university of groningen

faculty of arts

Evaluating Translation Quality in NMT
A Dutch–English Case Study

J.G.J. Oving

Bachelor thesis
Informatiekunde
J.G.J. Oving
s5251370
June 9, 2025

ABSTRACT

This study investigates the performance of different neural machine translation (NMT) models on the Dutch-English language pair, with a focus on zero-shot and few-shot translation. Using the Europarl corpus, we compare three NMT models: MarianMT, Facebook-NLLB-200, and TOWER. Each model is evaluated using four automatic evaluation metrics: BLEU, BLEURT, chrF, and COMET. Each metric highlights different aspects of translation quality, including fluency and adequacy. The results show that TOWER in the 5-shot setting, consistently outperforms across all the evaluation metrics, showing the benefits of instruction-tuning and in-context learning. Surprisingly, Facebook-NLLB-200 also shows a strong zero-shot performance, especially in the English-Dutch translation. MarianMT performs acceptable on surface-level metrics like chrF, but underperforms compared to the other models. These findings suggest that few-shot prompting can significantly improve performance in NMT, as well as the importance of the models architecture. Limitations of the study include reliance on automatic evaluation metrics and potential overlap between the Europarl corpus and model pretraining data. Future research should consider human evaluation and more diverse datasets.

CONTENTS

PREFACE

This thesis marks the end of my Bachelor's degree in Information Science at the University of Groningen. I wrote this during the final semester of my last year as a 21 year old Information Science student. The topic of my thesis contains different skills and aspects that I learned in the 3 years that I followed the Information Science course. I am happy too be finally writing this part of the thesis since it is late at night, once again. I would like to extend my thank to my supervisors, Kyo Gerrits, Xiaolu Wang, and Nastja Shaboltas, for their support, guidance and feedback throughout the whole thesis process. I would also like to thank Andreas van Cranenburgh for providing his guidance and feedback on the coding part of my thesis.

1 INTRODUCTION

Neural machine translation (NMT) has revolutionized the field by using deep learning to improve translation adequacy (how well the meaning is preserved) and fluency (how natural the translation sounds) (Zhang, 2025). With the rapid advancement of large language models (LLMs), multilingual models and domain-specific models, it is crucial to assess how different NMT models perform in certain language pairs, such as Dutch-English. Given the constant evolution of machine translation, assessing whether NMT models remain effective in this ever-changing research field is crucial to understand their long-term use and applicability (Ashraf, 2024). We present a comprehensive evaluation of NMT models for machine translation, analyzing their performance and limitations. The current state-of-the-art NMT models are based on the transformer structure (Vaswani et al., 2017). This architecture has proven to be extremely efficient in capturing the linguistic structures of sentences, as well as improving translation quality. To explore how well these NMT systems perform on domain-specific data, we perform experiments to examine their translation abilities. Specifically, we investigate the performance of different NMT models for machine translation across a domain-specific Dutch-English parallel corpus. We use the Europarl Dutch-English parallel corpus (Koehn, 2005) for conducting our experiments. This research examines and compares three NMT models: **MarianMT (Helsinki-NLP/opus-mt-nl-en Helsinki-NLP/opus-mt-en-nl)**, **TOWER (TowerInstruct-7b-v0.2)**, and **Facebook-NLLB-200 (distilled-600m)**. These models were selected based on their architecture, capabilities and relevance to our study. MarianMT serves as a strong baseline for high-resource language pairs. Facebook-NLLB-200 offers extensive multilingual support. TOWER is an instruction-tuned model which enables few-shot prompting. Each model offers us a unique perspective on translation quality. We explore zero-shot translation with all three models and few-shot translation only with the TOWER model, as it is instruction-tuned and built around the concept of few-shot translation. To evaluate these models, different automatic evaluation metrics are used for assessing translation quality. We use the following evaluation metrics: **BLEU**, **BLEURT**, **COMET**, and **chrF**. These evaluation metrics offer different perspectives, thus enhancing our ability to evaluate the results. By evaluating these different models with different translation approaches on translation quality, this study aims to provide insights into their strengths and weaknesses. We also explore the performance of zero-shot and few-shot translation on these models. Thus, looking at the goals stated above, our main research question is:

- **How do different neural machine translation (NMT) models perform on Dutch-English translation, as measured by automatic evaluation metrics?**

We derive the following sub-question:

- *What is the impact of zero-shot and few-shot learning on the translation quality of NMT models?*

The thesis is structured as follows: Chapter 2 contains the background section. In chapter 3, we explain the data, pre-processing and processing that was required to answer our questions. In chapter 4, we describe the methods used and in chapter 5 we discuss the results we gathered. Lastly, we conclude our thesis in chapter 6 by a reflection on our research.

2 BACKGROUND

2.1 INTRODUCTION TO MACHINE TRANSLATION

Machine translation (MT) is the task of a machine automatically translating a text from one language to another language. Over the past decade, Neural Machine Translation (NMT) has replaced the earlier statistical and rule-based approaches (Wang et al., 2022), offering huge improvements in translation fluency and adequacy. These current models are based on the transformer architecture, which was introduced by Vaswani et al. (2017). This architecture is based around the idea of attention mechanisms, which allow the model to weigh the importance of different words in a sentence when generating a translation. Earlier models processed sequences sequentially however, the transformer architecture processes the input data in parallel, which leads to an improvement in efficiency. More recently, the emergence of large language models (LLMs) such as GPT and multilingual transformers have introduced new capabilities (Lyu et al., 2023). These capabilities include zero-shot and few-shot translation (Brown et al., 2020). Zero-shot translation refers to a situation where a pre-trained model translates text between language pairs on which it has not been fine-tuned. With zero-shot translation the models rely on the knowledge learned during training without requiring additional examples. This approach can be useful for low-resource languages. On the other hand, few-shot translation involves feeding the pre-trained model a prompt with a few translation examples, to illustrate:

- *Translate from Dutch to English: Dutch: De zon schijnt vandaag. English: The sun is shining today. Dutch: Vertaal deze zin. English: ...*

This prompting technique allows the model to adapt to the task quickly, making it a useful technique in certain contexts or low-resource settings.

2.2 STATE-OF-THE-ART MODELS

As neural machine translation continues to evolve, different state-of-the-art models emerge. Among these models is TOWER (Alves et al., 2024): TOWER is a recent model that has gained attention in the machine translation domain. The paper mentions that TOWER is not like traditional NMT models, which are trained on vast corpora of parallel texts. TOWER is fine-tuned using task-specific prompts, such as "Translate from Dutch to English", allowing it to generalize better across translation and summarization tasks. This enhances TOWER's ability to perform tasks like translation and summarization by giving it explicit instructions during training. The paper mentions a few key features of TOWER:

- Multilingual capabilities: TOWER is built on LLaMA-2 (Touvron et al., 2023). However, the problem with LLaMA-2 is that it was exposed to relatively little non-English data during pretraining, limiting its use for multilingual capabilities. Alves et al. (2024) improve TOWER by continuing LLaMA-2's pretraining on a highly multilingual corpus. This makes TOWER highly suitable for translating between languages with varying amounts of parallel corpora.

- Instruction-tuning: TOWER is built on the idea that training models with clear task instructions enables the model to generalize better across different tasks.

This makes TOWER useful in a zero-shot or few-shot setting, which will be used in our research.

These features make TOWER highly relevant to the goals of our study, since our research aims to evaluate the performance of different NMT models on the translation of Dutch-English language pairs, especially in a zero-shot or few-shot setting. By evaluating TOWER's translation performance on the Europarl Dutch-English parallel corpus, we can assess its usefulness for translating domain-specific data. Another state-of-the-art model is the MarianMT model mentioned in Junczys-Dowmunt et al. (2018). We specifically use the Helsinki-NLP/opus-mt-nl-en model and Helsinki-NLP/opus-mt-en-nl model, which are designed for high-quality translation between Dutch and English. These models are developed as part of the Opus-MT project, which is an extension of the MarianNMT framework (Tiedemann and Thottingal, 2020). The models are based on the transformer architecture, making them very efficient. Multiple versions of the MarianMT model can be found in the Opus project, and so we make use of the specific opus-mt-nl-en and opus-mt-en-nl models, which are specifically built for the Dutch-English translation task. Key features of the MarianMT are:

- Multilingual translation: The opus-mt-nl-en and opus-mt-en-nl model are specifically trained to handle the Dutch-English translation pair, but the model is part of a broader multilingual collection that covers other language pairs. The many types and versatility of the model allows it to perform well across different texts and domains.

- Pre-trained on large dataset: The opus-mt-nl-en and opus-mt-en-nl model have been pre-trained on the OPUS corpus, which consists of millions of sentences from different domains. This pretraining on a large corpus makes these models useful for the translation task we will perform.

The Helsinki-NLP/opus-mt-nl-en and Helsinki-NLP/opus-mt-en-nl models are a great candidate for our research, since they are lightweight and efficient models, as well as their zero-shot capability. Another model we will be using in this study is Facebook's NLLB-200 (distilled-600m) (Costa-jussà et al., 2024), which is part of the No Language Left Behind project by Meta AI. The NLLB project by Meta AI focuses on "Delivering evaluated, high-quality translations directly between 200 languages—including low-resource languages like Asturian, Luganda, Urdu and more. It aims to give people the opportunity to access and share web content in their native language, and communicate with anyone, anywhere, regardless of their language preferences."[1] The NLLB-200 model has support for 200 languages and is trained using self-supervised learning and multilingual data. We use the distilled-600M version, which is a smaller and more efficient version of the original model. Key features of Facebook-NLLB-200 include:

- Multilingual translation: NLLB-200 is one of the first models to provide high-quality support for 200 languages.

- No specific task prompt: Unlike the other two models Facebook-NLLB-200 is not instruction-tuned, instead it relies on language tokens to specify the language. This makes it very simple to use. The model instead uses ISO 639 language tags (e.g. eng-Latn, nld-Latn) to identify the soruce and target language.

Similar to the other two models, Facebook-NLLB-200 is an efficient multilingual model. Facebook-NLLB-200 is useful for high- and low-resource contexts, hence we will use it in our research. Facebook-NLLB-200's ability to handle Dutch-English translation directly, and being a light, efficient model makes it an ideal benchmark

[1] https://ai.meta.com/research/no-language-left-behind/

$$\text{BLEU} = \underbrace{\min\Big(1, \exp\big(1 - \frac{\text{reference-length}}{\text{output-length}}\big)\Big)}_{\text{brevity penalty}}\underbrace{\Big(\prod_{i=1}^{4} precision_i\Big)^{1/4}}_{\text{n-gram overlap}}$$

Figure 1: BLEU score formula

suitable for our research. Its performance in zero-shot settings is one of the interests in this research, as this study aims to assess how well these models generalize without additional fine-tuning.

2.3 EVALUATION METRICS

To assess these models properly in terms of adequacy and fluency, we will make use of multiple automatic evaluation metrics (Kocmi et al., 2021). Adequacy refers to the degree to which the translation retains the same meaning as the source text, while fluency concerns the grammatical correctness and naturalness of the output in the target language. First, we will use the BLEU evaluation metric mentioned in Papineni et al. (2002). BLEU is one of the most used metrics for evaluating machine translation. BLEU works by measuring the n-gram overlap between the translation of the machine and the reference translation as seen in figure 1.

The BLEU score ranges from 0 to 1. A higher score indicates greater similarity. However, BLEU has a well-known issue: it primarily detects surface-level similarity and not the actual adequacy or fluency of a translation (ble, 2006).

To address this limitation, we also employ other automatic evaluation metrics such as COMET. Rei et al. (2020) introduce COMET as a learned evaluation metric that uses pre-trained language models and is fine-tuned on human judgments. COMET evaluates machine translation by comparing the source and hypothesis sentences in a shared embedding space. This offers a better alignment with adequacy and fluency than the n-gram overlap method of BLEU. This makes COMET useful for evaluating zero-shot and few-shot translation, since semantic adequacy is preserved. The COMET score ranges from 0 to 1, like BLEU.

In addition to the COMET and BLEU metric we also make use of the BLEURT automatic evaluation metric. Sellam et al. (2020) mention that BLEURT is a learned evaluation model just as COMET. Unlike BLEU, BLEURT is sensitive to meaning, word choice and phrasing, making it useful to detect improvements in adequacy. The BLEURT evaluation metric is trained on synthetic data and is fine-tuned on datasets containing human annotations. This allows it to generalize well across different domains and languages. The BLEURT score ranges from -2 to 1. A negative score indicates a very bad translation and a positive higher score indicates a good translation in terms of adequacy and fluency.

Finally, we also make use of the chrF metric (Popovic, 2015) in our evaluation. ChrF is a character-level F-score metric that measures the precision and recall of character n-grams between the machine translation and the reference translation. In contrast to the other evaluation metrics, chrF operates on the character level. ChrF balances both recall and precision, which can provide a more subtle view of translation quality than the other evaluation metrics alone. By combining BLEU, COMET, BLEURT and chrF, we aim to obtain a consistent and well-generalized evaluation of translation quality. While recent multilingual models like MarianMT, Facebook-NLLB-200 and TOWER have shown promise in zero- and few-shot translation, direct comparable evaluations of neural machine translation models on a specific language pair like Dutch-English remain limited. This research aims to fill that gap. Table 1 shows the metrics we will use in our study, including the focus of each metric.

Metric	Type	Focus	Score Range
BLEU	Surface-based	n-gram overlap (precision)	0 to 100
BLEURT	Learned (neural)	Semantic adequacy, fluency	-1 to 1
COMET	Learned (neural)	Meaning, adequacy, reference + source aware	0 to 1
chrF	Surface-based	Character n-gram F-score (precision + recall)	0 to 100

Table 1: Overview of the evaluation metrics used in this study.

2.4 RELEVANCE

Previous work like van Egdom et al. (2023) and Arenas and Toral (2022) have evaluated Dutch-English translation quality in terms of style, creativity, partially adequacy and fluency. However, they made use of literary texts. Mohammed et al. (2024) did evaluate NMT translation quality on non-literary Dutch-English data, however this was part of a shared-task as well as that the models were fine-tuned on the dataset. We instead propose to use a real-world domain-specific dataset. Koehn (2005) introduced the Europarl corpus. The corpus consists of professionally translated sentences from the European Parliament. This dataset offers us a high-quality parallel Dutch-English corpus in a specific domain. By using Europarl, this study aims to assess not only the grammatical correctness but also the domain-specific adequacy of translations, which is often overlooked in evaluations based on literary texts. The Europarl dataset consists of parliamentary/political language which tends to be more formal. Evaluating translation quality in this domain can reveal different strengths and weaknesses of translation models. This study is the first, to our knowledge, to benchmark zero- and few-shot performance of these multilingual models on the Europarl Dutch-English corpus using multiple automatic evaluation metrics.

3 DATA AND MATERIAL

3.1 COLLECTION

For our dataset, we used a Dutch-English parallel corpus from the Europarl dataset, which can be found and downloaded on the Europarl website.[1] The Europarl dataset is widely used in machine translation and contains sentences from European parliament proceedings, making it a suitable corpus for a domain-specific machine translation task. We obtained the data from the Europarl website and downloaded the following file:

- parallel corpus Dutch-English, 190 MB, 04/1996-11/2011

The corpus is 190MB in size and covers a period from April 1996 to November 2011. This corpus provides us with around 2 million sentences and each language contains around 50 million words. This corpus is in a pre-aligned format, where each sentence in Dutch corresponds to a sentence in English. This makes it well-suited for our machine translation task. We chose this version of the corpus because it provides a large and consistent dataset covering a 15-year span, ensuring both size and topical consistency in the political domain.

3.2 PROCESSING

We ensured the dataset was consistently encoded in UTF-8 and removed any empty lines. Before we apply the data to the models, we first need to make sure that the data is properly pre-processed. We applied the following steps for pre-processing the data:

1. Text cleaning: We will remove untranslatable and irrelevant items to make sure the data is properly formatted and has readable sentences. We specifically removed the following:
   (a) URL's
   (b) Email-addresses
   (c) Unwanted characters (e.g. @,)
   (d) Whitespace
2. Lower-casing: we will lowercase all data so the models do not treat the same words in different cases as different
3. Sentence tokenization and subword tokenization: We will tokenize the corpus into sentences and also process the text into subword units. The models rely on sentence-level alignment and can also learn more effectively when subwords are used. For subword tokenization, we applied Byte Pair Encoding (BPE).

[1] https://www.statmt.org/europarl/

4 METHOD

To answer our research question, we adopt a methodology centered on neural machine translation (NMT) models. These models form the foundation of our experiments and are evaluated under different translation conditions, including zero-shot and few-shot settings. Given that NMT has demonstrated impressive progress in translation quality, it is essential to explore how various models perform under different conditions such as zero and few-shot translation. The methodology involves several steps:

1. NMT models
2. Execution of the code
3. Pre-processing of the dataset
4. Zero-shot translation
5. Few-shot translation
6. Assess the performance of the models using automatic evaluation metrics

4.1 MODELS

For this experiment we make use of three different models: MarianMT, TOWER, and Facebook-NLLB-200. All three of these models are openly available on the Hugging Face website[1]. For the experiments, we make use of these exact models:

1. **MarianMT (Helsinki-NLP/opus-mt-nl-en) and (Helsinki-NLP/opus-mt-en-nl)**[2]
2. **Facebook-NLLB-200 (distilled-600m)**[3]
3. **TOWER (TowerInstruct-7B-v0.2)**[4]

These models were selected to represent different types of translation architectures, allowing us to compare a diverse set of models. Figure 2 shows a comparison of the parameter size of the different models.

4.1.1 MarianMT

MarianMT is a standard encoder-decoder NMT model (Junczys-Dowmunt et al., 2018), which will serve as a strong baseline for a high-resource language pair like Dutch-English. We specifically use the Dutch-English **(opus-mt-nl-en)** and **(opus-mt-en-nl)** model from the OPUS project Tiedemann and Thottingal (2020). MarianMT is evaluated solely in a zero-shot setting in our experiment without any additional fine-tuning. The input text is pre-processed using the model's own tokenizer, which is imported from the Huggingface transformer library (Wolf et al., 2019). The translations MarianMT produces are generated by the greedy decoding strategy.

[1] https://huggingface.co/

[2] https://huggingface.co/Helsinki-NLP/opus-mt-nl-en https://huggingface.co/Helsinki-NLP/opus-mt-en-nl

[3] https://huggingface.co/facebook/nllb-200-distilled-600M

[4] https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2

4.1.2 Facebook-NLLB-200

Facebook-NLLB-200 is a model designed for multilingual translation (Costa-jussà et al., 2024), with support for low-resource languages. Although Dutch-English is not a low-resource language pair, this model allows us to observe how well a multilingual system is able to perform on a high-resource language pair. For this study, we specifically make use of this version: **Facebook-NLLB-200 (distilled-600m)**. The input text is pre-processed using the model's own tokenizer: NllbTokenizer[5] Similar to MarianMT, the Facebook-NLLB-200 model is evaluated solely in a zero-shot setting. Neither of these models are suitable for evaluation in the few-shot setting, since both models are not designed for providing few-shot prompting.

4.1.3 TOWER

TOWER is a recently developed instruction-tuned NMT (Alves et al., 2024), which is optimized for translation using few-shot prompting. Unlike traditional sequence-to-sequence models, TOWER makes use of an embedding space where task instructions and examples are embedded in the input. We evaluate TOWER in a zero-shot setting, 1-shot setting, 3-shot setting, and finally a 5-shot setting, which is common practice in few-shot learning research to observe how performance scales with the amount of context provided (Brown et al., 2020). We construct few-shot prompts containing 1-5 Dutch-English sentence pairs from the Europarl dataset, followed by a new Dutch input sentence for the model to translate. For this study, we make use of the following TOWER model: **(TowerInstruct-7B-v0.2)**. We use TowerInstruct-7b.v0.2, as our hardware was not able to support the larger `TowerInstruct-13B` model.

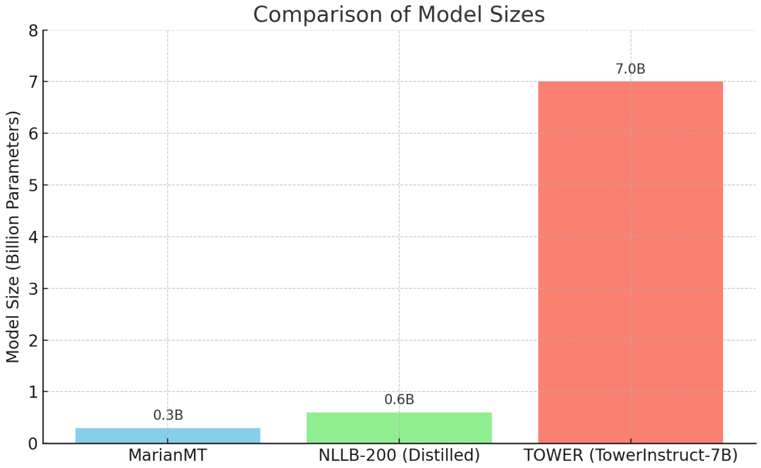

Figure 2: Comparison of model sizes based on parameters.

4.2 EXECUTION OF THE CODE

The Hábrók high-performance computing cluster was used to execute Python scripts given its GPU capabilities[6]. We used Hábrók in the following way:

- Access to the Hábrók cluster was obtained through the University of Groningen's student application process[7].

[5] Documentation can be found here: https://huggingface.co/docs/transformers/model_doc/nllb

[6] Habrok documentation can be found here: https://wiki.hpc.rug.nl/habrok/start

[7] This is only available for Students or Teachers at the University of Groningen, https://www.rug.nl/society-business/center-for-information-technology/research/services/hpc/habrok

- A virtual environment was created for running our different files.
- We assigned GPU(s) to our files.
- We ran a bash script through the Hábrók terminal to execute our code.

Because Hábrók allocates a powerful GPU to our files, we could run our experiments efficiently.

4.3 PRE-PROCESSING OF THE DATASET

For the pre-processing of our Europarl dataset, we developed a Python script called `preprocess-data.py`[8], which takes the Dutch-English parallel corpus and prepares it for zero and few-shot evaluation. To implement and pre-process our dataset, we utilized a list of different Python libraries:

- `transformers`
- `torch`
- `nltk`
- `re`
- `tokenizers`
- `tqdm`

The transformer library from Huggingface was used to load and process the NMT models, this includes encoder-decoder models `AutoModelForSeq2SeqLM` and decoder-only models `AutoModelForCausalLM`. For the tokenization of the models, we used `AutoTokenizer`. The `torch` library was essential for handling tensor operations, model loading and utilizing GPU acceleration through Hábrók. This allowed us to efficiently run our experiments on the Hábrók high computing cluster. For sentence tokenization, we made use of the `nltk` library. Regular expressions were handled with Python's built-in `re` module, which was especially useful for cleaning model outputs during evaluation. For the subword tokenization we used the `tokenizers` library from Huggingface. This library allowed us access to Byte-Pair encoding (BPE), which enabled subword segmentation. Lastly we made use of the `tqdm` library, this library provided a real-time progress time-bar. Tqdm helped us monitor our experiment's runtime, and allowed us to debug errors that occurred in the code. We first executed a bash file called: `preprocess-data.sh` to execute our code in the Hábrók terminal. The `preprocess-data.py` file then first loads the `europarl-v7.nl-en.nl` and `europarl-v7.nl-en.en` files. The code then proceeds to pre-process the data according to the steps described in Section 3.2. `preprocessdata.py` returns two pre-processed files called: `preprocessed-dutch.txt` and `preprocessed-english.txt`, which contain the pre-processed data from the Dutch-English parallel corpus.

4.4 ZERO-SHOT TRANSLATION

To evaluate the zero-shot translation performance of the NMT models, we implemented a Python script called `zero-shot.py`, which loads the pre-processed Dutch-English parallel corpus and applies the MarianMT, Facebook-NLLB-200, and TOWER model. These models were evaluated for their ability to translate from Dutch to English and English to Dutch without additional fine-tuning. The Python script relies on the following Python libraries:

[8] Files can be found in this GitHub repository`https://github.com/JoostOving/Scriptie`

- `transformers`
- `torch`
- `tqdm`
- `nltk`
- `evaluate`
- `re`
- `sacrebleu`

These libraries are explained in Section 4.3. Apart from the `tokenizer` library, we make use of the same libraries with the addition of the Huggingface `evaluate` library and the `sacrebleu` library. These libraries will be explained in Section 4.5. The `zero-shot.py` script first loads in the different models[9], it then performs inference by batching input sentences, and then processing these sentences through each model. The models received 5000 sentences as input for their translation. We chose 5000 sentences as this number of sentences yields consistent scores across the different models used in our study. Given the large size of these models (especially TOWER), we utilized two separate GPUs on the Hábrók cluster. We assigned MarianMT and Facebook-NLLB-200 to one GPU, and TOWER to another GPU. This configuration significantly improved the runtime and avoided memory issues during inference. After inference each model's output is then post-processed to remove prompt-related padding tokens. For the TOWER model specifically we implemented a function to clean the responses, since TOWER tended to respond in the following manner:

- Tower: English: as environment alists , we do not want an excuse for a motion . Dutch:

Evaluation metrics would penalize TOWER for the prompt words in the output (i.e. English:, Dutch:). TOWER required structured prompts to generate the translations in the zero-shot setting. The prompts used in our experiments go as follows:

- `Translate the following from Dutch to English.\nDutch: {s}\nEnglish:`
- `Translate the following from English to Dutch.\nEnglish: {s}\nDutch:`

This prompt yielded the most consistent results in our experiment, hence we used it.

4.4.1 Few-shot translation

The steps taken for translation in the few-shot setting in our experiment are the same as in the zero-shot setting. We developed a separate Python script called: `few-shot-tower.py`. This script only loads the TOWER model and makes use of a single GPU, since the use of two separate GPUs is redundant here. The script proceeds to take the same steps as in the `zero-shot.py` file, however only the TOWER model is loaded. The script then performs inference on the TOWER model by batching 5000 input sentences for translation. In our experiments we ran the TOWER model in three different few-shot settings:

- 1-shot setting
- 3-shot setting
- 5-shot setting

Shot Setting	EN → NL Prompt	NL → EN Prompt

1-shot	`English: the commission must track down...`
	`Dutch: de commissie moet onwettige steun...`
	`Translate the following...`
	`English: {s}`
	`Dutch:`
		`Dutch: de commissie moet onwettige steun...`
		`English: the commission must track down...`
		`Translate the following...`
		`Dutch: {s}`
		`English:`

3-shot	`3 English-Dutch example pairs...`
	`Translate the following...`
	`English: {s}`
	`Dutch:`
		`3 Dutch-English example pairs...`
		`Translate the following...`
		`Dutch: {s}`
		`English:`

5-shot	`5 English-Dutch example pairs...`
	`Translate the following...`
	`English: {s}`
	`Dutch:`
		`5 Dutch-English example pairs...`
		`Translate the following...`
		`Dutch: {s}`
		`English:`

Table 2: Few-shot prompts for EN ↔ NL Translation Tasks. The {s} variable represents the input sentence.

The prompts 2 seen above were provided for the different few-shot settings. The prompts that were used in our study can be found in the appendix, section A. Once translations were generated, we evaluated model performance using automatic evaluation metrics, as detailed below.

4.5 ASSESS THE PERFORMANCE OF THE MODELS USING AUTOMATIC EVALUATION METRICS

After post-processing the model outputs, the script proceeds to evaluate translation quality using multiple automatic evaluation metrics. We evaluated each model's performance on the Dutch->English and English->Dutch translation task using the following metrics:

- BLEU
- BLEURT
- COMET
- chrF

These evaluation metrics were imported using the `evaluate` Huggingface library and the `sacrebleu` library. The `evaluate` library offers a wide variety of commonly used NLP metrics, including BLEURT, chrF, and COMET. On the other hand, `sacrebleu` is a library designed specifically for the computation of BLEU. A detailed explanation of how these evaluation metrics work can be found in Section 2.3. We evaluated the models in the following translation scenarios:

1. Zero-shot translation: MarianMT, Facebook-NLLB-200, and TOWER
2. 1-shot translation: TOWER
3. 3-shot translation: TOWER
4. 5-shot translation: TOWER

After processing all the input sentences through the different models, the `zero-shot.py` and `few-shot-tower.py` files return the output sentences in the following structure:

Sentence:	839
English:	the competition report 1998 is not a bad foundation for this but , in fact , there is nothing that could not be further improved upon .
Reference:	het mededingings verslag 1998 is daarvoor geen slecht uitgangspunt , maar is over de gehele lijn voor verbetering vatbaar .
MarianMT:	het mededingingsverslag 1998 is hiervoor geen slechte basis , maar er is in feite niets dat niet verder zou kunnen worden verbeterd .
Facebook NLLB:	Het mededingingsrapport 1998 is hier geen slechte basis voor , maar er is in feite niets dat niet verder kan worden verbeterd .
TOWER:	;het verslag over de mededinging van 1998 is hiervoor geen slechte basis, maar er is in feite niets dat niet verder zou kunnen worden verbeterd.

Table 3: Example of translation output for English → Dutch (zero-shot)

Sentence:	6
Dutch:	ik wil u vragen deze minuut stilte staande in acht te nemen .
Reference:	please rise , then , for this minute ' s silence .
MarianMT:	I would ask you to observe this minute's silence.
Facebook NLLB:	I 'd like to ask you to observe this minute of silence .
TOWER:	to ask you to observe a minute's silence.

Table 4: Example of translation output for Dutch → English (zero-shot)

The `few-shot-tower.py` file returned the results in the same way as seen in Tables 3 and 4, but without the MarianMT and Facebook-NLLB-200 in the output. Some samples of the output can be found in the appendix A. After collecting the translated outputs, we applied the same evaluation pipeline using the automatic evaluation metrics. Table 5 below presents the evaluation output format.

Metric	**MarianMT**	**Facebook-NLLB-200**	**TOWER**
BLEU	–	–	–
BLEURT	–	–	–
COMET	–	–	–
chrF	–	–	–

Table 5: Evaluation scores across all models in the zero- and few-shot settings

[9] Documentation on how to load the models can be found here: https://huggingface.co/docs/transformers/main/en/index

5 RESULTS AND DISCUSSION

In this thesis, we use the following definitions for concepts related to our research question:

- **Performance** refers to how well a machine translation model translates Dutch-English in terms of output quality, as measured with automatic evaluation metrics.
- **Translation quality** is evaluated using four metrics: **BLEU**, **BLEURT**, **chrF**, and **COMET**. Each metric captures different aspects of translation quality.
- **Impact of zero-shot and few-shot learning** is defined as the change in performance of the TOWER model when given a number of example translation sentences as part of its input.
- **Comparison** between the models refers to a relative evaluation of the models scores on the same evaluation metric and dataset.

The results as seen in Tables 6 and 7, offer a detailed view of the performance of three different NMT models: MarianMT, Facebook-NLLB-200, and TOWER under both zero-shot and few-shot translation conditions. In this section, we interpret and compare the performance of these models on the Dutch-English and English-Dutch translation tasks. We also discuss why certain models may have performed better or worse, and reflect on some limitations of the experimental setup.

System	**COMET**	**BLEURT**	**chrF**	**BLEU**
MarianMT (opus-mt-nl-en) Zero-Shot	0.424	-1.000	42.914*	4.438
Facebook-NLLB-200 (distilled-600m) Zero-Shot	0.696*	0.000*	40.073	13.025
TowerInstruct-7b-v0.2 Zero-Shot	0.664	-0.110	40.636	15.104*
TowerInstruct-7b-v0.2 One-Shot	0.735	0.067	45.999	17.846
TowerInstruct-7b-v0.2 Three-Shot	0.568	-0.303	32.371	10.352
TowerInstruct-7b-v0.2 Five-Shot	**0.740**	**0.137**	**46.952**	**18.438**

Table 6: Evaluation Results for NL → EN Translation. The best scores across the models are marked bold. * denotes the best result in the zero-shot setting.

System	**COMET**	**BLEURT**	**chrF**	**BLEU**
MarianMT (opus-mt-en-nl) Zero-Shot	0.556	-0.498	43.336	3.427
Facebook-NLLB-200 (distilled-600m) Zero-Shot	**0.756***	-0.213*	**47.402***	14.974*
TowerInstruct-7b-v0.2 Zero-Shot	0.715	-0.304	44.094	14.652
TowerInstruct-7b-v0.2 One-Shot	0.722	-0.333	44.145	13.804
TowerInstruct-7b-v0.2 Three-Shot	0.667	-0.383	40.296	12.853
TowerInstruct-7b-v0.2 Five-Shot	**0.756**	**-0.184**	47.118	**16.735**

Table 7: Evaluation Results for EN → NL Translation

5.1 DUTCH → ENGLISH TRANSLATION RESULTS

When we look at Table 6 and figures 3 and 5, we can clearly see that the 5-shot TOWER model performs the best across all metrics. This might be due to TOWER's

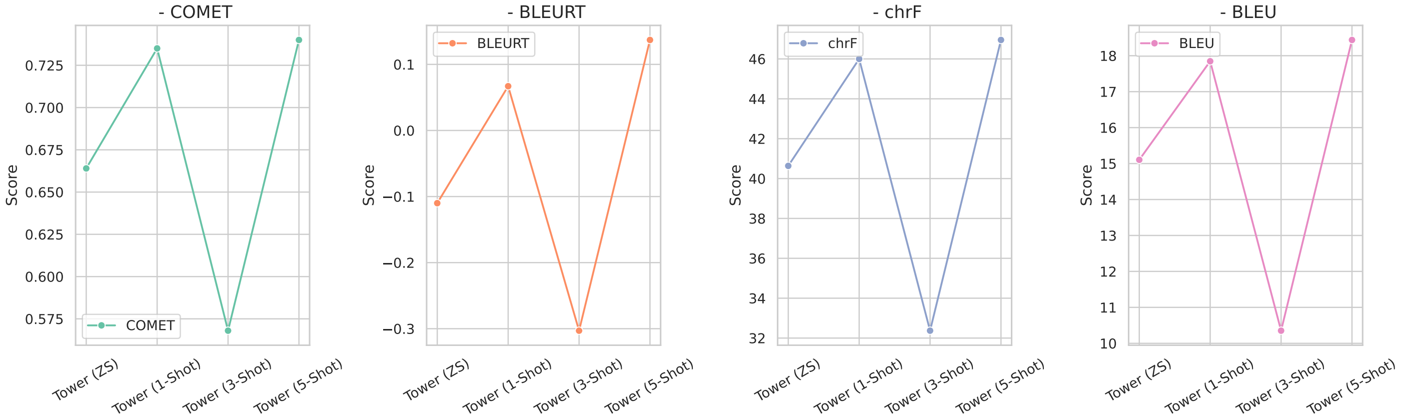

Figure 3: Performance of TOWER at different shot settings Dutch -> English.

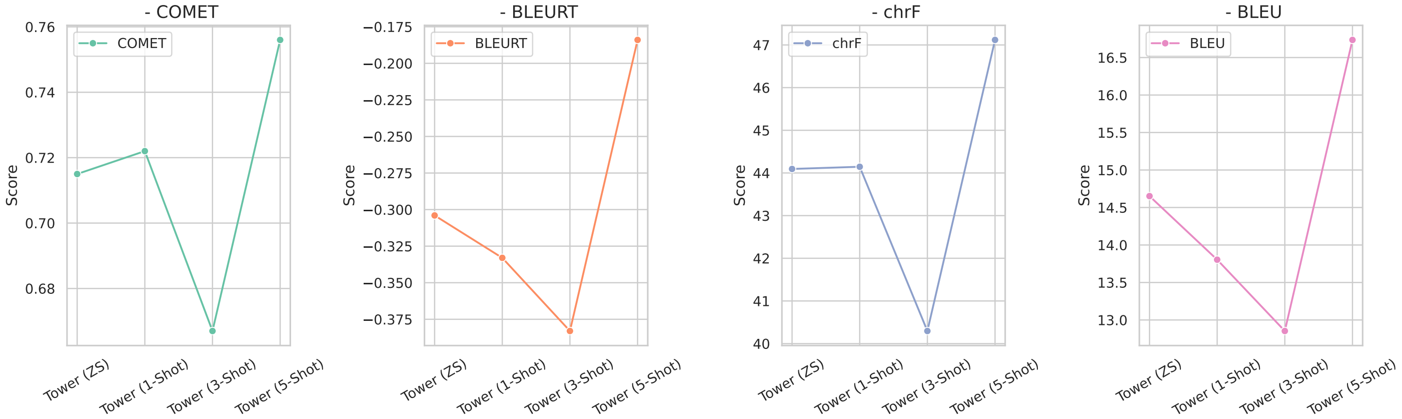

Figure 4: Performance of TOWER at different shot settings English -> Dutch.

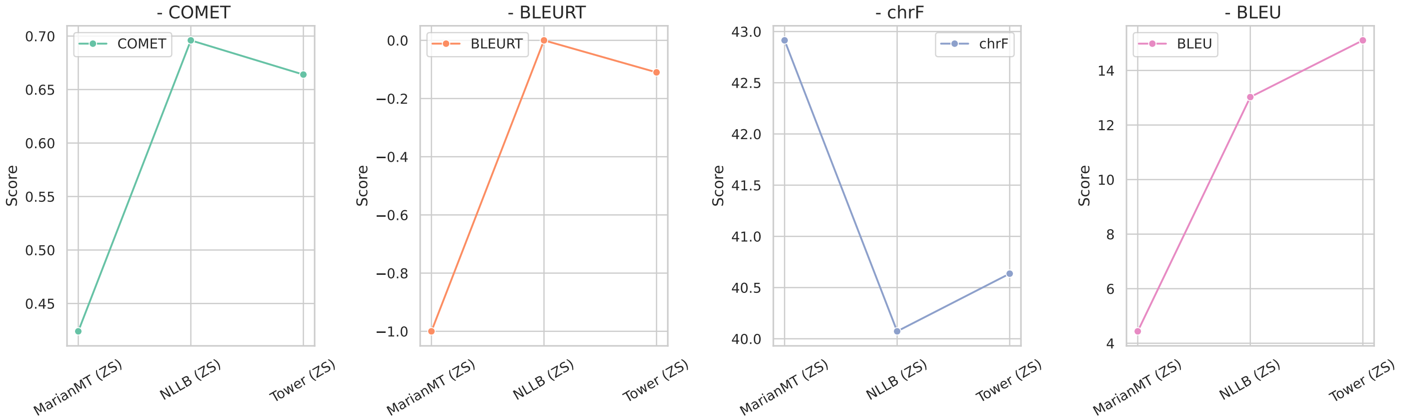

Figure 5: Performance of the models in zero-shot setting Dutch -> English.

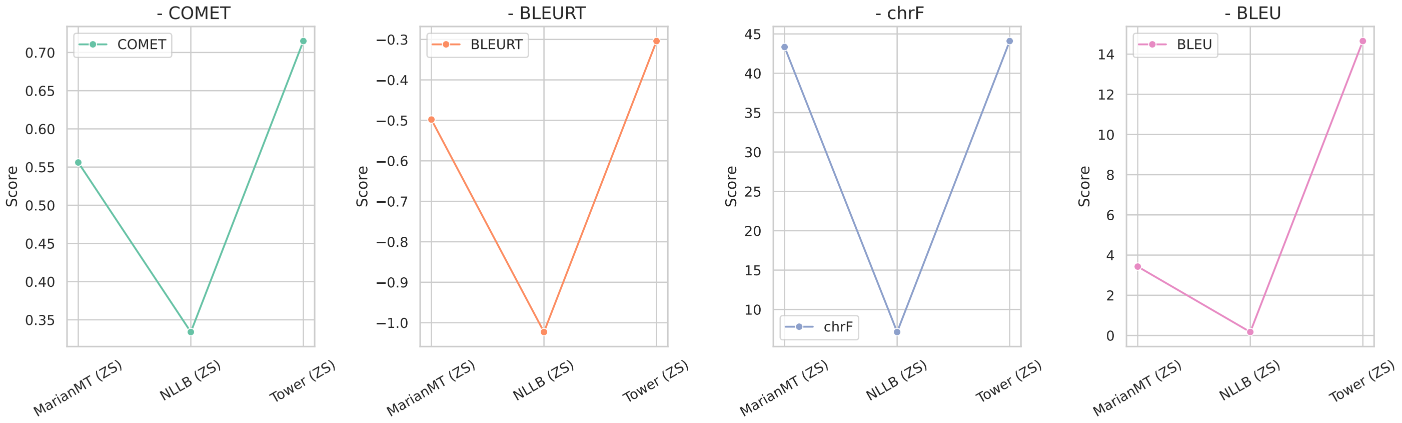

Figure 6: Performance of the models in zero-shot setting English -> Dutch.

instruction-tuning and in-context learning. It achieves the highest COMET score (0.740), BLEURT score (0.137), chrF score (46.952), and BLEU score (18.438), indicating that more few-shot examples significantly improves translation quality. The improvements are consistent across both surface-level metrics and reference-based metrics.

MarianMT overall performs the worst across the models, however it is interesting to mention that MarianMT scores the highest chrF of all the models in a zero-shot setting, although the other models follow very closely. Interestingly, while MarianMT achieves the highest chrF score in the zero-shot setting among the models, it scores the lowest in BLEU for zero-shot Dutch to English translation. This might suggest that although the surface-level word overlap may be low, the model still captures morphological or lexical similarity, as reflected in the chrF score. This large gap in performance illustrates the limitations of BLEU, especially in morphological rich languages like Dutch and English.

Facebook-NLLB-200 performs better than MarianMT on all metrics except chrF, if we look at figure 5. This aligns with the multilingual capabilities of Facebook-NLLB-200, which was trained on many languages including Dutch and English. Facebook-NLLB-200 has the highest COMET score (0.696) and BLEURT score (0.000) across all the models in zero-shot setting. Suprisingly, TOWER's zero-shot BLEURT score is lower (-0.110), despite TOWER having the highest BLEU score (15.104) across the models in the zero-shot setting. This highlights the importance of using multiple automatic evaluation metrics, since each metric captures different aspects of translation quality.

BLEURT scores across models were generally low, with several negative values, especially for MarianMT and NLLB-200. This is expected, as BLEURT often produces negative scores when translations are semantically weak or only slightly aligned with the reference. However, BLEURT is designed to be used comparatively rather than absolutely, its real value lies in distinguishing between systems. In our study, 5-shot TOWER achieved the highest BLEURT score on the Dutch to English translation, suggesting better adequacy and fluency compared to the other models. The low BLEURT scores of MarianMT and Facebook-NLLB-200 indicate a poor alignment with the reference translation.

Interestingly, the TOWER 3-shot setting results in a significant drop in performance across all metrics compared to 1-shot and 5-shot. This dip might suggest an instability in the prompt sensitivity of TOWER. This highlights a notable challenge in few-shot learning: the performance can vary depending on the example selection and order. Overall, this confirms our hypothesis that few-shot prompting improves translation performance, with TOWER in the 5-shot setting yielding the best results. It also suggests that TOWER is more adaptable to contextual input than MarianMT or NLLB-200.

5.2 ENGLISH → DUTCH TRANSLATION RESULTS

When we look at Table 7 and the corresponding figures 4 and 6, we can see that the TOWER 5-shot setting performs the best overall for the English -> Dutch translation task. TOWER's 5-shot achieves the highest scores, with COMET score (0.756) shared with Facebook-NLLB-200, BLEURT score (-0.184), and BLEU score (16.735). The consistent increase in scores with more few-shot examples once again confirms the importance of in-context learning in the TOWER model. When we compare these results to the Dutch -> English translation task, we notice that 5-shot TOWER's COMET score and chrF score are slightly higher for the English -> Dutch translation task than the Dutch -> English translation task. However, the BLEURT score is significantly lower with the English -> Dutch translation compared to the Dutch -> English translation, and the BLEU score is also a bit lower for the English -> Dutch translation. Despite the BLEURT score being negative, it is still a lot higher than the other models BLEURT score, which might suggest improved adequacy and fluency.

Our baseline MarianMT performs relatively poor across all metrics, scoring a low BLEU and BLEURT score, although it achieves a decent chrF score just as with the Dutch -> English translation. This performance gap between chrF and BLEU reinforces that BLEU fails to capture lexical or morphological overlaps in morphological rich languages. When we compare MarianMT's performance to the Dutch -> English translation task, we see that the evaluation metrics scores generally increase for English -> Dutch translation. Only the BLEU score for English -> Dutch translation is lower than the BLEU score for Dutch -> English translation.

However, what stands out in this translation direction is the strong zero-shot performance of Facebook-NLLB-200. To our surprise, it outperforms all models in the zero-shot setting across evaluation metrics, with a COMET score of (0.756) which is equal to TOWER's 5-shot score, and a chrf score of (47.402). When we look at the results of Facebook-NLLB-200 in table 6 and compare them to table 7, we see that Facebook-NLLB-200 performs generally well across both translation directions, but is slightly better in translating from English to Dutch. This may suggest that the model has been trained with more English to Dutch data, or that it is more capable in generating Dutch text than English text. It could also indicate that because Facebook-NLLB-200 has a multilingual architecture, that it benefits more from English prompts.

TOWER scores consistent across all evaluation metrics. Its zero-shot COMET score (0.715) is strong, while the BLEU score (14.652) is very competitive. These results show the consistency of TOWER even while being used in a zero-shot setting. However, as with the Dutch to English translation results, the 3-shot setting of TOWER unexpectedly causes a crash in performance across all metrics. This gap in performance again might suggest the prompt sensitivity. TOWER again performs at its best in the 5-shot setting, indicating that it might benefit from more in-context examples. One slightly notable difference is that if we compare the results of the Dutch to English translation and English to Dutch translation, we can see that the improvement in TOWER's performance from the zero-shot setting to the 5-shot setting is larger in the Dutch -> English direction than in the English -> Dutch direction. This might indicate that TOWER benefits more from few-shot prompting when translating into English, possibly because its instruction-tuning and training data are more optimized for English as the target language. It may also suggest that generating fluent and adequate English sentences is easier for TOWER compared to generating Dutch sentences.

5.3 DISCUSSION

Overall, the results for the English -> Dutch translation task confirm the patterns seen in the Dutch -> English translation results. When we look at our first research question: "How different NMT models perform on Dutch-English translation", we found that TOWER consistently outperforms all other models in terms of translation quality. This can be observed in table 6 and 7, with more few-shot examples resulting in better results. These results also answer our second research question, regarding the impact of few-shot prompting. A 5-shot prompt leads to a big improvement in translation quality across all metrics for TOWER, displaying the effectiveness of in-context learning. On the contrary, MarianMT shows a relatively weaker performance, specifically for the Dutch to English translation. Interestingly, Facebook-NLLB-200 performs quite well in the English -> Dutch translation task, achieving the highest zero-shot scores across all evaluation metrics. These results show that even without instruction tuning or few-shot examples, multilingual models like Facebook-NLLB-200 can be surprisingly strong and competitive, and in some cases even outperform a model that has received few-shot examples. The performance difference between translation directions suggests that some multilingual models may be more effective translating into a dominant or high-resource target language. TOWER's strong performance across all results suggests that instruction-tuning and in-context learning make it more adaptable to different translation tasks.

6 CONCLUSION

This study aimed to evaluate the performance of different neural machine translation (NMT) models, MarianMT, Facebook-NLLB-200, and TOWER on the Dutch-English translation task, with a specific focus on the performance of these models in the zero-shot and few-shot settings. Our main research question was: "*How do different neural machine translation (NMT) models perform on Dutch-English translation, as measured by automatic evaluation metrics?*". We derived the following sub-question from the research question: "*What is the impact of zero-shot and few-shot learning on the translation quality of NMT models?*".

Our experiments on the Europarl corpus showed us that TOWER consistently achieved the best overall performance, with Facebook-NLLB-200 following closely behind. TOWER's performance improved significantly as more in-context examples were provided, showing the effectiveness of few-shot learning. Our study shows that models optimized for in-context learning profit the most from few-shot prompting. Facebook-NLLB-200, although not being instruction-tuned performed surprisingly well in the zero-shot setting, specifically when translating from English to Dutch. It outperformed TOWER in the zero-shot setting across all evaluation metrics and even came close with TOWER's 5-shot scores in some evaluation metrics. This suggests that multilingual models, even without explicit prompting, can generalize and perform well.

MarianMT, while being a solid baseline, was not in the same level as the other models in most metrics. However, it still generated a decent output in terms of surface-level similarity (chrF score), which also underlines the limitation of only relying on BLEU for translation evaluation. The results shown in this study show how the translation direction matters. TOWER benefited more from few-shot examples when translating into English, however Facebook-NLLB-200 showed stronger performance when translating into Dutch.

This study made use of the smaller versions of Facebook-NLLB-200 and TowerInstruct, since these transformers models are computationally expensive and require a GPU to run efficiently. We suggest using the the small models if you do not have the computational power to run the larger models, otherwise make use of the larger models. A larger model would probably generate better translation and results. This study also relied on the use of automatic evaluation metrics (BLEU, BLEURT, chrF, and COMET) and a single domain-specific dataset (Europarl). These automatic evaluation metrics all have their own issues, and we should not rely on them too much. Human evaluation could clear up the differences in fluency and adequacy between models in future research. Lastly, a limitation of this study lies within the Europarl dataset. Many translation models have been exposed to Europarl during pre-training, which could lead to overly optimistic performance results due to data overlap. Future research could consider testing on more diverse and unseen datasets, to better evaluate the generalization of the models.

A APPENDIX

A.1 PROMPTS USED FOR ZERO- AND FEW-SHOT TRANSLATION DUTCH→ENGLISH

A.1.1 Zero-shot prompt

```
"Translate the following from Dutch to English.\nDutch: {s}\nEnglish:"
```

A.1.2 1-shot prompt

```
"Dutch: de commissie moet onwettige steun en steun die de interne markt
    daadwerkelijk dwarsboomt, zien te traceren."
"English: the commission must track down the illegal aid and the aid which
    actually hinders the internal market."
"Translate the following from Dutch to English.\nDutch: {s}\nEnglish:"
```

A.1.3 3-shot prompt

```
"Dutch: de commissie moet onwettige steun en steun die de interne markt
    daadwerkelijk dwarsboomt, zien te traceren."
"English: the commission must track down the illegal aid and the aid which
    actually hinders the internal market."
"Dutch: in de huidige situatie sta ik daarom sceptisch tegenover het idee van
    een europese openbare aanklager, dat zeer waarschijnlijk niet kan worden
    uitgevoerd binnen het kader van de huidige verdragen."
"English: in the present situation, I am therefore sceptical about the idea of
     a European prosecutor, which it is scarcely possible to implement within
    the framework of the present treaties."
"Dutch: de heer nielson kan hier vandaag niet aanwezig zijn omdat hij in Zuid-
    Afrika is."
"English: Mr. Nielson cannot be present today since he is in South Africa."
"Translate the following from Dutch to English.\nDutch: {s}\nEnglish:"
```

A.1.4 5-shot prompt

```
"Dutch: de commissie moet onwettige steun en steun die de interne markt
    daadwerkelijk dwarsboomt, zien te traceren."
"English: the commission must track down the illegal aid and the aid which
    actually hinders the internal market."
"Dutch: in de huidige situatie sta ik daarom sceptisch tegenover het idee van
    een europese openbare aanklager, dat zeer waarschijnlijk niet kan worden
    uitgevoerd binnen het kader van de huidige verdragen."
"English: in the present situation, I am therefore sceptical about the idea of
     a European prosecutor, which it is scarcely possible to implement within
    the framework of the present treaties."
"Dutch: de heer nielson kan hier vandaag niet aanwezig zijn omdat hij in Zuid-
    Afrika is."
"English: Mr. Nielson cannot be present today since he is in South Africa."
```

```
 "Dutch:  hieruit volgt dat elk voorstel dat een ingrijpende hervorming van de toepassing van de concurrent ieregels beoogt , grondig moet worden getoetst ."
"English:  it therefore follows that any proposal which suggests major reform of the machinery for competition policy enforcement must be closely and carefully examined ."
"Dutch: mijnheer de voorzitter , beste collega ’ s , met uw goed vinden wil ik kort het woord nemen om een tweetal punten te belichten die ons in dit verslag op vallen en die van essentieel strategisch belang zijn voor onze visie op de toekomst van de europese unie ."
"English: mr president , i would like to say a few words in order to highlight two points made in these reports which are of fundamental strategic importance to the way we see the union ."
"Translate the following from Dutch to English.\nDutch: {s}\nEnglish:"
```

A.2 PROMPTS USED FOR ZERO- AND FEW-SHOT TRANSLATION ENGLISH→DUTCH

A.2.1 Zero-shot prompt

```
"Translate the following from English to Dutch.\nEnglish: {s}\nDutch:"
```

A.2.2 1-shot prompt

```
"English: the commission must track down the illegal aid and the aid which actually hinders the internal market."
"Dutch: de commissie moet onwettige steun en steun die de interne markt daadwerkelijk dwarsboomt, zien te traceren."
"Translate the following from English to Dutch.\nEnglish: {s}\nDutch:"
```

A.2.3 3-shot prompt

```
"English: the commission must track down the illegal aid and the aid which actually hinders the internal market."
"Dutch: de commissie moet onwettige steun en steun die de interne markt daadwerkelijk dwarsboomt, zien te traceren."
"English: in the present situation, I am therefore sceptical about the idea of a European prosecutor, which it is scarcely possible to implement within the framework of the present treaties."
"Dutch: in de huidige situatie sta ik daarom sceptisch tegenover het idee van een europese openbare aanklager, dat zeer waarschijnlijk niet kan worden uitgevoerd binnen het kader van de huidige verdragen."
"English: Mr. Nielson cannot be present today since he is in South Africa."
"Dutch: de heer nielson kan hier vandaag niet aanwezig zijn omdat hij in Zuid-Afrika is."
"Translate the following from English to Dutch.\nEnglish: {s}\nDutch:"
```

A.2.4 5-shot prompt

```
"English: the commission must track down the illegal aid and the aid which actually hinders the internal market."
"Dutch: de commissie moet onwettige steun en steun die de interne markt daadwerkelijk dwarsboomt, zien te traceren."
```

```
"English: in the present situation, I am therefore sceptical about the idea of a European prosecutor, which it is scarcely possible to implement within the framework of the present treaties."
"Dutch: in de huidige situatie sta ik daarom sceptisch tegenover het idee van een europese openbare aanklager, dat zeer waarschijnlijk niet kan worden uitgevoerd binnen het kader van de huidige verdragen."
"English: Mr. Nielson cannot be present today since he is in South Africa."
"Dutch: de heer nielson kan hier vandaag niet aanwezig zijn omdat hij in Zuid-Afrika is."
"English:  it therefore follows that any proposal which suggests major reform of the machinery for competition policy enforcement must be closely and carefully examined ."
 "Dutch:  hieruit volgt dat elk voorstel dat een ingrijpende hervorming van de toepassing van de concurrent ieregels beoogt , grondig moet worden getoetst ."
"English: mr president , i would like to say a few words in order to highlight two points made in these reports which are of fundamental strategic importance to the way we see the union ."
"Dutch: mijnheer de voorzitter , beste collega ’ s , met uw goed vinden wil ik kort het woord nemen om een tweetal punten te belichten die ons in dit verslag op vallen en die van essentieel strategisch belang zijn voor onze visie op de toekomst van de europese unie ."
"Translate the following from English to Dutch.\nEnglish: {s}\nDutch:"
```

A.3 SAMPLE OUTPUT OF ZERO-SHOT.PY AND ZERO-SHOT-ENGLISH.PY

A.3.1 Output sample zero-shot-english.py (English → Dutch)

```
"?? Sentence 2517
?? English:    i say they have a right to solidarity .
?? Reference:  ik zeg dat ze recht op solidariteit hebben .
?? Marian:   Ik zeg dat ze recht hebben op solidariteit.
?? NLLB:     Ik zeg dat ze het recht hebben op solidariteit .
?? Tower:    Ik zeg dat ze recht hebben op solidariteit.

?? Sentence 2518
?? English:    indeed the european parliament must make it its business to uphold this right in hours of need .
?? Reference:  het europees parlement moet er in deze noodsituatie voor zorgen dat ze die solidariteit ook krijgen .
?? Marian:   Het Europees Parlement moet er inderdaad toe overgaan dit recht in uren van nood te handhaven .
?? NLLB:     Het Europees Parlement moet zich er inderdaad voor zorgen dat dit recht wordt gehandhaafd in tijden van nood .
?? Tower:    , het Europees Parlement moet er inderdaad voor zorgen dat dit recht wordt geerbiedigd in tijden van nood.

?? Sentence 2519
?? English:    i urge the commission not to keep having to be asked and to provide assistance for the victims of the storms .
?? Reference:  ik wil dat de commissie de slachtoffers van het nood weer onmiddellijk helpt .
?? Marian:   Ik dring er bij de Commissie op aan om niet steeds gevraagd te worden en hulp te bieden aan de slachtoffers van de stormen.
?? NLLB:     Ik vraag de Commissie om niet langer gevraagd te worden en om hulp te verlenen aan de slachtoffers van de stormen .
?? Tower:    Ik dring er bij de Commissie op aan om niet steeds opnieuw om hulp voor de slachtoffers van de stormen te moeten vragen.

?? Sentence 2520
?? English:    it knows the ins and outs of aid provision better than any
    local organisation or authority .
?? Reference:  de commissie weet , beter dan welke plaatselijke organisatie of
     offici le instantie dan ook , hoe en via welke weg ze hulp kan bieden
    .
?? Marian:   zij kent de ins en outs van de steunverlening beter dan welke
    lokale organisatie of autoriteit dan ook.
?? NLLB:     Het kent de details van de steunverlening beter dan elke lokale
    organisatie of overheid .
?? Tower:    the EU kent de ins en outs van hulpverlening beter dan welke
    lokale organisatie of instantie dan ook.

?? Sentence 2521
?? English:    i call upon you , ladies and gentlemen , to support me in
    bringing home to the commission the fact that what most brussels offices
    are lacking is not so much the where with al for providing aid as the good
     will .
?? Reference:  dames en heren , ik vraag uw steun om de commissie duidelijk te
     maken dat ze wel degelijk hulp kan bieden , maar dat in sommige bureaus
    in brussel daartoe de wil ontbreekt .
?? Marian:   Ik roep u op , dames en heren , om mij te steunen bij het naar
    huis brengen van de commissie het feit dat wat de meeste brussels kantoren
     ontbreken is niet zozeer het waar met al voor het verstrekken van hulp
    als de goede wil .
?? NLLB:     Ik roep u op , dames en heren , om me te steunen bij het brengen
    van het feit aan de commissie dat wat de meeste Brusselse kantoren
    ontbreken niet zozeer de plaats is waar Al hulp verleent ... als de goede
    wil .
?? Tower:    Ik roep u, dames en heren, op mij te steunen om de Commissie
    duidelijk te maken dat de meeste Brusselse bureaus niet zozeer gebrek
    hebben aan de middelen om hulp te verlenen, maar aan de goede wil.

?? Sentence 2522
?? English:    permit me one further comment .
?? Reference:  ik wil nog het volgende opmerken .
?? Marian:   Ik wil nog     n     opmerking maken.
?? NLLB:     Laat me nog een opmerking maken .
?? Tower:    Toestaan mij nog een opmerking.
"
```

A.3.2 Output sample zero-shot.py (Dutch → English)

```
"?? Sentence 5
?? Dutch:    nu wil ik graag op verzoek van een aantal collega ’ s een minuut
    stilte in acht nemen ter nagedachtenis van de slachtoffers . ik doel
    hiermee met name op de slachtoffers van het nood weer dat verschillende
    lidstaten van de unie heeft geteisterd .
?? Reference:  in the meantime , i should like to observe a minute ’ s silence
     , as a number of members have requested , on behalf of all the victims
    concerned , particularly those of the terrible storms , in the various
    countries of the european union .
?? Marian:   Now I would like to observe a minute’s silence in memory of the
    victims at the request of a number of colleagues . I am referring in
    particular to the victims of the weather that has plagued several Member
    States of the Union .
?? NLLB:     I would like to address the issue of the situation in the Member
    States , particularly in the context of the disaster that has affected
    several Member States of the Union .
?? Tower:    I would now like to observe a minute’s silence, at the request of
     a number of colleagues, in memory of the victims, and I am referring in
    particular to the victims of the natural disasters that have affected
    several Member States of the Union.

?? Sentence 6
?? Dutch:    ik wil u vragen deze minuut stilte staande in acht te nemen .
?? Reference:  please rise , then , for this minute ’ s silence .
?? Marian:   I would ask you to observe this minute’s silence.
?? NLLB:     I ’d like to ask you to observe this minute of silence .
?? Tower:    to ask you to observe a minute’s silence.

?? Sentence 7
?? Dutch:    het parlement neemt staande een minuut stilte in acht
?? Reference:  the house rose and observed a minute ’ s silence
?? Marian:   Parliament observes a minute’s silence
?? NLLB:     Parliament observes a minute of silence while standing
?? Tower:    the parliament observes a minute’s silence.

?? Sentence 8
?? Dutch:    mevrouw de voorzitter , ik wil een motie van orde stellen .
?? Reference:  madam president , on a point of order .
?? Marian:   Madam President, I would like to raise a point of order.
?? NLLB:     Madam President , I would like to put forward a motion .
?? Tower:    angela merkel madam president, I would like to raise a point of
    order.

?? Sentence 9
?? Dutch:    u zult via de media hebben vernomen dat er zich in sri lanka een
    aantal bom explos ies en schiet partijen hebben voorgedaan .
?? Reference:  you will be aware from the press and television that there have
     been a number of bomb explos ions and killings in sri lanka .
?? Marian:   You will have heard through the media that there have been some
    bomb explosions and shootings in sri lanka.
?? NLLB:     You must have heard from the media that there have been a number
    of bombing and shooting parties in Sri Lanka .
?? Tower:    n you will have heard via the media that there have been a number
     of bomb explosions and shootings in Sri Lanka.

?? Sentence 10
?? Dutch:    een van de mensen die zeer recent in sri lanka is vermoord , is
    de heer ku mar pon nam bal am , die een paar maanden geleden nog een
    bezoek bracht aan het europees parlement .
?? Reference:  one of the people assass inated very recently in sri lanka was
    mr ku mar pon nam bal am , who had visited the european parliament just a
    few months ago .
?? Marian:   One of the people who was murdered very recently in Srilanka is
    Mr kumar ponnam bal am , who visited the European Parliament a few months
    ago .
?? NLLB:     One of the people who was murdered in Sri Lanka very recently is
    Mr Ku Mar Pon Nam Bal Am , who a few months ago made another visit to the
    European Parliament .
?? Tower:    One of the people who was murdered very recently in Sri Lanka is
    Mr Ku Mar Pon Nam Bal Am, who visited the European Parliament a few months
     ago.
"
```

BIBLIOGRAPHY

2006. Re-evaluating the role of bleu in machine translation research. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Ana Guerberof Arenas and Antonio Toral. 2022. Creativity in translation: machine translation as a constraint for literary texts. *ArXiv*, abs/2204.05655.

Mudasir Ashraf. 2024. Innovations and challenges in neural machine translation: A review. *International Journal of Science and Research (IJSR)*, 13:656.

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

M. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, K. Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, L. Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, C. Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alex Mourachko, C. Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841 – 846.

Gys-Walt van Egdom, Onno Kosters, and Christophe Declercq. 2023. The riddle of (literary) machine translation quality. *Tradumàtica tecnologies de la traducció*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu T. Hoang, Kenneth Heafield, Tom Neckermann, F. Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. In *Annual Meeting of the Association for Computational Linguistics*.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Conference on Machine Translation*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Chenyang Lyu, Jitao Xu, Longyue Wang, and Minghao Wu. 2023. A paradigm shift: The future of machine translation lies with large language models. In *International Conference on Language Resources and Evaluation*.

Wafaa Mohammed, Sweta Agrawal, M. Amin Farajian, Vera Cabarrão, Bryan Eikema, Ana C. Farinha, and José G. C. de Souza. 2024. Findings of the wmt 2024 shared task on chat translation. *ArXiv*, abs/2410.11624.

Kishore Papineni, Salim Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *WMT@EMNLP*.

Ricardo Rei, Craig Alan Stewart, Ana C. Farinha, and A. Lavie. 2020. Comet: A neural framework for mt evaluation. *ArXiv*, abs/2009.09025.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Annual Meeting of the Association for Computational Linguistics*.

J. Tiedemann and Santhosh Thottingal. 2020. Opus-mt – building open translation services for the world. In *European Association for Machine Translation Conferences/Workshops*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, J. Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wei Zhang. 2025. Applications of deep learning in natural language processing: A case study on machine translation. *Journal of Computer, Signal, and System Research*, 2:80–90.