Learning to Live with Machine Translation

Hoyt Long

> *"I have been relying on the Universal Translator far too much. Before I left Earth, I learned thirty-eight languages, and now all I do is push a button and the computer does all the work."*
>
> --Hoshi Sato, *Star Trek: Enterprise* (2002)

Recent advancements in AI technologies, specifically those engineered for text and image generation, increasingly put the perceived autonomy of human creativity under threat. Although the hype around these technologies usually outruns the reality, the ability of neural-based large language models to generate complex prose, poetry, and even narrative elements mean they are no longer just a thing to laugh at. Each advancement narrows the uncanny valley between human and machine writing, heightening attention to the latter's potential social and cultural effects. Whether these systems are aware of what they are doing is beside the point. As they spread, warn Leif Weatherby and Brian Justie, "the ability to dissent from their conclusions begins to fade, and the gap between their signification and ours closes."[1] Many have begun theorizing the space that remains, arguing for a hermeneutics of automatic writing or revisiting ingrained assumptions about the originality of aesthetic language and writing as creative process.[2] It feels imperative to do so before the systems fade from view and just as society is "in the process, not yet complete, of plugging these powerful index machines into our social-conceptual icons, allowing their expressions to adopt the ring of truth, the benefit of the doubt."[3] And all at the press of a button.

This essay is not about this new family of generative language models, significant as their existential threat may be. It is about their quirky cousins – a family branch that once ruled over the utopian dreams invested in AI but long since dethroned: machine translation (hereafter MT).

Where literary and other creative texts are concerned, the gap between what a human and neural model can do still looks impossibly, often comically, wide. Even as the technology has advanced alongside innovations in language modeling, critics are quick to point out its failures when directed at anything more challenging than a dinner menu or business email.[4] As a consequence, MT has rarely been taken seriously as an object of theorizing except as a naïvely idealistic, mechanical attempt to substitute culture-blind engineering for the messiness of cross-linguistic exchange and the subtle layers of human interpretation it requires. Or as Brian Lennon describes the rise and fall of MT in the 1950s and 1960s, "the story of MT is the story of an attempt to assert the independence of computation from culture and, at the same time, to assert computation's dominion over culture."[5] MT only ever merits dissent in this story because it is seen as not really translation at all, insofar as the search for equivalent expressions between source and target languages is understood to entail an individual, and also ethical, decision about how far to narrow the semantic gaps that linguistic and cultural differences introduce.[6]

It is time that we update this story, however, fixated as it is on an ideological vision of MT (i.e., as purely automatic) that has evolved substantially since those heady early days when it symbolized the future of AI. The scholar Erga Heller has suggested that the evolution loosely maps onto how translation technology is represented in *Star Trek* (1966) and its subsequent spinoffs, specifically in terms of the need (or not) for human intervention. The original series imagined a "Universal Translator" capable of simultaneously translating all human and alien languages, both spoken and written. It took the shape of a computer on the chief deck that ran with minimal assistance from Lieutenant Uhura, a translator with expertise in linguistics and cryptology who did little more than operate the device. By *Star Trek: The Next Generation* (1987-1994), the UT had morphed from a piece of hardware to a software program that ran on

portable devices. Yet the ship's crew increasingly encountered situations where human mediation was necessary to bridge cultural conceptual gaps that the UT could not cross by itself. In *Star Trek: Enterprise* (2001-2005), the UT is for the first time accompanied by a human specialist, Hoshi Sato, a professor of linguistics who is essential to making the technology function to its full potential and who, as in the epigraph, recognizes the value of her own expertise.[7] If the fantasy of a general-purpose automatic translation device persists, it is more and more imagined as entangled with human input rather than as an autonomous tool, a shift that has the effect of expanding ideas of translation itself. This essay considers what a critical response to MT might look like if framed around this more recent vision. How do the stories we tell about MT change when we see it as a potential collaborator, not mere mechanical replacement?

As much as critical theorists have clung to the latter view, the notion of MT as inherently hybrid technology has been around, in fact, from the start. Well before the dreams of a pure MT came crashing down with the infamous ALPAC report in 1966, Yehosua Bar-Hillel, one of the field's founding researchers, advocated for a "machine-post-editor partnership." In 1980, Martin Kay, also a prominent figure in the field and vocal critic of pure MT proposed the "Translation Amanuensis," an interactive editor that would assist the human translator by offering suggested translations based on MT output or memory banks of past translations. It became a template for the forms of computer-assisted technology that eventually took over the translation profession, so much so that Douglas Robinson declared in 2003 that "all translators are cyborgs" because all translation falls in-between the hypothetical abstractions of purely machine and human translation.[8] The increasing acceptance of this cyborg vision is related to real improvements in MT technology, at least for certain domains and certain types of texts. When MT becomes "good enough" in a particular translation setting, where this might mean good enough to be post-edited

or good enough to convey the gist of a source text, the potential "partnerships" between human and machine start to proliferate. Indeed, the expectation that human ingenuity will always fill in for "less than intelligible" MT has, argues Michael Cronin, legitimated the roll-out of online translation services so that anyone with a smartphone or access to social media can interact with MT at the push of a button.[9] The future imagined in *Star Trek* is here, all of us like Hoshi Sato learning to live with MT by helping it work while forgetting what it was to live without it.

It is here not as the science fiction version of the technology, both perfect and perfectly opaque, but in the sense of being inescapable even in its imperfection. Researchers in the broader field of translation studies have studied the implications of this ubiquity for years now, including the integration of MT tools into professional translation settings, their adoption by multinational corporations for managing information flow and content localization, and the impact of online translation services – facilitating as they have new forms of crowd-sourced and collaborative translation – on the global circulation of news and cultural material.[10] Others have investigated the inevitable repercussions of MT for the value of translation work and explored ways to nurture a critical literacy of the technology and the ideologies about language that it encodes.[11] In light of this research, the lack of attention given to MT by translation theorists is all the more striking. How good must the machine translation of literary and other creative content be before it is taken seriously and not simply as oppositional strawman? Or to ask the question in reverse, what critical interventions go missing by refusing to traverse those edges where the technology's gaps are most visible? To rule out the study of MT is to be bound by an obsolete dichotomy between purely mechanical and purely human translation. And, as Lawrence Venuti says to those who would stigmatize or rule out the study of translation altogether, it is "to abdicate to the status quo by withdrawing from the areas where social struggles can occur."[12]

It will be hard to reckon with these sites of social struggle so long as MT is imagined as what Cronin calls "pure, allographic practice." That is, as a "machine materially executing in another language the script of the source language," and as opposed to the "autographic," characterized by an artisanal or handmade quality.[13] This opposition does not reflect how the technology is actually used, which is within hybrid assemblages that harbor no false pretenses about the possibility of perfect translation. Nor how it actually operates, which is as the entangled learning of humans with algorithms whose authorship, following Louise Amoore, must be read as "multiple, continually edited, modified, and rewritten through the algorithm's engagement with the world."[14] If MT's actual existence is one of dynamic entanglement with human translation and socio-technical systems, it is in our theoretical interest to learn how it mediates cross-lingual exchange even as it fails to be a universal translation device. If, as Venuti writes, "no translation can be understood as providing direct or unmediated access to its source text," and if all translation "is an interpretative act that necessarily entails ethical responsibilities and political commitments," we should try to understand how MT is implicated in this act as a mediating agent and as already wrapped up in its ethical-political entailments.[15]

To do so will require taking the technology seriously, which means first grasping how it works in its recent neural instantiation. This essay begins with an overview of the technology and the theory of translation underwriting it. It then proceeds to survey the edges where gaps in the technology are most visible, using literature as a limit case to build a stronger intuition about where current MT is prone to expose these gaps and where it effectively conceals them. This can only be done through the close, qualitative analysis of machine translated literary texts, which I do here by applying evaluative methods used by MT researchers to a small sample of Japanese fiction. This analysis affords a granular assessment of MT "errors" that avoids the absolutist

logic of right and wrong, or the essentialist logic that says no translation is ever enough, opting instead for a continuous, contextualized understanding of translation quality that leaves conceptual room for good enough. Reading MT output in this way means redirecting attention to what translation can do despite all the ways it is imagined to fail.

The final section of the essay takes a speculative turn and considers what good enough machine translation of literary texts might be good for in a future of ubiquitous and ever more accessible MT. Not a future, to be clear, where machines translate literature without us. But a future where, according to Alan Liu, it is conceivable that neural-network translation will allow us "to analyze text in multiple languages based on their intersecting 'interlingua' — machine-generated, emergent, and transitional language forms that are a kind of pure comparatism."[16] I test this assertion by reading across the machine translations of several hundred works of fiction originally written in German, Spanish, and Japanese. The results point to more immediate ways that MT invites inquiry into the present conditions of world literature, but also a future where the entanglement of human translation and agency with the material agency of the technology bring forth potentials in both – for how we think about the task of literary translation and how we learn to live with what MT can and cannot do for this task.[17]

The Probabilities of Machine Translation

Neural machine translation (NMT) became the reigning paradigm in MT research around 2017, though its emergence signaled less an epochal shift than an extension of statistical machine translation (SMT). The latter revolutionized the field in the early 1990s, transforming it from a largely rules-based to a probabilistic endeavor.[18] This revolution was catalyzed by the growing availability of digitized text and, specifically, parallel or aligned corpora that paired source text

segments with authoritative human translations. Such corpora were already used by professional translators in the form of "translation memories," essentially searchable databases of prior translations that increased translation consistency and efficiency in organizational settings. The growth and consolidation of such corpora made it possible to imagine translation as a process of predicting the likeliest target language sequence given a source language sequence, at least within specific narrow domains. In short, translation as a collective voting mechanism whereby past translation choices carry more predictive weight the more often they are repeated.[19]

Such a theory of translation arguably amounts to no theory at all, especially as meaning and interpretation seem to fall by the wayside. "Few translation theorists," Dorothy Kenny has observed, have tried to draw serious parallels between how SMT and human translation operate, or to argue for a common approach to meaning within the two fields."[20] This is surely because SMT (and now NMT) can feel so alien. These approaches have no way to account for linguistic and cultural contexts external to the source text; treat the sentence as primary unit of translation, eclipsing longer range dependencies; and assume that past translation choices can predict future ones, an assumption that seems tenuous where past examples are few and highly variable. Yet as statistical systems train on ever larger collections of translations (the Internet has been crucial in increasing the size and diversity of aligned corpora) and as translators further fold translation memories, MT output, and other such technical appendages into their workflows, one might wonder if there isn't some aspect of translation that is probabilistic after all, at least up to a point. Past translation theorists have wondered as much, questioning the inherent redundancies in language and the degree to which translation is a norm-governed behavior.[21] NMT raises these questions anew, retaining all that is problematic about a statistical theory of translation (i.e., that it is a collective voting mechanism; that this collective can be sufficiently representative; that its

votes stand for authoritative translations) while operationalizing the theory in ways that change how we read and reason with its output.

Although NMT models come in all shapes and sizes, they share a core architecture that maps the relation of a set of inputs to predicted outputs using artificial neural networks. Where classical SMT uses the frequency of discrete words and short phrases to perform this mapping, NMT uses neural networks to create distributed representations of individual words that take their immediate context into account. That is, it "sees" the meaning of a word in a given source text (and an aligned target text) as a function of the words surrounding it, which then informs the statistical mapping NMT creates between input and output. The end result is a mapping that better accounts for how the probability of a translation will vary with the presence or absence of certain words and where probability is assessed nonlinearly, such that multiple words can be similarly plausible according to the model. In effect, NMT provides a more robust model of polysemy and ambiguity within source and target texts as well as between them.[22]

This gets us slightly ahead of ourselves, however. For these distributed representations are merely the building blocks of a more complex process that learns these representations and selectively feeds off them as it models a language or language pair. In the monolingual case, where the primary task is predicting the word that comes next in a given sequence of words, the probability of a word is based not just on the words before and after it, but on parameters that can control how much of this context is seen and how much weight to assign closer versus more distant words. As this process moves from one word to the next, it is thus feeding off multiple inputs that include the immediately previous word and some controlled representation of the words before and after it in the sequence. This basic process is also part of how NMT systems model language pairs, but now with the added task of predicting target words based on a source

sequence and the challenge of controlling for context when it includes information about source and target sequence alike.

This modeling process can be broken down into three major components: an “encoder,” “decoder,” and “attention mechanism.” The first creates a representation of the input or source sentence; the second carries out the translation prediction task with information about the input sentence and any parts already translated; and the third regulates how information flows from encoder to decoder. Given a source-target pair from the training data, the model first encodes the source sentence as described above, representing the words as a function of their left and right contexts. This representation is passed to the decoder, which steps through the target sentence trying to predict the next word. It does so using what it knows about the source sentence and any information it has gleaned from the target sentence up to that point (e.g., the previous predicted word; a representation of the predicted sentence thus far). The attention mechanism is used to control what the decoder “knows” about the source sentence by learning which words are most relevant to predicting the next target word. In essence, it varies the attention given to words in the source sentence according to the decoder’s position in the target sentence. Together, these components try to set their values such that they maximize the probability of predicting the “correct” target word as defined by the source-target example they are learning from. Repeat this across millions of examples and the result is a base translation model for a given language pair.

Now begins the actual process of translation (or “decoding”), which works similarly to the decoder, plugging information about the input sentence and any already translated words into the model to obtain a probability distribution of the words most likely to come next in the output. That is, at each step, the model returns a list of words ranked by their probability of being a suitable continuation of the sentence being translated. In many cases there will be a clear winner;

in other cases, there may be several equally likely candidates. Yet even in the case of a clear winner, what is highly probable at one moment in the process may be less so later, and vice versa. How, then, is the model to decide? One way is to keep track of multiple paths through the garden of possibilities, computing at each step the combined probabilities of selected words and keeping the highest ranked sequences. These sequences are then fed into the next prediction. At the end of the sentence, a model will simply choose the sequence with the highest probability.

Engineers have explored any number of ways to vary the decoding process, constrained as it is by a sequential logic that so heavily depends on the choices made prior to each successive prediction. One way of doing this is to have multiple models weigh in at each decision point, averaging across them to generate a combined probability distribution. Others have sought ways to diversify the final list of best sentences by penalizing the most probable sequence or by reranking the candidates according to alternative models.[23] However, the underlying principle remains the same. For NMT, the act of translating is one of memorizing a vast repository of past human translations, funneling whatever patterns it learns through the narrow task of word-by-word translation, and selecting a final translation from some subset of possible candidates.

By design, NMT suffers some of the same drawbacks as SMT, most notably a reliance on massive parallel corpora presumed to be authoritative. Indeed, NMT models are even more data hungry than their predecessors, which means the world's dominant languages and language pairs are the most likely to benefit.[24] NMT also struggles, if to different degrees, with issues that have dogged statistical approaches from the start: anaphora, idioms, various kinds of word ambiguity (semantic, morphological, structural), and a blindness to long range contextual dependencies in a text and to real world knowledge outside it. With NMT also come new drawbacks. For instance, it is known to privilege fluency over accuracy, creating translations that sound natural but distort

the content of the source text. Neural models are also engineered in such a way that their decisions are largely opaque to interpretation and understanding, making it hard to trace errors back to specific phrases or subsegments in the source.[25]

In other respects, however, NMT has shown enough improvement to warrant its broader adoption as state-of-the-art. The gains are hard to generalize since "the results of any comparison between an SMT and an NMT system may vary depending on how similar the training data are to the actual texts to be translated, the language pair, and the specific machine translation configurations used."[26] Yet evaluative studies thus far indicate that it does better with resolving word sense ambiguity (which is expected given the distributed representations it uses); generates fewer errors with word order, but also fewer morphology and lexical errors; and is generally better at creating more fluent translations which, as already noted, can be a double-edged sword when accuracy is sacrificed in the process.[27] This may be cold comfort, however, to those who doubt that statistical approaches can approximate anything like what human translation achieves. Neither does it tell us how such approaches fare differently across language pairs, knowledge domains, and genres of writing. To make any claims for the value of NMT, we must step back to a more basic question: what makes for a "good" translation, machine or otherwise?

Lost and Found in Machine Translation

The problem of how to evaluate machine translation output has long vexed researchers in the field. They know evaluation can be highly subjective and vary with translation context, but they also recognize the need for consistent measures to compare a model's performance against human translation and other models. Qualitative approaches to the problem have resulted in rich taxonomies of error centered around the accuracy and fluency of MT output, but these tend not

to scale well as the number of texts and evaluators increases. Quantitative approaches have tried to scale up evaluation with automatically calculated metrics such as the BLEU score (Bilingual Evaluation Understudy), but they also reduce "good" translation to a measure of overlapping substrings between model output and some human reference translation.[28] Until recently, both approaches generally treated translation quality as an absolute goal with a single optimal solution for any given translation task. Yet with the increasing accessibility and ubiquity of MT there has been a noticeable shift to thinking of quality as a judgment dynamically realized according to knowledge domain, intended use and audience, the availability of resources, monetary concerns, and similar situational constraints. Quality is evaluated contextually and on a continuum, in other words, which in practice has meant an acceptance of varying degrees of "rawness" in MT output and a recognition that even "gist" or "indicative" translations can sometimes be good enough.[29] Scholars in translation studies have raised concerns about the consequences of this conceptual shift for how the intellectual labor of human translation is itself (de)valued in specific corporate or institutional settings.[30] But it is just as important to pursue the implications of this shift for the imagination of human-machine partnerships in translation and the ways MT might be brought to bear on literary and other creative texts.

Such texts were long understood to lay outside the scope of MT research proper, which was also a way to delimit certain kinds of "technical" translation as free from aesthetic concerns. This hardline stance has loosened, however, with the shift in attitudes toward quality. Some have approached the issue from the literary translator's perspective to ask what, if any, benefit MT offers, looking at how output from different systems affects the translation process or how post-editing of this output can alter a translator's style.[31] Others have tried to adapt existing metrics and assessment criteria to the evaluation of literary MT or, more commonly, have developed new

taxonomies to identify accuracy and fluency errors across different language pairs and literary genres.[32] This section builds on these efforts by adapting and applying one taxonomy to the machine translation (into English) of a handful of Japanese-language texts in order to test several hypotheses about how NMT models carry out the task of translation. A careful reading of the output can strengthen our intuition about what the specific affordances of this technology entail for literary texts. What sort of partial, fractured representations of a text come from forcing the elephantine memory of prior human translations through the eye of a probabilistic needle? When do we find ourselves giving MT the benefit of the doubt? When does it fail to convince?

Four texts were selected for evaluation: Natsume Sōseki's *Kokoro* (1914), a novel about a young student's awakening to the harsh realities of modernity; Hayashi Fumiko's "Shitamachi" (Downtown, 1949), the story of a young widow struggling to survive in Tokyo's postwar ruins; Murakami Haruki's *Noruwei no mori* (Norwegian Wood, 1987), in which the male narrator recounts his coming-of-age during the 1960s student movement and his formative relationships with two women; and Kawakami Mieko's *Heben* (Heaven, 2009), about a student mercilessly bullied by his peers and which presents readers with a deeply philosophical account of the everyday violence underpinning social relationships. The temporal range of these works is meant to test for historical bias in NMT models, which are likely less well suited to older texts given the contemporary data they are trained on. I also selected works which vary in narrative structure and style to evaluate how NMT handles first- versus third-person, different narrative frames (e.g., description of events, interior thought, dialogue), and different levels of semantic richness. Although hardly a representative sample, the hope is that such temporal and formal variation will provide different openings onto where NMT models fall short with literary texts, how the models might be modified, and where they expand our thinking about translation. When do errors help

clarify the interpretative work that human translators perform? Conversely, when does their absence reveal how far probabilistic strategies can carry this act?

Before delving into these details, however, it is useful to gain a more holistic picture of how and how often NMT models stray from what would be expected of a human translator. For each of the sample texts, the first 100 sentences were translated with models available through Google Translate and Mirai Translate. The latter is a subscription service designed and marketed for Japanese corporate clients and is thus presumably better optimized for Japanese-to-English translation.[33] The output was then evaluated for fluency with the SCATE taxonomy, which uses the following general categories – coherence, grammar and syntax, lexicon, style and register, and kinds of mistranslation – to identify 21 kinds of error.[34] A few key observations emerge from this evaluation, carried out by myself and a second native-English speaker fluent in Japanese. First, there are far fewer sentences with identified errors in the more contemporary texts, and generally less in the Mirai output than in Google. Table 1 shows what percent of sentences were marked as having some sort of error. Second, the bulk of errors are generally of the following type: co-reference, logical problem, lexical choice, and disfluent construction. The first is used when entities in a sentence are mismatched or misidentified (e.g., he becomes she); the second for output that is illogical or confusing given the rest of the sentence; the third for words whose meaning is clear in context but are not idiomatic; and the last is similarly used to identify grammatically correct sentences that are otherwise difficult to read or could be rendered more idiomatically. Figures 1 and 2 show the distribution of all error types in one set of evaluations. Despite considerable variation across the four texts and two models, overall interrater agreement was in the range of 60-70%.[35] In the majority of cases, then, there was agreement at the sentence

level about the broad category of error that applied, suggesting the variation cannot be explained by subjective judgment alone.

Text	Model	Evaluator 1	Evaluator 2
Kokoro	Google	86%	91%
Kokoro	Mirai	75%	85%
Downtown	Google	84%	89%
Downtown	Mirai	86%	89%
Norwegian	Google	58%	58%
Norwegian	Mirai	53%	49%
Heaven	Google	68%	70%
Heaven	Mirai	61%	61%

Table 1: Percentage of sentences containing at least one identified error for each NMT model.

To evaluate accuracy, defined here as the degree to which information in the source text is reflected in the target output, each translated sentence was judged against its source on a ten-point scale. A 10 means the output is determined to convey 100% of the source text information, regardless of how well stylistic elements are preserved. Table 2 gives the average score for each text by three evaluators along with the *average deviation* index, which provides a measure of interrater disagreement.[36] Only the Mirai output was evaluated for accuracy after observing its generally higher quality in the fluency evaluation. The scores reinforce the finding that newer texts contain fewer errors, but also that the model had the most difficulty with the one work not narrated in the first person. They also reveal how surprisingly accurate a model can be with some texts (i.e., *Norwegian Wood*), raising the question of what makes some texts more amenable to machine translation than others.

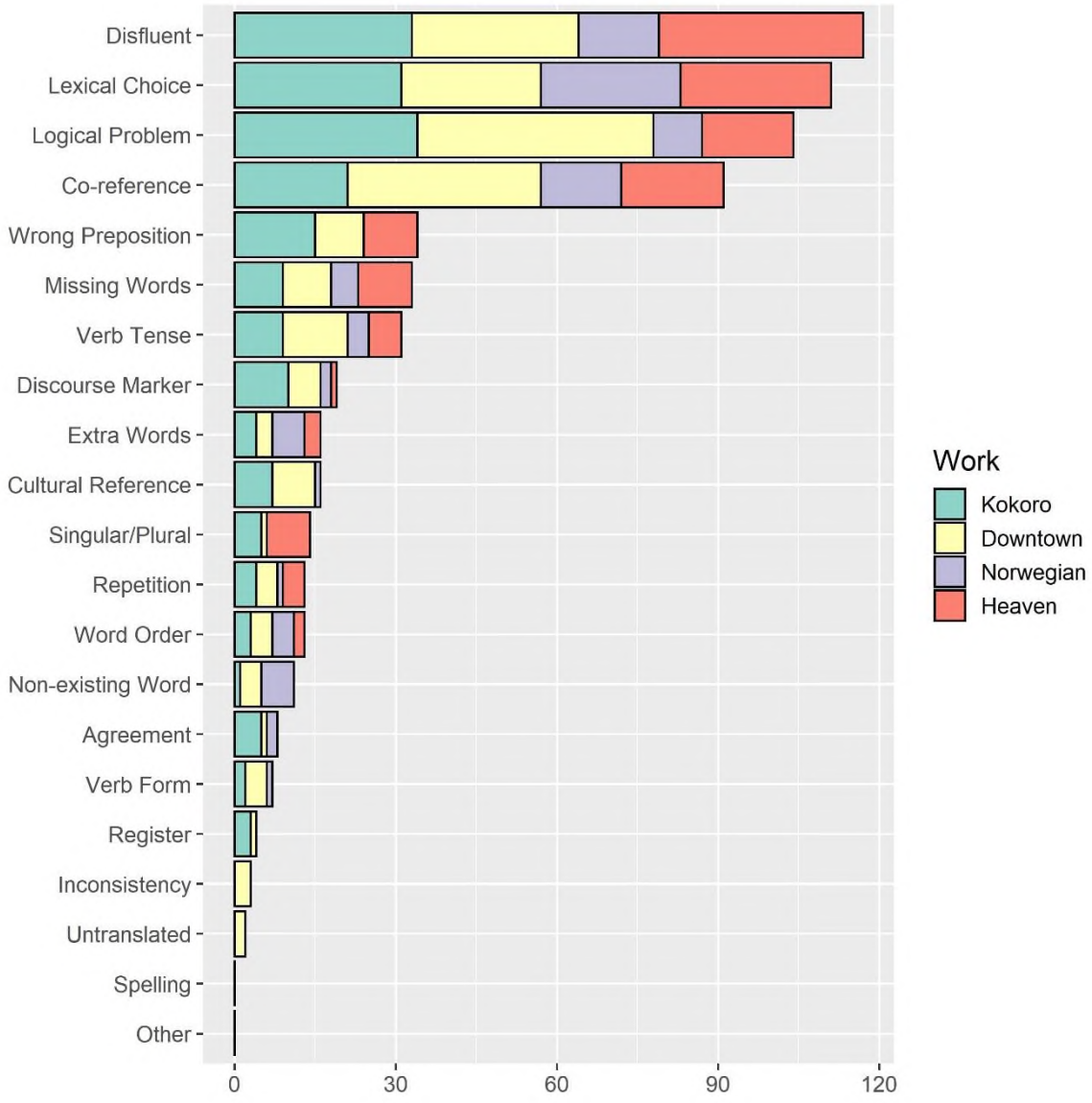

Figure 1: Number of translation errors by type (Google output)

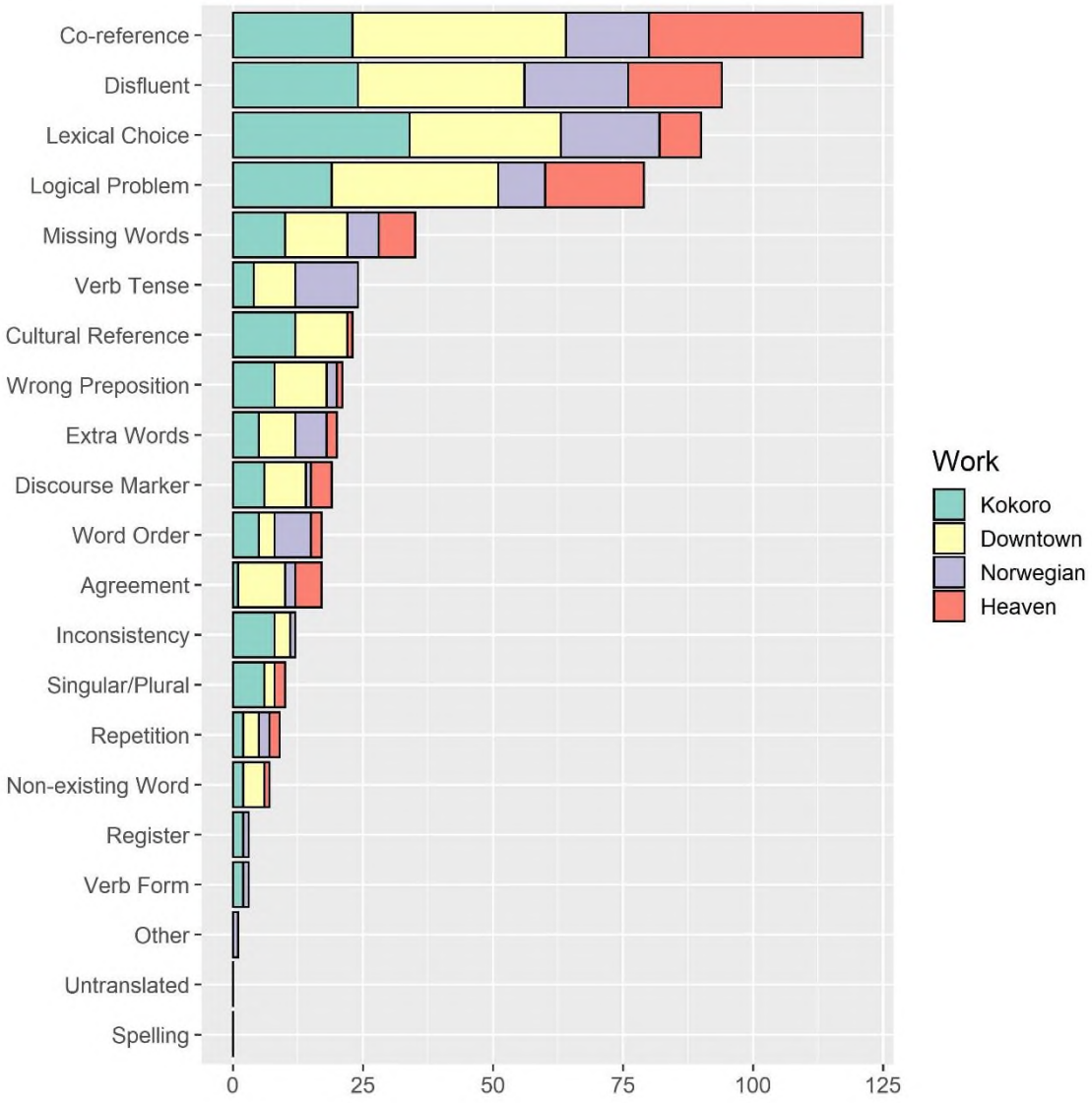

Figure 2: Number of translation errors by type (Mirai output)

Text	Evaluator 1	Evaluator 2	Evaluator 3	AD Index
Kokoro	7.20	7.80	7.70	0.35
Downtown	5.46	7.68	6.33	0.47
Norwegian	8.27	9.28	7.75	0.33
Heaven	7.70	8.84	8.02	0.30

Table 2: Average rating of evaluators for all sentences in each work. The AD index measures the relative level of interrater disagreement, with lower values indicating that more similar ratings were given by evaluators.

This high-level analysis gives us some intuition about the type and frequency of errors to expect with NMT models for different kinds of texts. Sharpening this intuition requires a closer reading of the errors and the effects they produce vis-a-vis the source text and target language. A common source of error are rare or unseen words that are missing from, or have been filtered out of, a model's training data. This is especially pronounced with *Kokoro*, in part owing to its age. In the following output from Google Translate, for instance, the model is at a loss for how to deal with 猿股 (*sarumata*), a traditional male undergarment in Japan. In this scene, the young narrator is entranced by a Western man at the beach sporting one of these: "Wearing a pure Japanese yukata, he stood facing the sea with his arms folded, leaving it thrown on the floor. He didn't wear anything other than one of our *monkey crotch*."

Having no knowledge of the word, the model divides it into its constituent characters and translates them separately, creating a clothing item whose indecorous name is all the more ridiculous for its apparent cultural representativeness ("one of our"). It could have taken the approach it does with *yukata*, offering a transcription that leaves the reader to glean the meaning from context and/or assumes the word has acquired the status of a common loanword. This is the approach taken by Mirai Translate, though *sarumata* is not likely to be recognized by readers unfamiliar with Japanese. Indeed, past translators have opted for "thongs" (which casts a similarly primitivizing gaze) and "drawers" (which preserves semantic distance with a word more befitting the time of the novel).[37] Both choices highlight the complex calculus involved in trying to align an interpretation of the narrator's own perspective with what is likely to produce a similar effect for target readers, especially in the case of uncommon words. The model, however, has only past precedent to go on.

There are, however, ways to override the model's best predictions for specific words and phrases with hard-coded equivalents. There are also ways to split words into smaller units so that rare sequences of characters found in the training data (typically proper nouns and numbers) can be identified and handled separately. Such sequences can then be copied over as is or else passed to a dictionary.[38] The latter strategy is clearly part of the Mirai model, as it sometimes includes a parenthetical definition for culturally specific terms (e.g., 掛茶屋 is transliterated *kakejaya* and then glossed as "tea house where geisha entertain their guests"). Naturally, such hard-coded rules and dictionaries create other issues in that they have to be constantly updated or otherwise adapted for different translation contexts. In the case of literary translation, where the incidence of rare words and culturally specific terms is likely to be greater than with other genres, a literal approach may be more useful, signaling that a word requires further inspection and that its non-translation is worth deliberating over.[39] We could even imagine having a lever that adjusts how much non-translation the model is allowed to perform.

Such a lever could only be pushed so far, however, and at some point one has to follow the model's compulsion to make decisions in the face of uncertainty and insufficient contextual knowledge. In the case of "Downtown," this compulsion propels the model off the narrative rails to the point that it loses track of who is who. What emerges here is the problem of co-reference, a longstanding challenge for MT. Japanese is a stark reminder of how acute the problem is given the ease with which subjects can be elided. It does not help that Hayashi's text is particularly loose in its specification of subjects and in its narrative focus, often sliding fluidly from one narrative situation to another with little warning. In one early scene, the main protagonist Riyo is asked about her husband's whereabouts, which sends the narration wandering across the emotional turbulence of the past few years before it snaps into focus with her answer over a

dozen sentences later. Both Google and Mirai struggle to keep up as subjects quickly shift and the narration flips between third person and snatches of free indirect discourse. Here are several lines from the Google ouput:

> "Where is he in Siberia?" There was a message from Baikal's Suchin, autumn has passed, and this winter has passed. It's become a habit for Riyo to wake up every morning and feel depressed. *I* don't really feel *it* because *it's* too far away, but *I'm* getting used to the fact that *I* don't feel *it* anymore. *He* said that a song called "Foreign Hill" was popular, so *he* asked Tomeyoshi to sing it, but as *he* listened to the song, Riyo became lonely. *He* wondered if there was still a feeling of war around *him*.

The italicized words indicate co-reference errors where the model has attempted to fill in for the lack or under-specification of the subject. In the second sentence, there is no subject specified at all in the Japanese, allowing for a seamless merging of Riyo's and the narrator's voice as we are jolted away from the conversation at hand. Although the tense is off, Google's version hues close to the original in neglecting to give the sentence a subject. When Riyo drops out again in the fourth sentence, however, the model handles the ambiguity with the first-person singular before it switches to a third-person account that is incorrectly gendered. It also leaves us to guess about the object of her feelings (i.e., her husband), which is elided in the original but obvious from the larger context.

While the basic facts of the passage are conveyed, then, the model loses track of who is doing what to whom, a result of how it sequentially processes text at the sentence level. There has been work to overcome these limitations by distributing model attention over larger contexts or translating multiple sentences at once, but strategies of this sort can be computationally expensive.[40] And they do not get around the more basic problem that literature is unique in the

density and cohesiveness of its coreference chains relative to other types of text.[41] There are not only more entities mentioned within any given stretch of text, but the links between them are sustained over longer spans. With languages like Japanese, where these references are elided at higher rates compared to English, NMT models are hampered both by an inability to see beyond the sentence *and* the higher possibility that a sentence has no coreference information at all.

Coreference errors need not always lead to nonsense, however, as we see with *Norwegian Wood*. Murakami is well known for his habit of writing Japanese as if writing English, and in this novel he over specifies subjects such that the model has more information to work with, thus producing fewer coreference errors.[42] In fact, there are fewer errors across all categories. What does it say that Murakami's style is apparently well suited to machine translation into English? Is this relative "translatability" indicative of his predisposition to write with the global translation market already in mind? It would be interesting to pursue these questions on their own, but here I want to use this apparent translatability to delve into higher order aesthetic concerns. When an MT model is less encumbered by grammatical structure, how else does its probabilistic logic and translation memory guide it to make decisions that human translators might deem unreasonable or insufficient? Can its output still be aesthetically pleasurable despite these choices?

One category where *Norwegian Wood* does not shine above the rest is "lexical choice," where the meaning of a word is clear in context but deemed not idiomatic in English. This can be a highly subjective determination, entangled with an evaluator's own unique relation to English and their aesthetic expectations for the text in question. These are part of the translator's internal calculus too, calibrating as they must how diction and other stylistic elements can best be aligned with the imagined norms of a translation's intended audience. Such sensitive calibration cannot be expected from MT models keyed to the normative patterns in their training data. Yet when a

word does feel wrong (or right), what might this say about probability as a way of resolving the semantic ambiguity unleashed by translation?

Consider the following passage, in which the male narrator digs through his memory for images that he eventually welds together into a composite vision of a past lover. Here I use the Mirai model output as the base translation while also showing how the model's lexical choices align with the Google output (second position) and translations by Jay Rubin and Alfred Birnbaum (third and fourth positions).

> She had *little / small / tiny / small* cold hands, *fair / clean / black / sleek* hair that was *smooth / smooth / smooth / silky* and straight, *soft / soft / soft / full-fleshed* and round, a *little / small / microscopic / tiny* mole on her earlobe and just below it, the elegant camel coat she used to wear in the winter, the habit of always looking into the other person's eyes when asking questions, and the occasional *tremulous / tremble / trembling / tremble* voice (as if she were talking on a windy hill), and so on, and as [such images] *piled / piled / joined / built* up one after another her face began to rise spontaneously. [Her] profile first appeared. This is probably because I and Naoko always walked side by side. Then she turned to me, smiled, *turned / shook / tilts / tilts* her head slightly to one side, talked to me, and looked into my eyes, as if she were *looking for / looking for / trying to catch / gazing after* the shadow of a *little / small / -- / tiny* fish which *flitted / glanced / darted / darted* across the bottom of a *clear / clear / limpid / crystal clear* spring.

Reading the choices of machine and human translators together illuminates how a probabilistic approach often reproduces the latter's choices but is also more constrained in its semantic range. It is at once more literal and more repetitive, deciding always to translate the same word in the

same way. Thus, for instance, the word 小さな (*chiisana*) is always rendered "little" by Mirai and "small" by Google, whereas Birnbaum and Rubin add "tiny" and "microscopic" to the mix. Adding to this conservatism is the model's adherence to the page, so that it lacks the liberty to turn "fair" into "sleek," or "soft" into "full-fleshed." Presumably the models could make such semantic leaps if exposed to enough examples in the training data, but such unusual and creative solutions never rise to the top. Still, there are times when literalness proves just as evocative a choice, as with the images of Naoko that "pile" up and the little fish "flitting across" the bottom of the clear spring.

The model's literalness is a function of the translations it has seen, the way it attends to other words in the sentence, and how all this gets computed into a vector of probabilities from which a single choice must be made. If one had access to all this machinery, one might alter the model's final calculus by comparing multiple models or creating more interactive systems that display the most probable translations of a word much as with predictive text.[43] Another option would be to expand the model's training data with specialized parallel corpora, using statistical patterns learned from this "in-domain" data to temper the more dominant patterns learned by generalized models. Hypothetically, one could collect a corpus of Japanese to English literary translations, perhaps with an emphasis on works in the romance or YA genres, and condition word predictions more heavily on patterns in this corpus. Doing so might then push the model towards the more diminutive, feminized language found in the Rubin and Birnbaum translations, like applying a style filter to the output in order to redirect its literal-mindedness.[44]

Such a scenario assumes ready access to large amounts of literary parallel corpora and the computing power necessary to build models from scratch. Lacking any incentive to marshal such resources for highly particular adaptations, it's likely that we will continue to have to live with

the flat-footed aesthetic of generalized models.[45] But this may not always be a bad thing. A final error type suggests how it could be instructive, revealing aspects of the source language easily domesticated by human translators but which can remain in MT output when it fails to *solve* for higher order grammatical, semantic, and other complexities. These often show up as "logical problems" in the fluency analysis, an open-ended error category meant to capture confusing or illogical output. It thus becomes a grab bag for instances when the literal rendering of idiomatic words and phrases, or a stubborn reliance on the original word order, produces expressions at once non-sensical and surrealist. Yet it is not always a straightforward matter to distinguish MT output that confuses the meaning of the source text from that which produces confusion because it sticks too close to it. This is especially true with Kawakami's *Heaven*, whose spare, abstract style already leaves readers to second-guess the reality of the world inhabited by the narrator and the female student, Kojima, who befriends him. When the two social outcasts have their first face-to-face meeting in a neighborhood park, the narrator's description of the scene feels comically absurd. Here is how Mirai translates it: "There was a bench made of sideways tires and a concrete whale between which there was a sandbox of about three tatami mats filled with boxes of sweets and plastic bags."

Lacking access to the source text, or a familiarity with urban parks in Japan, one might reasonably question the logicality of this sentence. But the translation by Sam Bett and David Boyd reveals that Mirai has not missed all that much. "There was a kind of bench made from tires on their sides, and a concrete whale, and between them a sandbox not much bigger than a mattress, littered with candy wrappers and plastic bags."[46] The Mirai version is less polished in how it strings the objects together, but it does capture them all while producing a foreignizing effect by retaining the traditional unit for measuring small areas (e.g., tatami mat). Something

similar happens later on, when the narrator pulls back to comment on the overall flatness of the world around him. Here is Google's rendering of this moment: "It was a flat landscape as usual, with no depth. And as I always do, I cut the scenery in front of my eyes into squares like a picture-story show, and every time I blinked, I flipped each one under my feet."[47] Here is Bett and Boyd's translation: "As usual, the world was flat and lacking depth. My eyes took in the scenery like a postcard, but when I blinked, it slipped from view, replaced by a new scene." In this case, the Google output sticks closer to the source in a way that retains the very intentional acts of cutting and flipping conveyed in the Japanese, and which sustain the sense of willful detachment the narrator exerts on the world. In using "picture-story show" to translate *kami-shibai*, instead of "postcard," it also preserves the cultural reference to an artistic media that has very particular associations with children, oral story-telling, and public space. Here, as elsewhere in the text, the NMT output retains something of the non-idiomatic and defamiliarizing qualities of Kawakami's prose, which are often precisely the point.

MT is not necessary for such literal-minded translation, of course, which is just as easily obtained from less experienced translators everywhere. Nevertheless, it is worth distinguishing between errors of misrepresentation and those stemming from stubborn fidelity to the source. The latter invite us into the epistemological underpinnings of the source language and are not something to engineer away with more data or by the explicit hard coding of idiomatic and colloquial expressions. Instead, we could let them flourish and proliferate in MT systems to help dispel naïve fantasies about allographic translation machines. To suggest as much, ironically, is already to admit that these systems are becoming fluent and accurate enough to produce a stable background against which such errors are noticeable. But how stable must they be to begin to trust MT as a viable assistant in professional literary or other creative translation, let alone to

process such material by itself?[48] How long before we see cracks in the dam that has for so long walled off MT from the literary? Where we find these cracks and how far they extend will surely vary with the translation context, but even being able to recognize them will require more of the close analysis done in this section carried out across many more aesthetic domains and language pairs. It will also require more rigorous assessment of how NMT models impact lexical diversity, cohesion, syntactic structure, and style at the level of both the sentence and complete text.[49] The point at which we can recognize these cracks, however, is also the point at which we can start to debate what "good enough" machine translation for literary and creative texts looks like. And, just as critically, what good enough might be good for.

When the Gist is Enough

If the previous evaluation results are any indication, it is too early yet to accept machine translated literature as good enough for public consumption or even as a translator's aid. It might be good enough for fan communities eager to access new content, or for users of social writing platforms like Archive of Our Own and Wattpad who wish to read some of the millions of non-English works uploaded to these sites.[50] With literary fiction, however, the need and tolerance for machine translations of any quality is obviated by the higher value placed on fidelity to the source and on the aesthetic pleasure that only skilled translators can be trusted to deliver. Yet this presumes a particular mode of literary consumption, and of knowledge creation, wherein benefits accrue by attending to and reconstructing every detail of a work. Might there be occasions when a rough mockup will suffice? This final section argues that one such use case is precisely to peer into the shadows cast by the global marketplace for literary translation. Gist translations can be a

means to read more of what counts for fiction around the world and, I contend, to reason anew about world literature itself.

This is not the place to wade into heady debates about present and future forms of literary production around the world; about world literature as a historical and ever-mediated construct; or about global capitalism's impact on the variety of worlds that get written and their potential social uses in the contemporary media landscape.[51] I call attention to them, however, to consider what they might gain from an expanded comparative perspective. If there is one thing scholars of world literature can agree on, it is that the market for translated literature is characterized by a trade imbalance that sees far more English-language works exported to other languages than imported into English. The little that is translated represents what publishers, literary agents, and translators deem desirable to English readers, and their choices can in turn create a feedback loop that encourages authors writing in other languages to adapt their work for this market. The market's uneven structure thus circumscribes what readers and scholars working in an English-language context will see or choose to write about. Even when they look beyond English, they are typically limited to specific languages and a handful of authors. Could gist translations be a way to see around and through these structural impediments by bypassing the myopia of the English publishing market and expanding our evidentiary horizons? What if readings of world literature were organized not around the circulation of translated texts first, nor on the insistence of translation's impossibility, but by the enlarged, if algorithmically warped, windows onto non-English literary production afforded by MT?

To test the potential of an MT-assisted comparative analysis, I take a corpus of recent literary fiction in Spanish, German, and Japanese and translate it into English using Google Translate. The idea is that this output can serve as a kind of "pivot language" to explore thematic

connections across this corpus in ways less constrained by English hegemony. Acknowledging that the effects of this hegemony can manifest further upstream, this approach at least reduces the filtering effects of what the Anglophone publishing market deems worthy of translation and the time it takes for these translations to appear. To wit, the corpus consists of roughly 100 works of fiction published in the last five years and randomly sampled from longer lists of works reviewed in major print publications and/or literary journals in Argentina (AR), Germany (DE), and Japan (JP).[52] By reading this sample of world literary output all at once, what can we learn of the contemporary field of literary production across these three sites?

Having developed an intuition for the distortions that NMT models are susceptible to, we know that any reading will have to rely on lower-level stylistic features. Diction, for instance, is one that these models do fairly well with as contextual word embeddings are part of their design. From an exploratory data perspective, "bag-of-words" methods like topic modeling or vector-space models should, in principle, be viable for seeing how our gist translations relate. As these methods use the most frequent words in a corpus to identify lexical or topical similarity, they are less likely to amplify translation errors stemming from rarer words. With this in mind, I decided to topic model the machine translated corpus after segmenting each translation into 1,000-word passages. The model uses a vocabulary of 13,049 words (or about 8% of the total 154,881 words in the corpus) – after filtering out several dozen words common to all the passages and a larger number that appear in a small fraction of passages – to derive 60 topics across all three language collections.[53]

With these 60 topics we can identify regions of thematic divergence and convergence in the collections, as if they shared a common language. Prior to creating the topic model, I was sure that MT would flatten linguistic differences such that works would readily cluster based on

their thematic content or genre. In fact, when I compared how similar all 290 works were based on their relative share of each topic, averaged across all passages of a work, I was surprised by how little topic overlap there was across languages.[54] Figure 3 visualizes as a network the 284 works that are similar at the .6 level (on a scale of 0 to 1). The average similarity of Japanese to Argentine works is a mere .16, compared with .18 to German works. Between Argentine and German works, the average is only slightly higher at .27. At this resolution, whatever common language these works now share, it is not enough to dilute the lexical particulars of each corpus. Some of this is the result of culturally specific references latent in the source texts. But it may also partly be a function of the MT process itself, which is known to leave linguistic traces when it translates semantically similar items differently.[55]

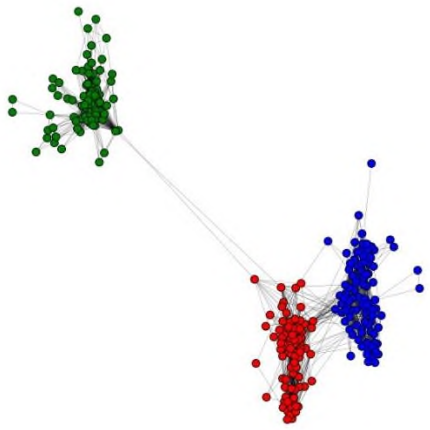

Figure 3: Novels connected by topic similarity at the 0.6 level.
Blue = Argentine, Red = German, Green = Japanese.

Probing the topic model results further, we can identify topics unique to each collection and whose higher proportions form a semantic wall overshadowing any similarities. Table 3 lists these distinctive topics as determined by a pointwise mutual information measure.[56] Based on the words associated with these topics, there appear to be two kinds of topics separating the collections. One consists of reference points (i.e., people, places, events) that assume a shared

cultural context between authors and audience or else can be traced to specific works of historical fiction (e.g., topics 34, 57, 21, 54, 9, and 51). A second kind of topic is not obviously bound by cultural context, but nevertheless has a unique association with just one of the three collections. Topics 41, 48, 55, 4, and 59 fall into this latter camp and suggest that the Argentine novels have relatively more scenes of women and men coming and going; German novels spend more time discussing visual art and family interactions; and the Japanese novels give greater space to bodies in motion or struggle as well as extended informal dialogue. Assuming that these collections are representative of reviewer and reader preferences in each regional context, we can imagine using these distinctive topics to assess the degree to which writers in these places are willing (or not) to dilute the presence of local elements, and the degree to which such dilution is valued.[57] Closer inspection of the works that accentuate these topics might reveal whether they are serving generic imperatives or even aiding in the building of worlds designed to stall ready uptake in the global marketplace for translated fiction.

Topic	Distinctive In	Top Words
34	AR	aires buenos country perón general war military juan government argentine argentina most colonel its president against city political days josé
41	AR	girl woman began men leave hands walked herself women night returned voice tried although arms continued arrived door young lost
57	AR	mercedes clara don monkey diego antonio josé tell luis continued redhead los doctor buenos aires added immediately walked insisted silence
21	DE	german war germany charlotte franz hilde berlin alexander germans russian camp american wilhelm jews men country paul stern comrades comrade
48	DE	days year later evening home morning parents summer together night father family months weeks during week sometimes friend apartment christmas

54	DE	ida uncle august anna papa friedrich karl kurt von wilhelm otto doctor mama nice herself madame really aunt course actually
55	DE	picture photo art painting ada pictures photos drawing wall camera paintings portrait paint painter white exhibition light museum paper artist
4	JP	voice body sound suddenly side hands stood moment stopped immediately tried hit mouth move shook both ran arm fell floor
9	JP	master seems sword year young samurai edo tea temple tatami play money name rice person osaka brother family story mansion
51	JP	castle nobunaga clan oda battle family army takeda lord chief ieyasu soldiers former war tokugawa vassals year nobunaga's general military
59	JP	person seems isn't really words may i'll doesn't i've story sorry feel you're understand voice wasn't okay wonder saying talk

Table 3: Distinctive topics in each collection and the top 20 words associated with each.

Here I defer such inquiries in order to spend more time peering over the initial semantic wall. To only seek out lines of cultural difference and the extra-ordinary would be to reinforce the literary translation market's obsession with otherness, as Angie Chau has described it.[58] The gist translations are an opportunity to investigate areas of shared literary attention across regions and bring them into comparative dialogue, or even to explore thematic convergence as cause or consequence of the global market. To peer over this semantic wall we might think first of trying to narrow the vocabulary to only the most frequent terms. But even at 4% and 2% of the total vocabulary, the overall structure of the similarity network stays the same.[59] If we instead filter out the most distinctive topics, we get a network with far more cross-lingual connection (figure 4), but it is still unclear which topics drive this connectivity. For that it helps to identify the most "indistinctive" topics, which I define here as those topics distributed nearly equally across the three collections. There are 15 such topics, from which I have selected 9 that reflect particularly quotidian activities and objects. Table 4 shows the words heavily associated with these topics.

These topics are interesting precisely because they point to aspects of the contemporary human condition for which writers anywhere might exhibit common concern.

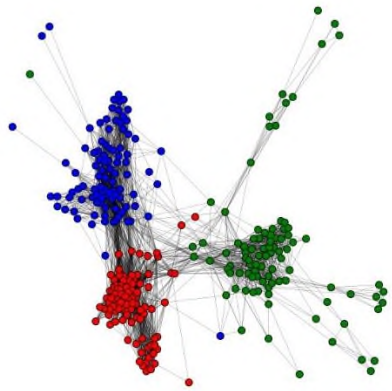

Figure 4: Novels connected by topic similarity at the 0.6 level, after excluding distinctive topics. Blue = Argentine, Red = German, Green = Japanese.

Topic	Top Words
0	body water hands blood mouth neck skin legs hair fingers against chest under lips open feet arms finger cut smell
1	door floor window wall open light opened stairs glass small apartment kitchen inside box table closed living side building windows
16	phone call cell office number desk morning message voice door answer name i'll minutes home talk ask tomorrow tell please
18	car station seat train driver bus road window street cars taxi door side parking stop truck parked passenger drive traffic
29	night bed sleep morning asleep sleeping woke light dream fell slept hours wake body door lying window days bathroom dark
31	glass drink beer table bar wine bottle counter glasses drank iris drinking drunk alcohol guests hotel men evening sip restaurant
42	book read letter write wrote books letters written reading writing novel story paper name writer page pages newspaper library poetry

56	love woman women beautiful sex men girl relationship herself feel young night friends friend loved met together married most happy
58	eat kitchen table food water coffee eating meat bread ate tea soup plate mouth rice cup small taste bowl delicious

Table 4: Indistinctive topics and the top 20 words associated with each.

Using these as a wedge into the topic model output uncovers a host of novel connections. If we take the 25 novels with the highest overall average across the 9 topics (i.e., which express them the most) and group them according to topic overlap, the resulting clusters are no longer sharply divided by source language. Figure 5 illustrates these clusters and their overlap in the form of a heatmap. Cluster 1 has 7 German and 2 Argentine novels that overlap along topics 1 (household interiors), 16 (communication), and 18 (transportation); cluster 2 has 2 German novels and 1 Japanese very loosely joined by topic 29 (sleep); cluster 3 has 3 Argentine novels, 1 German, and 1 Japanese mostly linked by topic 0 (the body) and to a lesser degree by topics 29 (sleep) and 1 (household interiors); cluster 4 contains 1 novel each from Germany and Argentina, and 2 Japanese novels, all sharing a strong focus on topic 42 (writing); and cluster 5 contains 3 Argentine novels and 1 German with a shared interest in topic 0 (the body). Out of the 25 novels most attentive to the ordinary and everyday emerges a composite image of something like the liberal humanist subject – eating, sleeping, drinking, communicating, loving, and reading their way through life. Further investigation of how these topics manifest at the narrative level might reveal how writers in these regions are collectively mapping and/or redrawing the borders of this subject. It could also invite us back to the source texts to see how this shared interest splinters off under the constraints of more local concerns, whether cultural, linguistic, or stylistic.[60]

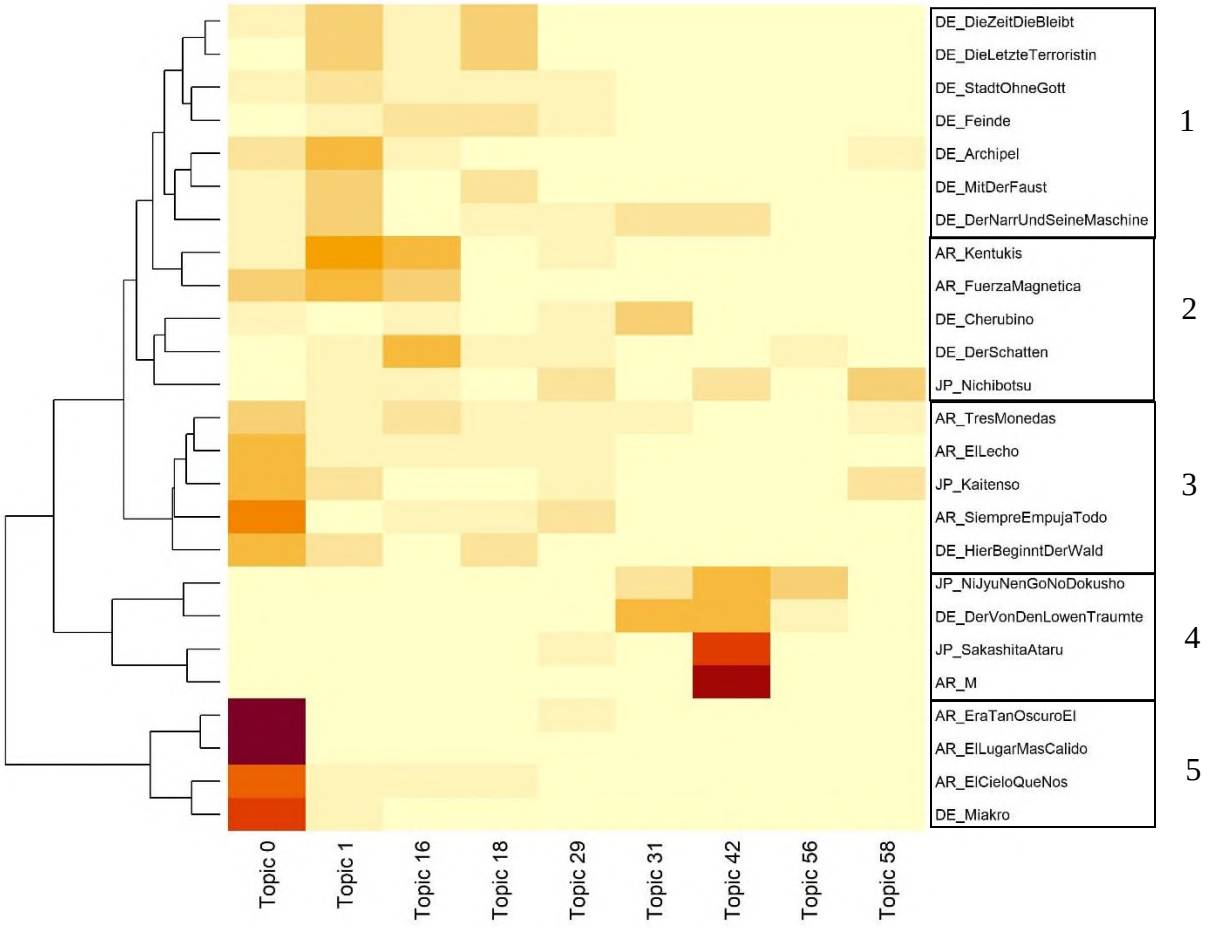

Figure 5: Heatmap of select works grouped by "indistinctive" topic similarity. Dendrogram on left indicates groupings. Work titles on right are ordered according to these groupings.

To illustrate where such investigation might lead, let us examine cluster 4, comprised of works that problematize writing as a way of being in the world while also addressing broader philosophical questions around creative autonomy. The novel *M* (2019), by Argentine writer Eric Schierloh, is a mélange of short archival fragments (e.g., letters, reviews, newspaper articles, underlined passages, biographical notes) that together imagine the life of Herman Melville as he navigated the vicissitudes of literary fortune. *Der von den Löwen träumte* (Who Dreamed of the Lions, 2019), by the German novelist Hanns-Josef Ortheil, is a fictionalized account of Ernest Hemmingway's long stay in Venice in the late-1940s, tackling themes of writer's block,

depression, and authorial originality whilst watching the old man – through his own eyes and those of some native informants – grasp after the shadows of his former glory. Machiya Ryōhei's *Sakashita Ataru to, shijō no uchū* (Sakashita Ataru and the Poetic Universe, 2019), is about a high-school student and aspiring poet who struggles with the lack of recognition his poems receive. He finds inspiration in, but is also jealous of, classmate Sakashita Ataru, whose preternatural talent has earned him awards and a sizable fan community on the online writing platform where he uploads his fiction. Their relationship comes to a head when an AI bot replicates Sakashita's account (and all his writing) using a nearly identical username, tweaking his words and phrasing in ways only a generative language model can, and often for the better by both his estimation and that of his fans. As this literal autofiction gains in popularity on the platform, the two aspiring writers wrestle with existential questions about the nature of creative autonomy, aesthetic originality, and readership.

There is nothing novel about writers writing about writing. But there is something novel about how these specific writers have been brought together. Without MT as an intermediary, it is hard to imagine them being associated with one another. None of the authors have appeared in English translation before, and all occupy distinct positions in their respective literary universes: Schierloh is a minor poet and artisan who runs a small publishing house; Ortheil is a stalwart fixture in contemporary German letters who writes across multiple genres; and Machiya is an up-and-comer who has earned some of Japan's most coveted literary awards since his debut novel was published in 2016. What might seem like strange bedfellows from the standpoint of existing discriminative channels, particularly those which run through mainstream US publishing houses and agents, is here a suggestive arrangement that could expand our ideas about autofiction and its variants around the world. Further analysis is needed to see how well this particular comparative

arrangement holds up under inspection of the works in their original language. That said, it is a compelling enough arrangement to inspire investigation of other clusters in figure 5 to see how else the picture of the liberal humanist subject is filled out in works from these three regions. We can even imagine expanding this search to machine translated corpora from other languages to see how prevalent these topics are in literary fiction from elsewhere, and indeed how essential they may be to the project of the contemporary novel.

A literary critical future that heartily welcomes MT into the study of world literature is, I suspect, still a ways off. In the near term, we need to continue to explore the use of generalized MT models as a kind of search engine, using gist translations to more readily scan the spaces of world literary production and generate novel arrangements of texts for further close analysis. As I have shown, we can do so while acknowledging the model's imperfections, understanding that what they are capable of seeing or transmitting varies with the languages being translated, the formal and stylistic qualities of the source texts, and the diversity of the past human translations that inform their probabilistic calculus. Longer term, we will need to continue to strengthen our intuitions about MT as mediator by conducting qualitative analysis across more text types and language pairs. There is also incentive to wrest control of the tool from corporate entities in order to create more specialized and finely tuned models informed by translations of literary and other creative material.[61] In this future we carry with us not a single universal device, but one hacked for specific needs and use cases and always open to tinkering. Let 1,000 MT models bloom, each one entangled in human translation and linguistic expertise in its own way.

Yet we cannot all be Hoshi Sato. And for the moment, we are stuck with these much less than universal translators – imperfect devices that only fade into the background when we fail to attend to their specific ways of generating static in the message, of masking their uncertainties,

of relying too much on past precedent. But has it not always been thus with translation? Learning to live with MT as an imperfect device means, paradoxically, to be more self-reflexive about the task of translation itself, not less. Deciding when and where to use it, and how much to trust its transformations of literary texts, will naturally entail various ethical responsibilities and political commitments. But this should not prevent us from using the technology responsibly to hear what we can of literary voices elsewhere, even if just well enough to realize how it distorts them, or to recognize the new human-machine creoles it invents as it traverses translation's ever-present gaps.

---

[1] Leif Weatherby and Brian Justie, "Indexical AI," *Critical Inquiry* 48, no. 2 (January 2022), 383.
[2] Matthew Kirschenbaum, "Spec Acts: Reading Form in Recurrent Neural Networks," *English Literary History* 88, no. 2 (2021): 361-386; Katherine Elkins and Jon Chun, "Can GPT-3 Pass a Writer's Turing Test," *Journal of Cultural Analytics* 5, no. 2 (September 2020); and Meghan O'Gieblyn, "Babel," *n+1* 40 (Summer 2021).
[3] Weatherby and Justie, 414.
[4] Douglas Hofstadter, "The Shallowness of Google Translate," *The Atlantic* (January 30, 2018).
[5] Brian Lennon, "Machine Translation: A Tale of Two Cultures," in *A Companion to Translation Studies*, eds. Sandra Bermann and Catherine Porter (Oxford: John Wiley & Sons, Ltd, 2014), 137.
[6] Haun Saussy argues that this notion of translation is itself historically and conceptually narrow, and has had the effect of reducing the key terms of recent translation study to a single axis, "that of self and other." See *Translation as Citation: Zhuangzi Inside Out* (Oxford: Oxford University Press, 2018), 14.
[7] Erga Heller, "The Evolution of the 'Universal Translator': Technical Device and Human Factor in *Doctor Who* and *Star Trek* from the 1960s to the Present," in *Representing Translation*, ed. Dror Abend-David (London: Bloomsbury Academic, 2019), 2-20.
[8] Spence Green, et al., "Natural Language Translation at the Intersection of AI and HCI," *Queue* 13, no. 6 (June 2015): 30-42; John Hutchins, "Machine Translation: A Concise History," *Journal of Translation Studies* 13, no. 1-2 (2010): 29-70; Michael Cronin, *Translation in the Digital Age* (New York: Routledge, 2013), 117; Douglas Robinson, "Cyborg Translation," in *Translation, Translation*, ed. Susan Petrilli (Amsterdam: Rodopi, 2013): 369-386.
[9] Cronin, 120.
[10] On MT in professional settings, see Stephen Doherty, "The Impact of Translation Technologies on the Process and Product of Translation," *International Journal of Communication* (February 2016): 947-970; Akiko Sakamoto, "The Value of Translation in the Era of Automation: An Examination of Threats," in *When Translation Goes Digital*, eds. Renée Desjardins et. al (New York: Palgrave MacMillan, 2021), 231-255; and Dorothy Kenny,

"Technology and Translator Training," in *The Routledge Handbook of Translation and Technology*, ed. Minako O'Hagan (London: Routledge, 2020), 498-515. On new forms of collaborative translation facilitated by online tools, see Minako O'Hagan, "Massively Open Translation: Unpacking the Relationship Between Technology and Translation in the 21st Century," *International Journal of Communication* 10 (2016): 929-946; Miguel Jiménez-Crespo, *Crowdsourcing and Online Collaborative Translations: Expanding the Limits of Translation Studies* (Amsterdam: John Benjamins Publishing Company, 2017); and Ning Ding et. al, "Where Translation Impacts: The Non-Professional Community on Chinese Online Social Media," *Global Media and China* 6, no. 2 (2021): 171-190.

[11] See Lucas Nunes Vieira, "Automation Anxiety and Translators," *Translation Studies* 13, no. 1 (January 2020): 1-21; Lynne Bowker and Jairo Buitrago Ciro, *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community* (UK: Emerald Publishing Limited, 2019); and David Gramling, "Supralingualism and the Translatability Industry," *Applied Linguistics* 41, no. 1 (February 2020): 129-147.

[12] Lawrence Venuti, *Contra Instrumentalism: A Translation Polemic* (Lincoln, NE: University of Nebraska Press, 2019), 78-79.

[13] Cronin, 88.

[14] Louise Amoore, *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others* (Durham, NC: Duke University Press, 2020), 22.

[15] Venuti, *Contra Instrumentalism*, 3, 6.

[16] Alan Liu, "Toward a Diversity Stack: Digital Humanities and Diversity as Technical Problem," *PMLA* 135, no. 1 (January 2020), 136-137.

[17] This idea is inspired by Maeve Olohan, "Translators and Translation Technology: The 'Dance of Agency'," *Translation Studies* 4, no. 3 (2011): 342-357.

[18] See Dorothy Kenny, "Sustaining Disruption? The Transition from Statistical to Neural Machine Translation," *Revista Tradumàtica*, no. 16 (December 2018): 59-70; and Philipp Koehn, *Neural Machine Translation* (Cambridge: Cambridge University Press, 2020), 39-40.

[19] On the rise of SMT, see Thierry Poibeau, *Machine Translation* (Cambridge, MA: The MIT Press, 2017), 121-145.

[20] Dorothy Kenny, "Machine Translation," in *Routledge Encyclopedia of Translation Studies*, 3rd edition (London: Routledge, 2021), 307.

[21] See, for example, J.C. Catford, *A Linguistic Theory of Translation* (London: Oxford University Press, 1965), 29-31; Eugene A. Nida, *Toward a Science of Translating*, originally 1964 (Leiden: Brill, 2003), 253; and Gideon Toury, *Descriptive Translation Studies – and Beyond* (Amsterdam: John Benjamins Publishing Company, 1995).

[22] Extending the voting metaphor, Anthony Pym suggests that NMT corrects for SMT's simple majority rules by sorting "votes" according to translation context, eliminating those candidates that are contextually aberrant. See "Quality," in *The Routledge Handbook of Translation and Technology*, 441.

[23] Koehn, 148-154.

[24] It should be noted, however, that some benefits do accrue for less dominant languages simply by virtue of these model's massive size. In a process called "zero-shot translation," for instance, models can transfer knowledge from language pairs they have seen (e.g., Japanese and English) to pairs they pairs they have not (e.g., Japanese and Korean).

[25] My description of NMT relies heavily on Bowker and Ciro, 44-49; Mikel L. Forcada, "Making Sense of Neural Machine Translation," *Translation Spaces* 6, no. 2 (December 2017): 291-309; Koehn, 103-163; and Alan K. Melby, "Future of Machine Translation: Musings on Weaver's Memo," in *The Routledge Handbook of Translation and Technology*, ch. 25.
[26] Forcada, 14.
[27] Ibid., 13.
[28] On the history of quality evaluation in MT, see Poibeau, 199-208. The BLEU score was first proposed in 2001 by Kishore Papineni, et al., "BLEU: A Method for Automatic Evaluation of Machine Translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (July 2002): 311-318.
[29] Jiménez-Crespo, 123-124. See also Pym, 439, on translation "quality" as a *relation* rather than an absolute value.
[30] See especially Vieira, "Automation Anxiety and Translators"; and Claire Larsonneur, "Neural Machine Translation: From Commodity to Commons?," in *When Translation Goes Digital*, eds. R. Desjardins, et al. (Palgrave Macmillan, 2021): 257-280.
[31] See Joss Moorkens et al., "Translators' Perceptions of Literary Post-Editing Using Statistical and Neural Machine Translation," *Translation Spaces* 7 (November 2018): 240-262; Antonio Toral et al., "Post-editing Effort of a Novel with Statistical and Neural Machine Translation," *Frontiers in Digital Humanities* 5, no. 9 (May 2018); and Dorothy Kenny and Marion Winters, "Machine Translation, Ethics and the Literary Translator's Voice," *Translation Spaces* 9, no. 1 (August 2020): 123-149.
[32] Emma Boumans, "Evaluating Machine Translation for Literary Texts," Thesis, University of Amsterdam, 2016; Antonio Toral and Andy Way, "What Level of Quality can Neural Machine Translation Attain on Literary Text?" (January 2018): http://arxiv.org/abs/1801.04962; Evgeny Matsuov, "The Challenges of Using Neural Machine Translation for Literature" (2019): https://aclanthology.org/W19-7302.pdf; Arda Tezcan et al., "When a 'Sport' is a Person and Other Issues for NMT of Novels" (2019): https://aclanthology.org/W19-7306.pdf; Margot Fonteyne et. al., "Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level," *Proceedings of the 12th Conference on Language Resources and Evaluation* (2020): 3790-3798; and Rebecca Webster et al., "Gutenberg Goes Neural: Comparing Features of Dutch Human Translations with Raw Neural Machine Translation Outputs in a Corpus of English Literary Classics," *Informatics* 7, no. 32 (August 2020): 1-21. These studies generally focus on just a handful of texts and address the following languages: German, Catalan, Russian, and Dutch.
[33] Each set of 100 sentences was roughly equivalent in length (~ 4,300 characters). The DeepL online translation tool is another popular resource competitive with Google Translate. When it was applied to this corpus, however, it often failed to recognize Japanese quotation marks and simply dropped quoted material.
[34] I adapted the SCATE taxonomy and evaluation guide used in Fonteyne et al., available here: https://github.com/margotfonteyne/StylesNMT/blob/9bdfaed2f167134671a906fc778bcb16f43e578e/AnnotationGuidelines.pdf. Errors were labeled using INCEpTION, an open-source tool for shared annotation tasks (https://inception-project.github.io/).
[35] Interrater agreement was calculated as the number of times that evaluators agreed on an error category label for a given sentence over the sum of instances where they did and did not agree.

[36] The average deviation (AD) index is computed by calculating, for every sentence, how much each evaluator deviates from the median rating for that sentence. These values are then summed and their absolute value is divided by the number of evaluators. Averaging these scores across all sentences gives us the overall AD index for a particular work.
[37] The first is from Edwin McClellan's 1957 translation and the second from the 2010 translation by Meredith McKinney.
[38] Koehn, 224-233.
[39] On transcription and other forms of non-translation as essential (and yet overlooked) practices in the history of translation, see Saussy, *Translation as Citation*, ch. 1.
[40] Koehn, 212.
[41] See Rob Voight and Dan Jurafsky, "Towards a Literary Machine Translation: The Role of Referential Cohesion," *Workshop on Computational Linguistics for Literature* (2012): 18-25.
[42] Nobel laureate Ōe Kenzaburō famously quipped, "Murakami Haruki writes in Japanese, but his writing is not really Japanese. If you translate it into American English, it can be read very naturally in New York." Cited in Matthew Carl Strecher, *Dances with Sheep: The Quest for Identity in the Fiction of Murakami Haruki* (Ann Arbor, MI: University of Michigan Press, 2020), 1.
[43] On attempts to develop interactive and adaptive translation systems, see Samuel Laübli and Spence Green, "Translation Technology Research and Human-Computer Interaction (HCI)," in *The Routledge Handbook of Translation and Technology*, ch. 22.
[44] Koehn, 239-248.
[45] Matusov experiments with domain adaptation for literary translation in his "The Challenges of Using Neural Machine Translation for Literature," but finds only minimal improvement with the results from generalized models like Google Translate, at least according to automatic measures.
[46] Kawakami Mieko, *Heaven*, trans. Sam Bett and David Boyd (New York: Europa, 2021), 12.
[47] Ibid., 13.
[48] In addition to already cited works on MT's viability as a translation assistant, see Kristiina Taivalkoski-Shilov, "Ethical Issues Regarding Machine(-Assisted) Translation of Literary Texts," *Perspectives: Studies in Translation Theory and Practice* 27, no. 5 (October 2019): 689-703.
[49] See especially Webster et al., on the development of quantitative measures for assessing more complex stylistic and document-level features.
[50] Minako O'Hagan has done pioneering work on the role that fan-subbing cultures have played in redefining non-professional translation, especially with the advent of Web 2.0 platforms. The impact of NMT models on these communities is a topic in need of further research, but consider that a Chrome extension exists for automatically translating *manga* from Japanese, Chinese, and Korean (https://rebrand.ly/tcf912g).
[51] Insightful recent interventions in this debate that have informed my own thinking include Alexander Beecroft, *An Ecology of World Literature: From Antiquity to the Present Day* (London: Verso, 2015), 243-299; Pheng Cheah, *What is a World?: On Postcolonial Literature as World Literature* (Durham, NC: Duke University Press, 2016), 191-215; B. Venkat Mani, *Recoding World Literature: Libraries, Print Culture, and Germany's Pact with Books* (New York: Fordham University Press, 2017), 9-48; and Sarah Brouillette, *UNESCO and the Fate of the Literary* (Stanford, CA: Stanford University Press, 2019).

[52] These collections were compiled as part of the World Literature Data Collective. For more on this project and the sampling rationale, see https://worldliteraturedatacollective.wordpress.com/.
[53] After arriving at this strategy, I subsequently learned that it had been proposed in the field of political science in 2015. See Christopher Lucas et al., "Computer-Assisted Text Analysis for Comparative Politics," *Political Analysis* 23, no. 2 (2015): 254-277. Several follow-up studies have validated the robustness of this approach by comparing topic model output from machine translated texts and gold-standard human translated texts. See, for example, Erik de Vries et al., "No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications," *Political Analysis* 24, no. 4 (2018): 417-430. Needless to say, further such studies need to be carried out on literary texts from different languages and genres to see if the method is similarly robust.
[54] Similarity was calculated using cosine similarity on the averaged topic distributions.
[55] See Lucas et al., 270, for a discussion of this issue. As an example, they note that Google will translate the Chinese word for liquor as "wine," and the Arabic word for liquor as "spirits." They propose using structural topic models to minimize such systematic differences, as these allow for topics to co-vary with source language. Thus, they can extract a general liquor or alcohol topic to which source-specific words contribute based on their co-occurrence with other mutually shared terms related to alcoholic consumption.
[56] This measure was developed by Lauren Klein and Jacob Eisenstein. Code is available at this website: https://github.com/laurenfklein/dimensions-of-scale.
[57] Matt Erlin et al., explore these questions in 200 works of recent fiction translated into English from 22 countries, finding little support for Pascale Casanova's hypothesis that minor literatures are more nationalistic. It would be interesting to see how this result holds up for works that have not been favored by the English translation market. See "Cultural Capitals: Modeling 'Minor' European Literature," *Journal of Cultural Analytics* 2 (2021): 40-73.
[58] Angie Chau alludes to this obsession and possible critical responses to it in "Healing Through Ordinary Stories," *Public Books* (January 12, 2022): https://www.publicbooks.org/healing-through-ordinary-stories/.
[59] Separate topic models were built using vocabularies of 6,515 and 3,040 words. The networks produced from these models showed no discernible increase in cross-lingual connection, which suggests that semantic differences overshadow commonalities even when considering the most frequent terms.
[60] Structural topic models could even aid in such inquiry by highlighting terms associated with a shared topic but more prevalent in texts from a given source language. For example, experiments with a 60-topic model (and a vocabulary of 12,914 words) reveal a "body" topic much like Topic 0 in Table 4. But they also show how the topic is inflected differently in each collection. Thus in the AR corpus it is inflected by terms like caressing, nipples, caress, erection, and orgasm; in the DE corpus by terms like Schiller, bared, twitch, pubic, and thighs; and in the JP corpus by terms like eyeball, eyeballs, abdomen, horrifying, and crotch.
[61] On the possibilities and challenges of de-corporatizing these tools, see Larsonneur, "Neural Machine Translation: From Commodity to Commons?," 269-272.