Using Language Models to Disambiguate Lexical Choices in Translation

Josh Barua **Sanjay Subramanian** **Kayo Yin** **Alane Suhr**

University of California, Berkeley
{joshbarua,sanjayss,kayoyin,suhr}@berkeley.edu

Abstract

In translation, a concept represented by a single word in a source language can have multiple variations in a target language. The task of lexical selection requires using context to identify which variation is most appropriate for a source text. We work with native speakers of nine languages to create DTAiLS, a dataset of 1,377 sentence pairs that exhibit cross-lingual concept variation when translating from English. We evaluate recent LLMs and neural machine translation systems on DTAiLS, with the best-performing model, GPT-4, achieving from 67 to 85% accuracy across languages. Finally, we use language models to generate English rules describing target-language concept variations. Providing weaker models with high-quality lexical rules improves accuracy substantially, in some cases reaching or outperforming GPT-4.

1 Introduction

Resolving ambiguity in translation is a fundamental challenge (Weaver, 1952) that remains unsolved (Campolungo et al., 2022). This paper focuses on *lexical selection*, a key aspect of translation that requires using context in a source sentence to determine the best translation for an ambiguous source word from several target-language options. Figure 1 shows variations of the concept *date* (fruit) in Farsi and an example of the lexical selection task.

Our work has two main goals. First, we investigate the capabilities of language models in disambiguating lexical choices in translation by comparing instruction-tuned language models with high-performing neural machine translation systems. Second, we test whether language models can be used to extract useful natural language rules that accurately describe how to translate ambiguous words based on source-side context.

We work with native speakers to introduce the Dataset of Translations with Ambiguity in Lexical Selection (DTAiLS), a test set of 1,377 sentence pairs spanning nine languages where concept variation can be explained by context in the source sentence. Evaluating five models on DTAiLS reveals that, without rules provided in-context, only the best-performing LLM, GPT-4, is competitive with NMT systems.

Concept: date (fruit)

Variations and Generated Rules
Khorma refers to the fruit of the date palm when it is fully ripe and dried. It is commonly consumed as a sweet, chewy snack or used in various dishes, particularly desserts.

Rotab refers to fresh, soft dates that are partially ripe. These dates are moister and sweeter than fully ripe, dried dates (khorma). Rotab is often eaten as a fresh fruit or used in cooking where a softer, sweeter texture is desired.

Kharak refers to dates that are unripe and are in a semi-dried state. They are less sweet compared to rotab and khorma and are often used in cooking or further processed into other forms.

Lexical Selection
Source Sentence: she brings me dried dates and tells me old stories
Correct Variation: khorma

Figure 1: Generated rules for English *date* with lexical variations *khorma*, *rotab*, and *kharak* in Farsi.

We also present a simple method for generating rules for lexical selection from LLMs, which native speakers verify are highly accurate. Figure 1 shows rules generated for three Farsi variations of the concept *date*. We observe improvements in performance across all LLMs when prompted with accurate self-generated rules. In addition, while open-weight LLMs lag behind both NMT systems and GPT-4, providing rules from GPT-4 can help substantially to bridge the gap in performance. This suggests that parametric knowledge of concept variation poses a greater challenge to models than the ability to apply such knowledge in-context. Our work demonstrates that LMs can generate high-quality rules, and further leverage these rules to rival specialized NMT systems on lexical selection, but still fall short of native speakers.[1]

[1] Code and data are publicly released here: https://github.com/Berkeley-NLP/Lex-Rules

2 Task and Data

To evaluate a model's ability to understand concept variations, we study lexical selection in translation. For example, the noun *date* has multiple lexical variations in Farsi, which distinguishes variations by fruit ripeness and dryness (Figure 1). We collect a dataset of translation pairs that require understanding and appropriately applying concept variation, where sufficient context is provided to distinguish between variations.

Lexical Selection In translation, lexical selection is the task of selecting the most appropriate lexeme in the target language that maps from a single lexeme in the source language, in the context of a source sentence (Apidianaki, 2009). Formally, let $(\bar{x}, \bar{y})$ be a sentence pair where $\bar{x} = \langle x_1, \ldots, x_{|\bar{x}|} \rangle$ is a sequence of words in the source language and $\bar{y} = \langle y_1, \ldots, y_{|\bar{y}|} \rangle$ is its translation in the target language. For a source word x_i, we define the set of possible translations of x_i as $\bar{v} = \langle v_1, \ldots, v_{|\bar{v}|} \rangle$ where $\exists\ j$ such that $v_j \in \bar{y}$. The task of lexical selection is to identify the most appropriate translation v_j conditioned on the source sentence $\bar{x}$.

Source Data Despite the existence of large-scale datasets for low-resource languages through bitext mining techniques (Schwenk et al., 2021), we focus on datasets curated by human translators to mitigate the potential for incorrect translations due to misalignment. We use OpenSubtitles (Lison and Tiedemann, 2016; Lison et al., 2018), TED2020 (Reimers and Gurevych, 2020), PMIndia (Haddow and Kirefu, 2020), and TEP (Pilehvar et al., 2011) to acquire parallel data for English paired with 7 low-resource and 2 high-resource languages (Japanese and Farsi).[2] All datasets are downloaded from the digital platform OPUS[3] (Tiedemann, 2009).

Expert Recruitment We work with bilingual speakers to ensure our methods and data faithfully represent the processes associated with translation under concept variation. For each language, we recruit from Prolific[4] three annotators who are fluent English speakers and native speakers of the target language. All annotators are paid $16 USD / hour.[5]

Language	# Concepts Extracted	Precision	Recall
Afrikaans	17	99.2	82.4
Armenian	18	85.4	77.8
Farsi	100	96.2	72.3
Galician	24	95.8	91.7
Hindi	41	96.2	89.4
Japanese	202	97.6	78.9
Latvian	16	99.1	87.5
Tamil	18	92.4	79.6
Telugu	21	95.9	87.3

Table 1: Details for identifying concepts with variations, including the number of extracted concepts and the precision and recall of the extracted variations. On average, we identify 2.3 variations per concept.

2.1 Identifying Concepts with Variations

We first identify concepts that are represented as a single word in our source language (English) but have several variations in a target language. We build upon Chaudhary et al. (2021)'s approach to identify fine-grained lexical distinctions that arise due to concept variation. Given a parallel corpus, we lemmatize all words using Stanza (Qi et al., 2020) and compute word alignments for each sentence pair with the AWESOME aligner (Dou and Neubig, 2021). Using these alignments, we create a one-to-many mapping from source-language lexemes to target-language lexemes. Lastly, we remove source lexemes that do not map to at least two target lexemes, exhibit low entropy, or correspond to target lexemes that arise due to polysemy.[6] While this approach was originally designed and applied to Spanish and Greek parallel corpora, we apply it to nine additional languages. Table 1 lists the total number of extracted concepts for each language.

We also perform comprehensive analysis of this approach's precision and recall in identifying target-language variations of concepts. All three expert annotators for each language provide feedback on the extracted variations, including whether each variation matches the meaning of the English lexeme (for computing precision) and whether any key variations are missing from the set (for computing recall). Precision is measured as the proportion of accurate variations; recall is measured as the proportion of concepts with all key variations recovered.[7] In general, the precision of identified variations is very high, even for low-resource lan-

[2] Table 7 lists the data sources and number of sentence pairs available per language.

[3] https://opus.nlpl.eu/

[4] https://www.prolific.com/

[5] More details on annotator recruitment and task design are available in Appendix A.1.

[6] See Appendix B.1 for additional details on the pipeline.

[7] Appendix A includes more details on this analysis, including inter-annotator agreement.

Language	Full Dataset	% Sentences w/ Sufficient Context	Expert Dataset
Afrikaans	4,123	78.9	180
Armenian	9,610	63.9	176
Farsi	43,911	59.0	127
Galician	13,393	67.0	164
Hindi	17,417	61.0	145
Japanese	99,741	76.0	149
Latvian	6,944	81.3	184
Tamil	5,833	65.3	134
Telugu	9,167	63.3	118

Table 2: DTAiLS dataset statistics. A sentence is determined to have sufficient context when the majority of annotators select its ground truth variation.

Language	% Correct Rules
Afrikaans	99.2
Armenian	86.2
Farsi	95.1
Galician	99.4
Hindi	98.5
Japanese	99.1
Latvian	100.0
Tamil	99.2
Telugu	98.0

Table 3: Mean accuracy of GPT-4 generated rules.

guages, but our approach is less consistent in identifying all possible variations of a source concept, which could be due to the limited size of datasets used or the use of domain-specific data.

2.2 Dataset Construction

Our goal is to collect a dataset of sentence pairs that require understanding target-language concept variation for accurate translation. Expert annotators help us curate this dataset by performing the lexical selection task, provided only source-language sentences and target lexemes.

All annotators for a given language are presented with the same set of concepts and source-language sentences. We shuffle the order in which concepts, sentences, and target lexemes are shown to each annotator. Our resulting dataset, DTAiLS, includes sentences for which the majority of annotators selected the variation used in the original sentence pair, which indicates that there is sufficient context for consistent lexical selection. Although there could be cases of context-dependent translation where there isn't a single optimal lexical variation, for our dataset we rely on majority agreement to select examples that are clearly evaluable. Table 2 includes dataset statistics.[8]

3 Rules for Lexical Selection

We experiment with generating human-readable English rules that describe all extracted concepts and their variations, and analyze how these rules influence model performance on lexical selection when provided as input.[9]

For each target-language variation, we find the 50 longest source-language sentences with less than 50 tokens in length where the variation appears in the target translation. Motivated by work showing LLMs are useful for describing differences between sets of text (Zhong et al., 2022), we construct one prompt per concept including all target-language variations and their respective lists of source-language sentences, and prompt each model to provide a natural language description of each target-language variation.[10] We generate rules from three instruction-tuned models: Gemma-1.1 (Gemma Team et al., 2024), LLaMA-3 (AI at Meta, 2024), and GPT-4 (OpenAI, 2023); Figure 1 provides an example rule set from GPT-4. For each language, we ask all three native speakers to label every GPT-4-generated rule for correctness. Table 3 shows these rules are overwhelmingly correct according to native speakers.

4 Experiments

We evaluate three instruction-tuned models, GPT-4, LLaMA-3-8B-Instruct, and Gemma-1.1-7B-IT; and two high-quality NMT models MADLAD-400-10B-MT (Kudugunta et al., 2023) and NLLB-200-3.3B (NLLB Team et al., 2022), sampling with temperature of 0.

We hypothesize that models performing lexical selection will benefit from access to rules describing concept variations (Section 3). Thus, we evaluate instruction-tuned LLMs in 3 settings: (1) no rules provided, (2) with self-generated rules, and (3) with rules generated by GPT-4. For each setting, we instruct the model to explain its reasoning prior to selecting a lexical choice (Wei et al., 2022).[11]

Accuracy is measured as the proportion of sentences for which the model selects the correct lexi-

[8] Appendix A.2 contains more dataset construction details.

[9] We refer to these natural language descriptions as "rules" to be consistent with prior work studying lexical selection (Chaudhary et al., 2021).

[10] The Appendix (Figure 9) shows the prompt template.

[11] Appendix B includes prompts used for lexical selection and a description of lexical selection with NMT systems.

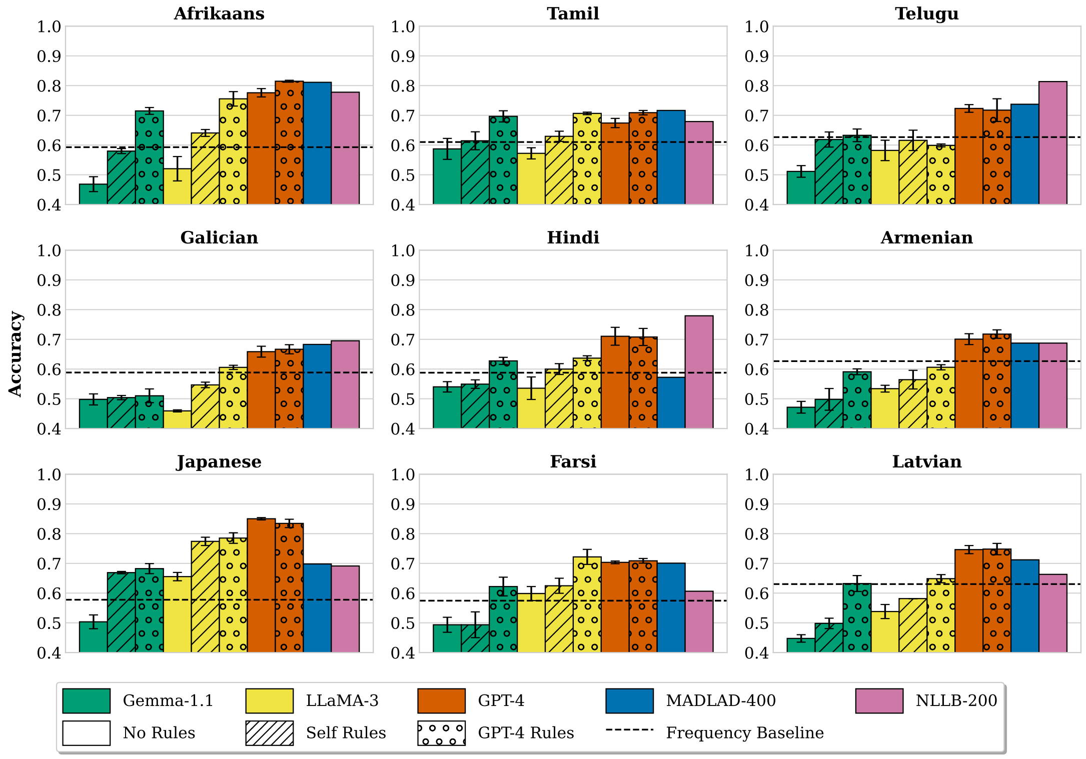


Figure 2: Comparisons between LMs with and without rules to NMT systems on lexical selection. We report $\mu_{\pm\sigma}$ across 3 runs for LM experiments.

cal variation. We include a simple frequency baseline that always predicts the most common target-language variation in our dataset for each concept during lexical selection. Figure 2 shows results for each model.

First, we find that SOTA NMT systems trained on our low-resource language pairs outperform all LLMs on Telugu and Hindi, while being comparable on Afrikaans, Galician, Tamil, and Farsi. For the remaining three languages, the best-performing LLM, achieves a 4-15% absolute improvement in performance over the NMT systems. We find a large gap in performance between open- and closed-weight LLMs, with the frequency baseline outperforming open-weight LLMs without rules in seven languages. LLaMA-3-8B-Instruct outperforms Gemma-1.1-7B-IT for all nine languages.

Self-generated lexical rules improve model performance on nearly all languages, with improvements being even more significant for the weaker open-weight models. This suggests that models can better reason over lexical choices if their parametric knowledge of concept variation is explicitly included as context in the prompt. While lexical selection with GPT-4 is significantly more expensive than with open-weight models, rule generation has a one-time cost. When providing the open-weight models with rules acquired from the strongest model (GPT-4), we see total improvements up to 23% in accuracy, with these models performing close to or even exceeding GPT-4 on several languages. However, even when these high-quality (Table 3) rules are provided, there is still a significant gap to human performance. We hypothesize that while the generated rules are accurate, they fail to enumerate all possible contextual factors that could influence lexical choice in all translation settings.

5 Related Work

The most closely related work is by Chaudhary et al. (2021), who study lexical selection for English to Spanish and Greek. They present a pipeline for collecting lexical selection data, train models to perform lexical selection, use a linear SVM model to extract features as interpretable rules, and evaluate the efficacy of these rules in a second-language acquisition setting. In contrast, we use modern LMs, generate natural language rules, and evaluate on several low-resource languages. We also curate a test set for lexical selection validated by native speakers of our target languages.

Lexical selection is closely related to the problem of ambiguity in machine translation, where context is essential for disambiguating various possible translations. Fernandes et al. (2023) investigate such ambiguity that arises from discourse and grammar, while Campolungo et al. (2022) explore ambiguity due to polysemy. Iyer et al. (2023) evaluate LLMs on translations under polysemy, demonstrating that in-context learning and fine-tuning on ambiguous datasets improves translation. We study the potential for LLMs to resolve ambiguity arising from target-language concept variation, focusing on low-resource languages.

Prior work has shown improvements in LLM translation quality by incorporating ground truth dictionary entries into prompts (Ghazvininejad et al., 2023). We further demonstrate that models can accurately describe concept variations in low-resource languages using only parametric knowledge and example usages from source-language sentences. Our experiments follow a line of work showing that modern LLMs exhibit non-English language capabilities, though these LLMs are often trained primarily on English data (Robinson et al., 2023; Asai et al., 2023).

6 Conclusion

We introduce DTAiLS, a dataset of 1,377 sentence pairs with 9 language pairs that exhibit ambiguity in translation due to concept variation. Using this dataset, we evaluate 3 popular instruction-tuned LLMs and 2 high-performing NMT systems on the task of lexical selection. Out of nine languages tested, the strongest LLM outperforms the NMT systems on three languages, has comparable performance on four languages, and fall short of these systems on two languages. No model is able to disambiguate the full set of sentences that native speakers can.

We also present a simple approach to extract high-quality rules from language models, demonstrating improvements on lexical selection when LMs are given access to rules. We find that providing weaker open-weight models with rules from a stronger LLM can effectively bridge the gap to or even outperform the stronger model for several languages. Future research could investigate additional applications of lexical rules in NMT and assess how these human-readable rules can assist L2 learners in vocabulary acquisition.

Limitations

Because our focus is on low-resource languages, the parallel corpora we use are small; thus, we are only able to extract roughly 20 concepts for six out of nine languages. Further, due to the time and effort required to collect human judgements on lexical selection,[12] our test sets curated by experts are just 120-180 examples per language and 1,377 examples overall. Developing automated methods for example selection is an interesting direction for future work that will enable larger-scale evaluation. We also note that the recall of the pipeline for identifying concepts with variations might be inaccurate due to the challenges annotators face brainstorming all possible variations in the semantic space. Lastly, due to a lack of available models for WSD, dependency parsing, and POS tagging for low-resource languages, we are only able to evaluate on language pairs where English is the source language. In theory, the methods we present can work for any arbitrary language pair.

Acknowledgments

The authors are grateful to the anonymous reviewers for helpful feedback. This work was partially supported by AI2 Young Investigator Grant and a Gemma Academic Program Award. KY is supported by the Future of Life PhD Fellowship. During this project, JB was supported by the Berkeley SURF program. We would also like to thank Amartya Bose, Niloofar Mireshghallah, Prithviraj Ammanabrolu, Yanai Elazar, Mihran Miroyan, PV Subramanian, and Elzette van Rensburg for helpful discussions on generated lexical rules and dataset construction. Finally, we are deeply thankful to all of the crowdworkers who participated in our study, without whom this work would not have been possible.

References

AI at Meta. 2024. Llama 3 model card.

Marianna Apidianaki. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *EACL*.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila B Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. BUFFET: Benchmarking large lan-

[12] Annotating the full Japanese dataset would require roughly 5 thousand total hours of work.

guage models for cross-lingual few-shot transfer. *arXiv preprint*.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *ACL*.

Ted Byrt, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring word sense disambiguation biases in machine translation. In *ACL*.

Aditi Chaudhary, Kayo Yin, Antonios Anastasopoulos, and Graham Neubig. 2021. When is wall a pared and when a muro?: Extracting rules governing lexical selection. In *EMNLP*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *EACL*.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? A data-driven, multilingual exploration. In *ACL*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint*.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint*.

Barry Haddow and Faheem Kirefu. 2020. PMIndia - A collection of parallel corpora of languages of India. *arXiv preprint*.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation, WMT*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *NeurIPS*.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *LREC*.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC*.

NLLB Team, Marta Ruiz Costa-jussà, James Cross, Onur Celebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint*.

Mohammad Taher Pilehvar, Heshaam Faili, and Abdol Hamid Pilevar. 2011. TEP: Tehran English-Persian parallel corpus. In *Computational Linguistics and Intelligent Text Processing*.

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

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP*.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *ICLR*.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation, WMT*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *ACL-IJCNLP*.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing, 2009*.

Warren Weaver. 1952. Translation. In *Proceedings of the Conference on Mechanical Translation*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *ICLR*.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. 2022. Describing differences between text distributions with natural language. In *ICML*.

A Additional Annotation Details

A.1 Expert Annotation

The filters applied for participants before joining include (1) highest level of degree earned as technical/community college or above and (2) fluency in English and native proficiency in one of the nine languages we study. First, a pilot study was conducted to vet annotators for fluency and comprehension of the task. We then published two studies to the group of annotators who qualified from the pilot study. The first study required annotation of the rules generated by GPT-4 and feedback on concepts and lexical variations extracted by the pipeline (Section 2.1) for computing precision and recall. The interface for the first study can be found in Figure 3. The second study required annotators to complete the lexical selection task. The interface for the second study can be found in Figure 4. We ensure that the same annotator doesn't participate in both studies. We remove all personal identifying information from all data collected. Prior to taking part in any study, annotators were informed of the purpose of the study and how their data would be used.

A.2 Example Selection

Collecting human judgments of lexical selection for all parallel sentences is infeasible; for example, fully annotating Japanese would require labeling 99,741 examples (Table 2) and roughly 5 thousand total hours of work. Due to limited resources, we sample up to 20 concepts for each target language and gather 10 sentence pairs per concept, ensuring that each variation is represented at least once. This results in a test set of up to 200 sentences per language. For languages with more than 20 extracted concepts, we first filter for concepts that have a roughly uniform distribution over variations. Specifically, for each concept we compute the relative frequency of each lexical variation. Concepts are discarded if any individual variation deviates by more than 20% from a uniform distribution. After filtering, we uniformly sample 20 concepts to be included in the lexical selection task.

A.3 Inter-Annotator Agreement

In this section we report statistics on inter-annotator agreement for all studies conducted with native speakers. We include 3 metrics: *total agreement* is the proportion of questions for which all 3 annotators were in agreement, *Fleiss' kappa* (Fleiss, 1971) is a popular metric for measuring reliability of agreement between more than 2 annotators, and *prevalence-adjusted bias-adjusted kappa* (PABAK) (Byrt et al., 1993) is a modified kappa that addresses problems that arise from an imbalanced distribution of data over classes. We include Fleiss' kappa for completeness, but note that it is often misleading due to the very high prevalence of class 1 (correct). For example, in the first row of Table 4 we find that despite annotators being in total agreement for over 97% of questions, the Fleiss' kappa measure suggests poor agreement. This is a common paradox for measures of inter-annotator agreement that are "chance-corrected." Table 4 presents inter-annotator agreement statistics for evaluation of generated rules, while Tables 5 and 6 display the same for precision and recall of the concept variation extraction pipeline.

Language	Total Agreement	Fleiss' κ	PABAK
Afrikaans	0.975	-0.008	0.950
Armenian	0.707	0.181	0.415
Farsi	0.857	-0.021	0.713
Galician	0.982	-0.006	0.964
Hindi	0.954	-0.016	0.908
Japanese	0.973	-0.009	0.945
Latvian	1.000	NaN	1.000
Tamil	0.977	-0.008	0.955
Telugu	0.959	0.319	0.918

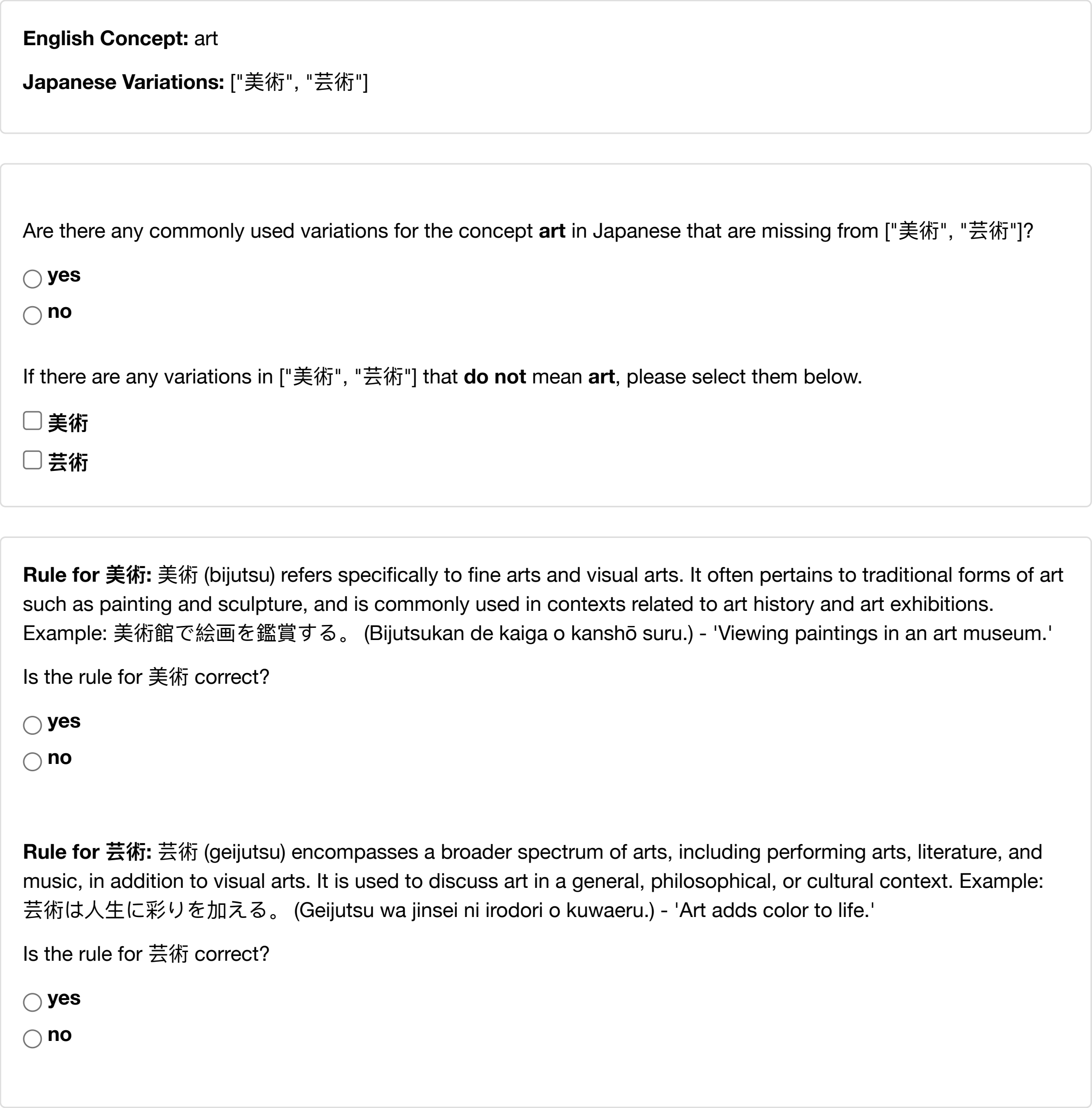
Table 4: Inter-annotator agreement statistics for evaluation of generated rules.

Language	Total Agreement	Fleiss' κ	PABAK
Afrikaans	0.975	-0.008	0.950
Armenian	0.634	0.024	0.268
Farsi	0.917	0.241	0.835
Galician	0.909	0.254	0.818
Hindi	0.931	0.376	0.862
Japanese	0.951	0.307	0.903
Latvian	0.974	-0.009	0.947
Tamil	0.818	0.134	0.636
Telugu	0.959	0.652	0.918

Table 5: Inter-annotator agreement statistics for precision of concept variation pipeline.

Language	Total Agreement	Fleiss' κ	PABAK
Afrikaans	0.588	0.056	0.176
Armenian	0.444	-0.071	-0.111
Farsi	0.360	-0.066	-0.280
Galician	0.792	0.091	0.583
Hindi	0.683	-0.118	0.366
Japanese	0.510	0.019	0.020
Latvian	0.625	-0.143	0.250
Tamil	0.611	0.201	0.222
Telugu	0.667	-0.002	0.333

Table 6: Inter-annotator agreement statistics for recall of concept variation pipeline.



English Concept: art

Japanese Variations: ["美術", "芸術"]

Are there any commonly used variations for the concept **art** in Japanese that are missing from ["美術", "芸術"]?

○ **yes**

○ **no**

If there are any variations in ["美術", "芸術"] that **do not** mean **art**, please select them below.

□ **美術**

□ **芸術**

Rule for 美術: 美術 (bijutsu) refers specifically to fine arts and visual arts. It often pertains to traditional forms of art such as painting and sculpture, and is commonly used in contexts related to art history and art exhibitions. Example: 美術館で絵画を鑑賞する。 (Bijutsukan de kaiga o kanshō suru.) - 'Viewing paintings in an art museum.'

Is the rule for 美術 correct?

○ **yes**

○ **no**

Rule for 芸術: 芸術 (geijutsu) encompasses a broader spectrum of arts, including performing arts, literature, and music, in addition to visual arts. It is used to discuss art in a general, philosophical, or cultural context. Example: 芸術は人生に彩りを加える。 (Geijutsu wa jinsei ni irodori o kuwaeru.) - 'Art adds color to life.'

Is the rule for 芸術 correct?

○ **yes**

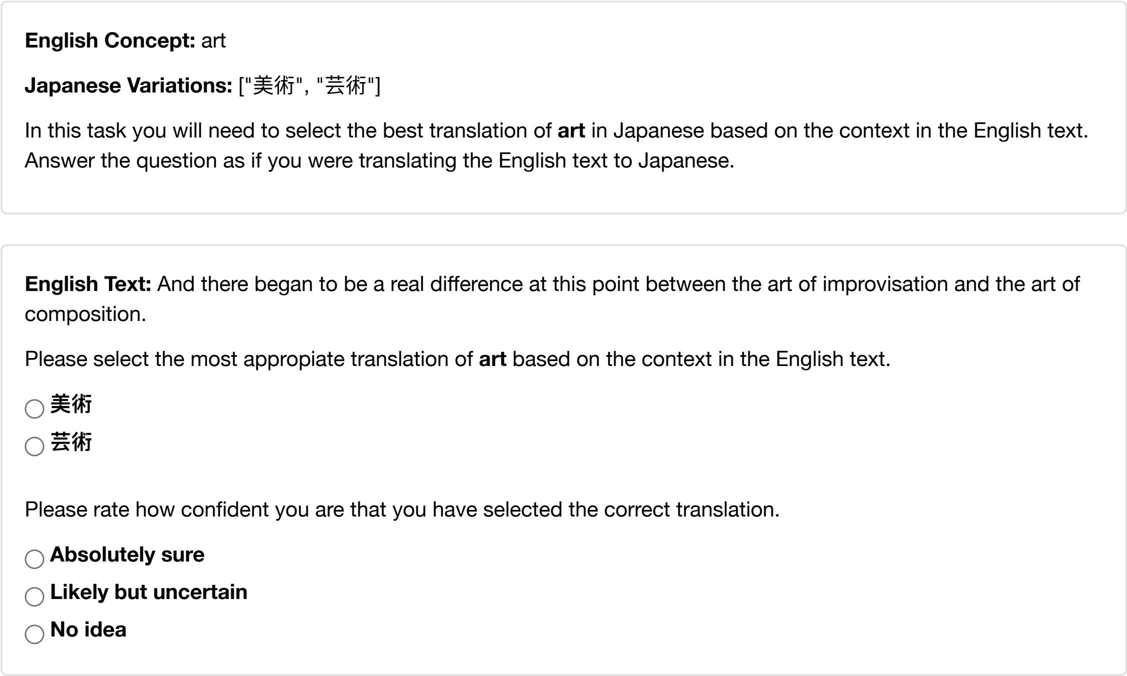
○ **no**

Figure 3: Interface for annotating rules and extracted concepts and variations.

B Additional Experimental Details

B.1 Pipeline for Identifying Concepts with Variations

In this section, we formally describe the pipeline for identifying concepts with variations, which we adopt from Chaudhary et al. (2021). Let $\mathcal{D} = \{(\bar{x}_1, \bar{y}_1), \ldots, (\bar{x}_{|\mathcal{D}|}, \bar{y}_{|\mathcal{D}|})\}$ be a parallel corpus where $(\bar{x}_i, \bar{y}_i)$ is a source- and target-language sentence pair. For each sentence pair, we compute word alignments and lemmatize all words in $\bar{x}_i$ and $\bar{y}_i$ using the AWESOME aligner and Stanza respectively. Furthermore, for source sentences only, we perform automatic part-of-speech (POS) tagging and dependency parsing using Stanza and word-sense disambiguation (WSD) using EWISER (Bevilacqua and Navigli, 2020). Source words are characterized by tuples of their lemmatized form and POS tag $\langle l_x, t_x \rangle$ to avoid conflating different meanings across POS tags. First, we enumerate all word alignments across the corpus and create a one-to-many mapping from each source word to the lexical variations it is aligned with. Second, we remove all source words that do not map to at least two lexical variations at least 50 times. We require 50 occurrences for each variation to prevent incorrect translations being extracted due to noisy alignments. Next, we describe a process for filtering out source words based on entropy. For a given source word tuple $\langle l_x, t_x \rangle$ with lexical variations $\bar{v} = \langle v_1, \ldots, v_{|\bar{v}|} \rangle$, let n_i be the number of occurrences of variations v_i. We compute the conditional



English Concept: art

Japanese Variations: ["美術", "芸術"]

In this task you will need to select the best translation of **art** in Japanese based on the context in the English text. Answer the question as if you were translating the English text to Japanese.

English Text: And there began to be a real difference at this point between the art of improvisation and the art of composition.

Please select the most appopiate translation of **art** based on the context in the English text.

○ **美術**
○ **芸術**

Please rate how confident you are that you have selected the correct translation.

○ **Absolutely sure**
○ **Likely but uncertain**
○ **No idea**

Figure 4: Interface for lexical selection task.

Language	Data Source	# Parallel Sentences
Afrikaans	OpenSubtitles	44,703
Armenian	TED2020	37,122
Farsi	TED2020, TEP	916,975
Galician	TED2020	34,385
Hindi	OpenSubtitles, TED2020	140,649
Japanese	TED2020	366,661
Latvian	TED2020	55,488
Tamil	OpenSubtitles, TED2020	43,741
Telugu	OpenSubtitles, PMIndia, TED2020	72,860

Table 7: Details for parallel corpora of all nine language included in our study, including data sources and the number of parallel sentences.

probability of each variation v_i as

$$p(v_i \mid l_x, t_x) = \frac{n_i}{\sum_{j=1}^{|\bar{v}|} n_j}$$

and the entropy of a source word tuple as

$$H(l_x, t_x) = \sum_{j=1}^{|\bar{v}|} -p(v_j \mid l_x, t_x) log_e(p(v_j \mid l_x, t_x))$$

We remove all source word tuples with an entropy below 0.69. Lastly, we remove lexical variations that arise due to polysemy in the source word. In particular, since source words are only characterized by their lemmatized form and POS, it is possible that we extract variations that correspond to different senses of l_x. For example, the source word tuple $< plane, noun >$ is translated in Spanish as *avión* when referring to an aircraft and *plano* when referring to a geometric plane. When each variation is mapped to a source word, we store the word sense of l_x as it is used in the source sentence. This allows us to compute the most frequent word sense during post-processing and remove variations belonging to other word senses.

B.2 Lexical Selection with NMT Systems

To perform lexical selection with NMT systems, we pass the source sentence as input and parse the translated text for the predicted lexical variation. First, we check for an exact substring match in the translated text with all lexical variations. If an exact match is found, we take that to be the predicted variation. If not, we tokenize the translated text with Stanza and computing the Levenshtein ratio (Levenshtein, 1965) between every word and every lexical variation. We identify the variation with the highest ratio to any word in the translated text for fuzzy matching. If this ratio exceeds 0.7, we select it as the predicted variation; otherwise, we conclude that no variation is found and the prediction is labeled as incorrect.

B.3 Prompts

Since the Gemma family of models do not take a system prompt as input, we prepend the system prompt to the user prompt with the role *user* for all experiments involving Gemma-1.1. Figures 7, 8, and 9 show the prompts that we use to evaluate LMs on lexical selection and generate rules. For lexical selection with LLMs, we apply the same fuzzy matching scheme as Section B.2 to match the generated answer to a target-language variation. Qualitatively, the instruction-following capabilities of GPT-4 were greater than that of Llama-3 and Gemma-1.1. If any LM failed to generate an answer according to the provided template, we append "Please enclose your selected translation from <Translations> with 3 back ticks." to the prompt and resample once. If the LM fails to follow the template a second time, the prediction is labeled as incorrect.

B.4 Position Bias in Lexical Selection

We acknowledge that with our prompt, the lexical selection task is similar to a multiple choice question (MCQ) setup. While humans tend to be order-invariant when answering MCQs, several prior works have examined position bias in LLMs when solving MCQs (Robinson and Wingate, 2023; Zheng et al., 2024). To ensure our evaluation is not affected by position bias we take three steps. First, we shuffle the order of translations in the prompt for every example during lexical selection to reduce bias. Second, we report how often the LMs select an answer choice at each position across the Afrikaans, Latvian, and Japanese subsets of DTAiLS. Figures 5 and 6 show a roughly uniform distribution over selected positions for concepts that have 2 and 3 lexical variations, respectively. Lastly, we plot the mean and standard deviation of each LM experiment across 3 runs in Figure 2. We find that the test accuracy is consistent for all models despite each run being initialized with a unique seed for shuffling the order of translations. Based on these findings, we conclude that LMs are approximately order-invariant when doing lexical selection with our prompting setup.

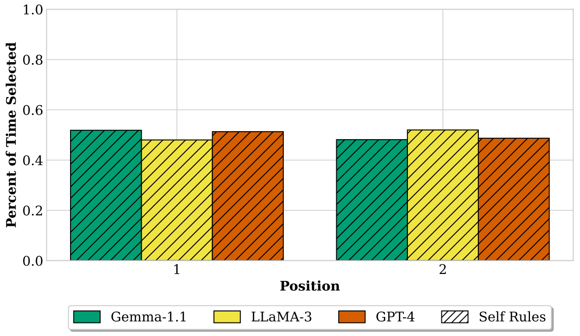


Figure 5: Percent of time each model selects an answer at each position when there are 2 lexical variations.

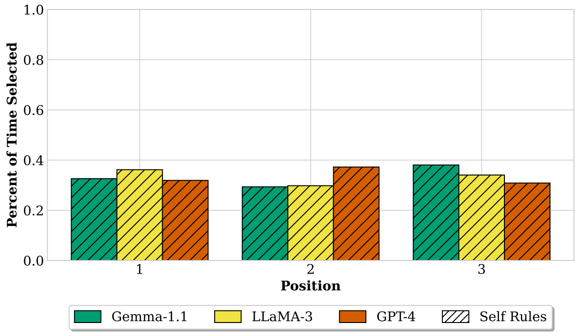


Figure 6: Percent of time each model selects an answer at each position when there are 3 lexical variations.

B.5 OpenAI Model Used

In our call to the OpenAI API, we use the model name *GPT-4-Turbo* which at the time of writing is a pointer to *gpt-4-turbo-2024-04-09*.

B.6 Computational Requirements

All experiments in this paper that do not involve models from OpenAI require approximately 50 GPU hours on an NVIDIA RTX A6000 GPU.

C Software and Licenses

The TED2020 dataset uses the CC-BY-NC-4.0 License. All models are utilized from Hugging Face; LLaMA-3-8B-Instruct uses the Llama3 License, Gemma-1.1-7B-IT uses the Gemma License, MADLAD-400-10B-MT uses the Apache License 2.0, and NLLB-200-3.3B uses the CC-BY-NC-4.0 License. Our use of datasets and models is consistent with their intended use.

————————— System Prompt —————————-

————————— User Prompt —————————

Please select the best translation of "*<Concept>*" in "*<Source Text>*" from the following list: *<Translations>*. Carefully explain your reasoning first and then enclose your final answer like this ```answer```.

Figure 7: Full prompt for the lexical selection task.

————————— System Prompt —————————-

Here are rules for how to translate "*<Concept>*" in *<Target Language>*:*<Rules>*

————————— User Prompt —————————

Based on the provided rules, please select the best translation of "*<Concept>*" in "*<Source Text>*" from the following list: *<Translations>*. Carefully explain your reasoning first and then enclose your final answer like this ```answer```.

Figure 8: Full prompt for the lexical selection task with self-generated rules.

————————— System Prompt —————————-

Please only return a json with the following keys *<Translations>* and no other text. For each key the value should be a string in English explaining how the meaning and usage of that *<Target Language>* word is different from the others. The string should also include a brief example in *<Target Language>* of the word being used with an English translation. Please include the transliteration from *<Target Language>* to Latin characters if necessary.

————————— User Prompt —————————

When translating the concept "*<Concept>*" from English to *<Target Language>*, what is the difference in meaning between *<Translations>* and in which contexts should they be used? Here are sentences where each word is used in-context to help you: *<Sentences>*

Figure 9: Full prompt for generating rules from LMs.