FAME-MT Dataset: Formality Awareness Made Easy for Machine Translation Purposes

Dawid Wiśniewski[1,2]**, Zofia Rostek**[1]**, Artur Nowakowski**[1,3]

[1] Laniqo, Poznań, Poland
[2] Faculty of Computing and Telecommunications, Poznań University of Technology, Poland
[3] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland
`name.surname@laniqo.com`

Abstract

People use language for various purposes. Apart from sharing information, individuals may use it to express emotions or to show respect for another person. In this paper, we focus on the formality level of machine-generated translations and present **FAME-MT** – a dataset consisting of 11.2 million translations between 15 European source languages and 8 European target languages classified to formal and informal classes according to target sentence formality. This dataset can be used to fine-tune machine translation models to ensure a given formality level for each European target language considered. We describe the dataset creation procedure, the analysis of the dataset's quality showing that **FAME-MT** is a reliable source of language register information, and we present a publicly available proof-of-concept machine translation model that uses the dataset to steer the formality level of the translation. Currently, it is the largest dataset of formality annotations, with examples expressed in 112 European language pairs. The dataset is published online[1].

1 Introduction

Motivation Different situations require using different language depending on whether it is a formal meeting or a casual talk with friends. Frequently, when speaking to an older or distinguished person, e.g., an owner of a company or a university professor, we use appropriate language forms to show respect. However, machine translation models may struggle with choosing an appropriate language form due to lack of context – often, we only have one sentence to translate, there are cultural differences between source and target language speakers, and models may be trained on parallel corpora, which do not focus on formality aspect of language enough.

For this reason, it is important to find ways of enforcing machine translation models to use a required formality level in translated sequences. While there are methods that can be used to achieve this goal, the existing datasets are scarce, focusing either on formality classification for a few selected languages (e.g., German or English) or providing formality-annotated examples of translations between pairs of languages, which are limited in size and include a narrow selection of target languages with English as the source language. What we found missing is a large-scale dataset providing formality-annotated examples of translations for a much broader set of languages, enabling easy fine-tuning of pre-trained machine translation models not only for directions including English but also other European languages.

Contribution In this paper, we present **FAME-MT**, the biggest to date dataset of translation pairs annotated with formality level. This dataset introduces 100,000 annotated translation examples for each of the 112 European language pairs considered. With 8 target languages (English, German, French, Italian, Dutch, Polish, Portuguese, and Spanish) classified according to formality and 15 European languages considered as the source lan-

[1]`https://github.com/laniqo-public/fame-mt/`

guages, the dataset introduces an 18-fold increase of the European language-pair coverage over the most diverse dataset **CoCoA-MT** (Nadejde et al., 2022) and over 100-fold increase of the dataset size considering the biggest formality-annotated datasets available (Rao and Tetreault, 2018).

Providing translation examples for a wide selection of source and target languages, **FAME-MT** allows for simple fine-tuning of machine translation models for the most popular European language pairs.

In this paper, we discuss the dataset creation process and analyze the data quality using various metrics to prove the usefulness of **FAME-MT**. We show how proof-of-concept models for formality-aware machine translation can be trained using Marian (Junczys-Dowmunt et al., 2018), provide examples of outputs of those models, and publish models online along with the dataset.

Research questions   We formulated the following research questions to be answered in this paper:

- **RQ1**: Is it possible to create a good quality large-scale dataset for formality-aware machine translation automatically based on available resources?

- **RQ2**: Is 100,000 translation examples for a given language pair enough to fine-tune a pretrained machine translation model to become formality-aware?

- **RQ3**: Considering translation pairs coming from **MTData** (Gowda et al., 2021), are formal sentences always translated into formal ones, and informal sentences translated into informal ones?

2   Related work

The idea of identifying formality level in texts is a widely analyzed area of (socio)linguistics (Biber and Conrad, 2019). In linguistics, there are 5 main language registers defined that can be used in particular situations. These are: frozen, formal, consultative, casual (informal), and intimate registers. Out of them, the formal and informal ones are most often analyzed in the context of machine translation.

Existing formality datasets   There are several datasets proposed that help incorporate formality awareness in various NLP tasks. The biggest one, Grammarly's Yahoo Answers Formality Corpus (or **GYAFC** corpus for short (Rao and Tetreault, 2018)), introduces 110,000 pairs of formal and informal sentences in English. The dataset is based on Yahoo Answers L6 corpus [2] and is proposed to be used for style transfer. The authors proposed an LSTM-based baseline showing that style transfer can be achieved using the machine translation approach and propose metrics for automatic evaluation of style-transfer models (e.g., fluency, meaning preservation).

Another dataset, consisting of 6,574 English sentences annotated with formality level, is proposed by Pavlick and Tetreault (2016). This work is aimed at the analysis of the formality level in English, considering humans' perceptions of formality in four different genres.

There are also datasets that do not focus on English, for example, a dataset consisting of 3,000 German sentences annotated by human experts on a continuous scale using comparative judgments (Eder et al., 2023). Each annotator was presented with several examples, and their goal was to rank the sentences according to formality level. This dataset covers a set of 12 diverse data sources (e.g., Twitter, Reddit, Wikipedia, or Springer Open Science articles).

Apart from the datasets that focus on a single language, there are several datasets with multilingual data. One of them is **XFORMAL** (Briakou et al., 2021). Similarly to **GYAFC**, the goal of **XFORMAL** is to provide a benchmark for style transfer by introducing pairs of informal sentences and their formal counterparts. The dataset provides examples in three languages: French, Italian, and Brazilian-Portuguese, and introduces 1,000 human-annotated examples for each of these languages. However, access to both **GYAFC** and **XFORMAL** is limited as it requires access approval as described on the website [3].

A publicly available dataset called **CoCoA-MT** (Nadejde et al., 2022) provides a set of triples consisting of English sentences translated into formal and informal versions for each of the supported languages, namely, German, Spanish, Italian, French, Dutch, Portuguese, Japanese, and Hindi. The dataset provides around 1,000 examples divided into train and test sets for each tar-

---

[2] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l
[3] https://github.com/Elbria/xformal-FoST/

get language. What distinguishes **CoCoA-MT** is the presence of phrase annotations that tell which phrases make a given example formal or informal. Even though **CoCoA-MT** provides translations between language pairs, they are always translations from English to one of the supported languages.

In general, although there are datasets focusing on formality, the large ones (**GYAFC**) cover only English or German, while those including other languages are relatively small (**XFORMAL**) and provide translations only from English to a given language (**CoCoA-MT**). No large-scale multi-language dataset exists up to date.

Formality control methods The tasks of style transfer and machine translation can be controlled using various methods. Here, we discuss the most relevant ones in the context of formality control.

The most straightforward method of controlling formality injects a special token in a source sequence, which tells us what level of formality should be achieved on the target side. This method was utilized in various approaches, e.g.: controlling honorifics in English to Japanese translation (Feely et al., 2019) using one of `{informal, polite, formal}` tags injected to the input of the Transformer model, controlling formality of French to English translations using one of predefined formality levels (Niu et al., 2017) `{low, neutral, high}`, controlling formality level by injecting a special token representing one of `{informal, formal}` classes to control output formality presented as part of **CoCoA-MT** evaluation (Nadejde et al., 2022), controlling politeness of the output text using predefined Latin tokens: `{vos, tu}` (Sennrich et al., 2016), or controlling formality by attaching a special token to both input and output sequences for better control of the output (Niu and Carpuat, 2020).

Alternatively, one can add a special embedding vector for each token (Schioppa et al., 2021) to represent the desired formality level, craft prompts (Garcia and Firat, 2022) for multilingual T5 (Xue et al., 2021) models, or use Bayesian factorization for constrained output generation (Yang and Klein, 2021).

3 Dataset

The **FAME-MT** dataset creation process was divided into three steps.

Step 1: Input data selection The aim of the **FAME-MT** project was to provide translations between pairs of languages that are annotated with formality information. For this purpose, the first step was to identify parallel corpora for languages of interest. We considered eight target languages, i.e., languages into which we want to translate input statements. These are: *English (EN), German (DE), French (FR), Italian (IT), Dutch (NL), Polish (PL), Portuguese (PT), and Spanish (ES)*. For each target language, our goal was to support a wide selection of source languages, i.e., those from which we translate into the target language.

We selected 15 source languages: *Czech (CS), Danish (DA), German (DE), English (EN), Spanish (ES), French (FR), Italian (IT), Dutch (NL), Norwegian* – including Norwegian Bokmål *(NO + NB), Polish (PL), Portuguese (PT), Russian (RU), Slovak (SK), Swedish (SV)*, and *Ukrainian (UK)*. This selection of source and target languages resulted in 112 potential language pairs, for each of which we needed a parallel corpus of translations between the given language pair. To collect such a dataset, we used **MTData** tool that provides an access to machine translation dataset collections, e.g., OPUS (Tiedemann, 2016), gives an access to popular datasets such as Europarl (Koehn, 2005), or Paracrawl (Bañón et al., 2020), and covers every language pair considered. For each language pair, we acquired a corpus of translation examples from **MTData** and applied the following postprocessing to increase the quality of the collected data: we rejected documents with more than 15% of characters being digits, having less than 5 characters in any sentence, having any token longer than 28 characters, having any sentence longer than 500 characters. We also rejected those with the number of tokens in any sentence higher than 100. Moreover, we applied FastText's LID-201 model (Burchell et al., 2023) to verify whether the source and target sentences are indeed expressed in the correct language (the expected language should have probability score of at least 10%).

For language pairs where English is present as a source or target language, we used Bicleaner AI package, which estimates the likelihood of a pair of sentences being mutual translations (Zaragoza-Bernabeu et al., 2022). We set the bicleaner-score threshold to 50%, rejecting everything below this score. Bicleaner cannot be applied to other languages as for now because it does not provide

open-source models for such language pairs.

Step 2: Formality classification The next step involved extracting two subsets for each language pair: those with formal and informal translations, respectively. For this reason, for each target language, we searched for formality classifiers or golden standard annotations that can be used to train a classifier. As discussed in Section 6, we only need to classify the target language according to its formality level. There were three scenarios to address:

- Formality classifier available: For English, there is a publicly available classifier available online [4], the quality of which is proven by an accompanying research paper (Babakov et al., 2023). As the classifier was pre-trained on the biggest dataset available (**GYAFC**) and produces the probability of a given example to be formal, we selected this classifier to process those pairs of sentences where the target language is English.

- No formality classifier, but golden standard dataset available: For German, French, Italian, Spanish, Portuguese, and Dutch, no formality classifier accompanied by a research paper could be found. For this reason, we decided to train a classifier using **CoCoA-MT** dataset, as it provides pairs of formal and informal forms of sentences for each of the aforementioned target languages.

- No formality classifier nor dataset available: For Polish, we did not even have a golden standard dataset to train a classifier on. For that language, we created a hold-out subset of sentences in Polish downloaded using **MTData**, and then the set was annotated by a group of six native speakers. This way, we collected examples of formal and informal sentences, which were of roughly the same size as golden standard examples provided in **CoCoA-MT** and were split into train and test sets to mimic the structure of **CoCoA-MT**.

We decided to include the Polish dataset in examples collected from **CoCoA-MT** and tried to fine-tune a single multilingual classifier to capture inter-lingual relations.

However, early experiments with models described later in this Section showed that fine-tuning a pre-trained model using **CoCoA-MT** combined with Polish leads to extreme probabilities assigned to most sentences collected from **MTData**. As the dataset contains only formal and informal examples, classifiers learn to treat every sentence as either formal or informal, while in many cases (especially sentences that are not related to an interlocutor) they are neither formal nor informal.

An analysis of annotations from **CoCoA-MT** showed that formality and informality are frequently expressed using appropriate personal pronouns. For this reason, we decided to generate an additional neutral dataset for each language generated as a random hold-out sample of target sentences from **MTData** that does not contain any phrase marked in **CoCoA-MT** as either formal or informal. For Polish, we continued the annotation task, asking natives to identify neutral examples in the hold-out dataset.

We fine-tuned several language models verifying how big a neutral sample size should be to maximize the scores. At each verification step, we changed only the size of the neutral set in the train set, while preserving a constant set of 600 neutral examples in the test set. We evaluated accuracy score for mDeberta-v3-base (He et al., 2021), XLM-RoBERTa-uncased (Conneau et al., 2019), and BERT-base-multilingual-uncased (Devlin et al., 2019). The results, summarized in Table 1 show that the best average accuracy score is achieved using mDeberta-v3-base with a neutral training sample size $= 500$. This comes in line with the trainset size of original **CoCoA-MT**, as it uses 400 formal and 400 informal examples for each language pair, so including a neutral sample of this size results in an approximately balanced training dataset.

Diving deeper into per-language scores, as presented in Table 2, we observe that all **CoCoA-MT** languages obtain very high accuracy scores. This may be due to the fact that frequently examples in **CoCoA-MT** distinguish formality based on personal pronouns. For Polish, the scores are lower as personal pronouns are often dropped and the form of a verb may be used to express a linguistic person implicitly. However, even for Polish, the scores are much higher than random guesses. For English, we report the scores for **GYAFC** obtained from a research paper that a given model accom-

NSS	mDeberta	XLM-RoBERTa	mBERT
100	0.9304	0.9301	0.9297
200	0.9475	0.9437	0.9395
300	0.9519	0.9465	0.9454
400	0.9531	0.9511	0.9455
500	**0.9552**	0.9523	0.9499
600	0.9551	0.9534	0.9492

Table 1: Neutral sample size (NSS) for training vs. average accuracy score (over all languages) of a given model. mDeberta stands for mDeberta-v3-base, XLM-RoBERTa stands for XLM-RoBERTa-uncased, and mBERT stands for BERT base multilingual uncased.

panies (Babakov et al., 2023).

Step 3: FAME-MT compilation The dataset compilation process was performed as follows: For each language src from source languages set and each target language tgt, we classified all targets among translations between src and tgt. When the target language was English, we used a model trained on **GYAFC** (Babakov et al., 2023). For other languages, we used mDeberta-v3-base, which we fine-tuned in the previous step. Since mDeberta-v3-base returns probability distribution over three possible classes: formal, informal, and neutral, we chose the class with the highest probability as the final model decision. However, since the model trained on **GYAFC** returns only the probabilities of formal and informal classes (that always sum up to one), we split the probability range into three equal parts, treating examples assigned with formal class probability in ranges: $< 0, \frac{1}{3} >, (\frac{1}{3}, \frac{2}{3} >, (\frac{2}{3}, 1 >$ as informal, neutral, and formal, respectively.

The classification process continued until we reached 50,000 informal and 50,000 formal examples for each language pair. Finally, for each language pair, we stored translations where the target sentence was considered formal or informal into separate files.

The dataset is published online[5] along with the scripts used for analysis, formality classifiers, and MT models. It has the following structure: Each target language has its own folder assigned, and inside these folders, there is a separate folder for each source and target language combination. In those folders, there is a pair of files: `informal.tsv` and `formal.tsv` each

---

[5] https://github.com/laniqo-public/fame-mt/

providing 50,000 translations between source and target languages, where the target was considered informal and formal, respectively.

4 Explorative analysis

To verify the quality of the dataset, we selected several approaches that would show key characteristics of the dataset.

Average length of sentences One of the basic metrics that can describe the difference between formal and informal language is the average sentence length in each of the categories. We may expect that formal sentences may be longer than informal, which are frequently used to share knowledge quickly. Figure 1 proves that for each target language, the formal translations are much longer than informal ones in the collected dataset.

Per-category key tokens To check which tokens characterize a given class, we aimed at understanding classifier decisions and feature importance. As the classifiers we used in the dataset creation process were too big to be handled by the popular explainable AI framework LIME (Ribeiro et al., 2016), we decided to fit a lightweight linear model to **FAME-MT** and interpret its features to understand the difference between classes.

Having a set of sentences classified as formal and informal for each target language, we applied the TF-IDF vectorizer to a class-balanced sample of target sentences to identify the 200,000 most important tokens among target texts. Then, we used those tokens as features for a logistic regression model with the intercept value set to 0. This way, we ensure that the coefficients assigned by the logistic regression model correlate with a given class. As we modeled formal sentences as a positive class and informal ones as negative, tokens that are assigned with high positive coefficient values are strongly related to the formal class, and those associated with the highly negative values are strongly related to the informal class. The coefficients selected during training were then sorted for each language, and the top 100 tokens and bottom 100 were selected as those describing formal and informal categories, respectively.

Table 3 presents the most important features for each category and each language considered. As can be seen, for languages with formal *you*, the most important words are personal pronouns (e.g., formal *Sie* vs. informal *du* in German or formal

Language	German	French	Italian	Dutch	Polish	Portuguese	English
Accuracy	0.9928	0.9926	0.9772	0.9962	0.7861	0.9789	0.9 (Babakov et al., 2023)

Table 2: Classification scores for each language using mDeberta-v3-base and neutral sample size = 500 in comparison to English classifier trained on **GYAFC**. The average score of mDeberta-v3-base is equal to 0.9552.
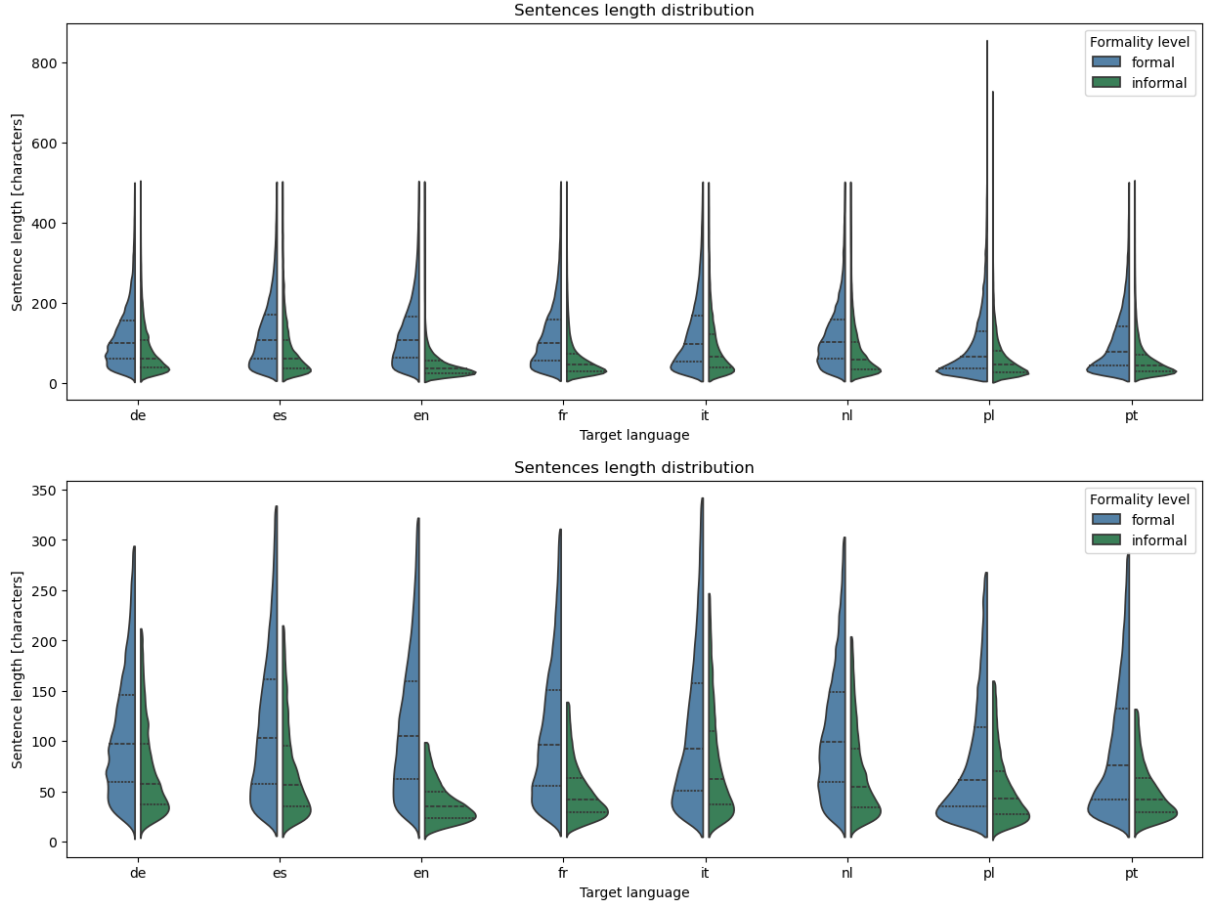

Figure 1: Violin plots representing the distributions of sentence lengths interpreted as the number of characters. The upper figure represents the distributions calculated over the original dataset. As it shows that there are some outliers with big values, we provide the lower figure generated over a subset of texts whose lengths are between Q1 - 1.5 IQR and Q3 + 1.5 IQR (Q1=first quartile, Q3=third quartile, IQR=inter-quartile range) to focus more on the most common scenarios.

usted and informal *tu* in Spanish). For languages, where personal pronouns may be omitted in informal scenarios, the grammar form of verbs may indicate informality (e.g., *jesteś* (informal *you are*) in Polish). For English, where there is no formal *you*, more sophisticated words are observed as the most important formal tokens, and contractions or swear words are found among the most characteristic informal tokens.

Textual complexity We used textacy[6] to measure the complexity of target sentences classified as formal and informal. The motivation for this step was the intuition that informal language may be easier to comprehend while formal texts should

be harder. For English, we used the automated readability index function, which calculates the relation between the number of characters in a given sequence related to the number of words and sentences and is calculated using the following formula: $4.71 \cdot count_{chars}/count_{words} + 0.5 \cdot count_{words}/count_{sents} - 21.43$. We calculated the score for all language pairs with English as a target and calculated the mean value. We obtained 11.54 for formal sentences and 4.28 for informal ones, which means that formal sentences are more complex than informal ones. Considering each language pair with the target set to English, formal translations are more complex every time. The biggest difference between formal and infor-

---

[6]https://pypi.org/project/textacy/

Language	formal words	informal words
German (DE)	sie, ihre, ihr, ihnen, ihres, …	du, deine, dir, dich, dein, …
English (EN)	distinctive, relations, moreover, obtain, refers, …	gotta, f*****g, gonna, ain, wanna, …
Spanish (ES)	su, sus, usted, le, está, …	tu, te, tus, estás, quieres, …
French (FR)	vous, votre, vos, pouvez, avez, …	tu, toi, te, ton, ta, …
Italian (IT)	sua, suo, suoi, lei, le, voi, …	ti, hai, tuo, tua, tuoi, sai, tu, …
Dutch (NL)	uw, kunt, wilt, bent, hebt …	je, jij, jullie, jouw, jou, …
Polish (PL)	pan, pani, wam, państwa, państwo, …	jesteś, ci, możesz, myślisz, musisz, …
Portuguese (PT)	você, sua, seu, suas, seus, …	te, teu, tua, tu, tens,…

Table 3: Words with the strongest relation to classes selected by analyzing feature coefficients in logistic regression model fitted on **FAME-MT**.

mal sentence scores is seen in the case of French-English (formal: 12.856, informal: 2.788) and the smallest in the case of Czech and English (formal: 11.175, informal: 7.005). We analyzed the Flesch score, which indicates the reading ease for sentences collected. We sampled 5,000 sentences for each source - target language pair, where target sentence was among languages supported by textacy (German, English, French, Italian, Spanish, Portuguese, Dutch). We present the full analysis in Table 10 provided in Appendix A. For each language pair, the informal texts are scored higher than formal ones, which indicates that informal texts are easier to read. This follows the general intuition. Averaging over all source languages, the average Flesch scores for formal and informal texts are (F and I represent formal and informal, respectively): German: 51.99 F and 72.57 I, English: 63.19 F & 90.89 I, Spanish: 68.93 F & 80.45 I, French: 78.33 F & 99.69 I, Italian: 77.65 F & 81.4 I, Dutch: 57.61 F & 76.32 I, Portuguese: 66.9 F & 76.32 I.

Source vs. target sentence formality To explore the relation between the formality level of source sentence and its translation in target language, we selected a subset of **FAME-MT** for which we had classifiers for both source and target languages. As these are eight languages, from all 112 language pairs, we selected 56 of them (8 times 8 possible language pairs minus 8 pairs where both source and target are the same). Then, for each target language considered, we analyzed formal translations and informal translations separately: for each formal translation, we used appropriate classifier to verify whether a given source sentence is also formal and for each informal translation, we used appropriate classifier to verify whether a given source sentence is also informal. Then, we calculated the percentage of cases where the agreement was observed. As presented

in Table 4, on average, 38.81% of examples for which target sentence was considered formal had the source sentence classified as formal. Analogously, 40.28% of examples classified as informal had the source sentence classified as informal. These results show that the formality level of the source does not determine the formality level of the target text. However, there is some correlation as the scores are higher than a random choice, which would achieve 33% (assigning a random class from: formal, informal, neutral). Diving into language pairs as presented in Table 11 in Appendix A and considering translations that are formal in target language, the strongest relation is between French sources and Dutch targets (82.8% of formal cases in Dutch are also formal in French), and the weakest between Polish sources and English targets (5.6% of formal cases in English are also formal in Polish). Considering translations that are informal in target language, the strongest relation is between Polish sources and French targets (73.4% of informal cases in French are also informal in Polish), and the weakest between German sources and English targets (5.6% of informal cases in English are also informal in German).

5 Application to Machine Translation

This section explores the application of the **FAME-MT** dataset to enable control over the formality tone of translated text. We conducted experiments on two translation directions: English-German and English-Polish, leveraging pre-trained models from the OPUS collection (Tiedemann, 2020) for fine-tuning.

Fine-tuning pre-trained MT models To facilitate the fine-tuning process, we augmented the existing model vocabulary with two new tokens: <FORMAL> and <INFORMAL>. Additionally, we expanded the existing model embeddings to accommodate these new vocabulary items by initial-

Language	Formal S+T	Informal S+T
German (DE)	42.82%	48.41%
English (EN)	17.71%	14.95%
Spanish (ES)	41.18%	40.82%
French (FR)	45.51%	54.4%
Italian (IT)	37.91%	37.36%
Dutch (NL)	55.56%	39.38%
Polish (PL)	37.16%	30.87%
Portuguese (PT)	35.67%	56.04%
Average	38.81%	40.28%

Table 4: Percentage of examples where formal source co-occurs with formal target (Formal S+T), and percentage of examples, where informal source co-occurs with informal target (Informal S+T). Formality of source and targets was determined using classifiers. The row *Average* represents the micro-average over all languages.

izing them with zeros.

To maintain consistency with the baseline models, we reused their Transformer-based (Vaswani et al., 2017) architecture and training hyper-parameters, except for batch size and learning rate. The English-German direction utilizes the Transformer-big architecture, while the English-Polish direction uses the Transformer-base architecture due to the lack of a larger model in the OPUS collection for this translation direction. Both fine-tuned models employed a small batch size of 5,000 target tokens and a learning rate of 1e-4. As the original models were trained using the Marian (Junczys-Dowmunt et al., 2018) framework, we utilized it for fine-tuning as well.

We pre-processed the **FAME-MT** dataset with pre-trained SentencePiece (Kudo and Richardson, 2018) tokenizers included in the model package, and then prepended <FORMAL> and <INFORMAL> tags to the formal and informal parts of the source input, respectively. To guarantee a diverse validation set during training, we randomly sampled 500 sentences from each formality category and added 500 random neutral samples not using personal pronouns, resulting in a total of 1500 validation samples per translation direction. The remaining data was used for training.

Automatic evaluation and contrastive examples Three decoding modes were evaluated for the fine-tuned models: standard, formal tone, and informal tone. Their performance was compared to the baseline OPUS model on the Flores (devtest) (Costa-jussà et al., 2022) and NTREX (Federmann et al., 2022) datasets, using BLEU (Pa-

pineni et al., 2002), chrF (Popović, 2015), and COMET[7] (Rei et al., 2020) metrics for evaluation. The chrF and BLEU results were computed with sacreBLEU[8][9] (Post, 2018). The evaluation results are presented in Tables 5 and 6 for Flores and NTREX datasets, respectively.

The results for English-German translation were encouraging. Fine-tuning on **FAME-MT** did not negatively impact overall translation quality, and automatic metrics even suggested a slight improvement. However, the English-Polish model exhibited a slight decrease in quality, potentially due to the smaller model size being more susceptible to the impact of embedding extension.

Tables 7 and 8 showcase various translations of the same sentence with different formality levels for each fine-tuned model. Interestingly, in languages like Polish, where gender neutrality is not present in singular pronouns, formal translations can introduce gender bias. For example, English *you* can be translated as either *Pan* (masculine) or *Pani* (feminine). In such cases, sentence-level machine translation might struggle to choose the correct form without additional context. As expected, formal translations generally tend to be more literal compared to their informal counterparts.

Released models To promote further research and facilitate accessibility, we have made the fine-tuned models publicly available as open-source resources [10].

This work demonstrates the potential of incorporating formality control datasets into machine translation pipelines. While further investigation is needed to refine the approach for different model sizes and language pairs, the results show promise for producing translations that accurately reflect the desired level of formality. The open-sourced models will enable researchers to build upon this work and explore applications in diverse scenarios.

6 Discussion

As we have shown in the previous section, the **FAME-MT** dataset can make pre-trained models formality-aware. Also, metrics used to analyze the dataset show that the characteristics of the dataset

---

[7] `wmt22-comet-da` COMET model was used

[8] BLEU signature: nrefs:1|case:mixed|eff:no|tok:13a |smooth:exp|version:2.3.1

[9] chrF signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:0 |space:no|version:2.3.1

[10] `https://github.com/laniqo-public/ fame-mt/`

Model		Flores					
		English → German			English → Polish		
		COMET	chrF	BLEU	COMET	chrF	BLEU
OPUS-Finetuned-50k	Standard	**0.8687**	66.14	39.98	0.8566	50.54	19.87
	Formal	0.8656	66.18	**40.02**	0.8523	50.47	19.70
	Informal	**0.8687**	65.85	39.60	0.8539	50.48	19.84
OPUS (baseline)	Standard	0.8675	**66.20**	39.90	**0.8624**	**50.85**	**20.15**

Table 5: Automatic evaluation results on the Flores dataset before and after fine-tuning the model on the FAME-MT dataset.

Model		NTREX					
		English → German			English → Polish		
		COMET	chrF	BLEU	COMET	chrF	BLEU
OPUS-Finetuned-50k	Standard	0.8295	**60.25**	**32.23**	0.8097	51.45	23.03
	Formal	0.8256	60.21	31.98	0.8049	51.32	22.83
	Informal	**0.8328**	60.15	31.95	0.8047	51.22	22.77
OPUS (baseline)	Standard	0.8251	59.91	31.64	**0.8125**	**51.89**	**23.41**

Table 6: Automatic evaluation results on the NTREX dataset before and after fine-tuning the model on the FAME-MT dataset.

English	Polish (Formal)	Polish (Informal)
You have to tip your cap.	Musi pan przechylić czapkę.	Musisz przechylić czapkę.
They think you're sad and will be pleased because they got to you.	Sądzą, że jest pan smutny i będą zadowoleni, bo do pana doszli.	Sądzą, że jesteś smutny i będą zadowoleni, bo cię dorwali.
They don't know you're furious.	Nie wiedzą, że jest pani wściekła.	Nie wiedzą, że jesteś wściekła.
But on Saturday, North Korean Foreign Minister Ri Yong-ho blamed US sanctions for the lack of progress since then.	Ale w sobotę, północnoko-reański minister spraw zagranicznych Ri Yong-ho obarczył sankcje USA za brak postępów od tego czasu.	Ale w sobotę, północnoko-reański minister spraw zagranicznych Ri Yong-ho obwiniał USA o sankcje za brak postępów od tego czasu.
Ring also settled a lawsuit with competing security company, the ADT Corporation.	Ring rozstrzygnął również sprawę z konkurencyjnym koncernem ochroniarskim, korporacją ADT.	Ring rozstrzygnął również sprawę z konkurencyjną firmą ochroniarską, firmą ADT Corporation.

Table 7: Examples of different machine translation results with formal/informal tone in English → Polish translation.

English	German (Formal)	German (Informal)
You just have to pay them the right amount of respect, he said..	Du musst ihnen nur den richtigen Respekt erweisen, sagte er.	Sie müssen ihnen nur die richtige Menge an Respekt zahlen, sagte er.
Sit down.	Setzen Sie sich.	Setz dich.
Would you like to go with me to the cinema next week?	Möchtest du nächste Woche mit mir ins Kino gehen?	Möchten Sie nächste Woche mit mir ins Kino gehen?
I'll be with you soon.	Ich werde bald bei dir sein.	Ich bin bald bei Ihnen.
I would like a word with your boss.	Ich hätte gerne ein Wort mit deinem Chef.	Ich möchte ein Wort mit Ihrem Chef.

Table 8: Examples of different machine translation results with formal/informal tone in English → German translation.

are consistent with intuition: formal documents are harder to read, they tend to be longer or use more formal words.

An interesting observation is that even though the **CoCoA-MT** dataset introduces a relatively small dataset of formal and informal examples per each language, the accuracy of the pre-trained mDeberta-v3-base model, which is fine-tuned on **CoCoA-MT** is very high when measured on **CoCoA-MT**'s test set with neutral samples added (0.9552). An analysis of the annotations provided in **CoCoA-MT** reveals that it focuses on expressing formality using personal pronouns. However, formality level, in general, can be expressed using various language constructs, e.g., contractions, formal greetings, slang words, and appropriate personal pronouns. On the other hand, e.g., **GYAFC** focuses more on constructs other than personal pronouns due to the lack of formal *you* in English. We think that this is the reason for the high scores observed for a small dataset (**CoCoA-MT**) and lower scores for a much bigger dataset (**GYAFC**). The focus of **CoCoA-MT** on pronouns leads to too optimistic quality estimates as compared to the evaluation provided for the English classifier.

However, since formal constructs co-occur with each other, classifiers trained using **CoCoA-MT** work well in practice. As we can see in Tables 7 and 8 – models trained with **FAME-MT** introduce subtle differences depending on the expected formality level (e.g., *hätte gerne* vs. *möchte* or *koncernem ochroniarskim* vs. *firmą ochroniarską*).

The motivation for using classifiers to classify only sentences in target languages is supported by the analysis of source vs. target sentence formality. As we have shown that the formality level of the source language does not determine the formality of the target language, in **FAME-MT**, we collect sentences of various levels of formality mapped to a given target sentence formality level. This way, we can inject a token representing the desired formality level (be it *formal* or *informal*) to steer the expected formality level. Thus, having only target sentences classified, we can fine-tune machine translation models to become formality-aware, regardless of the formality level of the source sequence.

7 Addressing research questions

In this paper, we stated three research questions that can be answered now. Regarding the first

one, we show that using a set of classifiers for the parallel corporas' target languages, we created a large dataset, which is useful for training formality-aware MT models. We proved it by fine-tuning general machine translation models and utilizing metrics confirming intuitions, thus, the answer to **RQ1** is positive. Also, the experiments show that fine-tuning pre-trained OPUS models with 50,000 formal and 50,000 informal examples is enough for fine-tuning English-German and English-Polish pairs of languages. The quality of the English-German model is higher but the quality of English-Polish is also satisfactory. For this reason, we can give the positive answer to **RQ2**. Finally, the classifiers used reveal that the formality level of the source language does not determine the formality level of the target language in the datasets collected using **MTData**. This observation justifies the need to inject a special formality token to ensure a given formality level of the translation. Thus, the answer to **RQ3** is negative.

8 Conclusions

In this paper, we introduce **FAME-MT** - a dataset consisting of 11.2 million translations between 112 European language pairs, where sentences in target languages are classified as formal or informal. As the dataset is a computer-generated silver standard, we used a set of metrics to prove the good quality of the data. Moreover, proof-of-concept models fine-tuned using formality data show that the dataset can be successfully utilized in problems requiring enforcing a given formality level of the system's output. Due to its size and large number of language pairs selected, **FAME-MT** is the largest and most diverse dataset available, introducing 18 times more European language pairs than biggest existing multilingual datasets (112 language pairs vs. 6 in **CoCoA-MT**) and providing over 100 times more sentences annotated with formality level than the biggest datasets (11.2 million vs. 110,000 in **GYAFC**). We made the dataset publicly accessible, and provided all the source codes for rerunning the analysis [11]. We hope that this dataset may help to produce better formality-aware machine translation models especially for pairs of languages that were not yet covered or underrepresented in existing datasets (e.g., Czech → French, Polish → German, or Danish → Dutch).

---

[11]`https://github.com/laniqo-public/fame-mt/`

References

Babakov, Nikolay, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't lose the message while paraphrasing: A study on content preserving style transfer. In Métais, Elisabeth, Farid Meziane, Vijayan Sugumaran, Warren Manning, and Stephan Reiff-Marganiec, editors, *Natural Language Processing and Information Systems*, pages 47–61, Cham. Springer Nature Switzerland.

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July. Association for Computational Linguistics.

Biber, Douglas and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Briakou, Eleftheria, Di Lu, Ke Zhang, and Joel R. Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In Toutanova, Kristina, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3199–3216. Association for Computational Linguistics.

Burchell, Laurie, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada, July. Association for Computational Linguistics.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Costa-jussà, Marta R, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Eder, Elisabeth, Ulrike Krieg-Holz, and Michael Wiegand. 2023. A question of style: A dataset for analyzing formality on different levels. In Vlachos, Andreas and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 568–581. Association for Computational Linguistics.

Federmann, Christian, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online, nov. Association for Computational Linguistics.

Feely, Weston, Eva Hasler, and Adrià de Gispert. 2019. Controlling japanese honorifics in english-to-japanese neural machine translation. In Nakazawa, Toshiaki, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Nobushige Doi, Yusuke Oda, Ondrej Bojar, Shantipriya Parida, Isao Goto, and Hidaya Mino, editors, *Proceedings of the 6th Workshop on Asian Translation, WAT@EMNLP-IJCNLP 2019, Hong Kong, China, November 4, 2019*, pages 45–53. Association for Computational Linguistics.

Garcia, Xavier and Orhan Firat. 2022. Using natural language prompts for machine translation. *CoRR*, abs/2202.11822.

Gowda, Thamme, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online, August. Association for Computational Linguistics.

He, Pengcheng, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13-15.

Kudo, Taku and John Richardson. 2018. Sentence-Piece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco, Eduardo and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Nadejde, Maria, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. Cocoa-mt: A dataset and benchmark for contrastive controlled MT with application to formality. In Carpuat, Marine, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 616–632. Association for Computational Linguistics.

Niu, Xing and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8568–8575. AAAI Press.

Niu, Xing, Marianna J. Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In Palmer, Martha, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2814–2819. Association for Computational Linguistics.

Papineni, Kishore, 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*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pavlick, Ellie and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Rao, Sudha and Joel R. Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In Walker, Marilyn A., Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 129–140. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Schioppa, Andrea, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6676–6696. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In Knight, Kevin, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 35–40. The Association for Computational Linguistics.

Tiedemann, Jörg. 2016. OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia, May 30–June 1. Baltic Journal of Modern Computing.

Tiedemann, Jörg. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online, November. Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and

R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In Toutanova, Kristina, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Yang, Kevin and Dan Klein. 2021. FUDGE: controlled text generation with future discriminators. In Toutanova, Kristina, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3511–3535. Association for Computational Linguistics.

Zaragoza-Bernabeu, Jaume, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France, June. European Language Resources Association.

A Detailed analysis of the FAME-MT dataset

To understand the **FAME-MT** dataset deeper, we include the extension of the explorative analysis by adding figures and tables that may help understand the dataset better.

Table 9 provides information about the size of each language pair subset in **FAME-MT**. For each language pair, 50% of examples (50,000) are examples with formal targets and 50% of examples with informal ones.

Table 10 extends the discussion on reading scores in terms of Flesch readability scores. Here, for each language pair, we see that informal translations are easier to read than formal ones.

Table 11 provides an in-depth analysis of the formality of source sentences in relation to the formality of target sentences extending information from Table 4. Here, we provide statistics for each language pair considering all kinds of disagreements separately, counting how many source sentences were classified as formal/informal/neutral when target sentence is formal or informal.

Finally, Figures 2 and 3 provide information about the punctuation distribution and word length distribution for formal and informal texts. While formal documents tend to use longer words, the distribution of punctuation marks is quite similar for both categories.

Source language \ Target language	German (DE)	English (EN)	Spanish (ES)	French (FR)	Italian (IT)	Dutch (NL)	Polish (PL)	Portuguese (PT)
Czech (CS)	✓	✓	✓	✓	✓	✓	✓	✓
Danish (DA)	✓	✓	✓	✓	✓	✓	✓	✓
German (DE)	NOT	✓	✓	✓	✓	✓	✓	✓
English (EN)	✓	NOT	✓	✓	✓	✓	✓	✓
Spanish (ES)	✓	✓	NOT	✓	✓	✓	✓	✓
French (FR)	✓	✓	✓	NOT	✓	✓	✓	✓
Italian (IT)	✓	✓	✓	✓	NOT	✓	✓	✓
Dutch (NL)	✓	✓	✓	✓	✓	NOT	✓	✓
Norwegian (NO+NB)	✓	✓	✓	✓	✓	✓	✓	✓
Polish (PL)	✓	✓	✓	✓	✓	✓	NOT	✓
Portuguese (PT)	✓	✓	✓	✓	✓	✓	✓	NOT
Russian (RU)	✓	✓	✓	✓	✓	✓	✓	✓
Slovak (SK)	✓	✓	✓	✓	✓	✓	✓	✓
Swedish (SV)	✓	✓	✓	✓	✓	✓	✓	✓
Ukrainian (UK)	✓	✓	✓	✓	✓	✓	✓	✓

Table 9: Language pairs in FAME-MT. The total size of FAME-MT = 11.2 million examples (14 source languages · 8 target languages · 100,000 examples). Each cell with ✓ represents a language pair with 100,000 examples (50% formal translations and 50% informal). Each cell with NOT represents a language pair that is not supported because source language equals target language.

Source language \ Target language	German (DE)	English (EN)	Spanish (ES)	French (FR)	Italian (IT)	Dutch (NL)	Portuguese (PT)
Czech (CS)	53.88/**77.84**	63.55/**83.98**	75.05/**88.91**	82.41/**103.41**	78.95/**88.3**	60.71/**84.91**	75.93/**92.6**
Danish (DA)	50.76/**72.02**	64.37/**86.56**	65.67/**78.72**	76.9/**100.84**	66.95/**80.14**	50.15/**73.15**	62.53/**89.26**
German (DE)	-/-	61.56/**83.35**	66.11/**76.03**	76.54/**95.52**	65.74/**75.74**	56.96/**72.34**	61.56/**85.69**
English (EN)	49.81/**63.57**	-/-	63.49/**72.15**	76.55/**95.33**	67.05/**75.14**	57.84/**72.29**	63.55/**83.46**
Spanish (ES)	49.85/**69.43**	59.98/**92.34**	-/-	76.91/**97.05**	72.54/**82.05**	57.41/**75.33**	63.89/**84.97**
French (FR)	50.4/**68.33**	58.58/**93.71**	66.47/**78.62**	-/-	68.77/**78.45**	57.6/**74.74**	66.96/**87.36**
Italian (IT)	49.9/**69.84**	61.37/**91.32**	66.9/**78.4**	76.55/**97.37**	-/-	56.66/**74.41**	64.4/**86.08**
Dutch (NL)	51.31/**71.75**	67.11/**94.6**	69.67/**81.12**	77.03/**101.03**	70.41/**82.1**	-/-	68.43/**90.56**
Norwegian (NO+NB)	51.88/**72.08**	68.05/**89.17**	69.1/**75.82**	75.55/**100.69**	74.42/**79.05**	54.4/**62.36**	66.94/**88.24**
Polish (PL)	53.12/**78.02**	65.36/**93.55**	71.68/**84.36**	80.53/**103.4**	76.89/**87.86**	60.76/**83.15**	74.61/**92.08**
Portuguese (PT)	52.12/**74.13**	62.01/**95.72**	70.07/**83.84**	79.54/**101.12**	72.92/**84.36**	60.29/**83.69**	-/-
Russian (RU)	54.71/**78.11**	60.37/**94.95**	69.43/**86.05**	82.92/**102.64**	79.77/**88.9**	62.07/**86.24**	72.61/**94.38**
Slovak (SK)	52.15/**73.61**	62.26/**90.41**	66.31/**81.35**	77.27/**99.28**	65.89/**78.68**	56.78/**76.96**	63.21/**88**
Swedish (SV)	52.4/**73.12**	65.26/**91.95**	69.56/**80.21**	77.06/**100.96**	68.14/**79.95**	57.67/**74.65**	67.58/**90.9**
Ukrainian (UK)	55.63/**74.16**	64.76/**90.88**	70.1/**80.79**	80.79/**96.98**	74.69/**78.82**	57.27/**74.29**	64.38/**84.51**
Average	51.99/**72.57**	63.19/**90.89**	68.93/**80.46**	78.33/**99.69**	71.65/**81.4**	57.61/**76.32**	66.9/**88.44**

Table 10: Flesch readability scores between each pair of languages (higher value → text easier to read). Each cell contains two numbers separated by a slash sign. The first number represents the Flesch score calculated for a random sample of 5000 formal sentences expressed in a given language pair. The second number represents an analogous score for informal target sentences. In each sample, for each language pair, informal texts are easier to read (higher scores marked in bold).

Source and target languages	Formal target			Informal target		
	Formal source	Neutral source	Informal source	Informal source	Neutral source	Formal source
English (EN) → German (DE)	35070 (**70.14%**)	12390 (**24.78%**)	2540 (**5.08%**)	5281 (**10.56%**)	14562 (**29.12%**)	30157 (**60.31%**)
Spanish (ES) → German (DE)	17555 (**35.11%**)	21012 (**42.02%**)	11433 (**22.87%**)	32253 (**64.51%**)	11288 (**22.58%**)	6459 (**12.92%**)
French (FR) → German (DE)	33929 (**67.86%**)	15242 (**30.48%**)	829 (**1.66%**)	18925 (**37.85%**)	9963 (**19.93%**)	21112 (**42.22%**)
Italian (IT) → German (DE)	5496 (**10.99%**)	31903 (**63.81%**)	12601 (**25.2%**)	30364 (**60.73%**)	16867 (**33.73%**)	2769 (**5.54%**)
Polish (PL) → German (DE)	6498 (**13.0%**)	25771 (**51.54%**)	17731 (**35.46%**)	35236 (**70.47%**)	12164 (**24.33%**)	2600 (**5.2%**)
Portugal (PT) → German (DE)	29499 (**59.0%**)	19213 (**38.43%**)	1288 (**2.58%**)	12310 (**24.62%**)	12770 (**25.54%**)	24920 (**49.84%**)
Dutch (NL) → German (DE)	21803 (**43.61%**)	15463 (**30.93%**)	12734 (**25.47%**)	35478 (**70.96%**)	10576 (**21.15%**)	3946 (**7.89%**)
German (DE) → English (EN)	8444 (**16.89%**)	39829 (**79.66%**)	1727 (**3.45%**)	2783 (**5.57%**)	42584 (**85.17%**)	4633 (**9.27%**)
Spanish (ES) → English (EN)	8489 (**16.98%**)	36997 (**73.99%**)	4514 (**9.03%**)	9080 (**18.16%**)	35870 (**71.74%**)	5050 (**10.1%**)
French (FR) → English (EN)	8760 (**17.52%**)	40154 (**80.31%**)	1086 (**2.17%**)	5765 (**11.53%**)	36896 (**73.79%**)	7339 (**14.68%**)
Italian (IT) → English (EN)	4204 (**8.41%**)	40544 (**81.09%**)	5252 (**10.5%**)	7748 (**15.5%**)	39355 (**78.71%**)	2897 (**5.79%**)
Polish (PL) → English (EN)	2821 (**5.64%**)	38964 (**77.93%**)	8215 (**16.43%**)	14810 (**29.62%**)	32751 (**65.5%**)	2439 (**4.88%**)
Portugal (PT) → English (EN)	13104 (**26.21%**)	35920 (**71.84%**)	976 (**1.95%**)	3234 (**6.47%**)	35328 (**70.66%**)	11438 (**22.88%**)
Dutch (NL) → English (EN)	5585 (**11.17%**)	37794 (**75.59%**)	6621 (**13.24%**)	8831 (**17.66%**)	39525 (**79.05%**)	1644 (**3.29%**)
German (DE) → Spanish (ES)	21142 (**42.28%**)	26750 (**53.5%**)	2108 (**4.22%**)	17301 (**34.6%**)	10018 (**20.04%**)	22681 (**45.36%**)
English (EN) → Spanish (ES)	36314 (**72.63%**)	10132 (**20.26%**)	3554 (**7.11%**)	8148 (**16.3%**)	14101 (**28.2%**)	27751 (**55.5%**)
French (FR) → Spanish (ES)	16087 (**32.17%**)	32477 (**64.95%**)	1436 (**2.87%**)	13417 (**26.83%**)	10920 (**21.84%**)	25663 (**51.33%**)
Italian (IT) → Spanish (ES)	15430 (**30.86%**)	27140 (**54.28%**)	7430 (**14.86%**)	27970 (**55.94%**)	18995 (**37.99%**)	3035 (**6.07%**)
Polish (PL) → Spanish (ES)	7099 (**14.2%**)	30481 (**60.96%**)	12420 (**24.84%**)	34870 (**69.74%**)	12181 (**24.36%**)	2949 (**5.9%**)
Portugal (PT) → Spanish (ES)	34107 (**68.21%**)	14891 (**29.78%**)	1002 (**2.0%**)	9874 (**19.75%**)	12240 (**24.48%**)	27886 (**55.77%**)
Dutch (NL) → Spanish (ES)	14564 (**29.13%**)	27806 (**55.61%**)	7630 (**15.26%**)	31427 (**62.85%**)	9942 (**19.88%**)	8631 (**17.26%**)
German (DE) → French (FR)	36371 (**72.74%**)	7966 (**15.93%**)	5663 (**11.33%**)	29805 (**59.61%**)	14752 (**29.5%**)	5443 (**10.89%**)
English (EN) → French (FR)	33675 (**67.35%**)	12696 (**25.39%**)	3629 (**7.26%**)	15148 (**30.3%**)	12471 (**24.94%**)	22381 (**44.76%**)
Spanish (ES) → French (FR)	17940 (**35.88%**)	15202 (**30.4%**)	16858 (**33.72%**)	30891 (**61.78%**)	13582 (**27.16%**)	5527 (**11.05%**)
Italian (IT) → French (FR)	5948 (**11.9%**)	27005 (**54.01%**)	17047 (**34.09%**)	29517 (**59.03%**)	16683 (**33.37%**)	3800 (**7.6%**)
Polish (PL) → French (FR)	8272 (**16.54%**)	18179 (**36.36%**)	23549 (**47.1%**)	36702 (**73.4%**)	10877 (**21.75%**)	2421 (**4.84%**)
Portugal (PT) → French (FR)	33897 (**67.79%**)	13684 (**27.37%**)	2419 (**4.84%**)	14252 (**28.5%**)	14079 (**28.16%**)	21669 (**43.34%**)
Dutch (NL) → French (FR)	23659 (**47.32%**)	7537 (**15.07%**)	18804 (**37.61%**)	34132 (**68.26%**)	14393 (**28.79%**)	1475 (**2.95%**)
German (DE) → Italian (IT)	17323 (**34.65%**)	31410 (**62.82%**)	1267 (**2.5%**)	15333 (**30.67%**)	8605 (**17.21%**)	26062 (**52.12%**)
English (EN) → Italian (IT)	34491 (**68.98%**)	10309 (**20.62%**)	5200 (**10.4%**)	8882 (**17.76%**)	12870 (**25.74%**)	28248 (**56.5%**)
Spanish (ES) → Italian (IT)	28522 (**57.04%**)	16438 (**32.88%**)	5040 (**10.08%**)	30705 (**61.41%**)	9555 (**19.11%**)	9740 (**19.48%**)
French (FR) → Italian (IT)	11666 (**23.33%**)	36215 (**72.43%**)	2119 (**4.24%**)	13321 (**26.64%**)	10083 (**20.17%**)	26596 (**53.19%**)
Polish (PL) → Italian (IT)	9932 (**19.86%**)	27943 (**55.89%**)	12125 (**24.25%**)	34979 (**69.96%**)	12180 (**24.36%**)	2841 (**5.68%**)
Portugal (PT) → Italian (IT)	31358 (**62.72%**)	17083 (**34.17%**)	1559 (**3.12%**)	10006 (**20.01%**)	11631 (**23.26%**)	28363 (**56.73%**)
Dutch (NL) → Italian (IT)	8003 (**16.01%**)	34786 (**69.57%**)	7211 (**14.42%**)	29321 (**58.64%**)	9900 (**19.8%**)	10779 (**21.56%**)
German (DE) → Polish (PL)	20330 (**40.66%**)	26744 (**53.49%**)	2926 (**5.85%**)	12171 (**24.34%**)	20941 (**41.88%**)	16888 (**33.78%**)
English (EN) → Polish (PL)	31044 (**62.09%**)	10617 (**21.23%**)	8339 (**16.68%**)	13598 (**27.2%**)	11868 (**23.74%**)	24534 (**49.07%**)
Spanish (ES) → Polish (PL)	16357 (**32.71%**)	28082 (**56.16%**)	5561 (**11.12%**)	18445 (**36.89%**)	18661 (**37.32%**)	12894 (**25.79%**)
French (FR) → Polish (PL)	19355 (**38.71%**)	27293 (**54.59%**)	3352 (**6.7%**)	13362 (**26.72%**)	21551 (**43.1%**)	15087 (**30.17%**)
Italian (IT) → Polish (PL)	11410 (**22.82%**)	32094 (**64.19%**)	6496 (**12.99%**)	20093 (**40.19%**)	26253 (**52.51%**)	3654 (**7.31%**)
Portugal (PT) → Polish (PL)	20408 (**40.82%**)	27196 (**54.39%**)	2396 (**4.79%**)	8348 (**16.7%**)	22168 (**44.34%**)	19484 (**38.97%**)
Dutch (NL) → Polish (PL)	11248 (**22.5%**)	28816 (**57.63%**)	9936 (**19.87%**)	22401 (**44.8%**)	22483 (**44.97%**)	5116 (**10.23%**)
German (DE) → Portugal (PT)	22322 (**44.64%**)	20169 (**40.34%**)	7509 (**15.02%**)	27060 (**54.12%**)	15757 (**31.51%**)	7183 (**14.37%**)
English (EN) → Portugal (PT)	33023 (**66.05%**)	11556 (**23.11%**)	5421 (**10.84%**)	14020 (**28.04%**)	12330 (**24.66%**)	23650 (**47.3%**)
Spanish (ES) → Portugal (PT)	21487 (**42.97%**)	14667 (**29.33%**)	13846 (**27.69%**)	31533 (**63.07%**)	14485 (**28.97%**)	3982 (**7.96%**)
French (FR) → Portugal (PT)	21263 (**42.53%**)	22366 (**44.73%**)	6371 (**12.74%**)	25245 (**50.49%**)	15808 (**31.62%**)	8947 (**17.89%**)
Italian (IT) → Portugal (PT)	10367 (**20.73%**)	24452 (**48.9%**)	15181 (**30.36%**)	29337 (**58.67%**)	17779 (**35.56%**)	2884 (**5.77%**)
Polish (PL) → Portugal (PT)	5696 (**11.39%**)	20486 (**40.97%**)	23818 (**47.64%**)	35852 (**71.7%**)	11792 (**23.58%**)	2356 (**4.71%**)
Dutch (NL) → Portugal (PT)	10915 (**21.83%**)	19712 (**39.42%**)	19373 (**38.75%**)	33186 (**66.37%**)	15465 (**30.93%**)	1349 (**2.7%**)
German (DE) → Dutch (NL)	40742 (**81.48%**)	7124 (**14.25%**)	2134 (**4.27%**)	16806 (**33.61%**)	11831 (**23.66%**)	21363 (**42.73%**)
English (EN) → Dutch (NL)	36558 (**73.12%**)	10737 (**21.47%**)	2705 (**5.41%**)	8698 (**17.4%**)	12496 (**24.99%**)	28806 (**57.61%**)
Spanish (ES) → Dutch (NL)	22760 (**45.52%**)	15556 (**31.11%**)	11684 (**23.37%**)	28258 (**56.52%**)	13826 (**27.65%**)	7916 (**15.83%**)
French (FR) → Dutch (NL)	41421 (**82.84%**)	7746 (**15.49%**)	833 (**1.67%**)	14766 (**29.53%**)	9928 (**19.86%**)	25306 (**50.61%**)
Italian (IT) → Dutch (NL)	6583 (**13.17%**)	28871 (**57.74%**)	14546 (**29.09%**)	25372 (**50.74%**)	20634 (**41.27%**)	3994 (**7.99%**)
Polish (PL) → Dutch (NL)	10729 (**21.46%**)	21618 (**43.24%**)	17653 (**35.31%**)	33024 (**66.05%**)	13191 (**26.38%**)	3785 (**7.57%**)
Portugal (PT) → Dutch (NL)	35708 (**71.42%**)	13140 (**26.28%**)	1152 (**2.3%**)	11072 (**22.14%**)	13441 (**26.88%**)	25487 (**50.97%**)

Table 11: Relations between source sentence formality and target sentence formality in **FAME-MT** determined using our classifiers. For each formal and informal target sentence, the classifier is used to determine the formality of the corresponding source sentence. Then, the number of source sentences classified as formal, informal, and neutral is reported for each target category.
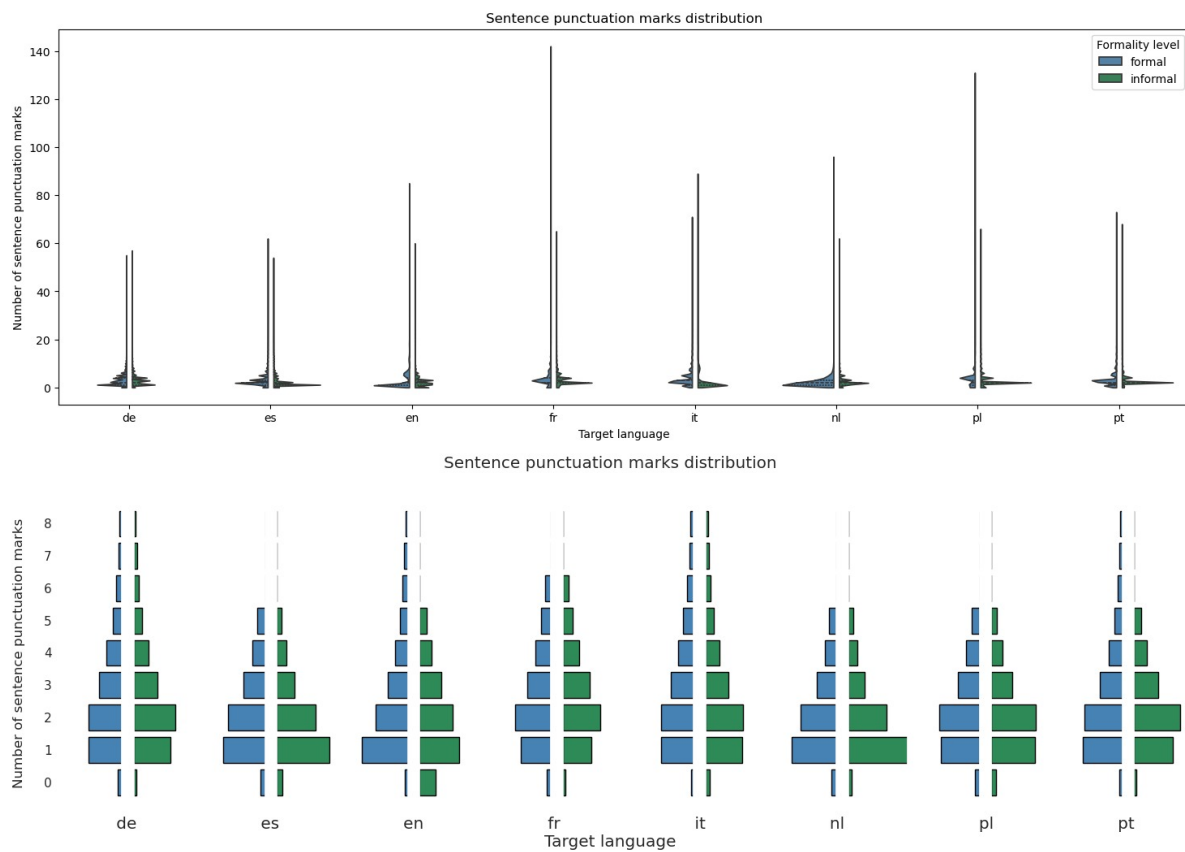
Figure 2: Plots representing the distributions of the number of punctuation signs in a sentence for a given language. The upper figure represents the distributions calculated over the original dataset. As it shows that there are some outliers, we provide the lower figure generated over a subset of texts whose lengths are between Q1 - 1.5 IQR and Q3 + 1.5 IQR (Q1=first quartile, Q3=third quartile, IQR=inter-quartile range) to focus more on the most common scenarios.
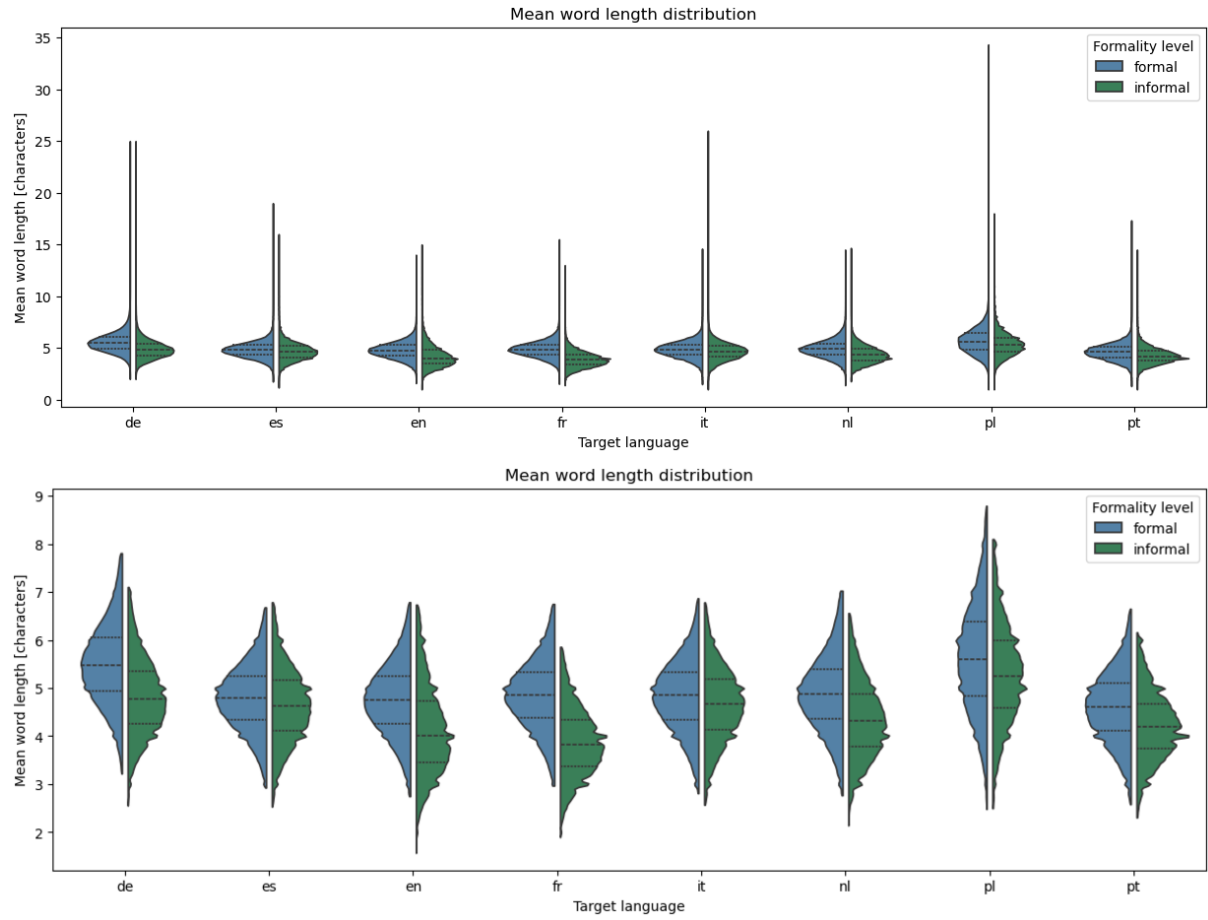
Figure 3: Plots representing the distributions of the mean word length in a given sentence per given language. The upper figure represents the distributions calculated over the original dataset. As it shows that there are some outliers, we provide the lower figure generated over a subset of texts whose lengths are between Q1 - 1.5 IQR and Q3 + 1.5 IQR (Q1=first quartile, Q3=third quartile, IQR=inter-quartile range) to focus more on the most common scenarios.