Simpson's Paradox and the Accuracy-Fluency Tradeoff in Translation

Zheng Wei Lim, Ekaterina Vylomova, Trevor Cohn,* and **Charles Kemp**
The University of Melbourne
z.lim4@student.unimelb.edu.au
{vylomovae,t.cohn,c.kemp}@unimelb.edu.au

Abstract

A good translation should be faithful to the source and should respect the norms of the target language. We address a theoretical puzzle about the relationship between these objectives. On one hand, intuition and some prior work suggest that accuracy and fluency should trade off against each other, and that capturing every detail of the source can only be achieved at the cost of fluency. On the other hand, quality assessment researchers often suggest that accuracy and fluency are highly correlated and difficult for human raters to distinguish (Callison-Burch et al., 2007). We show that the tension between these views is an instance of Simpson's paradox, and that accuracy and fluency are positively correlated at the level of the corpus but trade off at the level of individual source segments. We further suggest that the relationship between accuracy and fluency is best evaluated at the segment (or sentence) level, and that the trade off between these dimensions has implications both for assessing translation quality and developing improved MT systems.

1 Introduction

No translation can simultaneously satisfy all possible goals, and translation is therefore an art of navigating competing objectives (Darwish, 2008). Many objectives are discussed in the literature, but two in particular seem especially fundamental. The first is accuracy (also known as fidelity or adequacy), or the goal of preserving the information in the source text (ST). The second is fluency, or the goal of producing target text (TT) that respects the norms of the target language (TL) and is easy for the recipient to process (Kunilovskaya, 2023).

Here we study the relationship between accuracy and fluency and work with two operationalizations of these notions. The first relies on human judgments of accuracy and fluency collected in prior work on translation quality estimation (Castilho et al., 2018). The second relies on probabilities estimated using neural machine translation (NMT) models. Given a source-translation pair $(\boldsymbol{x}, \boldsymbol{y})$, $p(\boldsymbol{x}|\boldsymbol{y})$ corresponds to accuracy, and $p(\boldsymbol{y})$ corresponds to fluency (Teich et al., 2020). $p(\boldsymbol{x}|\boldsymbol{y})$ will be low if $\boldsymbol{y}$ fails to preserve all of the information in $\boldsymbol{x}$, and $p(\boldsymbol{y})$ will be low if $\boldsymbol{y}$ violates the norms of the target language. To highlight that model estimates $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ are related to but distinct from human ratings of accuracy and fluency, we refer to $p(\boldsymbol{x}|\boldsymbol{y})$ as accuracy$_M$ and $p(\boldsymbol{y})$ as fluency$_M$.

Some parts of the literature argue that accuracy trades off with fluency. In Figure 1a, the blue dots are translations of the same source segment, and Table 1 shows three translations that illustrate the same kind of tradeoff. A translator choosing between these alternatives cannot simultaneously maximize accuracy and fluency, because the most accurate translations are not the most fluent, and vice versa. Teich et al. (2020) argues that accuracy$_M$ and fluency$_M$ should trade off in this way, and the same view is implicitly captured by noisy-channel models of translation (Brown et al., 1993), which aim to generate translations $\boldsymbol{y}$ that maximize $p(\boldsymbol{y}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})$. Typically these models include weights for the two components $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ that can be interpreted as the extent to which accuracy$_M$ is prioritized over fluency$_M$, or vice versa (Yu et al., 2016; Yee et al., 2019; Yu et al., 2020; Müller et al., 2020).

An opposing view of the relationship between accuracy and fluency emerges from the literature on quality estimation. Here the common wisdom is that accuracy and fluency are highly correlated and practically indistinguishable to human annotators (Callison-Burch et al., 2007; Banchs et al., 2015; Mathur, 2021, but see Djiako 2019; Sulem et al. 2020). As a result, accuracy and fluency are conflated as a single assessment score in recent WMT General Machine Translation Tasks, with more emphasis given to accuracy than fluency (Farhad et al.,

*Now at Google.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 92–103
August 11-16, 2024 ©2024 Association for Computational Linguistics

Translation	accuracy	fluency	accuracy_M	fluency_M	$\log p(\boldsymbol{y}\|\boldsymbol{x})$
(i) Ich gab Ihnen eine Rückerstattung des Buches.	23.0	25.0	-10.81	-56.0	-10.31
(ii) Ich habe Ihnen eine Rückerstattung des Buches ausgestellt.	24.3	24.7	-6.13	-64.0	-12.13
(iii) Ich stellte Ihnen eine Rückerstattung des Buches aus.	25.0	23.0	-6.44	-70.0	-14.75

Table 1: Translations of "*I issued you a refund of the book.*" from English to German, which correspond to three of the orange dots in Figure 1. Human ratings of accuracy and fluency are derived from MQM scores, and accuracy_M ($\log p(\boldsymbol{x}|\boldsymbol{y})$) and fluency_M ($\log p(\boldsymbol{y})$) are estimated using an NMT model. Option (i) is acceptable but *gab* (past tense of give) is less accurate than the conjugations of *ausstellen* (issue) used in (ii) and (iii). Option (iii) is the least natural because *stellte ... aus* (Präteritum tense) is typically used only in formal writing.

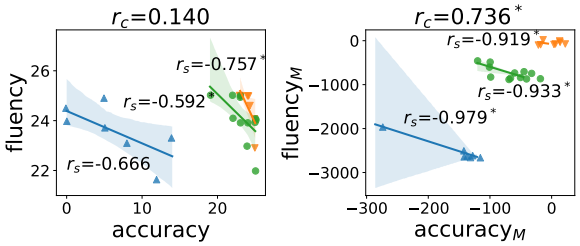


Figure 1: Simpson's paradox. Each panel shows translations of three source segments indexed by color and marker shape. At the source segment level, accuracy and fluency (left) and accuracy_M and fluency_M (right, probabilities plotted on log scales) both show negative correlations r_s. At the corpus level, both pairs of dimensions show positive correlations r_c (see panel labels). Significant correlations ($p < .05$) are marked with '*'. Source segments and translations are drawn from past WMT General Task submissions and data points have been jittered for clarity. The shaded areas show 95% confidence intervals based on 1000 bootstrapped samples. Full translations are included in Tables 2 (orange dots), 3 (green dots) and 4 (blue dots) of the appendix.

2021; Kocmi et al., 2022, 2023).

We argue that the conflict between these views is an instance of Simpson's paradox (Yuan et al., 2021), which occurs when a relationship at one level of analysis (e.g. the corpus level) disappears or is reversed at a different level (e.g. the segment or sentence level). Figure 1 shows how the correlation r_c between accuracy and fluency can be positive over a miniature corpus including translations of three source segments even though the correlation r_s for each individual source segment is negative. Of the two levels of analysis, the segment level is the appropriate level for understanding how humans and machine translation systems should choose among possible translations of a source segment. The central goal of our work is therefore to establish that the correlation between accuracy and fluency is negative at the level of individual source segments.

2 Tradeoff between $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$

Because accuracy_M and fluency_M have formal definitions, we start with these dimensions.

2.1 Theoretical formulation and simulation

Let $\boldsymbol{Y}$ be a finite set of translations of source segment $\boldsymbol{x}$, and let $\vec{p}_{x|y}$ and $\vec{p}_y$ denote log probability vectors that include accuracy_M and fluency_M scores for all $\boldsymbol{y} \in \boldsymbol{Y}$.[1] We use the Pearson correlation between the two vectors:

$$r_s = \text{corr}(\vec{p}_{x|y}, \vec{p}_y) \quad (1)$$

to quantify the tradeoff between accuracy_M and fluency_M across translations of $\boldsymbol{x}$. If $r_s > 0$ there is no tradeoff, and the translations with higher accuracy_M also tend to have higher fluency_M. If $r_s < 0$ the dimensions trade off, and improving a translation along one dimension tends to leave it worse along the other. Note that r_s is a correlation at the segment level, and should be distinguished from the corpus-level correlation r_c between $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ over an entire corpus of segments $\boldsymbol{x}$ and their translations $\boldsymbol{y}$.

Suppose that a translator is considering candidate translations $\boldsymbol{y}$ of source segment $\boldsymbol{x}$. There are a vast number of possible translations, including many nonsense translations, but we assume that the translator chooses among a small set of good translations that all have near-maximal values of $p(\boldsymbol{y}|\boldsymbol{x})$. Because $p(\boldsymbol{y}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})$ is roughly constant over this set of good translations, it follows that accuracy_M and fluency_M trade off within the set.

To validate this informal argument, we ran simulations to confirm that tradeoffs between $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ emerge when $\boldsymbol{x}$ and $\boldsymbol{y}$ are numeric vectors drawn from a Gaussian joint distribution $P(\boldsymbol{x}, \boldsymbol{y})$

[1] There are infinitely many possible translations, but here we consider a finite set generated by humans or machines.

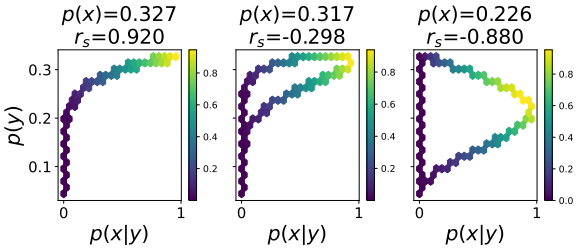
centered at zero.[2] We set an initial square matrix A with dimensionality equal to the total number of dimensions in $\boldsymbol{x}$ and $\boldsymbol{y}$ combined. Assuming all elements in $\boldsymbol{x}$ and $\boldsymbol{y}$ have $\sigma^2 = 1$ and pairwise positive covariance, all diagonal elements of A are set to 1 and other elements 0.7. To ensure the covariance matrix is positive semi-definite, we replace the initial matrix A with a final covariance matrix defined as $A^\top A$.

For each "source segment" $\boldsymbol{x}$ considered in our simulation, we generate 10,000 possible "translations" $\boldsymbol{y}$ by sampling from a distribution $q(\boldsymbol{y}) = \prod_i q(y_i)$, where each element y_i of $\boldsymbol{y}$ is sampled uniformly within two standard deviations of its mean. We then score each translation and compute $p(\boldsymbol{x}|\boldsymbol{y})$, $p(\boldsymbol{y})$ and $p(\boldsymbol{y}|\boldsymbol{x})$ using the known joint $P(\boldsymbol{x}, \boldsymbol{y})$.

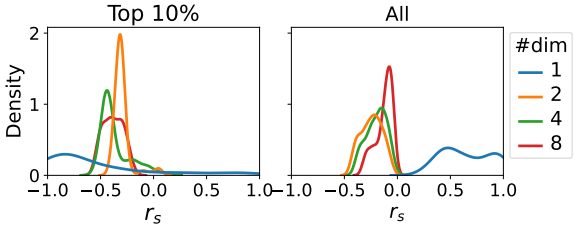
We initially assume that both $\boldsymbol{x}$ and $\boldsymbol{y}$ are one-dimensional vectors. Figure 2a shows the relationship between $p(\boldsymbol{y})$ and $p(\boldsymbol{x}|\boldsymbol{y})$ for 3 "segments" $\boldsymbol{x}$. Each point in each panel corresponds to a candidate translation $\boldsymbol{y}$, and candidates with highest $p(\boldsymbol{y}|\boldsymbol{x})$ are shown in yellow. The correlation above each panel results from applying Equation 1 to all translations with $p(\boldsymbol{y}|\boldsymbol{x})$ above the 90th percentile (i.e. all points in the brightest part of each plot). The first "segment" $\boldsymbol{x}$ (leftmost panel) has relatively high probability $p(\boldsymbol{x})$, and no tradeoff is observed in this case. The tradeoff emerges, however, and becomes increasingly strong as $\boldsymbol{x}$ moves away from the mode of the distribution $p(\boldsymbol{x})$. At the "corpus" level, $p(\boldsymbol{y})$ and $p(\boldsymbol{x}|\boldsymbol{y})$ are uncorrelated ($r_c = -.001, p = .970$) when the top 10% of translations for each of the three "segments" are combined.

Figure 2b shows that the tradeoff persists when the dimensionality of $\boldsymbol{x}$ and $\boldsymbol{y}$ is increased. The density plot for each dimensionality is based on a sample of 100 source "segments" (rather than the 3 in Figure 2a), and at all dimensionalities the majority of source "segments" induce tradeoffs. The tradeoffs are stronger (i.e. correlations more negative) when the candidate translations consist of the $\boldsymbol{y}$ with highest $p(\boldsymbol{y}|\boldsymbol{x})$ (top 10%), but for all dimensions except $n = 1$ most source "segments" still induce a tradeoff even if all candidate translations are considered. At the "corpus" level, $p(\boldsymbol{y})$ and $p(\boldsymbol{x}|\boldsymbol{y})$ of the top translations are positively correlated ($r_c = .399, .159, .113, .109$ for dimensionalities 1, 2, 4 and 8 respectively, $p < .001$).

[2]Code available at https://github.com/ZhengWeiLim/accuracy-fluency-tradeoff.



(a) Simulation with one-dimensional $\boldsymbol{x}$ and $\boldsymbol{y}$. The three panels correspond to three different source "segments" $\boldsymbol{x}$ of decreasing probability $p(\boldsymbol{x})$, and the points in each panel are candidate translations $\boldsymbol{y}$. Brighter colors indicate translations with larger $p(\boldsymbol{y}|\boldsymbol{x})$. Pearson correlations across translations ranked in the top 10% based on $p(\boldsymbol{y}|\boldsymbol{x})$ are shown at the top of each panel.



(b) Kernel density plots of tradeoffs across the top 10% (left) and across all translation choices (right). The tradeoff persists in higher dimensional space, and is stronger when selecting only $\boldsymbol{y}$ with the highest values of $p(\boldsymbol{y}|\boldsymbol{x})$.

Figure 2: Tradeoffs between $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ in synthetic data.

Although our simulations aim for simplicity rather than realism, they provide theoretical grounds for expecting tradeoffs at the segment level in real translations generated by humans and machines. They also suggest that the tradeoff may become stronger when only high-quality translations are considered, and that the strength of the tradeoff may depend on $p(\boldsymbol{x})$.

2.2 Human and machine translation

We now show that human and machine translations show the same tradeoff between accuracy$_M$ and fluency$_M$, which correspond to $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ estimated by an NMT model.

Data. We analyze 15 translation studies from CRITT TPR-DB (CRITT) that include 13 language pairs (Carl et al., 2016b). We also use a subset of the Russian Learner Translator Corpus (RLTC) that has been aligned at the sentence level by Kunilovskaya (2023). For machine translation, we use WMT test sets which include segments of (mostly individual) sentences that are annotated with Multidimensional Quality Metrics labels (MTMQM)

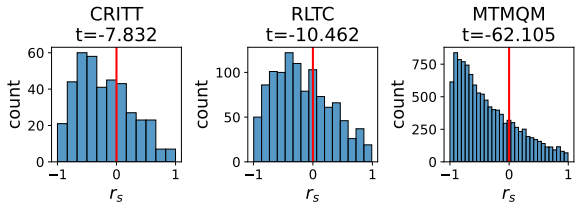


Figure 3: Tradeoffs between estimated $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ across source segments from three corpora. Paired-sample t-tests against randomly permuted $p(\boldsymbol{y})$ and $p(\boldsymbol{x}|\boldsymbol{y})$ are shown at the top of each panel.

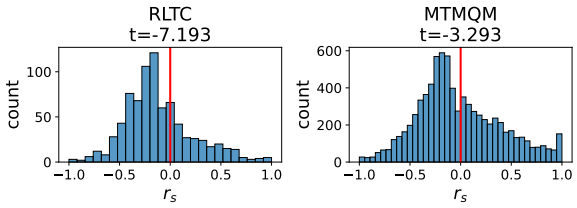


Figure 4: Tradeoffs between human ratings of accuracy and fluency across segments from two corpora. Paired-sample t-tests against randomly permuted scores are shown at the top of each panel.

(Freitag et al., 2021a,b; Zerva et al., 2022; Freitag et al., 2023). To reduce spurious correlations, we remove duplicate translations and source segments with fewer than four unique translations. Additional details are provided in the appendix.

Models. We use NLLB-200's 3.3B variant model (Costa-jussà et al., 2022) to estimate $p(\boldsymbol{y}|\boldsymbol{x})$ and $p(\boldsymbol{x}|\boldsymbol{y})$.[3] For consistency, we also extract $p(\boldsymbol{y})$ based on the same model, skipping all inputs except for special tokens (e.g., <eos> tags).[4] All probabilities are log scaled.

Results. Figure 3 is a histogram analogous to the densities in Figure 2b, and shows distributions of tradeoff scores for source segments in CRITT, RLTC and MTMQM. In all three cases most source segments induce tradeoffs (i.e. produce negative correlations). To test for statistical significance we compared the actual distributions against randomly permuted data. The results of all paired-sample t-tests are significant ($p < .001$), and are included in the figure.[5] When samples are aggregated at the corpus level, $p(\boldsymbol{y})$ and $p(\boldsymbol{x}|\boldsymbol{y})$ show significant positive correlations ($p < .001$) for CRITT ($r_c = .625$), RLTC ($r_c = .685$) and MTMQM ($r_c = .675$), revealing that Simpson's paradox applies in all three cases.

The simulation in Figure 2a suggests that segments with smaller $p(\boldsymbol{x})$ tend to show greater tradeoffs, which predicts that $p(\boldsymbol{x})$ and r_s (Equation 1) should be positively correlated. Our data support this prediction for CRITT ($r = .124$, $p = .013$), RLTC ($r = .225$, $p < .001$) and MTMQM ($r = .109, p < .001$).

[3] NLLB model card

[4] To ensure reproducibility across models, we repeat our analysis in the appendix using M2M100 (Fan et al., 2021).

[5] Each permuted data set is created by randomly shuffling the pairings of $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ within the set of possible translations of each source segment.

3 Tradeoff between accuracy and fluency

We now turn to human ratings of accuracy and fluency, and demonstrate that the two are again negatively correlated at the segment level.

Data. Only RLTC and MTMQM are rated by human annotators. The subset of RLTC released by Kunilovskaya (2023) includes accuracy and fluency scores derived from error annotations. For MTMQM, we follow Freitag et al. (2021a) where accuracy scores are aggregates of "Accuracy" and "Terminology" errors, and fluency scores are aggregates of "Fluency", "Style" and "Locale convention" errors. Targets that are labelled "Non-translation" receive scores of zero for both accuracy and fluency. Major and minor errors receive penalties of 5 and 1 respectively. Fluency/Punctuation is assigned a penalty of 0.1. We calculate the final rating as $s_c = \max(0, 25 - e_c)$, where e_c denotes the total penalty in error category c.[6] Because some systems submit the same translation but receive different ratings, we average these scores and remove the duplicate entries.

Results. Figure 4 shows correlations at the level of individual source segments. The majority of correlations are negative, and paired-sample t-tests reveal that both distributions are significantly ($p < .001$) different from distributions obtained from random permutations. The results therefore suggest that accuracy and fluency (as rated by humans) trade off at the level of individual segments. At the corpus level, accuracy and fluency are positively correlated for MTMQM ($r_c = .392, p < .001$), and are uncorrelated in RLTC ($r_c = -.085, p < .001$), suggesting again that Simpson's paradox applies to both cases.[7]

[6] The maximum score is set at 25 because the maximum MTMQM penalty score is 25.

[7] Fluency and accuracy may be uncorrelated in RLTC at the

Unlike the case for accuracy$_M$ and fluency$_M$, human ratings of accuracy and fluency do not induce a positive correlation between $p(\boldsymbol{x})$ and r_s ($r = -.150$ and $-.104$ for RLTC and MTMQM respectively). We therefore find no support for the simulation-based prediction that low-probability sentences are more likely to produce strong tradeoffs between accuracy and fluency.

Figure 4 is directly analogous to Figure 3, and we expected that source segments which showed strong tradeoffs (i.e. extreme negative correlations) in Figure 3 would also show strong tradeoffs in Figure 4. The two tradeoff measures, however, were uncorrelated,[8] which suggests that accuracy$_M$ and fluency$_M$ overlap only partially with human ratings of accuracy and fluency.

A similar conclusion is suggested by Figure 5, which shows Pearson correlations of translation probability ($p(\boldsymbol{y}|\boldsymbol{x})$; blue bars), accuracy$_M$ ($p(\boldsymbol{x}|\boldsymbol{y})$; brown bars) and fluency$_M$ ($p(\boldsymbol{y})$; green bars) with human ratings of accuracy and fluency for RLTC and MTMQM.[9] As expected, accuracy$_M$ shows a higher correlation with accuracy than fluency, and fluency$_M$ shows the opposite pattern. Figure 5 however, suggests that accuracy$_M$ is not superior to $p(\boldsymbol{y}|\boldsymbol{x})$ as a predictor of accuracy, and that fluency$_M$ is not superior to $p(\boldsymbol{y}|\boldsymbol{x})$ as a predictor of fluency. One reason why our model estimates of accuracy and fluency depart from human ratings is that accuracy$_M$ and fluency$_M$ are sensitive to segment length. For example, a longer segment will have lower fluency$_M$ than a shorter segment even if the two are both perfectly idiomatic.

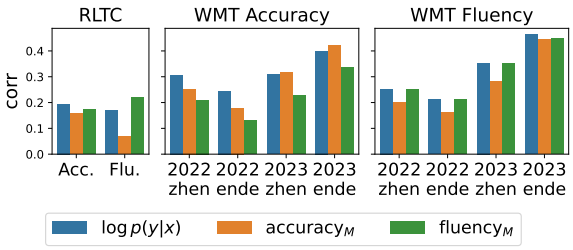


Figure 5: accuracy$_M$ and fluency$_M$ predict human accuracy and fluency ratings for RLTC and WMT submissions to the general translation task in 2022 and 2023. `zhen` and `ende` refer to Chinese-English and English-German language pairs. All correlations reported are significant ($p < .001$).

4 Conclusion

We showed that accuracy and fluency and $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ both trade off when translating individual source segments. This finding suggests that current protocols for assessing translation quality may need to be adjusted. Human assessments for recent WMT General Tasks are performed using Direct Assessment and Scalar Quality Metrics (DA+SQM) (Kocmi et al., 2022, 2023). This approach conflates meaning preservation and grammar into a single score indicative of overall quality of a translation. In contrast, MQM is much more costly, but produces highly detailed scores that use multiple sub-categories for both accuracy and fluency. Future approaches could therefore consider a middle ground that extends DA+SQM to include accuracy and fluency as independent aspects as in WMT16 (Bojar et al., 2016). This direction would allow automatic MT evaluation metrics such as BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2022) (both fine-tuned to DA scores) to be adapted to provide independent scores for accuracy and fluency.

Our results also suggest the value of developing MT models that navigate the accuracy-fluency tradeoff in human-like ways. In some settings (e.g. translating legal texts) accuracy is more important than fluency (Popović, 2020; Martindale and Carpuat, 2018; Vela and Tan, 2015; Specia et al., 2011; Martindale et al., 2019), but in others (e.g. translating informal conversation) fluency may take priority (Poibeau, 2022; Frankenberg-Garcia, 2022). One natural approach to navigating the accuracy-fluency tradeoff builds on noisy channel models (Yu et al., 2016; Yee et al., 2019; Müller et al., 2020), which incorporate both $p(\boldsymbol{y})$ and $p(\boldsymbol{x}|\boldsymbol{y})$ along with tradeoff parameters that specify the relative weights of the two. Tuning these parameters for specific registers may allow a model to find the right balance between accuracy and fluency in each case.

5 Limitations

Although we provided evidence for both accuracy-fluency and accuracy$_M$-fluency$_M$ tradeoffs in translation, we did not explore semantic and grammatical features that may predict which source segments produce the greatest tradeoffs. Outside of our simulation we do not have access to ground-

corpus level because of a ceiling effect – 63.5% and 70.6% of sentences receive maximum ratings for fluency and accuracy in RLTC compared to 55.6% and 58.4% for MTMQM.

[8] The Pearson correlations between the two tradeoff measures for RLTC and MTMQM are $r = .003, p = .933$ and $r = .022, p = .05$.

[9] Values are in log scale and are ranked by percentile.

truth values of $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$, and are only able to approximate these values using specific NMT models. Our work is also limited by the fact that MTMQM only includes translations generated by certain kinds of NMT models, and it is possible that our results do not generalize to translations generated by other types of models, such as statistical or rule-based MT systems. Finally, both RLTC and MTMQM have accuracy and fluency ratings derived from error annotations that are very similar in range. This constraint makes quality assessment and comparison at the segment level challenging.

Ethics Statement

We do not foresee any potential risks and harmful use of our work. Our analyses are based on licensed data which are freely available for academic use.

Acknowledgements

This work was supported by ARC FT190100200.

References

Fabio Alves and José Luiz Gonçalves. 2013. Investigating the conceptual-procedural distinction in the translation process: A relevance-theoretic analysis of micro and macro translation units. *Target. International Journal of Translation Studies*, 25(1):107–124.

Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating MT in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (WMT16). In *First conference on machine translation*, pages 131–198. Association for Computational Linguistics.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Chris Callison-Burch, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158.

Michael Carl, Akiko Aizawa, and Masaru Yamada. 2016a. 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 M Cristina Toledo Báez. 2019. Machine translation errors and the translation process: A study across different languages. *Journal of Specialised Translation*, 31:107–132.

Michael Carl, Moritz Schaeffer, and Srinivas Bangalore. 2016b. The CRITT translation process research database. In *New directions in empirical translation process research*, pages 13–54. Springer.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. *Translation quality assessment: From principles to practice*, pages 9–38.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ali Darwish. 2008. *Optimality in translation*. Writescope Publishers.

Gabriel Armand Djiako. 2019. *Lexical ambiguity in machine translation and its impact on the evaluation of output by users*. Ph.D. thesis, Saarländische Universitäts-und Landesbibliothek.

Barbara Dragsted. 2010. Coordination of reading and writing processes in translation: An eye on uncharted territory. In *Translation and Cognition*, pages 41–62. John Benjamins Publishing Company.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Ana Frankenberg-Garcia. 2022. Can a corpus-driven lexical analysis of human and machine translation unveil discourse features that set them apart? *Target*, 34(2):278–308.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. 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, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, et al. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774.

Kristian Tangsgaard Hvelplund Jensen, Annette C Sjørup, and Laura Winther Balling. 2009. Effects of L1 syntax on L2 translation. *Copenhagen Studies in Language*, 38:319–336.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45.

Maria Kunilovskaya. 2023. *Translationese indicators for human translation quality estimation (based on English-to-Russian translation of mass-media texts)*. Ph.D. thesis, University of Wolverhampton.

Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 233–243.

Nikita Mathur. 2021. *Robustness in Machine Translation Evaluation*. Ph.D. thesis, University of Melbourne.

Bartolomé Mesa-Lao. 2014. Gaze behaviour on source texts: An exploratory study comparing translation and post-editing. In *Post-editing of machine translation: Processes and applications*, pages 219–245. Cambridge Scholars Publishing.

Mathias Müller, Annette Rios Gonzales, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164.

Jean Nitzke. 2019. *Problem solving activities in post-editing and translation from scratch: A multi-method study*. Language Science Press.

Dagmara Płońska. 2016. Problems of literality in french-polish translations of a newspaper article. *New directions in empirical translation process research: exploring the CRITT TPR-DB*, pages 279–291.

Thierry Poibeau. 2022. On "human parity" and "super human performance" in machine translation evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6018–6023.

Maja Popović. 2020. Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 256–264.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Márcia Schmaltz, Igor AL da Silva, Adriana Pagano, Fabio Alves, Ana Luísa V Leal, Derek F Wong, Lidia S Chao, and Paulo Quaresma. 2016. Cohesive relations in text comprehension and production: An exploratory study comparing translation and post-editing. *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*, pages 239–263.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Annette Camilla Sjørup. 2013. *Cognitive effort in metaphor translation: An eye-tracking and key-logging study*. Frederiksberg: Copenhagen Business School (CBS).

Lucia Specia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz. 2011. Predicting machine translation adequacy. In *Proceedings of Machine Translation Summit XIII: Papers*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2020. Semantic structural decomposition for neural machine translation. In *Proceedings of the ninth joint conference on lexical and computational semantics*, pages 50–57.

Elke Teich, José Martínez Martínez, and Alina Karakanta. 2020. Translation, information theory and cognition. *The Routledge Handbook of Translation and Cognition*, pages 9781315178127–24.

Bram Vanroy. 2021. *Syntactic difficulties in translation*. Ph.D. thesis, Ghent University.

Mihaela Vela and Liling Tan. 2015. Predicting machine translation adequacy with document embeddings. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 402–410.

Lucas Nunes Vieira, Natalie Zelenka, Roy Youdale, Xiaochun Zhang, and Michael Carl. 2023. Translating science fiction in a CAT tool: Machine translation and segmentation settings. *Translation & Interpreting*, 15(1):216–235.

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5696–5701.

Lei Yu, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Tomas Kocisky. 2016. The neural noisy channel. In *International Conference on Learning Representations*.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

Fei Yuan, Longtu Zhang, Huang Bojun, and Yaobo Liang. 2021. Simpson's bias in NLP training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14276–14283.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José GC De Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, et al. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99.

A Appendix

A.1 Data specification

A.1.1 Corpora

The CRITT Translation Process Research Database (Carl et al., 2016b) is a collection of translation behavioural data in the area of Translation Process Research. From the public CRITT database we obtain 15 studies across 13 pairs of languages: RUC17 (enzh, Carl and Báez, 2019), ENJA15 (enja, Carl et al., 2016a), NJ12 (enhi, Carl et al., 2016b), STC17 (enzh, Carl and Báez, 2019), SG12 (ende, Nitzke, 2019), ENDU20 (ennl, Vanroy, 2021), BML12 (enes, Mesa-Lao, 2014), ACS08 (daen, Sjørup, 2013), MS13 (ptzh, Schmaltz et al., 2016), JLG10 (pten, Alves and Gonçalves, 2013), BD13 (daen, Dragsted, 2010), LWB09 (daen, Jensen et al., 2009), DG01 (plfr, Płońska, 2016), BD08 (daen, Dragsted, 2010) and CREATIVE (enzh, Vieira et al., 2023).[10] After deduplication and removing source segments with fewer than 4 unique translations, the total number of source segments included is 399, each with an average of 10.9 unique translations.

RLTC is a subset of the Russian Learner Translator Corpus that has been aligned at the segment level by Kunilovskaya (2023). We include a total of 1079 source segments from 5 genres: 'Essay', 'Informational', 'Speech', 'Interview' and 'Educational'. The average number of unique translations for each source segment is 10.5.

MTMQM is obtained from (Freitag et al., 2021a), which contains translations of TED talks and news data from the test sets of WMT General Tasks between 2020 and 2023.[11] The translations are annotated with MQM labels. After preprocessing we are left with 11219 source segments and an average of 9.9 unique translations per source segment.

A.2 Alternative result with M2M100 translation model

In Figure 6 and 7, we replicate our findings of accuracy$_M$ and fluency$_M$ in Section 2 and 3 with estimates based on M2M100 (1.2B variant) (Fan et al., 2021).[12]

A.3 Tradeoff examples

Tables 2, 3 and 4 include the full set of translations plotted in Figure 1. The tables specify accuracy, fluency, accuracy$_M$, fluency$_M$ and translation probability $p(\boldsymbol{y}|\boldsymbol{x})$ for each segment. All translations listed are submissions to the WMT General Task between 2020 to 2022.

[10] https://sites.google.com/site/centretranslationinnovation/tpr-db/public-studies

[11] https://github.com/google/wmt-mqm-human-evaluation

[12] https://huggingface.co/facebook/m2m100_1.2B

Ich gab Ihnen eine Rückerstattung des Buches. {accuracy: 23.0, fluency: 25.0, accuracy$_M$: -10.81, fluency$_M$: -56.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -10.31}
Ich habe dir eine Rückerstattung des Buches ausgestellt. {accuracy: 23.0, fluency: 25.0, accuracy$_M$: -5.84, fluency$_M$: -62.5, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -12.44}
Ich habe dir das Buch zurückerstattet. {accuracy: 23.0, fluency: 25.0, accuracy$_M$: -17.5, fluency$_M$: -44.25, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -7.63}
Ich habe Ihnen das Buch erstattet. {accuracy: 24.0, fluency: 25.0, accuracy$_M$: -15.19, fluency$_M$: -43.25, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -9.06}
Ich habe Ihnen das Buch zurückerstattet. {accuracy: 24.2, fluency: 25.0, accuracy$_M$: -17.25, fluency$_M$: -43.5, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -7.28}
Ich habe Ihnen eine Rückerstattung des Buches ausgestellt. {accuracy: 24.3, fluency: 24.67, accuracy$_M$: -6.13, fluency$_M$: -64.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -12.13}
Ich stellte Ihnen eine Rückerstattung des Buches aus. {accuracy: 25.0, fluency: 23.0, accuracy$_M$: -6.44, fluency$_M$: -70.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -14.75}
Ich habe Ihnen eine Rückerstattung für das Buch erteilt. {accuracy: 25.0, fluency: 24.0, accuracy$_M$: -11.56, fluency$_M$: -63.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -14.19}

Table 2: Translations of *I issued you a refund of the book.* (plotted in orange in Figure 1). Accuracy and fluency scores are derived from MQM ratings, and accuracy$_M$ and fluency$_M$ are estimates of $\log\, p(\boldsymbol{x}|\boldsymbol{y})$ and $\log\, p(\boldsymbol{y})$ derived from an NMT model.

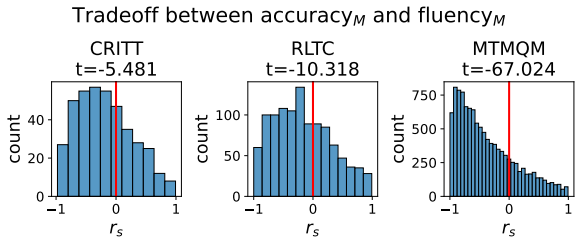


Figure 6: Histogram of tradeoffs between estimated $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ estimated by M2M100, which is analogous to Figure 3 in the main text. When analyzed at the corpus level, the correlations r_c for for CRITT, RLTC and MTMQM are .689, .703 and .801 respectively ($p < .001$ in all cases).

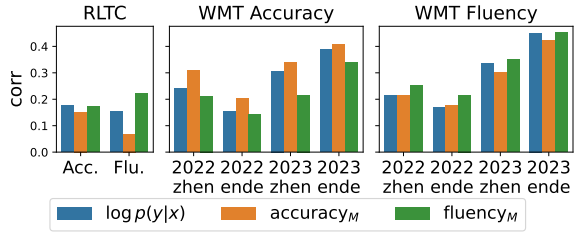


Figure 7: accuracy$_M$ and fluency$_M$ estimates based on M2M100 predict human accuracy and fluency ratings ($p < .05$). The figure is analogous to Figure 5.

Ashanti Development arbeitet seit fast 20 Jahren mit einer wachsenden Anzahl von Gemeinden in der Region Ashanti in Ghana zusammen und unterstützt sie in den Bereichen Wasser und sanitäre Einrichtungen, Bildung, Gesundheitsversorgung, Baumpflanzung und Landwirtschaft. {accuracy: 19.0, fluency: 25.0, accuracy$_M$: -120.5, fluency$_M$: -498.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -27.0}
Ashanti Development arbeitet seit fast zwanzig Jahren mit einer ständig wachsenden Anzahl von Gemeinden in der Region Ashanti in Ghana zusammen, engagiert sich mit Gemeinden und unterstützt Wasser und Sanitärversorgung, Bildung, Gesundheitsversorgung, Baumpflanzung und Landwirtschaft. Gemeinschaften erlangen das Wissen, um ihre eigene Entwicklung einzubetten und zu unterstützen. {accuracy: 22.0, fluency: 24.0, accuracy$_M$: -47.5, fluency$_M$: -748.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -47.25}
Ashanti Development arbeitet seit fast 20 Jahren mit einer ständig wachsenden Zahl von Gemeinden in der Ashanti-Region in Ghana zusammen, engagiert sich für Gemeinden und bietet Unterstützung in den Bereichen Wasser und sanitäre Einrichtungen, Bildung, Gesundheitsversorgung, Baumpflanzung und Landwirtschaft. Communities erwerben das Wissen, um ihre eigene Entwicklung zu verankern und zu unterstützen. {accuracy: 22.0, fluency: 25.0, accuracy$_M$: -46.5, fluency$_M$: -832.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -49.0}
Ashanti Development arbeitet seit 20 Jahren mit einer immer größeren Zahl von Gemeinden in der Region Ashanti in Ghana zusammen, engagiert sich mit Gemeinden und unterstützt Wasser und Sanitärversorgung, Bildung, Gesundheitsversorgung, Baumpflanzung und Landwirtschaft. {accuracy: 23.0, fluency: 24.9, accuracy$_M$: -101.0, fluency$_M$: -516.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -39.25}
Ashanti Development arbeitet seit fast 20 Jahren mit einer ständig wachsenden Anzahl von Gemeinden in der Ashanti-Region Ghanas zusammen, indem es sich mit Gemeinden beschäftigt und ihnen Unterstützung in den Bereichen Wasser und Sanitärversorgung, Bildung, Gesundheitsversorgung, Baumpflanzung und Landwirtschaft bietet. {accuracy: 23.0, fluency: 25.0, accuracy$_M$: -98.5, fluency$_M$: -652.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -29.625}
Ashanti Development arbeitet seit fast 20 Jahren mit einer ständig wachsenden Zahl von Gemeinden in der Ashanti-Region in Ghana zusammen, arbeitet mit Gemeinden zusammen und unterstützt sie in den Bereichen Wasser und Abwasserentsorgung, Bildung, Gesundheitswesen, Baumpflanzung und Landwirtschaft. Gemeinschaften erwerben das Wissen, um ihre eigene Entwicklung zu verankern und zu unterstützen. {accuracy: 23.0, fluency: 24.0, accuracy$_M$: -53.0, fluency$_M$: -828.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -42.5}
Ashanti Development arbeitet seit fast 20 Jahren mit einer ständig wachsenden Anzahl von Gemeinden in der Ashanti-Region in Ghana zusammen, engagiert sich für Gemeinden und unterstützt sie bei Wasser- und Sanitärversorgung, Bildung, Gesundheitswesen, Baumpflanzung und Landwirtschaft. Gemeinschaften gewinnen das Wissen, um ihre eigene Entwicklung einzubetten und zu unterstützen. {accuracy: 24.0, fluency: 23.0, accuracy$_M$: -47.5, fluency$_M$: -784.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -45.25}
Ashanti Development arbeitet seit fast 20 Jahren mit einer stetig wachsenden Anzahl von Gemeinden in der Ashanti-Region in Ghana zusammen, engagiert sich mit Gemeinden und bietet Unterstützung in den Bereichen Wasserversorgung und Abwasserentsorgung, Bildung, Gesundheitsversorgung, Baumpflanzung und Landwirtschaft. Gemeinden erwerben das Wissen, um ihre eigene Entwicklung zu verankern und zu unterstützen. {accuracy: 24.0, fluency: 24.0, accuracy$_M$: -49.5, fluency$_M$: -848.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -42.25}
Ashanti Development arbeitet seit fast 20 Jahren mit einer stetig wachsenden Anzahl von Gemeinschaften in der Ashanti-Region von Ghana zusammen, engagiert sich in den Gemeinschaften und bietet Unterstützung in den Bereichen Wasser und Sanitär, Bildung, Gesundheitswesen, Baumpflanzung und Landwirtschaft. Die Gemeinschaften erwerben das Wissen, um ihre eigene Entwicklung zu verankern und zu unterstützen. {accuracy: 25.0, fluency: 22.0, accuracy$_M$: -50.25, fluency$_M$: -828.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -43.0}
Ashanti Development arbeitet seit fast 20 Jahren mit einer ständig wachsenden Zahl von Gemeinden in der Ashanti-Region in Ghana zusammen, engagiert sich für Gemeinden und leistet Unterstützung bei Wasser- und Sanitärversorgung, Bildung, Gesundheitsversorgung, Baumpflanzung und Landwirtschaft. Gemeinschaften erlangen das Wissen, um ihre eigene Entwicklung zu verankern und zu unterstützen. {accuracy: 25.0, fluency: 24.0, accuracy$_M$: -45.75, fluency$_M$: -816.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -45.0}
Ashanti Development arbeitet seit fast 20 Jahren mit einer ständig wachsenden Zahl von Gemeinden in der Ashanti-Region in Ghana zusammen und unterstützt sie in den Bereichen Wasserversorgung und Abwasserentsorgung, Bildung, Gesundheitsversorgung, Baumpflanzung und Landwirtschaft. Die Gemeinden erlangen das Wissen, um ihre eigene Entwicklung zu fördern und zu unterstützen. {accuracy: 25.0, fluency: 24.0, accuracy$_M$: -74.0, fluency$_M$: -768.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -42.0}
Ashanti Development arbeitet seit fast 20 Jahren mit einer ständig wachsenden Zahl von Gemeinden in der Ashanti-Region in Ghana zusammen, engagiert sich für Gemeinden und leistet Unterstützung bei Wasser- und Sanitärversorgung, Bildung, Gesundheitswesen, Baumpflanzung und Landwirtschaft. Gemeinschaften erwerben das Wissen, um ihre eigene Entwicklung zu verankern und zu unterstützen. {accuracy: 25.0, fluency: 24.0, accuracy$_M$: -46.25, fluency$_M$: -812.0, $\log p(\boldsymbol{y}\|\boldsymbol{x})$: -46.25}

Table 3: Translations of *Ashanti Development has been working with an ever-expanding number of communities in the Ashanti region of Ghana for approaching 20 years, engaging with communities and providing support with water and sanitation, education, healthcare, tree planting and farming. Communities gain the knowledge to embed and support their own development.* These translations are plotted in green in Figure 1.

Interessanterweise war eine der 12 Galaxien in z66OD ein riesiges Objekt mit einem riesigen Gaskörper, bekannt als Himiko, das zuvor 2009 vom Subaru-Teleskop gefunden wurde. „Es ist vernünftig, einen Protohaufen in der Nähe eines massereichen Objekts wie Himiko zu finden. Wir sind jedoch überrascht zu sehen, dass Himiko nicht im Zentrum des Protohaufens lag, sondern am Rande 500 Millionen Lichtjahre vom Zentrum entfernt.“ Sagte Masami Ouchi, ein Teammitglied am Nationalen Astronomischen Observatorium von Japan und der Universität von Tokio, die Himiko im Jahr 2009 entdeckte, dass die Beziehung zwischen den Himiko und den Himiko-Klöstern noch immer nicht verstanden wird.
{accuracy: 0.0, fluency: 22.9, accuracy$_M$: -286.0, fluency$_M$: -1904.0, $\log p(\boldsymbol{y}|\boldsymbol{x})$: -139.0}

"""""""""Interessanterweise war eine der 12 Galaxien in z66OD ein riesiges Objekt mit einem riesigen Gaskörper, bekannt als Himiko, das zuvor vom Subaru-Teleskop im Jahr 2009 gefunden wurde. """""""""Es ist vernünftig, einen Protocluster in der Nähe eines massiven Objekts wie Himiko zu finden. Wir sind jedoch überrascht zu sehen, dass Himiko nicht im Zentrum des Protoclusters, sondern am Rand 500 Millionen Lichtjahre vom Zentrum entfernt war""""""""", sagte Masami Ouchi, ein Teammitglied am Nationalen Astronomischen Observatorium von Japan und der Universität von Tokio, der Himiko im Jahr 2009 entdeckte. Ironischerweise soll die mythologische Königin Himiko auch abgeschieden von ihrem Volk gelebt haben. Ouchi fährt fort: """""""""Es ist immer noch nicht verstanden, warum Himiko nicht im Zentrum liegt. Diese Ergebnisse werden ein Schlüssel für das Verständnis der Beziehung zwischen Haufen und massiven Galaxien sein"""""""""
{accuracy: 1.0, fluency: 23.4, accuracy$_M$: -125.0, fluency$_M$: -2624.0, $\log p(\boldsymbol{y}|\boldsymbol{x})$: -103.5}

"""""""""Interessanterweise war eine der 12 Galaxien in z66OD ein riesiges Objekt mit einem riesigen Gaskörper, bekannt als Himiko, das zuvor vom Subaru-Teleskop im Jahr 2009 gefunden wurde. „Es ist vernünftig, einen Protocluster in der Nähe eines massiven Objekts, wie Himiko, zu finden. Allerdings sind wir überrascht zu sehen, dass Himiko nicht im Zentrum des Protoclusters, sondern am Rande 500 Millionen Lichtjahre vom Zentrum entfernt war.“, sagte Masami Ouchi, ein Teammitglied am Nationalen Astronomischen Observatorium von Japan und der Universität von Tokio, der Himiko im Jahr 2009 entdeckte. Ironischerweise soll die mythologische Königin Himiko auch abgeschieden von ihrem Volk gelebt haben. Ouchi fährt fort: """""""""Es ist immer noch nicht verstanden, warum Himiko nicht im Zentrum liegt. Diese Ergebnisse werden ein Schlüssel für das Verständnis der Beziehung zwischen Haufen und und massiven galaxien sein."""""""""""""""""
{accuracy: 6.0, fluency: 24.0, accuracy$_M$: -121.0, fluency$_M$: -2688.0, $\log p(\boldsymbol{y}|\boldsymbol{x})$: -143.0}

"""""""""Interessanterweise war eine der 12 Galaxien in z66OD ein riesiges Objekt mit einem riesigen Gaskörper, bekannt als Himiko, das zuvor 2009 vom Subaru-Teleskop gefunden wurde. „Es ist vernünftig, einen Protocluster in der Nähe eines massiven Objekts wie Himiko zu finden. Wir sind jedoch überrascht zu sehen, dass sich Himiko nicht im Zentrum des Protoclusters befand, sondern am Rande 500 Millionen Lichtjahre vom Zentrum entfernt“, sagte Masami Ouchi, Teammitglied am National Astronomical Observatory of Japan und der Universität Tokio, der Himiko 2009 entdeckte. Ironischerweise soll die mythologische Königin Himiko auch abgeschieden von ihrem Volk gelebt haben. Ouchi fährt fort: """""""""Es ist immer noch nicht verstanden, warum Himiko nicht im Zentrum liegt. Diese Ergebnisse werden ein Schlüssel für das Verständnis der Beziehung zwischen Haufen und massiven Galaxien sein."""""""""""""""""
{accuracy: 6.0, fluency: 22.7, accuracy$_M$: -126.0, fluency$_M$: -2592.0, $\log p(\boldsymbol{y}|\boldsymbol{x})$: -123.0}

Interessanterweise war eine der 12 Galaxien in z66OD ein riesiges Objekt mit einem riesigen Gaskörper, bekannt als Himiko, das zuvor 2009 vom Subaru-Teleskop gefunden wurde. „Es ist vernünftig, einen Protocluster in der Nähe eines massiven Objekts wie Himiko zu finden. Wir sind jedoch überrascht zu sehen, dass Himiko sich nicht im Zentrum des Protoclusters befand, sondern am Rand 500 Millionen Lichtjahre vom Zentrum entfernt“, sagte Masami Ouchi, Teammitglied am National Astronomical Observatory of Japan und der Universität Tokio, der Himiko 2009 entdeckte. Ironischerweise soll die mythologische Königin Himiko auch abseits ihres Volkes im Kloster gelebt haben. Ouchi fährt fort: „Es ist immer noch nicht verstanden, warum Himiko sich nicht im Zentrum befindet. Diese Ergebnisse werden ein Schlüssel zum Verständnis der Beziehung zwischen Haufen und massiven Galaxien sein.“
{accuracy: 9.0, fluency: 22.0, accuracy$_M$: -131.0, fluency$_M$: -2512.0, $\log p(\boldsymbol{y}|\boldsymbol{x})$: -108.0}

"""""""""Interessanterweise war eine der 12 Galaxien in z66OD ein riesiges Objekt mit einem riesigen Gaskörper, bekannt als Himiko, das 2009 vom Subaru-Teleskop gefunden wurde. """""""""Es ist vernünftig, einen Protokluster in der Nähe eines massiven Objekts zu finden, wie z Himiko. Wir sind jedoch überrascht zu sehen, dass sich Himiko nicht in der Mitte des Protoklusters befand, sondern am Rand von 500 Millionen Lichtjahren vom Zentrum entfernt. """""""""" sagte Masami Ouchi, ein Teammitglied des Nationalen Astronomischen Observatoriums Japans und der Universität Tokio, das Himiko 2009 entdeckte. Ironischerweise soll die mythologische Königin Himiko auch im Kloster von ihrem Volk gelebt haben. Ouchi fährt fort: """""""""Es ist immer noch nicht klar, warum Himiko nicht im Zentrum liegt. Diese Ergebnisse werden ein Schlüssel zum Verständnis der Beziehung zwischen Clustern und massiven Galaxien sein."""""""""""""""""
{accuracy: 13.0, fluency: 20.7, accuracy$_M$: -132.0, fluency$_M$: -2688.0, $\log p(\boldsymbol{y}|\boldsymbol{x})$: -127.0}

"""""""""Interessanterweise war eine der 12 Galaxien in z66OD ein riesiges Objekt mit einem riesigen Gaskörper, bekannt als Himiko, das zuvor vom Subaru-Teleskop im Jahr 2009 gefunden wurde. """""""""Es ist vernünftig, einen Protocluster in der Nähe eines massiven Objekts wie Himiko zu finden. Wir sind jedoch überrascht zu sehen, dass Himiko nicht im Zentrum des Protoclusters lag, sondern am Rande 500 Millionen Lichtjahre vom Zentrum entfernt""""""""", sagte Masami Ouchi, Teammitglied am Nationalen Astronomischen Observatorium Japans und der Universität Tokio, der Himiko 2009 entdeckte. Ironischerweise soll auch die mythologische Königin Himiko von ihrem Volk abgeschottet gelebt haben. Ouchi fährt fort: """""""""Es ist immer noch nicht klar, warum Himiko nicht in der Mitte liegt. Diese Ergebnisse werden ein Schlüssel zum Verständnis der Beziehung zwischen Clustern und massiven Galaxien sein."""""""""""""""""
{accuracy: 16.0, fluency: 21.3, accuracy$_M$: -122.5, fluency$_M$: -2624.0, $\log p(\boldsymbol{y}|\boldsymbol{x})$: -111.0}

Table 4: Translations of *"""Interestingly, one of the 12 galaxies in z66OD was a giant object with a huge body of gas, known as Himiko, which was found previously by the Subaru Telescope in 2009. """""""""It is reasonable to find a protocluster near a massive object, such as Himiko. However, we're surprised to see that Himiko was located not in the center of the protocluster, but on the edge 500 million light-years away from the center."""""""""* said *Masami Ouchi, a team member at the National Astronomical Observatory of Japan and the University of Tokyo, who discovered Himiko in 2009. Ironically, the mythological queen Himiko is also said to have lived cloistered away from her people. Ouchi continues, """""""""It is still not understood why Himiko is not located in the center. These results will be a key for understanding the relationship between clusters and massive galaxies."""""""""""""* These translations are plotted in blue in Figure 1.