Dictionary-Aided Translation for Handling Multi-Word Expressions in Low-Resource Languages

Antonios Dimakis$^{\alpha,\beta}$**, Stella Markantonatou**$^{\alpha,\gamma}$**, Antonios Anastasopoulos**$^{\alpha,\delta}$
$^{\alpha}$Archimedes, Athena R.C.
$^{\beta}$Department of Informatics and Telecommunications, National and Kapodistrian University of Athens
$^{\gamma}$Institute for Language and Speech Processing
$^{\delta}$Department of Computer Science, George Mason University
andimakis@di.uoa.gr, stellamarks@athenarc.gr, antonis@gmu.edu

Abstract

Multi-word expressions (MWEs) present unique challenges in natural language processing (NLP), particularly within the context of translation systems, due to their inherent scarcity, non-compositional nature, and other distinct lexical and morphosyntactic characteristics, issues that are exacerbated in low-resource settings. In this study, we elucidate and attempt to address these challenges by leveraging a substantial corpus of human-annotated Greek MWEs. To address the complexity of translating such phrases, we propose a novel method leveraging an available out-of-context lexicon. We assess the translation capabilities of current state-of-the-art systems on this task, employing both automated metrics and human evaluators. We find that by using our method when applicable, the performance of current systems can be significantly improved. However, these models are still unable to produce translations comparable to those of a human speaker.[1]

1 Introduction

Multi-word expressions (MWEs) are defined as combinations of at least two words which exhibit lexical, morphological, syntactic, semantic or statistical idiomaticity (Baldwin and Kim, 2009). Processing MWEs has long been considered one of the most challenging tasks in natural language processing (NLP) (Ramisch et al., 2010), especially when it concerns translation systems.

Despite neural models outclassing more traditional NLP techniques in many areas, the distinct nature of MWEs makes it so that the generalizing abilities inherent in these approaches cannot significantly increase performance. There is also currently a significant lack of datasets and benchmarks in this domain, especially for low-resource languages. Faithful automatic translation of MWEs is particularly important in improving the performance of general translation systems, given that MWEs make up a significant amount (estimates range from 11% to 40%) of the vocabulary in any piece of text (Sag et al., 2002; Candito et al., 2021).

Standard 'Unaided' Greek-to-English Translation
src: Η γκλάβα του δεν κόβει καθόλου.
trg: His balaclava does not cut at all. ✗

Proposed 'Aided' Greek-to-English Translation
src: 'κόβει η γκλάβα σε κάποιον'→'to be smart'.
Η γκλάβα του δεν κόβει καθόλου.
trg: His mind is not sharp at all. ✓

Figure 1: Our proposed dictionary-aided translation better translates multi-word expressions. The Greek source sentence means "He is not sharp at all", but the (incorrect) word-for-word translation is "His head does not cut at all", with the specific Greek word used for "head" rarely being used outside of this expression.

In this work, we use a large corpus of human-annotated MWEs in Greek and we measure the translation ability of current state-of-the-art systems on sentences containing these MWEs, both automatically and with the help of human evaluators. We focus on two tasks; standard "unaided" translation, where the model is only provided with the source sentence, and an "aided" version, where we provide additional guidance on the translation of the MWE used in the source sentence, taking advantage of the available lexicon. An example of both tasks is shown in Figure 1.
In short, we make the following contributions:

- We propose a new method for machine translation of in-context MWE using an out-of-context lexicon.
- We evaluate the ability of current state-of-the-art translation systems in handling text containing MWEs, using a large and general test set.

[1] We publicly release all code and datasets produced for this work: github.com/andhmak/dictMWE_MT

Findings of the Association for Computational Linguistics ACL 2024, pages 2588–2595
August 11-16, 2024 ©2024 Association for Computational Linguistics

2 Related Work

Previous research has explored the complexities of processing MWEs in various tasks, including MWE detection (Lai et al., 2023) as well as translation-specific settings (Kabra et al., 2023), such as in contrastive examples or metaphors, highlighting the need for specialized approaches. While neural models have excelled in many NLP tasks, their performance with MWEs remains suboptimal. In certain sub-areas, such as detection, traditional non-neural statistical methods remain remarkably competitive (Constant et al., 2017; Pasquer et al., 2020). This is theorized to be due to MWEs having a lexical quality, meaning that they behave similarly to out-of-vocabulary words when they do not appear in their idiomatic usage in the training data (Savary et al., 2019). This problem is therefore exacerbated disproportionately more for lower-resource languages, such as Greek.

There has also been research in the domain of dictionary-aided neural machine translation in a general setting, concerning single one-word lemmas (Zhang et al., 2021). Older methods have also attempted to leverage MWE dictionaries by incorporating them into the non-neural, statistical translation pipeline (Bungum et al., 2013), while newer ones have tried adding the dictionary to the training data of neural models (Arthur et al., 2016; Zaninello and Birch, 2020). Our work borrows ideas from the efforts towards terminology translation (Alam et al., 2021), which usually incorporates hard (Dinu et al., 2019; Susanto et al., 2020), usually through constrained decoding (Post and Vilar, 2018; Hu et al., 2019), or soft constraints (Bergmanis and Pinnis, 2021) during decoding. Similar techniques have also been employed for translating rare words (Ghazvininejad et al., 2023) and terminology in specific domains (Moslem et al., 2023; Bogoychev and Chen, 2023).

In contrast, our method leverages the prompting capabilities of Large Language Models to imbue them with information from a MWE dictionary (somewhat similar to the soft constraints) while still taking advantage of their generalizing abilities.

3 Methodology

Our evaluation focuses on two tasks: standard ("unaided") and "aided" translation.

Unaided Translation In this task, the model generates an output solely based on the provided source sentence in a traditional manner.

Aided Translation This task is based on our proposed method. The model receives each lone MWE and its human-generated translations along with the source sentence.

We already know from previous work (Baziotis et al., 2023) that, at least in the current landscape, it is not possible to reasonably translate passages containing MWEs without them appearing in the training data or a lexicon. By providing this lexicon in a simple manner as part of the prompt, we aim to measure the improvement in the output.

3.1 Dataset

We rely on a comprehensive corpus of human-annotated Greek MWEs, providing a diverse and representative set for evaluating translation models on this specific task.

Specifically, we use the IDION dataset of Greek MWEs (Markantonatou et al., 2019). This dataset provides a large set of Greek verb MWEs in dictionary form, as well as (possibly multiple) context-unaware translations of these phrases. It also contains 5750 Greek sentences, without an English translation, matched to one of 916 MWEs. These sentences have all been validated by linguists as being grammatically correct and representative of each MWE, from a larger set of Greek-language data gathered from the internet. These 5750 sentences form our MWE-focused test set.

4 Experiments

We use two state-of-the-art models for our experiments. The first is No Language Left Behind (NLLB; NLLB Team et al., 2022), a state-of-the-art multilingual model specializing in low-resource languages. We specifically use the 600M parameter version due to computational resource constraints. The second is GPT-4-0613 (OpenAI et al., 2023), a large language model operating as a chatbot.

We then evaluate the various translations, both in absolute and comparative quality, in different ways, both manually and automatically.

Unaided Translation Under the unaided translation setting, we benchmark both models—NLLB and GPT-4.

The NLLB model was provided simply with the source sentence (in Greek), the source language (Greek), and the target language (English). GPT-4 was provided with the following prompt:

```
Can you translate the Greek text
"[source sentence]" into English?
```

Aided Translation Our aided translation proposed method relies on the prompting capabilities of LLMs, hence it is only applicable to GPT-4 and not to NLLB. We now provide GPT-4 with the following prompt:

```
Can you translate the Greek text
"[source sentence]" into English?  For
context, the Greek multi-word expression
"[MWE]" can mean "[English meaning]",
"[possible other English meaning]", ...,
or "[possible other English meaning]".
```

4.1 Human Evaluation

For human evaluation, we employ 4 human annotators,[2] all native Greek and fluent English-as-a-second-language (ESL) speakers, to annotate a subset (356 sentences) of the outputs.[3] We set up an annotation interface that provided the following information:

1. The Greek source sentence.
2. The MWE in the source sentence in isolation.
3. The "unaided" translation generated by GPT-4 or NLLB.
4. The "aided" translation generated by GPT-4.

We did not tell the annotators which translation was produced by which model. We asked them to (a) choose the better translation of the two; (b) to provide a quality rating from 1 (terrible) to 5 (perfect) for each one; and (c) to indicate whether the MWE was correctly translated in each of the MT hypotheses.

In cases where GPT-4 generated more than a single translation, or provided more context apart from the translated sentence, only the first output sentence was taken into account.

Note that; in that first study, the choice of ESL speakers might lead to untrustworthy results, as the annotators might be unconsciously biased to accept something closer to their native language as correct. Hence, we also repeat the annotation study with non-Greek-speaking English native speakers, asking them only to rate the quality of the English outputs, without any regard to its similarity to the Greek source (which they cannot understand).

4.2 Automatic Evaluation with Quality Estimation

Beyond human evaluation, we use COMET (Rei et al., 2022) to benchmark automatic quality estimation metrics. COMET tries to predict the quality scores provided by expert annotators. These are scores similar to the direct assessment ones, but in practice COMET is trained to predict an overall MQM score. More specifically, we obtained absolute and relative scores, using the `comet-score` and `comet-compare` functions, respectively, for every set of outputs produced, namely the ones generated by NLLB, and those generated by GPT-4, both in the "aided" and "unaided" variants.

	Human Evaluation	
Model	general	MWE
baselines		
`unaided GPT`	2.6/5	43%
`NLLB`	1.6/5	23%
ours		
`aided GPT`	3.7/5	75%

Table 1: Our method better handles MWEs and produces better translations overall. Average quality for the translations produced with each method, as given by Greek-speaking annotators. 'general' reflects overall translation quality, while 'MWE' is MWE-specific.

Note that reference translations are not available, so we make use of a quality estimation model.[4]

Since human evaluators showed significant preference for "aided" GPT-4 outputs, we also evaluated the other two translation sets using the former as a reference. We did this in two ways, evaluating first using all data, and then focusing on the subset whose "aided" translations were rated as exemplary (5 stars) by the annotators.

5 Results

Our human evaluations indicated that the "aided" method using GPT performed significantly better than the GPT-based "unaided" one, which in turn outperforms the NLLB-based "unaided" method. Quality estimation metrics generally agreed with the human evaluations when it comes to this order of quality.

In Table 1, we present the results obtained from the Greek-speaking annotators. They rank "aided" GPT as significantly better than "unaided" GPT, with NLLB worse than both. However, even the best model is subjectively rated noticeably worse than a theoretical correct translation. The first col-

[2] One of them is also an author of the paper.

[3] Due to budget constraints.

[4] The `Unbabel/wmt20-comet-qe-da` pretrained model, which is trained on scores provided by human annotators, which are then normalized into z-scores (hence they can be negative).

Model	Human eval
unaided GPT	3.4/5
aided GPT	3.8/5

Table 2: Average fluency score of the English translations, given by native English-speaking (non-Greek) annotators. Our method slightly improves fluency.

umn contains the quality of the translations as a whole, while the second one concerns the correctness of the translation of the MWEs exclusively, irrespective of the rest of the sentence.

Additionally, when asked to explicitly choose the best between the "aided" and "unaided" translations, humans judged "aided" GPT outputs to be better than the "unaided" (78% of the time) and than the NLLB (95% of the time).

In Table 2 we show the evaluations of the English-speaking (but not Greek-speaking) annotators, who only judged the quality of the output without taking into account the input. We skipped evaluating NLLB this way, as it was clearly worse than the others. Here we find that our method not only improves general translation quality, but it also slightly increases the fluency of the output.

In Table 3, we depict the results obtained from the COMET quality estimation model (without a reference translation). These results are of limited value, as the correct translations of sentences using MWEs would diverge significantly from anything a model with no lexical information could predict. They also indicate NLLB's significantly inferior quality to the other models, but the order between the two GPT methods is reversed.

We also attempt to better compare the two baselines, using high-quality (as judged by humans) "aided" GPT outputs as a reference. Table 4 shows the respective reference-based COMET scores. It too indicates the same conclusion as the humans. Under the same settings, we also explicitly found this difference to be statistically significant for both variations (all data and only high-quality ones), with the null hypothesis being rejected when taking $p = 0.05$. This was also the case when only using 100 samples, during initial testing.

Targetted versus Random Sentences Using any random translation examples in the prompt has been observed to improve translation quality in certain cases, because they can help the LLM focus on the translation task (Zhang et al., 2023). It is reasonable, then, that the improvements we observe in our "aided" setting are due to this particular phenomenon, and not due to helping the system translate specifically the target MWE. We therefore also conduct a "random" GPT experiment, which is similar to the "aided" one but instead uses an MWE that is unrelated to the sentence being translated. We found that it indeed scores worse than even the "unaided" variant under the above settings (see Table 4), indicating that the improvement in translation quality should indeed be attributed to the information provided by the lexicon.

All human scores are normalised as explained in Appendix B.

Model	COMET QE score
baselines	
unaided GPT	-0.0097
NLLB	-0.1865
ours	
aided GPT	-0.0266

Table 3: Quality estimation of the translations produced with each method, without using any reference translations (higher is better). Despite the shortcomings of this evaluation, it mostly agrees with the others and NLLB is confirmed to be significantly outmatched.

Model	COMET ("aided" as ref)
baselines, all data	
unaided GPT	0.77
NLLB	0.62
only high-quality references	
unaided GPT	0.81
random GPT	0.76
NLLB	0.49

Table 4: Automatic evaluation (COMET, higher is better) of the translations produced with each method, using the "aided" outputs as a reference. GPT translations are better than NLLB, agreeing with human evaluations.

6 Discussion and Future Work

Our experimental results (§5) show that, at least for the IDION dataset, our proposed "aided" method for MWE translation through GPT-4 by providing a pre-contextualized lexicon significantly outperforms the naive "unaided" GPT-4 usage, with both producing better results when compared to state-of-the-art MT models (NLLB). This is significant as the latter is tailored specifically for use cases similar to our experiment, namely translation of lower-resource languages.

However, under all settings, current models seem to underperform in MWE translation. As established previously, a lexicon is most likely required to achieve human-level translations, but having a context-agnostic lexicon and simply providing it to the model, even if it has been generated by humans and correctly mapped to each instance, appears to not be enough to reach this level, although it does significantly improve performance.

We hypothesize that English speakers seem to prefer the same system as English-Greek bilinguals because the model avoids awkward word-to-word translations. So while they cannot tell whether the translation is adequate, they can penalize cases where a literal translation is obviously nonsensical.

Moreover, it seems that GPT, on accessing its own evaluation, might amend its answer. We therefore also attempted further improving GPT-4's output by giving it the chance to evaluate its previous response and possibly change it (details provided in Appendix A). It did not seem to significantly help, as it was found to rate itself highly and barely, if at all, change its translation.

Future Work We believe that it would be worthwhile to professionally translate our dataset and extend it to more languages, thus providing the necessary context in bilingual settings, making it useful for training NMT models on MWEs.[5] Moreover, since current automatic evaluation methods present several shortcomings, our translated dataset could generally be used for benchmarking the MT abilities when it comes to MWEs.

In the era of LLMs, perhaps we should reevaluate how we construct dictionaries. LLMs learn more effectively from more context, so rather than providing simple definitions with minimal disambiguating example usages, lexicographers could focus on providing more and longer examples.

7 Limitations

Human and automatic evaluations seem to generally agree with each other, but we acknowledge that, due to the lexical nature of our task, without a reference translation, quality estimation models (such as the Unbabel one we use) cannot be relied upon to offer trustworthy results.

We also acknowledge that since the evaluators are all primarily native Greek speakers, they may have been influenced in their evaluation of word-for-word translations into English, as they would have made sense to them and it could occasionally have been difficult to know how natural they would sound to native speakers. We attempted to mitigate this issue with our English-speakers evaluation experiment, but ideally one would employ native English speakers that also understand Greek.

8 Ethics Statement

We believe that our work does not introduce any significant additional risks other than those inherent in the models used.

We have obtained permission from all annotators to publish the data they produced in the context of this paper. The annotators were volunteer co-authors and co-workers, and no monetary compensation was provided for their involvement.

The content of IDION is available under a CC-BY-NC license, in XML format. Their usage in this project is therefore consistent with its intended use. All models we use come with permissive licenses, at least when it comes to research.

Acknowledgments

The authors are grateful for the feedback from the anonymous reviewers and area chairs. This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. Antonios Anastasopoulos is also supported by the US National Science Foundation under award IIS-2327143.

This project was also supported by resources provided by the Office of Research Computing at George Mason University (https://orc.gmu.edu).

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the wmt shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

[5] Unfortunately we lack the funds to do so ourselves.

Timothy Baldwin and Su Nam Kim. 2009. Multiword expressions. In *Handbook of Natural Language Processing, Second Edition, CRC Press*, Boca Raton, USA. Nitin Indurkhya and Fred J. Damerau.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. 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.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Lars Bungum, Björn Gambäck, André Lynum, and Erwin Marsi. 2013. Improving word translation disambiguation by capturing multiword expressions with dictionaries. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 21–30, Atlanta, Georgia, USA. Association for Computational Linguistics.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2021. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2):415–479.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068.

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

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Indra Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

Stella Markantonatou, Panagiotis Minos, George Zakis, Vassiliki Moutzouri, and Maria Chantou. 2019. Idion: A database for modern greek multiword expressions. pages 130–134.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

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

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan

Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of NAACL-HLT*, pages 1314–1324.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. Multiword expressions in the wild? the mwetoolkit comes in handy. In *Coling 2010: Demonstrations*, pages 57–60, Beijing, China. Coling 2010 Organizing Committee.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. 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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Conference on Intelligent Text Processing and Computational Linguistics*.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with Levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online. Association for Computational Linguistics.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France. European Language Resources Association.

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

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.

A GPT-4 self-evaluation and translation amending

While maintaining the conversation context after each initial translation, we provided the following prompt:

> ’Would you evaluate your translation positively? Give a 1-5 score, and change your response to improve it if necessary. Format your response as: "X/5<new translation (if applicable)>", don’t include any other explanation.’

B Normalization of human annotations

Since not all human annotators reviewed the same sentences for each model (most only annotated "aided"/"unaided" GPT pairs (task I), while some also annotated "aided" GPT/"unaided" NLLB pairs (task II)), we normalized the results so that they are comparable.

Therefore, if the set of all participants is K and N_k is the set of all annotation scores by participant k on task I, the average for task I is defined as:

$$\sum_{k \in K} \sum_{i \in N_k} \frac{i}{|K| \cdot |N_k|}$$

If $L(\subset K)$ is the set of all participants that also annotated task II, and M_l is the set of all annotation scores by participant l on task II, the normalized average for task II is defined as:

$$\sum_{l \in L} \sum_{i \in M_l} \frac{i}{|L| \cdot |M_l|} \cdot \frac{\sum_{k \in K} \sum_{i \in N_k} \frac{i}{|K| \cdot |N_k|}}{\sum_{i \in N_l} \frac{i}{|L| \cdot |N_l|}}$$

These values, rounded to the nearest decimal, are what is displayed in the relevant tables in Section 5.