Using PPbMQM to evaluate Dutch to English Machine Translation

Ino van de Wouw
i.vande.wouw@student.vu.nl

1 Introduction

This research explores machine translation between Dutch and English, with a particular focus on challenging linguistic phenomena that are unique to Dutch, including diminutives, modal particles, and idiomatic expressions. While machine translation has made significant strides in recent years, these language-specific features often pose substantial challenges for automated translation systems. Dutch presents particularly interesting translation challenges due to its productive morphology, complex compound formations, and extensive use of modal particles that often lack direct English equivalents. The study utilizes both authentic parallel corpora from the Dutch Parallel Corpus (DPC) and synthetically generated data created using the Llama 3.1 language model to ensure comprehensive coverage of these linguistic phenomena. The primary contributions of this work are threefold. First, it presents a novel approach to synthetic data generation using large language models specifically targeted at challenging linguistic phenomena. Second, it introduces a hybrid evaluation framework that combines traditional BLEU scoring with LLM-based quality estimation to provide more nuanced assessment of translation quality. Third, it provides detailed analysis of how specific Dutch linguistic features such as diminutives, compounds, and modal particles are handled in machine translation, offering insights into potential areas for improvement in translation systems. This research not only advances our understanding of Dutch-English machine translation but also presents methodological innovations in the use of large language models for both data generation and translation quality assessment. All of the code, images and corpora are accessible through my GitHub repository.

2 Related Work

Machine translation evaluation has been a critical area of research, with Papineni et al. (2002) BLEU metric serving as a foundational contribution to automated translation assessment. BLEU revolutionized MT evaluation by introducing an automated metric that correlates strongly with human judgments while being quick and language-independent. This study builds upon BLEU's core principle that good translations share more n-gram matches with reference translations, while extending the evaluation framework to address specific challenges in Dutch-English translation. While BLEU provides a robust baseline metric, our research complements it by incorporating LLM-based quality estimation following Wang and Arnoult's (2024) Prompt-Pattern based MQM (PPbMQM) approach, which enables more nuanced analysis of translation errors.

The combined use of BLEU and LLM-based evaluation is particularly relevant for assessing challenging linguistic phenomena in Dutch-English translation. Where BLEU excels at capturing general translation quality through n-gram matching, it may not fully capture the nuanced handling of language-specific features like diminutives, modal particles, and idiomatic expressions. That is why additionally chrF (Popovic, 2015) character-level n-gram evaluation will be utilized as it can better capture morphological variations and subtle linguistic differences. Our hybrid evaluation framework addresses this limitation by supplementing BLEU's statistical and chrF's character n-gram approach with targeted quality estimation for specific linguistic phenomena. Additionally, our use of synthetic data generation through large language models represents a novel approach to ensuring comprehensive coverage of challenging linguistic phenomena, addressing one of the key limitations identified in traditional parallel corpus-

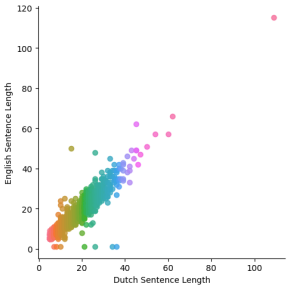

Figure 1: Dutch-English sentence length side by side

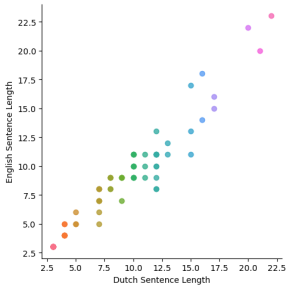

Figure 2: Dutch-English sentence length side by side for the augmented data

based approaches to machine translation evaluation.

3 Dataset Description

3.1 Dataset Analysis

In order to better understand the dataset, different properties were investigated.

3.1.1 Sentence length

The first is sentence length, as mentioned we are interested in translating diminutives and modal particles. Sentences containing these properties rely heavily on their context, it is therefore interesting to see if there are differences between the sentence pairs, and if so how big they are. The plot in figure 1 is the result of comparing the sentence length of both the Dutch and English sentences with each other, figure 2 incorporates this for the augmented data. These scatter plots have the length of Dutch sentences on its x-axis and English on its y-axis. These plots reveal a generally positive correlation between Dutch and English sentence length, indicating that longer Dutch sentences correspond to longer English sentences. There is some variation in the longer sentences, hopefully due to the presence of either diminutives or modal particles. Additionally, idiomatic differences contribute to this variation with some Dutch constructs needing more elaboration in English or vice versa. Understanding these differences is crucial for improving translation. Additionally concreteness level of the sentences on both sides of the translation could be compared to investigate how the presence of idioms change this.

3.1.2 Token endings

Diminutives are added to nouns in Dutch, to convey different nuances. To better understand their influence on our dataset we investigate the distribution of the different diminutive forms. Figure 3 shows the current distribution of diminu-

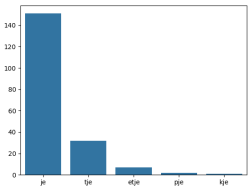

Figure 3: Diminutive distribution for the Dutch sentences

tives in the dataset. The suffix *-je* is by far the most commonly used, accounting for the majority of the data. This is due to its versatility and widespread application across Dutch. In contrast, other diminutive suffixes such as *-tje, -kje, -etje* and *-pje* occur much less frequently, indicating that their usage might be more context-specific depending on the preceding phonological components in the Dutch words. Another possibility is that it was thought to be unfit for the context of the original source texts, namely news paper articles. This uneven distribution suggests that the machine translation will excel in picking up *-je* forms compared to the other diminutives, but even the other forms contain the same *je* ending. It could therefore be hypothesized that the feature diminutive will even be recognized on novel and niche cases, as all diminutive forms contain that same *je* ending.

3.1.3 Sentiment influencing

Modal particles play an important role in Dutch by adding nuance or emphasis in a sentence. Figure 4 shows the specific distribution for our dataset, which shows that *maar, wel* and *even* are the most common modal particles. All have a significantly higher counts that *toch, misschien, gewoon, eens* and *nou*, whilst *soms* is hardly present at all. One way to investigate the impact this will have is

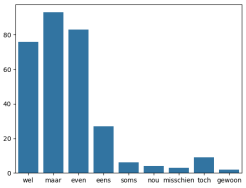

Figure 4: Modal particles distribution for Dutch sentences

to compare the sentiment of both the Dutch and English sentence, especially in those cases where modal particles are present. As a loss of nuance will be signalled by a deviating sentiment score, it might be interesting to include this rubric in the machine translation training stage.

4 Linguistic Analysis

4.1 Diminutives

Diminutives in Dutch are highly productive and formed by attaching suffixes such as *-je, -tje,-etje, -pje or -kje* to a noun. What suffix should be used fully depends on the phonological structure of the noun, which could be tricky to predict. Diminutives in Dutch are used to indicate size, affection or informality, but English often lacks direct translations and depends on adding adjectives or additional context:

huis-je	kijk dat hond-je!
house-DIM	look DEM dog-DIM
little house	look at that cute dog!

In order to evaluate machine understanding, Dutch sentences containing diminutives should be probed and translated to English, these translations will then be evaluated against an expected translation. A different way to evaluate machine understanding is probing English sentences and see how often it uses diminutives in its translation.

4.2 Compounding

Dutch is known for its highly productive, where multiple words can be combined into a single unit with a very specific meaning. Take into account the following examples:

keukentafelstoel	langetermijnplanning
kitchen-table-chair	long-term-planning
kitchen table chair	long-term planning

Compounds can be difficult to deal with for machine translation systems as English typically uses multi-word expressions and the meaning of a compound word might not be as straight forward as in our two examples. Human speakers often break compounds up into their components and translate those parts one by one to reconstruct meaning, that way *keukentafelstoel* becomes *a chair that stands at the table in the kitchen* in English. The way to test whether a machine translation model understand compounding is by showing it novel compounds and evaluate its performance on whether it manages to deduce meaning from that context, this could also be done by translating compound words to English and then back to Dutch to see whether the original form is preserved.

4.3 Modal particles

Modal particles like *maar, toch, wel, nou or even* are very common in Dutch, but just as we saw with diminutives, most have no direct translation to English. These modal particles are used to subtly convey emphasis or contrast but can also be used for negation. Speakers often translate these modal particles based on context by using multi word expressions, examples are:

Dat is toch mooi!	Dat kan je toch proberen?
DEM is MP nice!	DEM can you MP try?
That is nice, isn't it?	You can try that, can't you?

To test whether the model understands modal particles we can provide it with Dutch sentences that have modal particles in crucial positions that negate the meaning of a sentence and see whether this is included in the translation.

4.4 Polysemy

The Dutch language is full with words that have multiple meanings that fully depend on the context. The best example are *bank* and *steen*. Depending on the context it occurs in *bank* can indicate a *schooltable, monetary institution or sofa*,

where *steen* could refer *to a boulder, a tombstone, a stone or a gemstone*. The meaning is therefore derived from the context in which those polysemous words occur. In order to test model understanding it could be probed with polysemous words in sentence with and without distinctive context and evaluate how often those same words are used when translated back to Dutch.

4.5 Idioms

Idiomatic expressions in Dutch are culturally bound and often result in literal translations to English. *De appel valt niet ver van de boom - apple falls not far from the tree* means that the child resembles their parents, but is literally translated. There is no way to determine that the idiom has that specific meaning and for Dutch speakers the meaning is learned by heart. Translating such idioms to a different language relies on translation of the meaning or searching for idioms in the target language that hold a similar meaning. Idiomatic understanding can therefore be evaluated on whether the Machine Translation model manages to find fitting idioms in the target language or see whether it translates the sentence literally.

5 Prompting with LLM's

5.1 LLM language generation

In this translation task we created synthetic data by the use of Large Language Models (LLM's). These LLM's are valuable for this task because they are pre-trained on diverse data, equipping them with knowledge of linguistic phenomena such as diminutives, compounding, modal particles, polysemy and idioms. They enable efficient generation of synthetic sentences tailored to specific linguistic phenomena, significantly reducing manual effort. LLM's are also highly adaptable, allowing for custom prompts to focus on particular features such as diminutives or idioms. By leveraging LLM's, we can not only create targeted datasets but also evaluate machine translation models by comparing their outputs with LLM-generated translations or using the LLM as a probing tool to assess understanding.

A LLM, Llama 3.1, was used to generate sentence pairs that exhibit specific linguistic phenomena as this way we can ensure we have a balanced dataset and sufficient representation for these phenomena, Llama 3.1 was chosen over the newer version 3.2, as during an exploratory experiment Llama 3.1 created more accurate idiom translations. Most notably, both have not been optimized for Dutch, therefore there might be some mishandling of both the synthetic data as well as the Quality Estimation. It is important that the LLM is prompted in a correct and efficient way. Our prompting method involves the use of clear and explicit instructions that specify the linguistic phenomenon that has to be present, e.g. idioms, modal particles or diminutives, and in what format the LLM should present its output. To ensure the validity of the LLM outputs there were several requirements. Llama should only return the sentence pairs, it should generate at least 20 sentence pairs for each linguistic phenomena, the generated sentence should be between 10 and 15 words long, and the generated sentences should contain the requested phenomena. The first results were of decent quality, except for the sentences supposed to contain modal particles. The prompt to generate modal particles was altered from a zero-shot example to a few-shot example, containing the possible modal particles and one example for these particles. This improved the output of the LLM, but it contained excess information. In order to counter this the temperature for the LLM carrying out this prompt was lowered from 0.2 to 0.1. This resulted in sentence pairs that were manually deemed of sufficient quality. The final prompts can be found in Table 1. The distribution of the synthetic data follows the general trend of the data adapted from the Dutch Parallel Corpus, as can be seen in figure 2

Table 1: Synthetic Sentence Pairs Prompts and Quality Estimation Prompt

Idioms	Geef 20 Engelse zinnen en hun Nederlandse vertalingen. Elke Engelse zin moet tussen de 10 en 15 woorden lang zijn en een spreekwoord zijn. Jouw antwoord moet een lijst in een lijst zijn in Python. Het eerste element van de lijst moet de Nederlandse zin zijn, het tweede element de Engelse vertaling. Geef alleen de lijst terug! NIKS anders. Zie het volgende voorbeeld [["Nederlandse zin"], [Ëngelse vertaling], ...]

Continued on next page

Table 1: Synthetic Sentence Pairs Prompts and Quality Estimation Prompt (Continued)

Dim.	Geef 20 Engelse zinnen en hun Nederlandse vertalingen. Elke Nederlandse zin moet tussen de 10 en 15 woorden lang zijn en verkleinwoorden in zich hebben. Jouw antwoord moet een lijst in een lijst zijn in Python. Het eerste element van de lijst moet de Nederlandse zin zijn, het tweede element de Engelse vertaling. Geef alleen de lijst terug! NIKS anders. Zie het volgende voorbeeld [["Nederlandse zin"], [Ëngelse vertaling], ...]
Modal Particles	Geef 20 Engelse zinnen en hun Nederlandse vertalingen. Elke Nederlandse zin moet tussen de 10 en 15 woorden lang zijn en een modaal partikel in zich hebben. Modale partikels zijn: toch - Dat zou je toch moeten weten! wel - Ik heb het hem wel gezegd. nou/nu - Hou nou/nu op! maar - Doe maar wat je niet laten kunt. even - Kun je het raam even dichtdoen? eens - Bel hem eens op, hij zal nou wel thuis zijn. nu eenmaal - Het is nu eenmaal zo, daar is niks aan te doen. hoor - Dat is niet waar hoor. Jouw antwoord moet een lijst in een lijst zijn in Python. Het eerste element van de lijst moet de Nederlandse zin zijn, het tweede element de Engelse vertaling. Geef alleen de lijst terug! NIKS anders. Zie het volgende voorbeeld [["Nederlandse zin"], [Ëngelse vertaling], ...]
QE estimation	[System message] You are a professional Dutch-English translator.[Instruction message] You are a professional Dutch-English translator. Your task is to identify translation errors from a pair of the source Dutch sentence and the target English sentence. Please identify a maximum of 5 errors, assigning an error type for each with a severity scale from 1 (least sever) to 5 (most severe) using the MQM annotation scheme. Please consider the following criteria for identifying errors:Accuracy: when the target translation does not accurately represent the source Omission: when the target translation is missing content present in the source text, identify what was omitted.Fluency: issues with punctuation, spelling, grammar, register and inconsistency.Style: when the translation is grammatically correct but uses unnatural or awkward language.Terminology: inappropriate or inconsistent use of terms Locale convention: issues with formattingPlease provide the output in JSON format including the following keys for each error: ['error type'(value:Accuracy, Omission, Fluency, Style, Terminology, Locale convention), 'error span index'(start and end index of the marked sentence in the target sentence) ,'marked text'(the identified error in the target sentence,'severity'(1-5))] Do this for dutch_sent as source and translated_sentence as target. Not all sentence necessarily contain errors. [Requirement] output shall only consist of text in the prescribed JSON format

5.2 LLM evaluation

Llama was also used for Quality Estimation, within the context of machine translation that is done with the Multidimensional Quality Metrics (MQM) framework . This method can be used to identify quality issues in translation in a standardized manner. This is categorizing the errors by type, span and severity, the process started by adapting the prompt proposed by Wang and Arnoult (2024). This is a Prompt-Pattern based MQM (PPbMQM) metric that aims to supplement Quality Estimation such as BLEU by investigating what errors occur in sentences where the BLEU score is low. This proved to be a good start, additionally it was opted to leave out the few-shot explanations for the metrics and temperature was set to a relatively low number (0.1) in order to have deterministic results, as this encourages more factual and concise responses. Which worked well for 5 iterations, until it decided to add more unnecessary text. The few-shot explanations for the metric were added as well as during the process it became apparent that some error codes were wrongly assigned. A section requirement was added to the prompt which restated that the output should only be a JSON formatted string, this proved to be very helpful. Resulting in the following final version of the prompt can be found in table 1. It is important to note that this method of working, by using llama to evaluate all translated sentences raises some concerns. It is asked to find mistakes on sentences where there might be not mistake at all, which could result in hallucinations. Therefore, these PPbMQM Quality Estimations should be seen as a possible explanation for a translation with a low BLEU score, not as a leading metric.

6 Machine Translation Task

6.1 Experiments

Considering the machine translation task, we have to find an accurate way to evaluate how these translations were done. The translations themselves were made by using CTranslate2's NLLB-200-600M model (Klein et al., 2020). We check our translations by using an experimental Quality Estimation method, in which a LLM evaluates the source and translated sentence, and we use two robust metrics, BLEU and chrF, as baselines to check the quality of these Quality Estimations. The chrF metric is particularly valuable as it operates at the character level, making it more sensi-

tive to morphological variations and better suited for evaluating translations of languages with rich morphology like Dutch. While BLEU focuses on exact n-gram matches, chrF can capture partial matches and subtle linguistic differences that are especially relevant when evaluating translations of diminutives and compound words. Quality Estimation is included as it tries to encompass errors and mistakes that occurred during translation. By using an automated method to do this task it is hoped that less human intervention is necessary, but in order to investigate the current performance it is necessary to compare the Llama generated errors to quantitative metrics. This raises the first hypothesis h_1 When a LLM encounters a translation that has a high BLEU score it will manage to create incorrect errors. This will be done by manually checking the twenty highest rated translations on BLEU score and the errors a according to Llama corresponding to these sentences. This only checks whether the LLM is not hallucinating when asked to evaluate correct translations, but the Quality Estimations should also be investigated when the translations are deemed incorrect. This raises the second hypothesis h_2 When an LLM encounters a translation with a low BLEU score it will be able to correctly identify the error in the translation. This will evaluated similarly to hypothesis one, the twenty sentence pairs with the lowest BLEU score will be manually evaluated on the accuracy of the Quality Estimation, and whether the proposed errors are actually present in the source - translation pair. The two pervious hypotheses focus solely on the performance of the PPbMQM Quality Estimation, but it also interesting to look at the performance of the PPbMQM method when there are challenging linguistic phenomena present, this poses the following hypothesis h_3 When a Dutch source sentence contains one of the specified linguistic phenomena, the LLM's Quality Estimation will perform poorly.

6.2 Results

The analysis of the translation data revealed several interesting patterns. The performance analysis of the synthetic data revealed notable patterns in translation quality across different linguistic phenomena:

6.2.1 Poor Performance on Idioms

Of the twenty sentences with the lowest BLEU scores, fourteen contained idiomatic expressions.

Table 2: Corpus level scores

		DPC	Synthetic
BLEU	mean	32.09	31.39
	std	20.11	22.81
chrF	mean	59.17	55.43
	std	16.33	25.10

This suggests that the translation model struggles particularly with handling Dutch idioms. This challenge likely stems from the literal versus figurative translation problem inherent in idiomatic expressions.

6.2.2 Strong Performance on Modal Particles and Diminutives

The highest-performing translations (highest BLEU scores) had the following distribution: 10 sentences contained modal particles, 6 sentences contained diminutives and 4 sentences contained idioms. This indicates that the model handles modal particles and diminutives more effectively than idiomatic expressions.

6.2.3 Distribution Analysis

The BLEU-chrF score distribution shown in Figure 5 for the original corpus shows a wider spread of scores, and the synthetic corpus BLEU-chrF distribution (Figure 6) appears more concentrated, suggesting more consistent but potentially lower overall performance, but this is not reflected when you look at the corpus averages as can be seen in table 2. This difference in distributions might be attributed to the controlled nature of the synthetically generated data versus the more varied natural language in the original corpus.

6.2.4 Performance by Linguistic Phenomena

The modal particles show the most consistent high-quality translations, after which the diminutives demonstrate good translation quality but with more variation. Idioms show the most inconsistent performance with a tendency toward lower BLEU and chrF-scores. These results suggest that while the translation model performs adequately on straightforward linguistic features like modal particles and diminutives, it faces significant challenges when dealing with more complex phenomena like idiomatic expressions. This pattern is consistent across both the original and synthetic datasets, though more pronounced in the synthetic

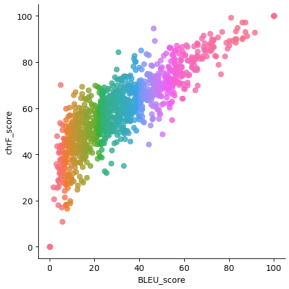

Figure 5: Corpus BLEU-chrF distribution

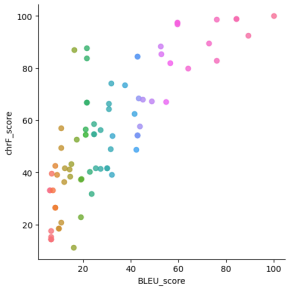

Figure 6: Synthetic Corpus BLEU-chrF distribution

data where these features were deliberately emphasized.

7 Discussion

In order to get a good look into the quality of the Llama evaluation we take a look at three of the sentence pair translations with the lowest BLEU score, as can be seen in Table 3 (in the appendix). The Quality Estimation hallucinated in five cases. For the first sentence pair, it raised an accuracy error on the first 3 words of the translated sentence whilst marking the whole sentence as wrong but only giving it a severity of 3 (out of 5), which is contradictory as if the whole sentence was wrongly translated this should have the highest severity possible. Additionally, it raised an terminology error for the sentence span 10 to 25, whilst the sentence does not contain that many words. For the second sentence pair an accuracy and fluency error are raised, the error span and marked text do not overlap and the fluency error is not the right error as the comma in that part is correct. The third sentence pair Llama managed to use the Dutch source text as marked text, whilst it should have been Engineers, the accuracy error that was raised was indeed correct.

When shown sentence pairs with near perfect translations, BLEU and chrF score of 100 out of 100, the Quality Estimation method still manages to find, or fabricate, errors as can be seen in table 4 (also in the appendix). For the first sentence pair it states that there is an accuracy error in the first five words, which is not true. Furthermore, a false terminology error is raised with a non existing span. The second sentence only received one error, which was erroneous. This proposed error had severity five and referred to *regulated* being an inadequate translation, whilst it is correct. The third sentence pair was flagged on both accuracy and fluency errors, which were both false. Even though the fluency error was wrong, it also had an incorrect error span.

As hypothesis 1 and 2 are now covered, what is left is investigating hypothesis 3. Table 5 in the appendix contains three examples of varying Quality Estimation performance, it managed to correctly pick out errors when sentences contained idioms but simultaneously made up errors. The best example is sentence 2, where it falsely identified a fluency error, as the first six words are correctly translated, and accurately identified a terminology error, as the translated sentence wrongfully translated deaf to dumb. Given that this sentence contained both an idiom and a modal particle the Quality Estimation was performed decent. This remains when you look at the first sentence pair in the same table, it partly identified that the translation was wrong by raising both an accuracy and translation error, but did not contain the right markings and span.

8 Conlusion

This research has provided valuable insights into the challenges and opportunities in Dutch-English machine translation, particularly regarding complex linguistic phenomena. The study's hybrid evaluation approach, combining traditional metrics with LLM-based quality estimation, has revealed several key findings. First, the translation performance varies significantly across different linguistic features. Modal particles and diminutives showed consistently strong translation quality, while idiomatic expressions proved more challenging, with 14 out of the 20 lowest-scoring translations containing idioms. This suggests that the machine translation system used handles structural linguistic features more effectively than culturally-bound expressions. The approach of using LLMs for both synthetic data generation and quality estimation showed promise but also revealed limitations. While Llama 3.1 suc-

cessfully generated contextually appropriate sentence pairs for specific linguistic phenomena, its quality estimation capabilities showed inconsistencies. The system sometimes hallucinated errors in high-quality translations (BLEU & chrF-score 100) while failing to accurately identify genuine translation issues in lower-scoring pairs. This suggests that while LLM-based quality estimation offers valuable insights, it should be used in conjunction with traditional metrics rather than as a stand-alone evaluation method. Future research should focus on improving translation quality for idiomatic expressions, perhaps through specialized training approaches or enhanced context modelling. Additionally, refinement of LLM-based quality estimation methods could address the current limitations in error identification accuracy. The methodology developed in this study, particularly the hybrid evaluation framework and synthetic data generation approach, provides a foundation for future work in machine translation evaluation, especially for language pairs with unique linguistic challenges.

References

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.

Lieve Macken, Orphée De Clercq, and Hans Paulussen. 2011. Dutch parallel corpus: A balanced copyright-cleared parallel corpus. *Meta*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu:a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

Sidi Wang and Sophie Arnoult. 2024. Large language models as annotators for machine translation quality estimation (under review). *ACL Anthology*.

9 Appendix

Data Statement The basis of the dataset used for this Machine Translation task lays in the Dutch Parallel Corpus (DPC). This corpus is freely available on the website of the instituut voor de Nederlandse taal. Curation rationale: The DPC is a corpus that features over 10 million parallel words for the language pairs Dutch-English and Dutch-French. The corpus contains pairs aligned on sentence level, but also contains Part of Speech tags and Lemma's, these were not taken into account whilst creating our dataset.
Language variety: For this task only the news articles that were paired on Dutch to English were taken into account. As the medium of the sentences in the corpus came from newspapers, it is rather difficult to say what the speaker demographic is.
Speaker Demographic: The corpus was collected from news articles, administrative texts, external communication, (non-)fictional texts and instructions. Therefore it is difficult to create one speaker demographic, as it contains a lot of different speakers.
Annotator Demographic: The translations provided were created by Macken et al. 2011 but they don't specify how they got the translations, therefore it is difficult to create an annotator demographic.
Synthetic data: next to the original 315 sentences that were collected from the DPC, and additional amount of 60 sentence pairs were created by using a Large Lange Model. To be more precise, Llama 3.1 with 8B parameters. The model was prompted in a manner that it only responded with the amount of asked sentence pairs containing a specific property, which after some persuading Llama managed to do for diminutives, modal particles and idioms. A total of 60 pairs added to the corpus, 20 containing diminutives, 20 containing modal particles and 20 idiomatic expressions.
The current version of the corpus is accessible through my personal GitHub repository.

Table 3: Evaluating Llama Evaluations on the lowest BLEU-scores

Dutch	Ze wonnen er de hoofdprijs mee van de KVIV-Ingenieursprijzen.		Degenen die sport en ontwikkelingssamenwerking aan elkaar koppelen, mikken op diverse werkingssferen.		Ingenieursprijs voor eindwerk rond schedelreconstructie	
Reference	Together, they won first place at the Engineering Awards organized by KVIV (Royal Flemish Society of Engineers).		Those who wish to couple sport together with development work aim at diverse spheres.		Engineering prize for thesis on cranial reconstruction	
Translated	They won the main prize of the KVIV-Engineersprize.		In order to connect sports and development cooperation with each other, we focus on diverse areas of work.		Engineers are awarded prizes for the final work on the re-construction of the skull.	
Llama Error	Accuracy	Terminology	Accuracy	Fluency	Accuracy	Fluency
Error Span	0-2	10-25	0-5	6-11	0-1	14-15
Marked Text	They won the main prize of the KVIV-Engineersprize.	KVIV-Engineeringsprize	connect sports and development cooperation with each other	we focus on diverse areas of work	Ingenieursprijs	re-construction
Severity	3	4	3	2	2	1
BLEU	1.777		2.954		3.125	
chrF	20.777		35.791		47.906	

Table 4: Evaluating Llama Evaluations on the highest BLEU-scores

Dutch	Hij is één van de redacteurs van het boek Sport and Development.		Sport is een gereguleerd conflict.	We gaan ervan uit dat een sterkere internationale positie van de Afrikaanse universiteiten positief is voor Afrika én Europa.		
Reference	He is one of the editors of the book Sport and Development.		Sport is a regulated conflict.	We assume that a stronger international position of African universities is positive for Africa and Europe.		
Translated	He is one of the editors of the book Sport and Development.		Sport is a regulated conflict.	We assume that a stronger international position of African universities is positive for Africa and Europe.		
Llama Error	Accuracy	Terminology	Accuracy	Accuracy	Fluency	
Error Span	0-4	5-17	2-3	0-1	5-6	
Marked Text	He is one of the editors	of the book Sport and Development.	regulated	We assume	and	
Severity	1	2	5	2	1	
BLEU	100		100	100		
chrF	100		100	100		

Table 5: Evaluation Llama Evaluations with linguistic phenomena

Dutch	Een lege maag is een slechte raadgever			Liefde is blind, maar gelukkig niet doof.			De kleine hond is erg lief voor zijn baasje thuis.	
Reference	An empty stomach is not a good advisor			Love is blind, but fortunately not deaf			The little dog is very sweet to his owner at home.	
Translated	A foolish man is a bad consultant.			Love is blind, but fortunately not dumb.			The little dog is very fond of his dog at home.	
Llama Error	Accuracy	Terminology		Fluency	Terminology		Accuracy	Terminology	Fluency
Error Span	0-1	3-8		0-5	6-11		0-3	4-8	9-12
Marked Text	foolish	consultant		blind	not dumb		The little dog	his dog	at home
Severity	2	4		2	3		2	3	1
BLEU	6.567			72.597			43.668		
chrF	14.366			89.544			57.738		