Alligators All Around: Mitigating Lexical Confusion in Low-resource Machine Translation

Elizabeth Nielsen **Isaac Caswell** **Jiaming Luo** **Colin Cherry**
Google
{eknielsen, icaswell, jmluo, colincherry}@google.com

Abstract

Current machine translation (MT) systems for low-resource languages have a particular failure mode: When translating words in a given domain, they tend to confuse words within that domain. So, for example, *lion* might be translated as *alligator*, and *orange* might be rendered as *purple*. We propose a recall-based metric for measuring this problem and show that the problem exists in a dataset comprising 122 low-resource languages. We then show that this problem can be mitigated by using a large language model (LLM) to post-edit the MT output, specifically by including the entire GATITOS lexicon for the relevant language as a very long context prompt. We show gains in average CHRF score over the set of 122 languages, and we show that the recall score for relevant lexical items also improves. Finally, we demonstrate that a small dedicated MT system with a general-purpose LLM as a post-editor outperforms a generalist LLM translator with access to the same lexicon data, suggesting a new paradigm for LLM use.

1 Introduction

Machine translation systems have recently expanded to cover many previously unsupported languages (Bapna et al., 2022b; NLLB et al., 2022). However, MT systems for low-resource languages (LRLs) still face many challenges. One particular difficulty is learning the correct mapping of words between two languages. This paper is motivated by the observation that some LRL MT models tend to confuse certain lexical items belonging to similar domains. This problem is first reported in Bapna et al. (2022b), who report this issue with unsupervised, sentence-level NMT, giving the following examples from their models. Examples from their paper are reproduced in Table 1.

These examples show that the model consistently errs by confusing lexical items that share similar distributions, such as using *crocodile* to translate other animal terms. This pattern is observed in the "next thousand languages" (NTL) MT models of Bapna et al. (2022b) over many language pairs and within relatively high-frequency lexical domains, including numbers, colors, animals, days of the week, and months. In this paper, we refer to the tendency to confuse words within a domain as the "alligator problem."[1] As we show in this paper, this pattern isn't only found in MT-specific models, but in translations produced by large language models (LLMs) as well.

Using a development set consisting of data from 122 LRLs, we show that this problem is widespread in translations of the NTL models, which are described in Bapna et al. (2022b). We then propose a method for prompting an LLM with lexical information to post-edit these translations, both translating into and out of English, leading to better performance on these frequently confused lexical items, as well as higher machine translation quality overall. The lexical information is provided by incorporating the GATITOS lexicon (Jones et al., 2023) into the LLM prompt. We further show that the LLM is able to improve its performance on these lexical items even when the lexicon entries presented in the prompt don't exactly match the source string because of morphological inflections.

This method combines the in-depth knowledge of the specialist NTL MT systems with the generalist abilities of the LLM. We show that the LLM is incapable of matching the MT system's performance on its own, even when given access to the lexicon, despite the fact that the MT system is much smaller, at only 850M parameters. However, given the specialist MT model's best hypotheses, the LLM can fix the MT model's persistent lexical confusions as a post-editor, making use of the infor-

[1] Not the "crocodile problem," because somewhere between encountering the crocodile-filled examples from Bapna et al. (2022b) and starting this work, we confused alligators and crocodiles. We kept the name, though, since our mistake is itself a nice illustration of the problem.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 206–221

Language	reference	translation
Meiteilon (mni)	I believe a **lion** is **stronger** than a **tiger**.	I believe a **snake** is **stronger** than a **crocodile**.
Twi (ak)	I would want to be a **dog** for a day.	I want to be a **crocodile** just one day.

Table 1: Examples from Bapna et al. (2022a) of the "alligator problem"

mation in the lexicon. Our primary contributions are:

- Demonstrating that the "alligator problem" (lexical confusion on distributionally similar words) is a failure mode not only in traditional MT, but also in LLMs.
- Developing a targeted evaluation for the alligator problem, and demonstrating a method for fixing the problem by using an LLM as post-editor with a lexicon as context.
- Revealing that specialist MT models still far outperform generalist LLMs on LRL translation, and introducing a new paradigm of generalist-LLM-as-post-editor.

2 Related work

MT for low-resource languages Before LLMs, for Very Low-Resource Language MT — i.e. anything beyond the most frequent hundred languages or so — there existed no parallel text at all outside of religious domains. In these cases, the only option was Unsupervised Machine Translation (UNMT), which uses only monolingual text to translate. This was pioneered in Lample et al. (2017); Artetxe et al. (2017); Song et al. (2019a), and eventually Bapna et al. (2022a) scaled up to 1000 languages in the NTL models. However, the unsupervised paradigm led to tell-tale mistakes, such as the "alligator problem" discussed here.

LLMs then barged in and changed all these paradigms, although they still perform poorly out of the box on LRLs (Kocmi et al., 2023). A common approach is in-context learning, or ICL (Brown et al., 2020; Agarwal et al., 2024) which gives examples in the prompt. ICL examples for LRLs have included diverse context like sentence pairs (Zhang et al., 2024; Tanzer et al., 2024), dictionaries (Elsner and Needle, 2023), the full GATITOS lexicon (Reid et al., 2024), and a full grammar of the Kalamang language (Tanzer et al., 2024). A popular variant of ICL is RAG, or Retrieval-augmented generation (Rubin et al., 2022), which draws only on examples for ICL that are relevant to the current sentence being translated. Despite its popularity, Vilar et al. (2023); Zhu et al. (2024); Zhang et al. (2023) find exemplar quality is more important than relevance.

LLMs as post-editors. Another less common approach for LRL MT has focused on automatic post-editing (APE) translations with LLMs, which is an approach often used in high-resource MT (Bhattacharyya et al., 2023; Zerva et al., 2024). Chen et al. (2024) let an LLM iteratively self-correct its translation, Lim et al. (2024) have a model post-edit its own translations from related higher-resource languages into the target language, and Xu et al. (2024) iteratively apply fine-grained error correction from an LLM. However, these efforts have focused on a base model and a post-editor that are the same size, and both large.

Rare word translation Many MT models struggle specifically with translating rare words, including MT models for high-resource languages. In our case, we study the inverse problem of difficulties with *common* words, but the approaches necessary to fix may be the same. Prior work includes placing soft constraints on the output terminology (Bergmanis and Pinnis, 2021) and augmenting parametric models with non-parametric datastores such as parallel corpora (Khandelwal et al., 2021) or lexica (Zhang et al., 2021). The latter is more similar to our approach, though we present a lexicon to the LLM as a part of a prompt, rather than using it during the training phase.

3 Methods

The approach we take to solving this problem is to (1) generate output for a set of LRLs using a specialist MT system; (2) create prompts for post-editing each segment that include the entire GATITOS lexicon, and (3) use these prompts to generate post-edited output using a generalist LLM. The example in Table 2 illustrates how a single Udmurt example passes through the pipeline of specialist MT system and LLM-posteditor:

3.1 Data

Evaluation data. To measure the magnitude of this problem, we evaluate the performance of the

Source	5:30 chasysen 2:30 chasoz' vordïs'konysen **kösnynaloz'**
Reference	between 5:30 am to 2:30 am from Mondays to **Saturdays**
MT output	from 5:30 a.m. to 2:30 a.m. Monday through **Friday**
Post-edit	from 5:30 a.m. to 2:30 a.m. Monday through **Saturday**

Table 2: An example of how the MT model and postediting step render a single example from Udmurt. The alligator problem is shown by the error highlighted in red, which is corrected by the postediting step.

models on 122 LRLs, translating into and out of English (complete list in Appendix C). The evaluation data comprises segments from FLORES-200 (NLLB et al., 2022), NTREX (Barrault et al., 2019; Federmann et al., 2022) and GATONES (Jones et al., 2023). For each language pair, there are 600-1000 segments.

Prompting data. This lexical information comes from the GATITOS lexicon (Jones et al., 2023). This is a 4000-entry multilingual lexicon with English segments, which have been translated by human translators into 170 very low resource languages. These lexical segments include frequent English tokens (including words for numbers, months, and days of the week), Swadesh wordlists (Swadesh, 1952), and some short English sentences.

3.2 Metrics

General MT quality: To measure general quality we report CHRF score (Popović, 2015).

Alligator recall: CHRF will not necessarily reflect wins or losses in the alligator problem. To directly measure this problem, we propose a recall-based metric over a set of predetermined lexical items with similar distributions, which we call *alligator recall*. The selected lexical items are shown in Appendix A, and are grouped into the domains of animals, colors, weekdays, months, common numbers, and rare numbers. They are restricted to terms that are in the GATITOS lexicon. For a given evaluation set, we find all references that have one of these words, and score the model hypotheses on whether they 1) produced the exact correct word (CORRECT); 2) produced a *different* in-domain word (CONFUSION, i.e., the alligator problem); or 3) produced neither a correct nor incorrect word (UNKNOWN). If a total of N alligator words appear in the set of all reference strings, and the model's hypotheses produce the corresponding correct alligator word R times and a different in-domain word W times, then we report the corresponding alligator scores as follows:

$$\text{CORRECT} = \frac{R}{N} \quad (1)$$

$$\text{CONFUSION} = \frac{W}{N} \quad (2)$$

$$\text{UNKNOWN} = \frac{N - R - W}{N} \quad (3)$$

We only report alligator recall for the into-English direction. Measuring the presence or absence of a word in the model output via simple string matching is problematic for more morphologically complex languages. For example, the Udmurt word for *April* is listed in citation form as *oshtolez'*. However, in one phrase in our evaluation data, "in April 2020," it is inflected to *oshtoleze* — with the final character of the citation form (transliterated as ') removed, and the suffix *-e* added. If we calculated alligator recall on Udmurt target data, we would count inflections like these as non-matches, unless we accounted for morphological inflection. However, accommodating the diverse morphologies of 122 languages is outside the scope of this paper. Therefore, for the out-of-English translation direction, we report only CHRF.

3.3 Models

We use the NTL MT models as our baseline (Bapna et al., 2022b). These are sentence-level, unsupervised transformer translation models, that are trained as follows: First, for each language in their training data (a set which includes our 122 evaluation languages), an encoder-decoder Transformer model with 6B parameters is trained. Because data is limited, this first phase uses a MASS de-noising task on monolingual data (Song et al., 2019b). The second phase of training consists of iterative back-translation, where the models are used to generate parallel data via online translation, and then trained on this synthetic data. Finally, these models are distilled into multilingual 850M parameter encoder-decoder models, and cover either the en > xx or xx > en direction.

For post-editing, we use the LLM Gemini 1.5 Pro (Reid et al., 2024), whose long context (up to 10M tokens) is ideal for our purposes. We perform greedy decoding to generate outputs.

4 Results and discussion

Tables 3 and 4 show that the best performance comes from using the LLM as a post-editor, and including the entire GATITOS lexicon in the prompt. The models we compare are (1), the MT models alone (our baseline), (2) the LLM model alone, and (3) the LLM as post-editor of the MT model output. The exact prompt templates are in Appendix B. The prompts given to the LLM include all 4000 entries from GATITOS for the given language, except when noted otherwise.

As shown in Table 3, lexical confusion is present in the initial MT system output, but when averaged over all evaluation languages, its severity is limited. When we subsample the 20% of languages with the highest level of lexical confusion, it becomes clear that this issue is much more severe for some languages than for others.[2] The highest quality output is consistently produced by prompting the LLM to postedit the MT system output. The lexical recall gains are particularly concentrated in the languages that had the highest rates of lexical confusion.

Other attempted methods fall short of the performance of LLM post-editing with access to the whole lexicon. The LLM on its own is a relatively poor translator, even given the entire GATITOS lexicon. On these high-confusion languages, we also experiment with presenting the LLM with a few different levels of lexical information: no lexical information, prompts with only the words in the given segment, and prompts with the whole lexicon. No lexical information is, as expected, a worse condition, but even limiting the prompt to include only the lexical items that are present in the source is unhelpful — this condition under-performs even the baseline.

As expected, the prevalence of lexical confusion correlates with the overall performance of the MT systems on a language, as shown in Table 3, where languages with higher confusion have lower CHRF score. For per-language scores, see Appendix C.

[2]For the list of languages constituting the high-confusion group, see the table in Appendix C.

4.1 Morphology and the shortcomings of string-match RAG

One reason why prompts with targeted lexical information fail may be that retrieving words from the lexicon for the prompt is difficult in languages with complex morphology: string matching can't retrieve words that don't appear in the *citation form* (the uninflected root form) in MT input. To measure how often a retrieval from the lexicon would fail, we identify times when an English word from our evaluation list (see Appendix A) appears in the gold reference in the xx→en direction. We then count how often the word is missing in the initial MT system output, but appears in the post-editing output of the LLM prompted with the whole lexicon. Of the cases where post-editing recovers the correct word, we measure how often the corresponding source language token (from GATITOS) appears in the source in citation form.

The citation form occurs in the source side only 56.1% of the total times that the post-editing procedure correctly recovered a lexical item. This suggests that the LLM was able to use information in the lexicon even when retrieval of the correct item from the lexicon would have required going beyond an exact match. A significant source of these retrieval failures is likely the morphological inflections in the source string that complicate retrieval. Recall the example given in Section 3.2: the Udmurt word for *April* is *oshtolez'*, but it appears in the evaluation data in an inflected form, *oshtoleze*, as part of a phrase meaning, "in April 2020." In this inflected form, the final character of the citation form (transliterated as ') is removed, and the suffix *-e* added. This makes direct retrieval of this item from the lexicon difficult. Additionally, the substitution of synonyms in the source string would affect this. Whether these retrieval failures are due to morphological inflection or synonymy, the LLM is able to recover the correct target word in many of these cases when simply given the entire lexicon and handles lexical variations itself.

5 Conclusion

This work is the first to document and quantify the *alligator problem* in Large Language Models for low resource languages, a systemic translation error mode that is not well captured in metrics like CHRF. This problem is much reduced, though not fully eliminated, by our proposed approach of lexicon-augmented post-editing. This also suggests

		Alligator recall scores			
		Correct (↑)	**Confusion (↓)**	**Unknown (↓)**	**ChrF (↑)**
All languages	**Baseline**	59.4	3.8	36.9	52.4
	Direct translation	2.9	4.1	93.0	48.5
	Post-edit, whole lex.	**62.4**	**2.8**	**34.8**	**53.2**
High-confusion languages	**Baseline**	49.6	8.3	42.2	45.3
	Direct translation	2.4	3.6	94.0	39.8
	Post-edit, whole lex.	**57.0**	**4.8**	**38.2**	**46.3**
	Post-edit, targeted lex.	53.0	6.0	40.9	43.7
	Post-edit, no lex.	51.1	7.2	41.7	44.9

Table 3: CHRF and lexical recall scores for the xx→en translation direction. High-confusion languages are the top quintile of languages by confusion score. "Post-edited" scores represent the output of the LLM that has been prompted to postedit the MT output.

		ChrF (↑)
All langs.	**Baseline**	43.5
	Direct translation	40.9
	Post-edit, whole lex.	**44.0**
High-conf. langs.	**Baseline**	37.6
	Direct translation	35.9
	Post-edit, whole lex.	**38.2**
	Post-edit, target lex.	34.4
	Post-edit, no lex.	36.5

Table 4: CHRF scores for the en→xx direction. High-confusion languages are the top 20% of languages by confusion in the xx→en direction. Alligator scores are not reported in this direction, since it can't be reliably calculated on non-English output.

a new paradigm for generalist models like LLMs, exploiting their better general-purpose reasoning and tool use to use them as post-editors. The small, specialized MT model provides a strong baseline for translation performance, one that the LLM cannot meet on its own, even when given access to a lexicon. However, the LLM can better extract and use information from a resource like GATITOS, and therefore improve upon its superior's work. The LLM is also able to overcome challenges such as complex morphology that would make it prohibitively difficult to use the lexicon directly to post-edit the MT output.

Limitations

One limitation of this work is the fact that exact string matching is used in the alligator recall evaluation, which doesn't account for morphological inflection or synonymy. So for example, if the word *twelve* appeared in the reference and the model output *a dozen*, this would fall into the UNKNOWN category of the metric rather than the CORRECT category, where it likely belongs. Likewise, if the reference word is morphologically inflected in such a way that the citation form doesn't appear in the output (e.g., *geese* instead of *goose*), it would fall into the UNKNOWN category. This is mitigated by the fact that the set of evaluation words we use have relatively few synonyms (weekdays, months, and common numbers, for example). All of them are also nouns with regular plurals, so even when they appear in an inflected form (plural being the only option for English nouns), the citation form should appear as a substring in the target output.

Other limitations include using a hand-picked set of words over which to evaluate the alligator problem. Finally, it would be preferable to be able to perform the alligator recall metric on non-English output. Addressing the English-only nature of this evaluation would require handling the morphology of 122 very low-resource languages, which would almost certainly require producing more resources for them, which lies outside the scope of this work.

References

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod,

Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022a. Building machine translation systems for the next thousand languages.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Nikhil Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Saldinger Axelrod, Jason Riesa, Yuan Cao, Mia Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apu Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Richard Hughes. 2022b. Building machine translation systems for the next thousand languages. Technical report, Google Research.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2023. Findings of the WMT 2023 shared task on automatic post-editing. In *Proceedings of the Eighth Conference on Machine Translation*, pages 672–681, Singapore. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models.

Micha Elsner and Jordan Needle. 2023. Translating a low-resource language using GPT-3 and a human-readable dictionary. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–13, Toronto, Canada. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. *ICLR 2021*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 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, Singapore. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Zheng Wei Lim, Nitish Gupta, Honglin Yu, and Trevor Cohn. 2024. Mufu: Multilingual fused learning for low-resource translation with llm.

Team NLLB, Marta Costa-jussa, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Gonzalez, Prangthip Hansanti, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. Technical report.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis

Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem W. Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomás Kociský, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, J Christopher Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Ying-Qi Miao, Lukás Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontan'on, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, A.E. Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Venkatesh Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matt Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara N. Sainath, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela de Castro Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, S'ebastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Joshua Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost R. van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya B Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, S'ebastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Michael B. Chang, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravichandra Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Luvci'c, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjosund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos L. Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Zhufeng Pan, Zachary Nado, Stephanie Winkler, Dian Yu, Mohammad Saleh, Lorenzo Maggiore, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Chung-Cheng Chiu, Zoe C. Ashwood, Khuslen Baatarsukh, Sina Samangooei, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabriel Barth-Maron, Craig Swanson, Dominika Rogozi'nska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Ren shen Wang, Dave

Lacey, Anastasija Ili'c, Yao Zhao, Woohyun Han, Lora Aroyo, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Raphael Lopez Kaufman, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, T. Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anais White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Kiran Vodrahalli, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, cCauglar Unlu, David Reid, Zora Tung, Daniel F. Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Rui Zhu, Ricardo Aguilar, Mai Gim'enez, Jiawei Xia, Olivier Dousse, Willi Gierke, Soheil Hassas Yeganeh, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Daniel Niels Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Daniel Toyama, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nicholas Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, Donghyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alexey Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Anna Bulanova, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Cl'ement Farabet, Pedro Valenzuela, Quan Yuan, Christoper A. Welty, Ananth Agarwal, Mianna Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Elnaz Davoodi, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, A. Ya. Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Alejandro Lince, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiří Šimša, Anna Koop, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Kalpesh Krishna, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas Fitzgerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel El Kaed, Jing Li, Jakub Sygnowski, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Poder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Junwen Bai, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, Oriol Vinyals, and Alexandra Chronopoulou. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019a. MASS: masked sequence to sequence pre-training for language generation. *CoRR*, abs/1905.02450.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019b. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to north american indians and eskimos. *Proceedings of the American Philosophical Society*, 96(4):452–463.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. Fine-grained llm agent: Pinpointing and refining large language models via fine-grained actionable feedback.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin

Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly.

Tong Zhang, Long Zhang, Wei Ye, Bo Li, Jinan Sun, Xiaoyu Zhu, Wen Zhao, and Shikun Zhang. 2021. Point, disambiguate and copy: Incorporating bilingual dictionaries for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3970–3979, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis.

A Lexical items for recall metric

B LLM prompts

B.1 Direct translation prompt with entire lexicon

You are asked to translate the text below into {target_language_name}.

Note the following translations:

{source_word$_1$} means {target_word$_1$}

{source_word$_2$} means {target_word$_2$}

...

{source_word$_n$} means {target_word$_n$}

Please output only the translation of the text without any other explanation.

{source_language_name}: {source_text}

{target_language_name}:

B.2 Post-editing prompt with no lexical information

You are asked to edit the following translation from {source_language_name} into {target_language_name}. The proposed translation is high-quality, but may have some incorrect words.

Please output only the translation of the text without any other explanation.

{source_language_name}: {source_text}

{target_language_name}: {MT_output}

B.3 Post-editing prompt with lexical information (whole lexicon or subset)

You are asked to edit the following translation from {source_language_name} into {target_language_name}. The proposed translation is high-quality, but may have some incorrect words.

Note the following translations:

{source_word$_1$} means {target_word$_1$}

{source_word$_2$} means {target_word$_2$}

...

{source_word$_n$} means {target_word$_n$}

Please output only the translation of the text without any other explanation.

{source_language_name}: {source_text}

{target_language_name}: {MT_output}

C Complete results

Animals	Common numbers	Colors	Rarer numbers	Weekdays	Months
cat	two	black	eighteen	Monday	January
chicken	three	white	eighty	Tuesday	February
frog	four	red	fifteen	Wednesday	March
bird	five	blue	fifty	Thursday	April
bee	six	yellow	forty	Friday	May
fish	seven	green	forty-two	Saturday	June
horse	eight	purple	fourteen	Sunday	July
goat	nine	orange	nineteen		August
elephant	ten	grey	ninety		September
butterfly	hundred		seventeen		October
dog	million		seventy		November
deer			sixteen		December
bear			sixty		
			ten		
			ten thousand		
			thirteen		
			twenty-one		
			zero		
			eleven		
			twelve		

Table 5: Words used for our recall metric for evaluating the prevalence of in-domain lexical confusion.

Table 6: Lexical recall and CHRF scores before and after post-editing for translation into English. The languages whose codes are highlighted in blue constitute the top 20% with the highest confusion scores, before editing. These are reported on as "high-confusion languages" elsewhere.

xx→en

	MT output				Post-edited			
	Correct	**Confusion**	**Unknown**	**ChrF**	**Correct**	**Confusion**	**Unknown**	**ChrF**
aa	22.3	3.8	73.9	24.5	21.0	4.2	74.8	25.1
ab	60.0	4.3	35.7	51.6	64.8	3.3	31.9	51.3
ace	74.2	0.7	25.1	60.7	74.9	0.5	24.6	61.9
ach	58.1	4.8	37.1	50.2	58.6	3.8	37.6	49.0
aii	40.0	7.1	52.9	40.8	51.9	3.3	44.8	44.7
alz	51.9	4.8	43.3	44.1	55.7	3.3	41.0	44.1
arz	70.1	1.4	28.5	62.3	70.9	1.6	27.5	63.3
av	62.6	2.9	34.5	50.7	69.6	2.3	28.1	54.6
awa	78.3	0.2	21.6	68.0	79.3	0.2	20.5	68.8
ayl	72.9	1.0	26.2	58.1	73.3	1.4	25.2	59.1
ba	65.6	2.9	31.5	47.6	68.0	2.5	29.5	49.0
bal	0.0	0.0	100.0	33.3	0.0	0.0	100.0	28.5
ban	63.0	4.9	32.1	51.3	64.5	4.4	31.1	52.6
bbc	59.7	2.9	37.4	49.8	62.6	2.5	34.9	51.3
bci	24.5	5.7	69.8	27.4	26.9	4.2	68.9	28.3
bem	62.9	5.0	32.1	54.5	64.0	7.0	29.0	55.9
ber	42.2	4.6	53.1	43.2	42.1	4.8	53.1	44.2
bew	66.4	0.0	33.6	56.5	67.2	0.0	32.8	57.7
bik	75.7	1.9	22.4	66.1	80.5	1.9	17.6	65.0

	MT output				Post-edited			
	Correct	Confusion	Unknown	ChrF	Correct	Confusion	Unknown	ChrF
bjn	75.4	0.4	24.2	65.3	76.6	0.4	23.0	66.5
bm-Nkoo	41.8	9.1	49.0	29.1	43.8	7.7	48.6	29.5
bo	57.8	2.2	40.0	47.5	55.9	2.8	41.3	48.0
br	76.2	2.9	21.0	62.3	77.6	3.8	18.6	62.2
brx	52.9	7.1	40.0	55.3	58.6	4.3	37.1	54.5
bts	67.6	3.8	28.6	57.6	71.0	2.4	26.7	57.3
btx	61.4	2.9	35.7	47.1	64.8	3.3	31.9	47.0
bua	64.8	2.9	32.4	50.8	68.1	1.9	30.0	51.8
bug	56.3	1.1	42.6	50.5	56.7	1.1	42.2	51.4
ce	53.4	3.4	43.3	48.4	60.5	2.5	37.0	53.9
cgg	64.8	4.8	30.5	53.0	65.7	4.3	30.0	52.2
ch	49.0	4.8	46.2	41.8	52.4	4.8	42.8	42.7
chk	51.9	3.8	44.2	47.4	60.6	2.9	36.5	48.0
chm	62.4	10.2	27.3	55.4	74.1	3.9	22.0	55.9
cnh	59.5	9.5	31.0	55.8	67.1	4.8	28.1	56.2
crh	67.1	3.8	29.0	57.5	71.9	1.0	27.1	58.7
crs	85.4	1.4	13.2	74.6	84.9	1.4	13.7	75.1
ctg	59.5	3.8	36.7	51.9	65.7	3.3	31.0	55.5
cv	62.6	2.9	34.5	53.6	63.0	2.5	34.5	54.0
din	34.0	3.6	62.4	36.1	33.3	4.3	62.4	36.8
dov	55.2	4.3	40.5	46.5	58.6	2.9	38.6	46.7
dyu	23.6	2.5	73.9	26.0	29.0	3.3	67.6	28.6
dz	50.0	3.4	46.6	41.3	50.7	2.7	46.6	41.8
fa-AF	74.7	2.2	23.1	62.0	75.2	1.9	22.9	63.5
ff	57.1	6.4	36.5	46.3	58.4	5.4	36.3	46.9
fj	72.5	2.0	25.5	58.8	72.4	1.5	26.2	56.1
fo	76.8	1.4	21.8	65.0	78.7	1.4	19.9	66.7
fon	37.1	4.6	58.3	38.9	38.3	3.9	57.8	39.9
fur	79.7	0.9	19.4	69.4	80.2	0.7	19.1	70.9
gaa	61.0	4.8	34.3	51.8	62.9	3.3	33.8	51.3
gv	19.2	13.5	67.3	27.6	20.7	15.9	63.5	28.3
hil	84.3	1.0	14.8	69.7	86.2	1.0	12.9	67.5
hne	81.1	0.2	18.7	74.8	82.4	0.5	17.1	75.6
hrx	68.6	1.9	29.5	65.4	74.8	2.4	22.9	65.7
iba	62.4	2.4	35.2	48.9	69.0	1.9	29.0	48.5
jam	86.2	0.5	13.3	77.7	90.5	0.0	9.5	78.9
kac	41.4	4.1	54.5	44.6	43.0	3.7	53.3	46.6
kbd	56.7	12.4	31.0	47.0	67.6	4.3	28.1	47.7
kek	43.8	5.2	51.0	39.2	48.1	4.3	47.6	39.6
kg	52.4	2.7	44.9	50.2	52.4	2.3	45.3	51.0
kha	51.9	12.5	35.6	55.0	67.8	4.3	27.9	57.8
kl	49.2	3.4	47.5	40.3	53.8	3.4	42.9	42.4
kr	57.8	2.0	40.3	45.8	58.5	2.1	39.4	46.5
ks-Deva	63.5	2.3	34.2	57.6	66.5	2.3	31.2	58.9
ks	62.6	2.5	34.9	58.7	63.3	2.3	34.4	60.2
ktu	77.6	2.9	19.5	57.6	77.1	1.4	21.4	57.1
kv	54.8	9.5	35.7	50.9	66.2	2.4	31.4	50.8
li	73.1	0.4	26.6	67.8	74.5	0.4	25.1	69.1
lij	79.7	1.1	19.3	71.9	81.8	0.9	17.3	73.5

	MT output				Post-edited			
	Correct	Confusion	Unknown	ChrF	Correct	Confusion	Unknown	ChrF
lmo	77.4	1.2	21.4	69.5	76.8	1.4	21.7	70.7
ltg	81.3	1.1	17.6	70.6	82.9	0.9	16.2	71.0
lu	48.1	7.1	44.8	40.5	50.0	7.1	42.9	39.9
luo	44.5	3.5	52.0	41.5	46.7	3.0	50.2	42.7
mad	65.5	3.8	30.7	55.8	71.0	3.4	25.6	56.9
mak	63.3	5.7	31.0	51.0	67.1	2.9	30.0	52.7
mam	43.3	2.9	53.8	35.6	47.1	2.4	50.5	36.8
mfe	82.9	1.4	15.7	71.2	83.3	2.9	13.8	70.0
mh	47.8	7.2	44.9	46.8	54.6	3.4	42.0	47.4
min	80.2	1.1	18.7	67.8	81.1	0.9	18.0	68.2
ms-Arab	86.2	1.9	11.9	69.4	84.8	1.9	13.3	68.5
mwr	72.4	1.9	25.7	54.6	77.6	1.0	21.4	55.8
nd	61.9	3.3	34.8	50.6	63.8	2.1	34.1	51.6
ndc-ZW	28.4	6.7	64.9	31.1	31.7	4.3	63.9	31.9
new	55.7	2.8	41.5	52.8	54.2	2.8	42.9	53.7
nhe	50.5	7.6	41.9	41.0	57.6	6.2	36.2	42.4
nr	73.8	4.3	21.9	64.3	77.1	1.9	21.0	62.8
nus	47.8	5.7	46.5	43.2	48.7	5.2	46.2	44.4
oc	87.3	0.4	12.3	78.8	87.7	0.2	12.1	79.5
os	53.3	10.5	36.2	53.0	67.1	4.3	28.6	54.1
pa-Arab	71.4	1.9	26.7	58.5	72.4	1.4	26.2	59.2
pag	58.6	0.9	40.5	56.2	60.2	0.5	39.2	57.4
pam	71.4	1.0	27.6	53.6	70.5	1.0	28.6	53.6
pap	82.2	0.2	17.6	76.5	81.6	0.0	18.4	77.0
quc	31.1	3.4	65.5	29.5	33.6	2.5	63.9	30.6
rhg-Latn	31.4	6.7	61.9	33.0	48.6	4.8	46.7	38.6
rn	61.1	3.9	34.9	52.7	63.3	2.5	34.2	53.9
rom	65.2	4.8	30.0	60.2	72.9	2.9	24.3	60.3
sah	62.4	8.6	29.0	52.5	68.6	3.8	27.6	52.9
sat-Latn	32.8	5.5	61.7	39.9	35.5	5.5	59.0	43.9
scn	78.6	1.1	20.3	67.7	78.3	1.1	20.7	68.3
se	65.2	7.6	27.1	59.9	74.8	2.9	22.4	60.0
sg	20.9	5.8	73.3	27.4	20.3	5.6	74.1	26.4
shn	60.8	3.6	35.7	53.5	62.0	2.9	35.1	54.7
ss	72.4	2.4	25.2	63.5	72.0	2.6	25.4	64.5
sus	54.3	5.7	40.0	41.1	54.3	4.8	41.0	41.2
szl	79.7	0.5	19.8	70.2	80.9	0.7	18.4	71.9
tcy	69.0	3.8	27.1	51.8	71.9	3.8	24.3	53.2
tet	76.7	3.8	19.5	64.7	77.1	3.8	19.0	64.8
tiv	18.5	3.4	78.2	20.1	19.3	4.2	76.5	20.4
tn	72.2	1.9	25.9	60.6	73.1	1.7	25.2	62.0
to	67.6	3.8	28.6	57.4	68.6	4.3	27.0	59.6
tpi	61.1	1.2	37.6	60.3	61.7	0.7	37.6	60.8
trp	37.0	5.8	57.2	37.4	52.4	2.4	45.2	39.4
tum	52.2	2.1	45.6	47.8	54.5	2.1	43.3	49.0
ty	65.7	2.3	32.1	50.0	65.2	2.6	32.2	50.3
tyv	60.0	5.2	34.8	52.0	71.4	2.9	25.7	53.2
udm	62.4	9.5	28.1	52.3	73.3	3.3	23.3	52.5
ve	67.8	6.9	25.3	60.3	72.4	2.9	24.6	61.7

	MT output				Post-edited			
	Correct	**Confusion**	**Unknown**	**ChrF**	**Correct**	**Confusion**	**Unknown**	**ChrF**
vec	79.3	1.2	19.4	69.9	79.1	1.2	19.6	71.4
war	71.7	0.4	28.0	75.5	72.4	0.2	27.5	75.5
wo	48.8	2.0	49.1	41.7	48.1	1.4	50.5	42.1
yua	52.1	2.1	45.8	42.7	52.9	2.9	44.1	44.4
zap	19.5	3.3	77.1	22.3	21.4	3.8	74.8	22.9
Average	**59.3**	**3.8**	**36.9**	**52.4**	**62.4**	**2.8**	**34.8**	**53.2**

Table 7: CHRF scores before and after post-editing for translation out of English. The languages whose codes are highlighted in blue constitute the top 20% with the highest confusion scores before editing, in the into-English direction. These are reported on as "high-confusion languages" elsewhere.

en→xx

	Pre-edit CHRF	**Post-edit CHRF**
aa	22.3	22.4
ab	41.7	43.0
ace	45.9	46.5
ach	42.3	39.9
aii	26.6	28.1
alz	36.8	38.7
arz	50.6	51.2
av	28.8	28.9
awa	54.0	50.1
ayl	51.3	51.6
ba	41.7	43.0
bal	21.1	21.3
ban	43.1	42.9
bbc	37.2	37.6
bci	29.3	29.1
bem	48.4	49.3
ber-Latn	21.4	34.5
bew	48.4	46.5
bik	59.4	60.1
bjn	53.8	56.5
bm-Nkoo	18.8	16.9
bo	42.1	43.1
br	51.4	52.3
brx	41.0	41.7
bts	48.5	48.5
btx	42.7	42.3
bua	40.5	41.0
bug	39.2	40.2
ce	25.3	25.8
cgg	43.9	44.9
ch	37.2	37.9
chk	37.4	40.8
chm	48.9	48.7
cnh	44.6	45.2
crh	47.8	48.7

	Pre-edit CHRF	Post-edit CHRF
crs	69.2	69.8
ctg	33.0	34.1
cv	49.6	48.5
din	25.6	26.4
dov	41.0	41.5
dyu	22.3	22.4
dz	43.0	43.8
fa-AF	48.2	46.7
ff	32.4	31.3
fj	60.5	60.2
fo	56.5	57.6
fon	26.1	25.9
fur	60.4	61.7
gaa	48.8	48.5
gv	22.9	24.0
hil	63.7	63.6
hne	57.2	56.2
hrx	47.5	51.3
iba	45.2	44.6
jam	60.7	55.2
kac	43.5	44.2
kbd	36.8	40.4
kek	31.9	35.1
kg	50.2	51.0
kha	54.3	57.0
kl	42.4	43.6
kr	32.8	33.3
ks-Deva	33.8	25.0
ks	24.0	34.7
ktu	63.2	64.7
kv	39.9	42.0
li	55.0	54.1
lij	57.4	58.0
lmo	39.2	40.2
ltg	64.0	63.8
lu	24.7	24.5
luo	41.2	41.5
mad	40.7	40.6
mak	44.9	46.3
mam	28.8	25.9
mfe	66.5	66.3
mh	42.1	41.4
min	58.6	59.4
ms-Arab	66.2	59.9
mwr	36.8	36.4
nd	41.8	43.2
ndc-ZW	27.9	29.6
new	37.4	36.9
nhe	38.6	41.2
nr	58.8	57.2

	Pre-edit CHRF	Post-edit CHRF
nus	32.5	30.6
oc	68.3	69.5
os	45.9	46.2
pa-Arab	43.3	45.1
pag	53.0	53.0
pam	47.7	47.3
pap	66.1	68.1
quc	24.7	25.3
rhg-Latn	20.6	24.0
rn	44.9	45.5
rom	37.0	36.4
sah	46.9	48.7
sat-Latn	22.8	24.4
scn	51.9	53.0
se	46.8	48.8
sg	30.5	31.1
shn	40.7	39.6
ss	56.2	55.9
sus	34.9	28.6
szl	59.2	59.5
tcy	39.1	40.9
tet	60.0	59.8
tiv	26.3	27.1
tn	55.8	55.7
to	52.0	54.6
tpi	51.9	52.3
trp	36.5	40.6
tum	44.7	45.0
ty	56.6	54.8
tyv	43.1	44.7
udm	45.9	46.2
ve	55.6	52.1
vec	55.4	54.7
war	61.8	63.0
wo	29.8	29.3
yua	38.5	39.5
zap	17.8	18.3
Average	**43.4**	**43.8**