक से अधिक वितरण-संबंधी विधाओं का इस्तेमाल किया जाता है"।

Table B.8: Numbers

Now let us consider a counterexample where an edit to the MT output was essential. Given the context and the domain, एक बिंदु चार चार does not seem fluent and is not used as such in the target language. We consider the post-editor's edit to be essential here.

Source	The standard size of a floppy is around one point four four MB.
MT	एक फ्लॉपी का मानक आकार लगभग एक बिंदु चार चार एमबी है।
PE1	एक फ्लॉपी की मानक साइज़ लगभग 1.44 एमबी होती है।

Table B.9: Numbers

B.3.1.6 Abbreviations

While Abbreviations are not a common phenomenon in Indian languages, at least in the data we have seen, we do find forms such as the one proposed by MT below - श्री which can be expanded as श्रीमान. But we do not consider them to be abbreviated usages but rather used independently and not in the sense such as the English "Mr." for "Mister", "Rs." for "Rupees" etc.

In example below, we find both the MT proposal and the post-editor's edit acceptable.

Source	For instance, Mr. A is more competent than ten percent of the teacher.
MT	उदाहरण के लिए, श्री ए शिक्षक के दस प्रतिशत से अधिक सक्षम है।
PE1	उदाहरण के लिए, मिस्टर ए दस प्रतिशत शिक्षक से अधिक सक्षम है।

Table B.10: Abbreviations

B.3.1.7 Mathematical Notation

Mathematical Notations are not translated. If any symbol is written as in its notational form, it is not translated.

Source	Ω is the last letter of Greek alphabet.
MT	ग्रीक वर्णमाला का अंतिम अक्षर है।
PE1	Ω ग्रीक वर्णमाला का अंतिम अक्षर है।

Table B.11: Mathematical Notation

If symbols are spelled out or written in words, then transliterate them.

Source	Omega is the last letter of Greek alphabet.
MT	ओमेगा ग्रीक वर्णमाला का अंतिम अक्षर है।
PE1	ओमेगा ग्रीक वर्णमाला का अंतिम अक्षर है।

Table B.12: Mathematical Notation

B.3.1.8 Equations and Formulae

Variables, equations and formulae are neither translated nor transliterated.

Note the first example below where the equation part is not transliterated.

Source	k Pressure P multiplied by volume V equals some constant k
MT	k दबाव k की मात्रा V से गुणा करने पर कुछ स्थिर k के बराबर हो जाता है
PE1	k दाब P और आयतन V का गुणनफल एक कॉन्स्टन्ट k के समान होता है
PE2	k दाब k और आयतन V का गुणनफल एक स्थिरांक k के बराबर होता है

Table B.13: Equations and Formulae

The second and third examples are for a Chemical Equation where it is copied as such from Source. The domain terms should be transliterated. The domain terms in the below examples are transliterated: एथिल फॉर्मेट for ethyl formate, हाइड्रोनियम हाइड्रोक्साइड for hydronium hydroxide, एमिनोलिसिस for aminolysis, फॉर्मैमाइड for formamide.

Source	Water is a weak solution of hydronium hydroxide - there is an equilibrium $2H2O\ H3O++OH-$, in combination with the solvation of the resulting hydronium ions.
MT	पानी हाइड्रोनियम हाइड्रॉक्साइड का एक कमजोर समाधान है - परिणामी हाइड्रोनियम आयनों के उत्थान के साथ संयोजन में एक संतुलन$2H2O3H3O++OHO$ है।
PE1	जल हाइड्रोनियम हाइड्रोक्साइड का एक दुर्बल विलयन है - परिणामी हाइड्रोनियम आयनों के सॉल्वैशन के साथ संयोजन में एक साम्य अवस्था है $2H2O\ H3O++OH-$

Table B.14: Equations and Formulae

Source	Formamide is also generated by aminolysis of ethyl formate: $HCOOCH2CH3+NH3{\rightarrow}HCONH2+CH3CH2OH$
MT	एथिल फॉर्मेट के एमिनोलिसिस द्वारा फॉर्मैमाइड भी उत्पन्न होता है: $HCOOCH2CH3+NH3{\rightarrow}HCONH2+CH3CH2OH$
PE1	एथिल फॉर्मेट के एमिनोलिसिस से फॉर्मैमाइड भी उत्पन्न होता है: $HCOOCH2CH3+NH3{\rightarrow}HCONH2+CH3CH2OH$

Table B.15: Equations and Formulae

B.3.1.9 Phrasal Ordering

Since Indian Languages are free word order languages multiple phrasal orderings are possible for a given target language segment. The guiding factor in these cases should be fluency of the target segment without introducing unnecessary edits to the MT proposal. Thus, an MT proposed order only needs to be changed if meaning is altered, or fluency impacted. Was the post-editor's proposed phrasal reordering necessary in the example below? We think not.

Source	So let's begin with these objectives in mind.
MT	तो चलिए शुरू करते हैं इन उद्देश्यों को ध्यान में रखकर।
PE1	तो चलिए इस उद्देश्य को ध्यान में रखते हुए शुरू करते हैं।

Table B.16: Phrasal Ordering

B.3.1.10 Verb Agreement

There might be confusion and disagreements related to gender agreement between target language verbs and borrowed foreign terminology or noun phrases. In the PE1 proposal below some might disagree with the proposed gender agreement.

Our advice in such cases is to go by what is conventionally acceptable in the target language if the term has been part of the language for a while. If it is a newly borrowed term, then the post-editor must go with their instincts. Some disagreement is expected and inevitable.

Source	We have seen unguided search methods.
MT	हमने बिना खोज के तरीके देखे हैं।
PE1	हमने अनगाइडेड सर्च मेथड्स देखी हैं।

Table B.17: Verb Agreement

At times the gender information could also be extraneous to the segment at hand, for example in a MOOC scenario. In the example below how do we know that the speaker is indeed a Male speaker as reflected by the gender agreement on the verb?

In these cases, typically it is the Translation Project Manager's job to provide such meta information.

Source	Now I present some sample listening activities.
MT	अब मैं कुछ नमूना सुनने की गतिविधियाँ प्रस्तुत करता हूँ।
PE1	अब मैं लिसनिंग की गतिविधियों के कुछ नमूने पेश करता हूँ।

Table B.18: Verb Agreement

B.3.1.11 Domain Terms

There are multiple possibilities when making decisions on how to deal with domain terms. Typically, a post-editor should have been made available with some guidelines by the translation project managers and a list of terms prepared in advance with their transliterations and/or translations provided.

Most CAT Tools and workbenches pick these domain terms from available domain dictionaries and highlight them for the translator, which makes their job easier. The absence of such domain dictionaries might lead to confusion about the correct choice. Consider the example below where 'substrate' was a domain term but had not been highlighted as such during the post-editing activity. We can see the various choices made by PE1 and PE2, whereas MT has just transliterated it. Such varied choices lead to confusion and lower content quality.

To avoid this, we must prepare a domain dictionary in advance and provide either transliterations or translations for the post-editor to use during the post-editing activity.

Source	Now, what are the different categories of fruits and vegetables which can be used as a substrate for fermentation.
MT	अब, फलों और सब्जियों की विभिन्न श्रेणियां क्या हैं जिनका उपयोग किण्वन के लिए एक सब्सट्रेट के रूप में किया जा सकता है।
PE1	अब, फल और सब्ज़ियों की कौन-कौन सी किस्में फरमेंटेशन के लिए आधार के बतौर इस्तेमाल की जा सकती हैं।
PE2	अब, फल और सब्ज़ियों की कौन-कौन सी किस्में फरमेंटेशन के लिए अधःस्तर के बतौर इस्तेमाल की जा सकती हैं।

Table B.19: Domain Terms

B.3.1.12 Style

This section lays out some examples of what we consider as style fixes and our recommendations on avoiding them following the principle of minimizing edits.

Disagreements over gender agreement between verbs and borrowed nouns are common, especially those stemming from recent or rare borrowings. These at times tend to be stylistic preferences when conventions are not yet established in the target language regarding their usage.

In the first example, PE1 and PE2 disagree as to the gender carried by the borrowed word 'dye'. We recommend not making edits like these in such cases.

Source	The dye reduces rotational angle between base pairs.
MT	डाई बेस जोड़े के बीच घूर्णी कोण को कम करता है।
PE1	डाई बेस पेयर्स के बीच घूर्णी कोण को कम करता है।
PE2	डाई बेस पेयर्स के बीच रोटेशनल एंगल को कम कर देती है।

Table B.20: Style

The second example is where there is disagreement about the aspectual information carried by the verb. We contend that it is merely a stylistic preference in this case and should not matter.

Source	Now, the question to you students is what is it that is common in these four morphological types?
MT	अब, आप छात्रों से प्रश्न है कि ऐसा क्या है जो इन चार रूपात्मक प्रकारों में सामान्य है?
PE1	अब, आप छात्रों से प्रश्न यह है कि ऐसा क्या है जो इन चार रूपात्मक प्रकारों में समान है?
PE2	अब, आप छात्रों से प्रश्न यह है कि ऐसा क्या है जो इन चारों रूपात्मक प्रकारों में समान होता है?

Table B.21: Style

B.3.1.13 Bad source language text

Very often, more so when working with text derived from Spoken language which contains disfluencies and repetitions etc. or even with Social media text, we come across imperfect or badly formed Source language segments. A bad Source leads to bad MT outputs and more post-editing is typically required for these kinds of segments.

It is not a post-editor's task to modify Source text for grammar, fluency or meaning. They are downstream in the translation activity and must work with the text as provided to them. In fact, modifying the Source is actively discouraged. The post-editor must make use of all the information at hand to infer the meaning in such cases and try to do their best.

See an example below of badly formed Source text. The post-editor has done a decent job in conveying the intended meaning.

Source	Neural network is come out of AI, is no longer part of AI actually, but still people talk about Neural Network when they talk about AI
MT	तंत्रिका नेटवर्क एआई से बाहर आता है, वास्तव में एआई का हिस्सा नहीं है, लेकिन फिर भी लोग न्यूरल नेटवर्क के बारे में बात करते हैं जब वे एआई के बारे में बात करते हैं
PE1	न्यूरल नेटवर्क्स एआई से बाहर आया है, अब एआई का हिस्सा नहीं है, लेकिन फिर भी लोग जब एआई के बारे में बात करते हैं, तो न्यूरल नेटवर्क के बारे में बात करते हैं

Table B.22: Bad Source

B.3.2 Avoiding Machine Translation Pitfalls

B.3.2.1 Reference Ambiguity

You might find co-references incorrectly inferred by the MT engine. See the example below on how 'they' has been inferred by the MT engine and later corrected to its intended meaning by the post-editor.

Source	So, here real listening has three basic steps. They are hearing, understanding and judging.
MT	तो, यहाँ वास्तविक सुनने के तीन मूल चरण हैं। वे सुन रहे हैं, समझ रहे हैं और निर्णय कर रहे हैं।
PE1	लिहाजा, यहाँ वास्तविक रूप से लिसनिंग के तीन बुनियादी चरण हैं। वे हैं- सुनना, समझना और राय बनाना।

Table B.23: Reference Ambiguity

B.3.2.2 Idioms and Internet Language

Machine Translation often is not very good with idiomatic usage of language. The repetition of the word 'technology' used for emphasis in the Source segment below is carried over literally into the target by the MT engine. But such usage rarely translates directly. The post-editor makes an excellent substitution in this case. (Ignore the slight ungrammaticality in the Source. This is taken from a transcript.)

Source	As we all know that this era is been revolutionized and the reason is technology, technology and technology.
MT	जैसा कि हम सभी जानते हैं कि इस युग में क्रांति हुई है और इसका कारण प्रौद्योगिकी, प्रौद्योगिकी और प्रौद्योगिकी है।
PE1	जैसा कि हम सभी जानते हैं कि इस युग में क्रांति हुई है और इसकी वजह सिर्फ़ और सिर्फ़ टेक्नोलॉजी है।

Table B.24: Idioms and Internet Language

B.3.2.3 Information missing from MT Output

A post-editor must be alert to MT engines missing out on some words, information, or phrases which were there in the Source. Take a look at the example below. Four almost synonymous entities are cited, MT only produces three of them in its output. The post-editor could have easily missed this but doesn't.

Source	In this type of scale, the attitude of an individual towards the matter, thing, object or system can be known from the score of his responses given on a questionnaire.
MT	इस प्रकार के पैमाने पर, किसी वस्तु, वस्तु या प्रणाली के प्रति व्यक्ति के रवैये को उसके स्कोर से जाना जा सकता है। एक प्रश्नावली पर दी गई उनकी प्रतिक्रियाएँ।
PE1	इस प्रकार के स्केल पर, किसी प्रश्नावली पर दिए गए उसके जवाबों के स्कोर से मामले, वस्तु, विषय या प्रणाली के प्रति एक व्यक्ति के दृष्टिकोण को जाना जा सकता है ।

Table B.25: Missing Information

Another example. Are you able to spot what happened here?

Source	It does not assume it to be that way.
MT	यह इस तरह से नहीं मान लेता है।
PE1	यह इसे उसी तरह से नहीं मान लेता है।

Table B.26: Missing Information

B.3.2.4 Extra Information in MT Output

It can also be the case that MT adds some extraneous information in its output that was not originally present in the Source.

In the first example below, do you find "ऐसे" to be extraneous going by the Source?

Source	Factors that hamper listening.
MT	ऐसे कारक जो सुनने में बाधा उत्पन्न करते हैं।
PE1	सुनने में बाधा उत्पन्न करने वाले कारक

Table B.27: Extra Information

And sometimes you see repetitions. In this example below, the additional extraneous repetition changes the meaning of the segment altogether. Post-editors must watch out for such cases.

Source	You might think how could the stripes be different for every zebra then again the fingerprints are different for every person.
MT	आप सोच सकते हैं कि आप कैसे सोच सकते हैं? पट्टीहर ज़ेबरा के लिए अलग होना चाहिए फिर फिर से उंगलियों के निशान हर व्यक्ति के लिए अलग-अलग होते हैं।
PE1	आप सोच सकते हैं कि हर ज़ेबरा के लिए धारियाँ अलग कैसे हो सकती हैं, फिर यह कि हर व्यक्ति के उँगलियों के निशान तो अलग-अलग होते हैं।

Table B.28: Extra Information

B.3.3 Minimizing Edits

B.3.3.1 Systematic Duplicates or Repetitions

Preserving cardinality or sequence information. For example, through use of: firstly, secondly, thirdly etc.

Recall the repetition example we mentioned earlier in Table B.24 and how it did not quite transfer literally to the target in Hindi. You will find below an example of systematic repetition or duplication, that should be conveyed in the target as far as possible.

Source	The first stage is the eggs, next you have the caterpillar, third you have the chrysalis or the pupa and the fourth stage is the beautiful colorful butterfly.
MT	पहला चरण अंडे का है, अगले में आपके पास कैटरपिलर है, तीसरा आपके पास क्रिसलिस या प्यूपा और चौथा चरण सुंदर रंगीन तितली है।
PE1	पहला चरण है; अंडे, अगला आपके पास कैटरपिलर है, तीसरा आपके पास है; क्रिसैलिस या प्यूपा और चौथा चरण सुंदर व रंगीन तितली है ।
PE2	पहली अवस्था है; अंडे, अगली आपके पास कैटरपिलर है, तीसरी आपके पास है; क्रिसैलिस या प्यूपा और चौथी अवस्था सुंदर व रंगीन तितली है।

Table B.29: Systematic Duplicates

B.3.3.2 Dates

The MT output at times might not match the date format in the source. If that's the case, then as before, go with the decision that minimizes the number of edits without compromising on meaning and fluency.

But remember that in some cases, where you are dealing with localization related texts, you might need to follow localization conventions regarding dates in the target locale or language.

All of these should be acceptable, but in the Indian context the *dd/mm/yyyy* convention might be more appropriate. When in doubt go with the Source format.

02/14/2006; 2-14-2006; 2.14.06; February 14, 2006; Feb. 14 '06; Feb/14/2006; 14/Feb/2006

Dates may also be spelled out in part or whole.

Source	This term that is blended learning was initially used in nineteen ninety seven in United Kingdom.
MT	यह शब्द जो मिश्रित शिक्षा है, शुरू में यूनाइटेड किंगडम में उन्नीस उन्नीस में इस्तेमाल किया गया था।
PE1	यह शब्द जो कि 'ब्लेंडेड लर्निंग' है इसका इस्तेमाल शुरुआत में यूनाइटेड किंगडम में वर्ष 1997 में किया गया था।

Table B.30: Dates

B.3.3.3 Punctuation

The target need not mirror the source punctuation. If the MT output has missing punctuation which affect the fluency and meaning of the target output, then insert those as per convention in the target language.

In this example below it was NOT necessary to insert a comma (,) in PE1. Post-editors should avoid introducing extra punctuation.

Source	But serious errors and limitations can be seen regarding e-learning.
MT	लेकिन ई-लर्निंग के संबंध में गंभीर त्रुटियां और सीमाएं देखी जा सकती हैं।
PE1	लेकिन, ई-लर्निंग से संबंधित गंभीर त्रुटियाँ और ख़ामियाँ देखी जा सकती हैं।

Table B.31: Punctuation

Bibliography

Rashid Ahmad, Priyank Gupta, Nagaraju Vuppala, Sanket Kumar Pathak, Ashutosh Kumar, Gagan Soni, Sravan Kumar, Manish Shrivastava, Avinash K Singh, Arbind K Gangwar, et al. 2018. Transzaar: Empowers human translators. In *2018 18th International Conference on Computational Science and Applications (ICCSA)*, pages 1–8. IEEE.

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Mona Baker. 1993. Corpus linguistics and translation studies—implications and applications. In *Text and Technology*, page 233. John Benjamins.

Srinivas Bangalore, Bergljot Behrens, Michael Carl, Maheshwar Gankhot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, and Annegret Sturm. 2015. The role of syntactic variation in translation and post-editing. *Translation Spaces*, 4(1):119–144.

Yehoshua Bar-Hillel. 1954. Can translation be mechanized? *American Scientist*, 42(2):248–260.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Douglas Bates, Martin Machler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardzic. 2016. A comparison between morphological complexity measures: typological data vs. language corpora.

In *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)*, pages 142–153.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, et al. The hindi/urdu treebank project. In *Handbook of Linguistic Annotation*. Springer Press.

Yuri Bizzoni, Tom S Juzek, Cristina Espana-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290.

Patrick Cadwell, Sheila Castilho, Sharon O'Brien, and Linda Mitchell. 2016. Human factors in machine translation and post-editing among institutional translators. *Translation spaces*, 5(2):222–243.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256.

Michael Carl, Akiko Aizawa, and Masaru Yamada. 2016. English-to-japanese translation vs. dictation vs. post-editing: Comparing translation modes in a multilingual setting. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4024–4031.

Michael Carl and Moritz Jonas Schaeffer. 2017. Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. *HERMES-Journal of Language and Communication in Business*, (56):43–57.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Andy Way, and Panayota Georgakopoulou. 2018. Evaluating mt for massive open online courses. *Machine translation*, 32(3):255–278.

Herbert H Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, 12(4):335–359.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Sheila CM De Sousa, Wilker Aziz, and Lucia Specia. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of dvd subtitles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 97–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bonnie Dorr, Bonnie Dorr, Matt Snover, and Nitin Madnani. Part 5: Machine translation evaluation.

Michael Farrell. 2018. Machine translation markers in post-edited machine translation output. In *Proceedings of the 40th Conference Translating and the Computer*, pages 50–59.

William Frawley. 1984. Prolegomenon to a theory of translation. *Translation: Literary, linguistic and philosophical perspectives*, 159:175.

Kent Fredholm. 2019. Effects of google translate on lexical diversity: Vocabulary development among learners of spanish as a foreign language. *Revista Nebrija*, 13(26):98–117.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 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.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. *arXiv preprint arXiv:2004.06063*.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Ignacio Garcia. 2011. Translating by post-editing: is it the way forward? *Machine Translation*, 25(3):217–237.

Federico Gaspari, Hala Almaghout, and Stephen Doherty. 2015. A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives*, 23(3):333–358.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Jesús Giménez and Lluís Màrquez. 2010. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24(3):209–240.

Christian Girardi, Luisa Bentivogli, M Amin Farajian, and Marcello Federico. 2014. Mt-equal: a toolkit for human assessment of machine translation output. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 120–123.

Attila Görög. 2014. Taus post-editing course. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 439–448.

Christian Hardmeier and Liane Guillou. 2018. Pronoun translation in english-french machine translation: An analysis of error types. *arXiv preprint arXiv:1808.10196*.

Ke Hu and Patrick Cadwell. 2016. A comparative study of post-editing guidelines. *Baltic Journal of Modern Computing*, 4(2):346–353.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The opennmt neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pages 102–109.

Věra Kloudová, Ondřej Bojar, and Martin Popel. 2021. Detecting post-edited references and their effect on human evaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 114–119.

Philipp Koehn. 2009. A process study of computer-aided translation. *Machine Translation*, 23(4):241–263.

Maarit Koponen. 2016. Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25:131–148.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1318–1326.

Hans P Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*, volume 5. Kent State University Press.

Samuel Läubli, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, and Martin Volk. 2019. Post-editing productivity with neural machine translation: an empirical assessment of speed and quality in the banking and finance domain. *arXiv preprint arXiv:1906.01685*.

Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, Martin Volk, Sharon O'Brien, Michel Simard, and Lucia Specia. 2013. Assessing post-editing efficiency in a realistic translation environment.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, Sivaji Bandyopadhyay, Mihaela Vela, and Josef van Genabith. 2020. English to manipuri and mizo post-editing effort and its impact on low resource machine translation. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 50–59.

M.T. Mir and S.R. Faruqi. 2022. *Ghazals: Translations of Classic Urdu Poetry*. Murty Classical Library of India Series. Harvard University Press.

Joss Moorkens, Sharon O'brien, Igor AL Da Silva, Norma B de Lima Fonseca, and Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation*, 29(3-4):267–284.

Peter Newmark. 1981. *Approaches to translation (Language Teaching methodology senes)*. Oxford: Pergamon Press.

Nils J Nilsson. 2009. *The quest for artificial intelligence*. Cambridge University Press.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Sharon O'Brien. 2006. Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures*, 7(1):1–21.

Sharon O'Brien. 2011. Towards predicting post-editing productivity. *Machine translation*, 25(3):197–215.

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. 2008. Translation universals: do they exist? a corpus-based nlp study of convergence and simplification. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Research Papers*, pages 75–81.

Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague bulletin of mathematical linguistics*, 93(1):7–16.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović. 2016. chrf deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504.

Maja Popović. 2020. On the differences between human translations. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 365–374.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Tiina Puurtinen. 2003. Genre-specific features of translationese? linguistic differences between translated and non-translated finnish children's literature. *Literary and linguistic computing*, 18(4):389–406.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Ritesh Shah, Christian Boitet, Pushpak Bhattacharyya, Mithun Padmakumar, Leonardo Zilio, Ruslan Kalitvianski, Mohammad Nasiruddin, Mutsuko Tomokiyo, and Sandra Milena Castellanos Páez. 2015. Post-editing a chapter of a specialized textbook into 7 languages: importance of terminological proximity with english for productivity. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 325–332.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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.

Sheila C. M. de Sousa, Wilker Aziz, and Lucia Specia. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 97–103, Hissar, Bulgaria. Association for Computational Linguistics.

Lucia Specia and Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with hter. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 33–43.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Midori Tatsumi. 2009. Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. *The Twelfth Machine Translation Summit (MT-Summit XII)*, pages 332–339.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. *arXiv preprint arXiv:1907.00900*.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018a. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*.

Antonio Toral, Martijn Wieling, and Andy Way. 2018b. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5:9.

Gideon Toury. 2012. Descriptive translation studies: And beyond. *Descriptive Translation Studies*, pages 1–366.

BOLT Activity A Machine Translation. Evaluation plan for phase 2.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. *arXiv preprint arXiv:2102.00287*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. 2020. Word order typology interacts with linguistic complexity: A cross-linguistic corpus study. *Cognitive science*, 44(4):e12822.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. *arXiv preprint arXiv:1906.08069*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.