Fine-tuning mBART50 with French and Farsi data to improve the translation of Farsi dislocations into English and French.

Behnoosh Namdarzadeh behnoosh.namdar@gmail.com
CLILLAC-ARP, Université Paris Cité, Paris, F-75013, France

Sadaf Mohseni sadafmohseni@gmail.com
CLILLAC-ARP, Université Paris Cité, Paris,F-75013, France

Lichao Zhu lichao.zhu@u-paris.fr
CLILLAC-ARP, Université Paris Cité, Paris, F-75013, France

Guillaume Wisniewski guillaume.wisniewski@u-paris.fr
Laboratoire de Linguistique formelle, Department of Linguistics, Université Paris Cité, Paris,F-75013, France

Nicolas Ballier nicolas.ballier@u-paris.fr
Laboratoire de linguistique formelle/ CLILLAC-ARP, Université Paris Cité, Paris, F-75013, France

Abstract

In this paper, we discuss the improvements brought by the fine-tuning of mBART50 for the translation of a specific Farsi dataset of dislocations. Given our BLEU scores, our evaluation is mostly qualitative: we assess the improvements of our fine-tuning in the translations into French of our test dataset of Farsi. We describe the fine-tuning procedure and discuss the quality of the results in the translations from Farsi. We assess the sentences in the French translations that contain English tokens and for the English translations, we examine the ability of the fine-tuned system to translate Farsi dislocations into English without replicating the dislocated item as a double subject. We scrutinized the Farsi training data used to train for mBART50 (Tang et al., 2021). We fine-tuned mBART50 with samples from an in-house French-Farsi aligned translation of a short story. In spite of the scarcity of available resources, we found that fine-tuning with aligned French-Farsi data dramatically improved the grammatical well-formedness of the predictions for French, even if serious semantic issues remained. We replicated the experiment with the English translation of the same Farsi short story for a Farsi-English fine-tuning and found out that similar semantic inadequacies cropped up, and that some translations were worse than our mBART50 baseline. We showcased the fine-tuning of mBART50 with supplementary data and discussed the asymmetry of the situation, adding little data in the fine-tuning is sufficient to improve morpho-syntax for one language pair but seems to degrade translation to English.

Keywords: mBART50, Farsi-French, Farsi-English, fine-tuning multilingual models

Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, pages 152–161
September 4–8, 2023, Macau SAR, China.
©2023 The authors. This article is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

1 Introduction

Farsi (Persian) is shown to be the language having least datasets in a survey of Neural Machine Translation for low-resource languages (Ranathunga et al., 2023). [1] Previous research on Neural Machine translation for Farsi shows clear limitations of existing systems (Ghasemi and Hashemian, 2016) and highlights the scarcity of NLP resources (Namdarzadeh et al., 2022) for Farsi, as shown in the contributions of the Proceedings of the Workshop on NLP Solutions for Under-Resourced Languages (Freihat and Abbas, 2021).Moreover, previous research on using mBART50 (Liu et al., 2020), a multilingual model trained on 50 languages, has shown its limitations for the translations of Farsi into English and even more so for the translations into French where hallucinations (Raunak et al., 2021) and English words were observed in the translated texts (Namdarzadeh et al., 2022). This paper is a follow-up on this initial series of observations and reports our first series of experiments to fine-tune mBART50 for the translation of Farsi into English and French. Farsi being an under-resourced language, we found that the translation into French gave way to hallucinations and English words, whereas most of the challenge set created for the occasion was translated into grammatical sentences from Farsi into English, but nevertheless failed to capture what we called 'pragmatic adequacy'. The translations into English like '*I hate the heat*' followed the English SVO canonical order and lost the pragmatic intention ('*as to the heat, I hate it*').

Focusing on a specific syntactic phenomenon, dislocation, may provide us with a better perspective on the nature of pragmatic inadequacies. Dislocation is a syntactic phenomenon that is productive in Farsi. Left dislocation constructions, like in French and, less frequently, in English, can have specific semantic and pragmatic functions . They can be used to promote a topic (Azizian et al., 2015) or to focus on a specific constituent of the sentence. Several constructions are available for dislocations in Farsi. Azizian et al. (2015) studied left dislocated construction as a marked construction in the framework of Construction Grammar (Goldberg, 1995). They suggest that "a syntactic two-place construction is responsible for preposing the oblique to the sentence-initial position by leaving a pronominal enclitic coreferential with it in its original place", show that it may apply to monovalent, divalent and trivalent verbs. In the construction they identify, almost all participant roles can be left-dislocated except agents and experiencers. The left-dislocated element can be marked with "-ra." In this example, the object (*Garma*) is initially stated, and it can transfer a semantic effect based on the context.

Garma ro azash motenaferam.

Heat- RA-OBJ of-OBJ hate-1SG-PRE

Reference translation: *As for the heat, I hate it* (Azizian et al., 2015: 104). Google Translation and Microsoft Bing: I hate the heat.

Ketab o Saman ferestad.

book RA-OBJ S send-1SG-PST

Reference translation: *The book, Saman sent.* (Azizian et al., 2015) Google: Saman sent the book.

To answer one of the reviewers' concerns on the centrality of dislocations in our study, the reason we focused our analysis on this construction is we first noticed in a previous paper that this structure was difficult to translate into French and into English for the mBART50 model (Namdarzadeh et al., 2022) and because this syntactic phenomenon is frequent in spoken data (Dabir-Moghaddam, 1992), and can manifest in various forms and functions. In other words, left dislocated constituents can co-occur with several types of markers such as reflexives, making their translations by NMT toolkits challenging. According to discourse configurations, dis-

[1] The use of Persian is more common in formal and academic settings, while Farsi is commonly used in informal and everyday conversations among native speakers.

location constructions can serve different functions, and capturing the corresponding pragmatic intention can be challenging. This is the reason why we assign such importance to investigating this construction.

The rest of the paper is structured as follows: Section 2 sums up previous research, Section 3 presents our testing set, our additional data used for fine-tuning and the parameters we used to fine-tune mBART50. Section 4 presents our results, we both discuss the Farsi to French and French to Farsi translations. Section 5 discusses them and outlines our future research.

2 Previous Research

To improve multilingual models, previous methods include adding data with back-translation from monolingual data (Sennrich et al., 2016), data from multimodal input (see for instance pictures and their descriptions for Bengali (Parida et al., 2021), data from languages of the same language family (Chronopoulou et al., 2022), using denoisers to add other languages incrementally (Üstün et al., 2021) or fine-tuning on small datasets (Smirnov et al., 2022). While other researchers have worked on more "massively multilingual NMT" (Aharoni et al., 2019) adding data to an initial subset of the TED talks for no less than 103 languages (including one million examples in Farsi), we have focused on controlling parallel data for Farsi to English and Farsi to French translations.

Among the research questions we had is the effect of language interference or at least the fact that we found English tokens in the translation of Farsi into French, probably because of this bootstrapping effect of the co-presence of the different languages. We were more interested in the change in one language, namely potentially English, when we added more data in French and especially we wanted to see if adding more French data would change the number of English tokens that we found in the translations into French.

3 Materials and Methods

For mBART50, the system does not include bilingual French and Farsi data, but the 25 and then 50 language pairs of languages with English as a pivot so that the system manages to learn from the different language pairs including English and another language and enable translation from one language to the other even if there are no training data for a specific language pair outside English. As indicated in (Tang et al., 2021), in mBART50, French belongs to the first group of training data, containing more than 10 Millions of training data in bitext pairs and Farsi corresponds to the middle tier (100k to 1M tokens). The appendix mentions 14,4895 sentences for Farsi (TED 58) used for training and 3,930 for validation and 4,490 for test. The training data for French is made up of the WMT14 training data, resulting in 36,797,950 (train), 3,000 (validation) and 3,003 (test) sentences.

3.1 Testing Set

As a testing set, we used a simple data set that suggested the research question in the first place (Namdarzadeh et al., 2022). The existence of English tokens in the translations into French by mBART50 and a discrepancy in the quality of the translation into English and into French triggered the investigation of fine-tuning with French and Farsi aligned data. For the translation into French, we had observed issues in syntactic adequacy, namely some sentences were incomplete and from the translation into English we analysed, we had observed syntactic adequacy, the sentences were grammatical, but we evidenced a form of pragmatic inadequacy: the order of the constituents was too close to the English canonical order (subject verb object), so that we regretted the lack of pragmatic adequacy to account for the focalisation effects in the Farsi original. For a sentence like, "as to the heat, I hate it", a "focus construction", as Huddleston and Pullum call them (Huddleston and Pullum, 2002), would be more suitable than

the simple, plain and somewhat inexpressive "I hate the heat". The data set is comprised of 57 sentences comprising a dislocation and has been compiled using several sources from typical textbooks to more elaborate grammars of Farsi.

3.2 Fine-tuning parameters for mBART50

For the fine-tuning, we followed the instructions from the scripts available from the HuggingFace implementation. [2] We retrained the model with three epochs, using gelu as an activation function. The process took 25 seconds on an NVIDIA A100 GPU and consumed 152.544 W for the GPU and 77.5 W for the CPU (we used the codecarbon library [3] (Schmidt et al., 2021) to measure our carbon footprint). For the data, we first use the translation of a short story from Farsi into French. We privileged literary material because we expected rarer structures to be relevant for the translation of the translation into French. We used the short story 'Zayandeh Roud's wounds', which is one of the short stories in a short story collection called 'Angel Cake Recipe', consisting of fifteen short stories written by Pooya Monshizadeh, an Iranian writer currently living in the Netherlands. This short story won several national literary awards in Iran. The short story was translated into French by the second author. For the translation into English, we found a translation by Sajedeh Asna'ashari published online in the "*Stockholm Review*". [4] Since we used the HuggingFace scripts for the fine-tuning as a starting point, this had a consequence on the output language after the fine-tuning. No possible predictions in another language than the one provided in the fine-tuning is then possible. As a consequence, we first fine-tuned with Farsi and French and then fine-tuned from scratch from the same mBART50 model with Farsi and English.

4 Results

This section summarises some of our findings in terms of the quality of the translations.[5] By paying attention to our distinction between pragmatic adequacy and syntactic adequacy, we discuss the differences in translation that were more satisfactory from a pragmatic standpoint.

4.1 Impact on the translation into French

Compared to the baseline of our initial mBART translation from Farsi into French, dramatic improvements were noted from a morpho-syntactic standpoint: the translation contained only one English token (suppressing English in the translations was our main expectation) and the subject/verbs agreements were correct except for two first plural person verbal agreements for a second singular person subject in a cleft sentence *C'est toi qui me connaissons mieux que lui. / Ce n'est pas toi qui l'avons vu, mais nous.* In spite of an error for gender (*ton lèvre*), this change was spectacular, given the small size of the fine-tuning data (116 sentences). Nevertheless, semantic issues were not resolved and other cropped up. For named entities, a regional football team was translated as *l'équipe de Napoléon [Napoleon's team]* and some sentences did not make sense at all. Dislocations tended to be more often translated by cleft sentences, which sounds like a relevant strategy for topic promotion. A case of catastrophic forgetting has been noticed: the NMT system lost its ability to translate towards any other language than the one it was fine-tuned with. As a consequence, we first tried to test the backtranslation abilities if the fine-tuning was operated with Farsi as the target language and French as the source Language. We report our main observations in the following subsection.

[2] https://github.com/huggingface/transformers/blob/main/examples/pytorch/translation/run_translation.py

[3] https://pypi.org/project/codecarbon/

[4] https://thestockholmreview.org/the-wounds-of-zayanderud-fiction-by-pooya-monshizadeh

[5] Data is available on https://github.com/Behnooshn/Summit2023

4.2 Impact on Back-translation

When using back-translation from French into Farsi, the following phenomena were observed: for some sentences where the reflexive is part of the dislocation construction, the translation hallucinated and repeated a reflexive pronoun. This seems to suggest that this construction is rarer in mBART50 training data. Some sentences were translated into English and some objects tended to be suppressed as well as reflexive pronouns. Lexical errors were spotted in the choice of verbs NER issues appeared, for example, confusions between family name (*Iradji*) and country (*Irak*). Analysing the fine-tuned back-translation output, from French to Farsi, we noted that, even though the fine-tuning bore on a very limited set of sentences, the number of tokens in English in the Farsi translation was limited to proper nouns, so that NER remains an issue. There are less hallucinations but more omissions could be noted: several phrases were not translated. Some translations were still impaired semantically, some verbs being translated by their very opposite ("*hate the heat*" becomes "*love the heat*"). Severe misunderstandings were spotted, sometimes leading to contradictions (ie the believer that does not believe).

4.3 BLEU scores

We compared with a reference translation before and after our fine-tuning with mBART50. Once more, the BLEU score is not a satisfactory instrument to measure progress in the textual output of the translation. Even though the difference in BLEU score is not really meaningful below five, the difference between and after fine-tuning is striking in terms of morphosyntactic quality and spectacularly so with so little data. The result is not necessarily more satisfying from a semantic point of view and definitely not from a pragmatic point of view, but the grammatical well-formedness has dramatically improved with fine-tuning with so few differences and so small BLEU scores. It is likely that for our BLEU scores below 5, the unigram matches can probably be attributed to punctuation or function words. The BLEU scores are inferior to five and show the mBART50 predictions are very different from our proposed translations. Named entities remain a crucial issue with first names in the reference translations often not being recognised in the translation.

4.4 Qualitative Analysis

Contrary to our findings for the baseline of the mBART50 translation from Farsi into French, no English tokens were found. Two cases of hallucination were observed, for one of the most complex sentences of the dataset and with the same token repeated over and over (*vieille*). The changes are quite dramatic for syntactic adequacy: contrary to our baseline all the agreements in person, number gender and even mode (for an imperative form) are correct and predictions for French are consistent with well-formedness constraints, even if some averbal sentences are debatable. Semantically, many issues remain unsolved, such as *"by bus"* translated by *"by train"* or self-contradictory sentences such as *"Je n'aime pas la chaleur, mais je l'aime"*, meaning *"I don't like the heat, but I like it"* instead of something like *"as to the heat, I hate it"*). Pragmatically, richer topic/focus constructions were observed with several *c'est* cleft constructions.

Regarding the English outputs after fine-tuning mBART50, we have observed critical points related to the translation of Farsi dislocation constructions into English. For instance, named entities have not been translated accurately, as they may not be (sufficiently) present in the training data. Additionally, there are instances where the English translations contain more information compared to the original Farsi source text, indicating over-translation (Wang, 2012).

Another frequent issue in NLP is gender bias, which involves a preference towards one gender over the other. This bias is prevalent in the outputs of the systems (Wisniewski et al., 2022a; Matthews et al., 2021). We have also noticed the same bias in the English outputs of

mBART50, where the system seems to prefer masculine genitives over feminine ones.

As Azizian et al. (2015) pointed out, some dislocations entail a complex interaction with argument structure, which may account for the introduction of unnecessary arguments in the mistranslation of reflexives. The use of reflexives holds high importance in Farsi, since they carry a pragmatic effect based on the discourse and the prosody effect applied by the speaker.

Man khodam yadam hast.

I-SUB-SG REF-SG remember-SG be-SG

Reference translation: *I myself remember*. mBART50: I remember you.

As shown in the above example, not only is the use of reflexive ignored, but mBART50 also adds the object **you** in its translation, which undermines the entire meaning of the Farsi source text. There are different examples of this kind in our test set for which the mBART50 outputs are not semantically and pragmatically adequate translations.

This failure may be due to the fact that such structures are often under-represented in the training data (Hovy and Prabhumoye, 2021) and more generally, in available reference data such as treebanks in Farsi (Namdarzadeh et al., 2022). In the next section, we will discuss the available data and materials, and integrate another model for transcription and translation of this low-resourced language.

5 Discussion

5.1 Asymmetry of Data

Contrary to expectations, our fine-tuning experiment with the same Farsi short story proved to be more efficient with French than for English, giving an edge to the lower resource for the initial training of mBART50, which contradicts the adage that NMT results in "worse quality in low-resource settings, but better performance in high-resource settings" Koehn and Knowles (2017). Such a discrepancy in the initial training data for mBART50 should be replicated for another language pair involving a lesser-resource language. We should also try to replicate the fine-tuning on a bigger scale, to see whether English needs more data than the other languages when fine-tuning mBART50.

An alternative strategy would be to fine-tune for English with the whole Farsi-English data available on OPUS. We have retrieved aligned Farsi-English TED Talks raw corpus from OPUS which is tokenized with Stanza NLP Pipeline and sub-tokenized by SentencePiece with a language model trained with CC-aligned Farsi data (5 million lines) for our future experiments with English and Farsi. The French-Farsi Opus data is available but a 2017 extraction from the most translated TED talks exists with their French and English translations. [6] We also aim to analyse the initial mBART50 training data, reported in the reference paper as being based on TED talks, but the source of the bilingual corpus is not mentioned (Farsi subtitles of American TED talks or English subtitles of Farsi talks).

5.2 Language Direction of the translated material

In our experiment, we have used data originally in Farsi translated into French. It should be noted that the mBART50 system is trained for Farsi with TED Talks. It is not clear whether the talks are in English and subtitled into Farsi, so that the direction of the language was also of interest to us. In our next experiments, we plan to control language direction for the fine-tuning data and use data from a translation of a French novel into Farsi, the translation of Simone de Beauvoir's *Journal de Guerre, septembre 1939-janvier 1941*. Because of its occasionally more informal style, we would expect the data to potentially include more dislocations in French.

[6] https://github.com/neulab/word-embeddings-for-nmt

5.3 Further Research

For Farsi, we have identified three types that we would like to translate better. And for French, we have identified the construction as being of paramount importance in the translation. We intend to search existing data using universal dependency annotation, making the most of the dislocated dependency relation label, see our previous paper that described the methodology. We intend to potentially observe a threshold of data that may enable a better translation of the dislocation construction. Possibly realised by using synthetic data, namely transforming, updating current authentic sentences by changing the noun, for example, by synonymous, so that we would easily expand our data sets by reduplicating the structure.

By investigating such challenging cases of dislocation, we aim to enhance the multilingual translation toolkits' accuracy and overall performance, especially for languages that exhibit similar characteristics (ie dislocation) and possess limited training data. Through a thorough analysis of the toolkits' errors, we can develop more robust models capable of effectively handling various types of dislocation.

The incorporation of the different typologies of dislocation and corresponding markers can help us expand our test sets, especially when it comes to less-resourced languages. To achieve this for Farsi, we will investigate three types of dislocation, including object markers *sh* and *râ*, and reflexive pronouns. For example, commercial toolkits still misfire when translating dislocation constructions such as the *sh* marker, the translation toolkits cannot identify the correct dependent (*mashin*) and produce incorrect translations (Google Translate: *Kiana's car hit the door.* Microsoft Bing: *Kiana knocked on the door*).

Mashino Kiana tup-o be dar- esh zad.

Car-râ Kiana ball-OBJ to door-OBJ OBJ-marker hit-3S

our reference Translation: As to the car, Kiana [only] hit its door with the ball. (After(Azizian et al., 2015))

One of the parameters that could be investigated is the distance between the two constituents *mashin* and *esh*, which may play a role in Natural Language Understanding and translations tasks. The frequency of the dislocated constituents can also be another issue that requires investigation. By training our own translation toolkit with calibrated training data enriched with specific examples of dislocations and their relevant translations, we hope to observe frequency thresholds: detect the frequency at which the translation system can "learn" the challenging constituent by fine-tuning.

6 Conclusion

In this paper we have reported a small-scale experiment trying to fine-tune mBART50 for the translations from French into Farsi. We have not tried yet to fine-tune on the Facebook implementation of Cross-lingual Language Model Pretraining (Lample and Conneau, 2019), which has Farsi as one of the 100 languages used for the multilingual training. The baseline results may be more satisfactory for this model than what we observed with mBART50. Our paper is just a small case study that contributes to cross-lingual language understanding (XLU) of these multilingual models. It remains to be seen whether this XLM-100 model developed for Cross-lingual Representation Learning (XLR), which outperforms mBERT (Conneau et al., 2020), would display similar tendencies, since more data was collected for Farsi (13,259 M tokens) than for French (9,870 M tokens) on Common Crawl and Wikipedia. Our main finding is that even a small amount of fine-tuning drastically reduces the number of foreign words in the translations and the number of hallucinations. An important drawback is that our fine-tuned model with French-Farsi parallel data only translates into French, as a direct consequence of the HuggingFace implementation. More research is needed to investigate whether more aligned data in the fine-tuning phase would allow predictions to be semantically more adequate. Taking into

account what we have managed to learn with so little data, we would like to be able to evaluate what a system is capable of learning during fine-tuning, and at what cost (at least in terms of the number of examples needed for learning, and – if we have a good idea to avoid that – in terms of "forgetting"). More specifically, we would like to compare the ability of fine-tuning to change :

- the translation of certain tokens (a sort of "lexical" capacity). Here, clearly, the absence of some tokens data accounts for some of the spotted errors, especially for very frequent place names in Farsi.

- the translation of certain grammatical constructions (a sort of "syntactic" capacity)

In both cases, the idea would be to use artificial examples to control exactly what is present in the fine-tuning data and how it impacts the resulting translations. For lexical capacity, we could, for example, create sentences with new words and watch when the system is able to produce them, or "change" the meaning of a word (typically, replace all "green" tokens with "peanuts" in a parallel corpus) and watch when the system's output changes. The assumption is there might be thresholds of frequency in the fine-tuning data to observe this. For syntactic capacity, we could start with the translation of dislocation: we would create an artificial corpus with well-translated dislocations and look at the size of the set that needs to be provided for the system to be able to translate dislocations correctly. To clarify what we mean by "artificial corpus", we have the intuition that we could define a "parallel pattern" like the pattern used to investigate gender bias (the N has finished PRON work) (Wisniewski et al., 2022b) to "model" the translation of a dislocation (in the same way a pattern like DET ADJ NOUN is translated into DET NOUN ADJ) that we would simply instantiate on a lexicon (without taking into account the semantics at the source sentence level, but guaranteeing that the "bilingual links" of the syntactic structure are well preserved). For both experiments, it would be interesting to see what happens on eng-fas, fra-fas, eng-fra pairs to see the impact of the pre-trained model. We have identified the parallel translations of the TED talks which could be used as a starting point for this triangulation.

Author Contributions

Behnoosh Namdarzadeh designed the dislocation test set in Farsi and the reference translations into English and into French and supervised the translation outputs. Sadaf Mohseni provided the aligned data for fine-tuning and contributed to the reference translations and to the analyses of the different translations. Lichao Zhu prepared the OPUS data for ulterior fine-tuning. Guillaume Wisniewski supervised the mBART50 fine-tuning experiments. Nicolas Ballier designed the study, and wrote the first draft of the manuscript. All authors contributed to the analysis of the outputs.

Acknowledgements

We thank the four anonymous reviewers for their input on a preliminary version of this paper. This publication has emanated in part from research supported by a PAUSE research grant to Sadaf Mohseni funded by Collège de France and Université Paris Cité under ANR grant ANR-18-IDEX-0001, Financement IdEx Université de Paris), which is gratefully acknowledged. Nicolas Ballier benefited from a CNRS research leave at LLF (Laboratoire de Linguistique Formelle), for which grateful acknowledgement is made.

References

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Azizian, Y., Golfam, A., and Kambuziya, A. K.-e. Z. (2015). A construction grammar account of left dislocation in Persian. *Mediterranean Journal of Social Sciences*, 6(6 S2):98.

Chronopoulou, A., Stojanovski, D., and Fraser, A. (2022). Language-family adapters for multilingual neural machine translation. *arXiv preprint arXiv:2209.15236*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Dabir-Moghaddam, M. (1992). *On the (In)dependence of syntax and pragmatics: Evidence from the postposition -rá in Persian*, pages 549–574. De Gruyter Mouton, Berlin, New York.

Freihat, A. A. and Abbas, M. (2021). Proceedings of the second international workshop on NLP solutions for under resourced languages (NSURL 2021) co-located with ICNLSP 2021. In *Proceedings of The Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*.

Ghasemi, H. and Hashemian, M. (2016). A comparative study of Google Translate translations: An error analysis of English-to-Persian and Persian-to-English translations. *English Language Teaching*, 9(3):13–17.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture Series. University of Chicago Press.

Hovy, D. and Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Huddleston, R. and Pullum, G. K. (2002). *The Cambridge Grammar of English Language*. Cambridge University Press.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 33.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Matthews, A., Grasso, I., Mahoney, C., Chen, Y., Wali, E., Middleton, T., Njie, M., and Matthews, J. (2021). Gender bias in natural language processing across human languages. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 45–54, Online. Association for Computational Linguistics.

Namdarzadeh, B., Ballier, N., Wisniewski, G., Zhu, L., and Yunès, J.-B. (2022). Toward a test set of dislocations in Persian for neural machine translation. In *The Third International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2022)*, pages 14–21.

Parida, S., Panda, S., Biswal, S. P., Kotwal, K., Sen, A., Dash, S. R., and Motlicek, P. (2021). Multimodal neural machine translation system for English to Bengali. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 31–39.

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., and Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Raunak, V., Menezes, A., and Junczys-Dowmunt, M. (2021). The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Schmidt, V., Goyal, K., Joshi, A., Feld, B., Conell, L., Laskaris, N., Blank, D., Wilson, J., Friedler, S., and Luccioni, S. (2021). CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Smirnov, A. V., Teslya, N., Shilov, N., Frank, D., Minina, E., and Kovacs, M. (2022). Comparative analysis of neural translation models based on transformers architecture. In *ICEIS (1)*, pages 586–593.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021). Multilingual translation with extensible multilingual pretraining and finetuning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Üstün, A., Berard, A., Besacier, L., and Gallé, M. (2021). Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6650–6662, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wang, M. (2012). An analysis of over-translation and under-translation in perspective of cultural connotation. *Lecture Notes in Information Technology*, 16:129–133.

Wisniewski, G., Zhu, L., Ballier, N., and Yvon, F. (2022a). Analyzing gender translation errors to identify information flows between the encoder and decoder of a NMT system. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 153–163, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wisniewski, G., Zhu, L., Ballier, N., and Yvon, F. (2022b). Analyzing gender translation errors to identify information flows between the encoder and decoder of a nmt system. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 153–163.