A Question of Style: A Dataset for Analyzing Formality on Different Levels

Elisabeth Eder and **Ulrike Krieg-Holz** and **Michael Wiegand**
Universität Klagenfurt, Klagenfurt, Austria
{elisabeth.eder | ulrike.krieg-holz | michael.wiegand}@aau.at

Abstract

Accounting for different degrees of formality is crucial for producing contextually appropriate language. To assist NLP applications concerned with this problem and formality analysis in general, we present the first dataset of sentences from a wide range of genres assessed on a continuous informal-formal scale via comparative judgments. It is the first corpus with a comprehensive perspective on German sentence-level formality overall. We compare machine learning models for formality scoring, a task we treat as a regression problem, on our dataset. Finally, we investigate the relation between sentence- and document-level formality and evaluate leveraging sentence-based annotations for assessing formality on documents.

1 Introduction

Textual style can be approached from various points of view. We focus on its inherent formality dimension stretching from informal to formal language use. See these two sentences, for example:

(1) We gave thorough thought to an adequate example.
Wir haben gründlich über ein adäquates Beispiel nachgedacht.

(2) racked our brains about a niiice example... :D
haben uns den kopf über ein schööönes beispiel zermartert... :D

While both sentences transport the same content, they differ in their degree of formality. (2) is less formal than (1). It may be suitable only for more informal discourse contexts and inappropriate in formal settings. Understanding these different nuances of formality is crucial for effective communication. Consequently, striking the right tone is relevant not only for humans but also for various NLP applications. May it be machine translation in need to transfer expressions of formality between different languages adequately (Niu and Carpuat, 2020; Anastasopoulos et al., 2022), chatbots aiming to produce contextually appropriate language to increase user satisfaction (Chaves et al., 2019; Elsholz et al., 2019), or writing assistance systems altering content to be more formal (Saberi et al., 2020). Hence, intra-lingual formality style transfer, which deals with generating a formal phrase given its informal version and vice versa, has recently also received increased attention (e.g., Shang et al., 2019 or Zhang et al., 2020).

Our paper addresses a prerequisite for this task: **assessing linguistic formality**. Rating the transferred style strength is necessary for evaluating formality style transfer models. Further, parallel corpora with formal and informal language pairs, often the basis for style transfer, are commonly built by automatically grading and extracting informal sentences first (Rao and Tetreault, 2018; Briakou et al., 2021b). For facilitating such formality assessments and analyzing linguistic formality in general, we make the following **contributions**:

1. We present the **first dataset of sentences from a wide range of genres with human formality assessments on a continuous informal-formal scale**. We ensure a comprehensive perspective on formality by collecting sentences from diverse domains. Formality annotations are obtained via a comparative annotation variant (annotators compare items to each other), which is not only more reliable than the rating scale method (Kiritchenko and Mohammad, 2017) but also satisfies the principle that a "continuum of formality" (Heylighen and Dewaele, 1999) exists rather than categorical distinctions. The dataset is the **first** to target **German** sentence-level formality unrestrictedly overall.

2. We **evaluate several machine learning models** for formality scoring on our dataset, which we treat as a **regression task**. Regression models have been found to be more suitable than classifiers for evaluating formality style transfer models since they grasp the broad spectrum of linguistic formality (Briakou et al., 2021a). Besides fine-tuning transformers on our dataset, we examine utilizing formality-informed corpora from different lan-

Findings of the Association for Computational Linguistics: EACL 2023, pages 580–593

guages with coarser or narrower representations of formality. Further, we employ feature-based approaches for formality scoring and analyze linguistic properties that constitute formality. For such analyses, we provide a tool with a variety of features for profiling characteristics of registers, genres, and author styles for various languages.

3. We **investigate the applicability of sentence-level formality annotations for the formality assessment of documents**. Lately, Jin et al. (2022) proposed extending formality style transfer, which so far exclusively focuses on the sentence level, to stylistically more complex documents. However, datasets targeting formality on this scope are rare and limited in size, probably because obtaining annotations is more expensive. Therefore, we analyze how sentence formality contributes to the formality of documents.

2 Related Work

With their continuous formality score based on frequencies of parts of speech, Heylighen and Dewaele (1999) established a milestone for the definition of formality. Lahiri et al. (2011) adapted this measure from the document to the sentence level. Most approaches targeting the lexical dimension of formality also regarded formality as a continuum (Brooke et al., 2010; Brooke and Hirst, 2014; Pavlick and Nenkova, 2015; Eder et al., 2021).

To the best of our knowledge, datasets comprising sentences with human formality assessments on a continuous informal-formal scale have not been constructed before. Pavlick and Tetreault (2016) built an English dataset collecting formality annotations on a 7-point Likert scale for sentences from only four sources (compared to the twelve in our dataset). They introduced formality detection as a regression task using features based on analyzing human perceptions of formality for a ridge regression model. Other datasets targeting sentence-level formality have binary labels since they primarily serve as parallel data for formality style transfer and contain formal and informal language pairs. They cover English (Rao and Tetreault, 2018; Cheng et al., 2020), Brazilian Portuguese, French and Italian (Briakou et al., 2021b), and Hindi, Bengali, Kannada and Telugu (Krishna et al., 2022).

Work on formality style transfer mainly used classification for measuring style strength and a handful of different classifiers (e.g., Lai et al. (2021) employed a CNN, Wang et al. (2019) an LSTM, and Krishna et al. (2020) transformers). Evaluating the style strength as a regression task, Rao and Tetreault (2018) borrowed the approach from Pavlick and Tetreault (2016), and Briakou et al. (2021b) relied on fine-tuning transformers.

For the German language, not yet considered for intra-lingual formality style transfer, two sentence collections with binary formality annotations based on formal and informal direct address exist (Faruqui and Padó, 2012; Nadejde et al., 2022). (Since these formality levels do not exist in English, they pose a problem for machine translation (Nadejde et al., 2022).) Hence, these datasets target a very constrained view of formality only.

Focusing on the document level, several works used traditional machine learning models for binary formality classification based on linguistic features. As training data, Abu Sheikha and Inkpen (2010) assumed binary labels for formality from the text genre, and Peterson et al. (2011) manually annotated emails from the English *ENRON* corpus (Klimt and Yang, 2004) with four formality classes. Treating formality assessment on documents as a regression task, Chhaya et al. (2018) employed linguistic features for formality scoring on *ENRON* emails, which have been rated on a 5-point Likert scale, whereas Eder et al. (2021) evaluated word formality scoring on emails from the German corpus *CodE Alltag* (Eder et al., 2020) based on continuous formality annotations. All these manually labeled document collections are small in size (∼1k) and built from a single domain only, i.e., emails. None of these works leverages formality-annotated sentences nor fine-tunes transformer models to assess the formality of documents.

3 Data

To build our dataset, we collected 3,000 German (DE) sentences from different domains and let crowdworkers assess their formality on a continuous formality scale via comparative annotations.

3.1 Collecting Sentences

We chose twelve different text sources, which we assumed to be related to diverse levels of formality, to cover the broad spectrum of linguistic formality best possible. From each source, we took 250 sentences. We picked these sentences randomly, but they had to consist of at least one word. Additionally, we attempted to enhance language variety by selecting a minimum number of sentences per au-

thor. We also tried spreading the data over different topics whenever such information was available.[1]

We utilized the following sources:

Tweets. We rehydrated tweets from a German *Twitter* snapshot (Scheffler, 2014).

Reddit. We extracted posts from the *GeRedE* corpus, which contains German communication on *Reddit* (Blombach et al., 2020).

Subtitles. To account for spoken language, we included German sentences from the *OpenSubtitles* collection of parallel corpora with movie and TV subtitles (Lison and Tiedemann, 2016).

Comments. 250 sentences were collected from the *One Million Posts Corpus*, which comprises comments on news articles (Schabus et al., 2017).

Emails. We took sentences from *CodE Alltag*, a corpus with German emails (Eder et al., 2020).

Blogs. Using the *DWDS* platform (Geyken et al., 2017), we obtained sentences from a blog corpus (Barbaresi and Würzner, 2014).

Fiction. Due to the lack of accessible corpora covering contemporary fictional texts, we reverted to an archive that, besides fan fiction, contains original work from nonprofessional writers.[2] We extracted 250 sentences from their short stories.

News. We gathered sentences from the German news corpus from 2020 provided in the *Leipzig Corpora Collection* (Goldhahn et al., 2012).

Wikipedia. From the *Leipzig Corpora Collection*, we also used sentences from the German *Wikipedia* corpus from 2021.

Political. For potentially more formal spoken language examples, we extracted sentences from German political speeches that are included in the parallel corpus *EuroParl* (Koehn, 2005).

Legal. We gained sentences from the legal domain by utilizing a dataset with German court decisions (Leitner et al., 2019).

Science. We used *Springer Link*[3] to manually collect sentences from scientific journals, proceedings, and books published between 2000 and 2022 under open access.

3.2 Human Assessment

We gathered human formality assessments for the resulting 3,000 sentences using Best-worst scaling (BWS) (Louviere et al., 2015), a form of comparative annotation. BWS delivers more reliable annotations than the rating scale method mitigating issues such as a scale region bias or inconsistent annotations (Kiritchenko and Mohammad, 2017). Further, it complies with the notion of formality as a continuum (Heylighen and Dewaele, 1999).

For BWS, annotators are presented with n items at a time (typically $n = 4$). They have to decide which item from the n-tuple is the *best* and which is the *worst* (i.e., the highest and the lowest regarding the property of interest). To get real-valued scores from these BWS annotations, the percentage of times the term is chosen as worst is subtracted from the percentage of times the term is chosen as best (Counts Analysis (Orme, 2009)). Thus, each item receives a score between $+1$ (most formal) and -1 (most informal).

We randomly generated $2N$ 4-tuples (where N denotes the number of sentences) under the premise that each term occurs only once in eight different tuples and each tuple is unique.[4] For the annotation process proper, we chose crowdsourcing to ensure the heterogeneity of annotators. Using the crowdsourcing platform *Clickworker*[5], German native speakers assessed each of the 6,000 tuples five times. Thus, we collected 30,000 annotations from 1,084 different annotators, with an average of 27.7 annotations per annotator.

All five annotators agreed in 19% of the annotations. In two-thirds, three or four annotators chose the same item, while only in 15% just two of the answers matched. The higher the difference between the real-valued formality scores of two sentences, the higher the agreement of the crowdworkers. For a score difference of just 0.1, the agreement is 64%. It rises to over 70% for higher score differences, with over 80% for differences higher than 0.4 and at least 90% for differences over 0.7.

We computed the split-half reliability[4] for our formality-assessed dataset by randomly splitting the annotations of a tuple into two halves, calculating scores independently for these halves, and measuring the correlation between the resulting two sets of scores. We got an average Spearman's ρ of 0.919 ($\pm$0.002) over 100 trials, which indicates a high reliability of the annotations.

3.3 The Final Dataset

Figure 1 displays the distribution of human-assessed formality scores for each of the twelve

[1] For some corpora, we subsumed subreddits, blogs, genres, or articles, to which comments refer, in place of topic.

[2] https://www.fanfiktion.de/

[3] https://link.springer.com/

[4] We employed scripts developed for emotion scaling by Kiritchenko and Mohammad (2016, 2017).

[5] https://www.clickworker.de

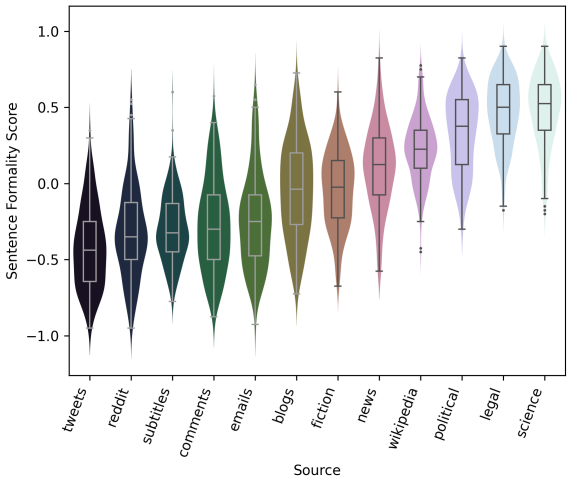

Figure 1: Distribution of formality scores for our 3,000 sentences per each of the twelve sources ordered by the average formality score of the source.

sources of the 3,000 sentences in our dataset. As expected, sentences from online communication or sources with more spontaneous language use, e.g., *tweets* or *comments*, tend to be linked to lower scores, while sentences with more elaborated language use, e.g., *legal* or *scientific* texts, have higher scores. However, sources scatter broadly, and assuming the same degree of formality per genre seems inappropriate.

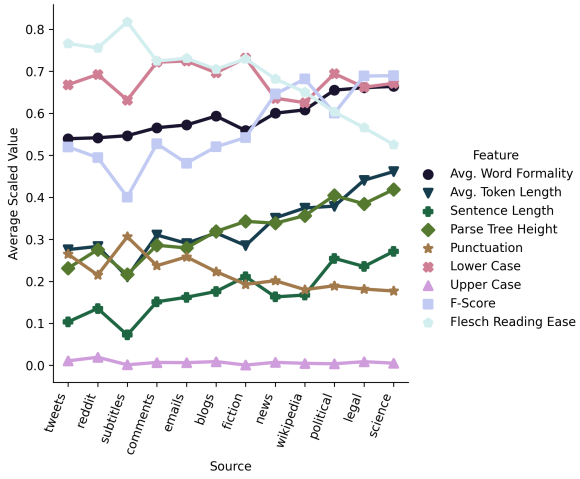

Figure 2: Averages of simple linguistic characteristics (scaled to a range between 0 and 1) of sentences for each source; sources ordered by their mean formality.

In Figure 2, we plot some simple linguistic features, which have been studied in relation to formality (Heylighen and Dewaele, 1999; Pavlick and Tetreault, 2016, i.a.) for each source. The mean word formality, token length, sentence length and parse tree height rise for sources with higher average sentence formality scores. The proportion of punctuation characters tends to sink, whereas the ratios of upper- or lower-case tokens are more stable. Heylighen and Dewaele's (1999) F-score indicates a higher formality and the readability score Flesch Reading Ease (Flesch, 1948) signals a lower readability for sources with higher mean formality.

In the following, we explore such properties for scoring the formality of the individual sentences.

4 Formality Scoring on Sentences

We compared different models for predicting formality scores for sentences on our dataset.

4.1 Within-Dataset Experiments

Transformers. We experimented with fine-tuning transformer models on our dataset. For that, we employed GBERT-base (Chan et al., 2020), a German BERT language model.[6] For all transformer-based experiments, we used the NLP library *FLAIR* (Akbik et al., 2019) as a framework.

Feature-based Models. We evaluated two feature-based models, which allowed us to examine the influence of linguistic characteristics more directly. The first ridge regression model employs eleven different feature groups and was developed for scoring the formality of English sentences (Pavlick and Tetreault, 2016). The second was created for English documents, more precisely emails (Chhaya et al., 2018). It borrows features from the first model and extends them with affect-based features. We adapted these feature sets to German and adjusted them to work on sentences and documents. We also employed a ridge regression model. Table 5 in the Appendix contains a detailed breakdown of the features we implemented.

4.2 Cross-Dataset Experiments

Learning from Other Languages. We examined using English sentences with formality scores determined via averaging over individual annotations on a 7-point Likert scale (Pavlick and Tetreault, 2016). This dataset (***PT16*** in the following) contains about 11k sentences from four sources: news and blogs from Lahiri (2015) extended by emails and Q&A sites. We evaluated three different settings. We fine-tuned GBERT-base transformers on *PT16* translated to German and tested them on

[6]Other German transformers (Chan et al., 2020; Minixhofer et al., 2022) either yielded no significant difference or performed worse (see Table 3 in the Appendix).

Training		Testing		Model	Spearman's ρ
sentences	ours (de)	*sentences*	ours (de)	GBERT	**0.919** ($\pm$0.009)
	ours (de)		ours (de)	feature-based ($\sim$ Pavlick and Tetreault, 2016)	0.857* ($\pm$0.007)
	ours (de)		ours (de)	feature-based ($\sim$ Chhaya et al., 2018)	0.830* ($\pm$0.018)
	PT16 (de)		ours (de)	GBERT	0.877* ($\pm$0.018)
	PT16 (en)		ours (de)	XLM-RoBERTa	0.847* ($\pm$0.017)
	PT16 (en)		ours (en)	BERT	0.844* ($\pm$0.022)
	XFORMAL (br-pt+fr+it)		ours (de)	XLM-RoBERTa	0.768* ($\pm$0.020)
	GYAFC (en)		ours (de)	XLM-RoBERTa	0.716* ($\pm$0.023)
	FP12 (de)		ours (de)	GBERT	0.595* ($\pm$0.042)

Table 1: Evaluation of different models for formality scoring on our sentences; '*' stands for a statistically significant difference of $p < 0.005$ with respect to **best** model (using two-sided Wilcoxon signed-rank test on Spearman's ρ); language(s) of datasets in brackets, translated data underlined.

our dataset. Consequently, we utilized BERT-base (Devlin et al., 2019) for fine-tuning on the original English *PT16* and testing on the English translation of our dataset. Further, we fine-tuned multilingual XLM-RoBERTa-base transformers (Conneau et al., 2020) on the English *PT16* and tested them on our German sentences. For the translations in both directions, we employed the models from Edunov et al. (2018) via the *fairseq* toolkit (Ott et al., 2019).

Formality Classifiers. Since there are huge datasets with binary formality annotations, we evaluated binary formality classifiers leveraging these data. We used the probability of the class determined by the classifiers as a prediction of a formality score. For the informal class, we took the probability as a negative number, thus ending up with scores from -1 to $+1$. Lacking more comprehensive German data, we experimented with a dataset from Faruqui and Padó (2012) that comprises 60k German sentences with binary formality annotations based only on formal and informal direct address (unmarked in English yet explicitly marked in German). We fine-tuned GBERT on this dataset, ***FP12*** in the following, for binary classification. As *FP12* is limited to this particular case of formality, we further utilized parallel datasets with formal and informal language pairs from languages other than German. These parallel datasets, containing informal sentences from a Q&A forum and their formal rewrites, are ***GYAFC*** with 110k English sentences (Rao and Tetreault, 2018) and ***XFORMAL*** with 23k Brazilian Portuguese, French and Italian sentences (Briakou et al., 2021b). We employed binary XLM-RoBERTa-based classifiers fine-tuned on *GYAFC*[7] and *XFORMAL*[8].

[7] https://github.com/martiansideofthemoon/style-transfer-paraphrase (Krishna et al., 2020)

[8] https://huggingface.co/SkolkovoInstitute/xlmr_formality_classifier

4.3 Evaluation

Table 1 reports the average Spearman's ρ for the different setups. Evaluated in a 10-fold cross-validation manner, the two feature-based models yielded high results. To explore their relation to formality, Figure 3 shows several linguistic features used by these models per the formality score of the sentences. While sentiment seems to be a relatively constant feature across the formality scale, other factors correlate better with formality. The punctuation ratio and the Flesch readability score tend to sink, whereas word formality, token length, constituency tree height, and the number of tokens rise with increasing sentence formality. According to *SHAP* (Lundberg and Lee, 2017)[9], among the most important features of the approach by Chhaya et al. (2018) are indeed the sentence length, the average word formality, the Flesch score and the average token length (already achieving 0.8 Spearman's ρ on their own). This shows that such simple linguistic properties are good indicators of formality, at least at the sentence level.

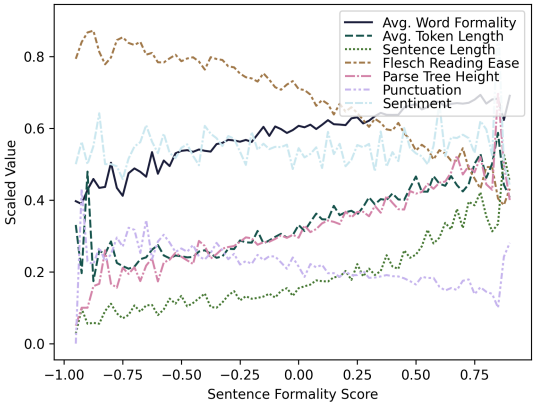

Figure 3: Relation between several linguistic features (scaled values) and the formality scores of the sentences.

[9] *SHAP* is a game theoretic approach that facilitates interpreting predictions of machine learning models.

However, fine-tuning transformers significantly outperformed the feature-based approaches (Table 1). In Figure 4, we plot the predictions of GBERT transformers fine-tuned on our dataset versus the human-assessed formality scores. The errors are lower on both ends of the scale. Sentences nearer to the scale's middle are more difficult to predict for the model since they carry fewer linguistic markers than sentences with extreme (in)formality scores. But in general, predictions are relatively accurate.

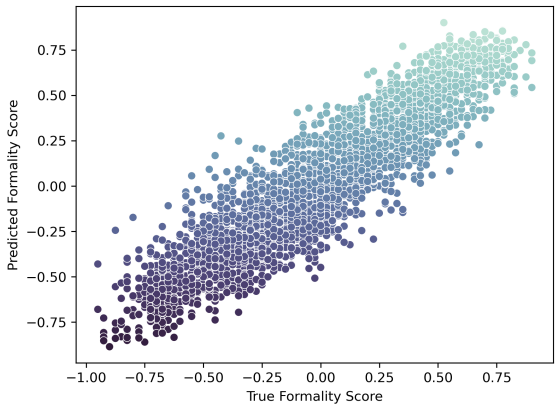

Figure 4: Predictions of the best model versus gold formality scores (brighter colors mean higher predictions).

Table 1 also shows that from the settings utilizing the *PT16* dataset, the model fine-tuned on *PT16* translated to German performed best. The formalization effect of machine translation (informal sentences get more formal through translation (Briakou et al., 2021b)) seems to influence the models using translated data since they tended to predict higher formality scores, especially for more neutral sentences. However, the results indicate that this is less critical when compared to the cross-lingual regression model fine-tuned on English and tested on German data. Contrasted to fine-tuning and testing on our dataset, the *PT16* models were still significantly worse, although *PT16* comprises over three times more sentences than our dataset. This may also be ascribed to its narrower scale of formality. *PT16* models tended to yield lower results on more formal domains of our dataset (*science*, *legal* and *Wikipedia*). Scoring these genres seems more challenging for those models since *news*, the most formal source in *PT16* (Pavlick and Tetreault, 2016), has only the fifth-highest average formality score in our dataset (see Figure 1).

The probabilities for being either formal or informal from the binary formality classifiers fine-tuned on *GYAFC* and *XFORMAL* in a cross-lingual setting also showed a correlation to the human assessments (Table 1). However, these models performed worse than regression models. Figure 5 exemplifies the class predictions of the binary formality classifier fine-tuned on *GYAFC* per formality score bin (formality scores rounded to one decimal place) on our dataset. It shows that sentences with lower formality scores tended to be classified as informal and sentences with higher scores as formal. However, formal and informal sentences were predicted in nearly every formality score bin. From that, we infer that a binary separation of formality into formal and informal sentences is not reasonable.

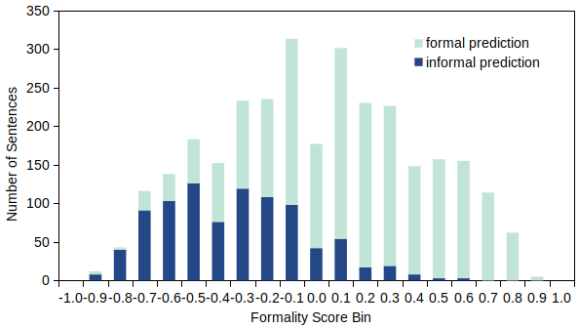

Figure 5: Formal and informal predictions of the *GYAFC* model per formality score bin of our sentences.

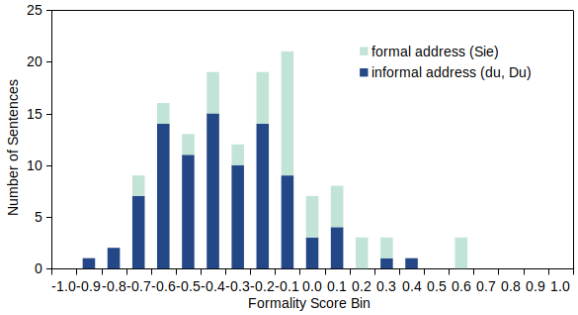

Figure 6: Distribution of formality scores of sentences with formal and informal address.

The monolingual binary classifier fine-tuned on *FP12*, which includes only formal and informal address sentences, performed significantly worse than all other setups. Figure 6 shows the number of sentences with formal and informal address in our dataset (only 137 in total) per formality score bin. Although they lean towards the lower end, even these sentences scatter broadly over the formality scale (average formality scores are -0.10 (± 0.30) for formal and -0.36 (± 0.25) for informal address). Formality is not only expressed via these different forms of address. (3) shows a sentence from our dataset with formal address but a

Training		Testing		Model	Spearman's ρ
d.	E21 (de)	*documents*	E21 (de)	GBERT	**0.891** ($\pm$0.059)
sent.	ours (de)		E21 (de)	GBERT	0.847 ($\pm$0.028)
	ours (de)		E21 (de)	feature-based ($\sim$ Pavlick and Tetreault, 2016)	0.686* ($\pm$0.039)
	ours (de)		E21 (de)	feature-based ($\sim$ Chhaya et al., 2018)	0.603* ($\pm$0.095)
d.	C18 (en)		C18 (en)	BERT	**0.827** ($\pm$0.041)
sentences	ours (en)		C18 (en)	BERT	0.729* ($\pm$0.059)
	ours (de)		C18 (en)	XLM-RoBERTa	0.703* ($\pm$0.054)
	ours (de)		C18 (de)	GBERT	0.674* ($\pm$0.054)
	PT16 (en)		C18 (en)	BERT	0.603* ($\pm$0.063)

Table 2: Results for formality scoring on documents; statistically significant differences (calculated with the two-sided Wilcoxon signed-rank test) are marked with '*' for $p < 0.005$ with respect to the **best** models; language(s) of datasets in brackets, translated data underlined.

low formality score because of other indicators. Consequently, formality is a much broader concept, and restricting it to this use case is insufficient for comprehensive formality analysis.

(3) Wollen *Sie*$_{formal\ address}$ nicht *reingucken*$_{informal}$?
Don't *you* want to *have a look*?

5 Formality Scoring on Documents

Documents may assemble an even more diverse range of clues for degrees of formality than sentences. Only recently, Jin et al. (2022) proposed extending style transfer to the more complex document level, but manual formality annotations of documents are more expensive to obtain than sentence-level assessments. Therefore, this section investigates how single sentences and linguistic properties contribute to the overall document formality. We examine if sentence-level formality annotations are useful for assessing formality on documents.

5.1 Evaluation on German Documents

We conducted experiments and analyses on German documents. For that, we utilized 800 emails with continuous formality scores (Eder et al., 2021). Sentences from emails show the highest standard deviation of formality of all domains in our dataset and the corpus from Pavlick and Tetreault (2016). Thus they possess a high stylistic variability. We denote the dataset ***E21*** in the following.

We compared transformers and feature-based approaches trained on our formality-informed sentences with transformer models fine-tuned on *E21* for predicting formality on this document collection. The upper half of Table 2 presents the average Spearman's ρ for these models. Fine-tuning GBERT on *E21* itself (10-fold cross-validation) performed best, but there is no statistically significant difference between utilizing the documents or our formality-assessed sentences as training data. The transformer models grasped the concept of formality more comprehensively since the feature-based ridge regression models yielded significantly worse results. It seems that linguistic features do not generalize well. Figure 7 shows some of the most predictive linguistic features for formality scoring on the sentence level for the documents. The average word formality and the Flesch Reading Ease correlate with document formality in a similar way than with sentence formality (Figure 3). However, the average sentence length and average token length are comparably more static across the formality scale of documents and thus less suitable features.

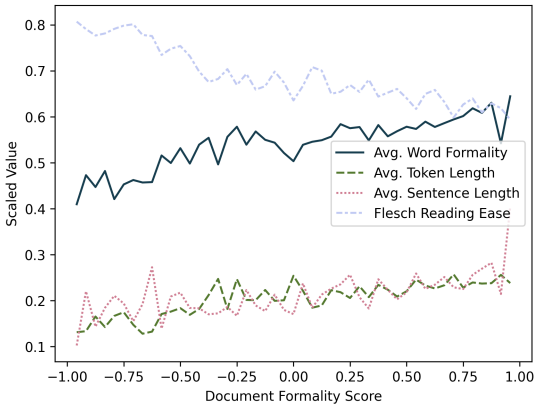

Figure 7: Linguistic characteristics (scaled values) of the documents per their formality scores.

To further understand how the formality of a document is affected by its sentences, we split the documents of *E21* into separate sentences. Then, we ran the GBERT model fine-tuned on our dataset on these sentences to determine their formality. Taking the average of the calculated scores as document score still returned a Spearman's ρ of 0.801. Although this result is significantly worse

($p < 0.01$) than running the model on the documents directly, it still shows a strong correlation between the scores of the sentences and the document formality score. In Figure 8, we plot the number of sentences per calculated formality score bin for each formality score bin of the corresponding documents. The sentence and document formality scores show some overlap. Nevertheless, the sentences in the documents have quite a range of formality scores.

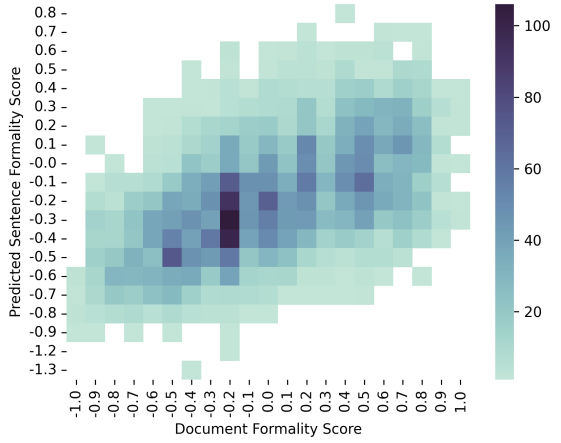

Figure 8: Frequency of calculated formality scores of sentences per formality score bin of documents.

How the formality of sentences changes throughout a document is shown in Figure 9, which depicts the mean sentence formality by position in the documents. The formality tends to decrease with increasing position in the text. This observation is in line with the assumption of Heylighen and Dewaele (1999) that formality is higher at the beginning of a text because of the lack of previous discourse to relate to. For threaded online discussions, Pavlick and Tetreault (2016) reported congruent findings.

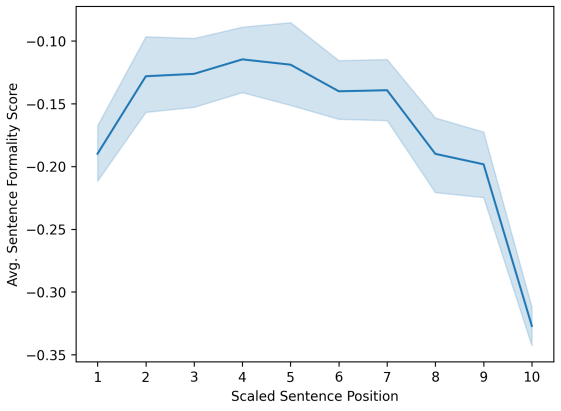

Figure 9: Average sentence formality score by position in documents; sentence positions are scaled to ten bins.

Concluding, fine-tuning transformers on sentences is applicable for assessing the formality of documents, as our results show. However, due to the variety of sentence formality scores, it may not be helpful to map formality assessments of documents to their sentences to save annotation efforts or assume mono-style documents regarding the formality dimension.

5.2 Evaluation on English Documents

To investigate the applicability of transformer models fine-tuned on our sentences for other languages, we evaluated them on English documents. We used 960 emails (***C18*** in the following) with formality annotations obtained via averaging over individual assessments on a 5-point Likert scale (Chhaya et al., 2018). The lower half of Table 2 shows the results. Fine-tuning on the documents (10-fold cross-validation) significantly outperformed sentence-based models. We ascribe this performance decline also to the manual annotations of *C18* since we only calculated an average split-half reliability of 0.573 ($\pm$0.015) Spearman's ρ over 100 trials. Given these conditions, a BERT model fine-tuned on our translated dataset still achieved a high correlation also compared to the English *PT16* model. Hence, we assume our dataset is beneficial for formality assessment of English-language documents too.

6 Conclusion

We presented the first dataset of sentences with highly reliable human formality assessments on a continuous informal-formal dimension obtained via Best-worst scaling. Our dataset comprises 3,000 sentences evenly distributed over twelve different domains to cover the broad spectrum of formality best possible. It is the first for the German language with a comprehensive perspective on sentence-level formality altogether.

We evaluated various machine learning models for the regression task of assessing formality on our dataset. We found that a transformer model fine-tuned on an existing German dataset including only sentences of formal and informal address (*Sie* vs. *Du*/*du*) yielded the worst results. Hence, this restricted view on formality is insufficient to capture a more comprehensive concept of formality. Cross-lingual settings utilizing transformer-based classifiers pre-trained on huge datasets with formal and informal language pairs not restricted

to a particular form of formality performed better. However, a binary categorization of formality strikes as inappropriate since ridge regression models employing simple linguistic features outperformed them. Fine-tuning transformers for regression on an English dataset produced similar (for the cross-lingual setting or the English translation of our dataset) or higher (for the German translation of the training data) results. In comparison, a transformer model fine-tuned on our dataset with its broader formality scale outperformed all other settings significantly.

Expanding the scope to longer texts, a requested future research direction of style transfer (Jin et al., 2022), we investigated the influence of the formality of sentences on a document's formality. We observed that the sentences included in the documents cover a wide spectrum of formality with higher formality scores at the beginning. Our results indicate that a transformer model fine-tuned for formality scoring on our sentences generalizes better across text levels than linguistic features and can be used to predict the degree of formality of German and English documents. We anticipate our dataset to facilitate future work on German formality style transfer and formality analysis in general on both the document and the sentence level. It may also be valuable for other languages.

Our **dataset** and a **tool** for analyzing styles with a wide range of linguistic features are **available** under https://github.com/ee-2/in_formal_sentences and https://github.com/ee-2/register.

Limitations

This work assesses the formality of texts in isolation, excluding any conventional and situational contexts. However, for different genres and situations different expectations have to be met. For example, an expression regarded as formal in one genre may be perceived as too informal in another. We also do not take forms of formality beyond the pure text level into account. Properties that contribute to formality besides the text itself may include the structure of a text (e.g., blank lines in emails (Chhaya et al., 2018)) or the volume, the pitch, the speech rate, or the rhythm of speech (Labov, 1972). For future research and downstream applications, it might be helpful to consider the contextual circumstances and non-textual varieties of formality too.

Our experiments on the document level include only emails due to the lack of other corpora with formality annotations on this text level. With their composition, often including greeting, signoff, and signature, emails present a particular genre. Potentially, the greeting provides already a good indication of the formality of the text that follows (e.g., *'Dear Mrs. Doe'* vs. *'Hi Jane'*). Although we anticipate congruent findings, future work should experiment with other types of documents, possibly more challenging to assess. Further, extending the cross-lingual experiments on the document level to languages other than English (e.g., languages with multiple forms of honorifics, such as Japanese) will be required.

Ethical Considerations

We ensured that our dataset can be made publicly available (sentences from *comments* are restricted to non-commercial use only). Since our data originates from several different domains, we gave careful consideration to finding a balance between copyright and data privacy regulations. Finally, we pseudonymized text spans containing personal information in user-generated content where necessary (*tweets*, *Reddit* posts, *comments* and *blogs*). This means we replaced sensitive text with automatically generated substitutes, e.g., female names with other female names or locations with other locations. We only release the IDs for *tweets*, *Reddit* posts and *comments*. For *blogs*, we follow the license requirements and publish the respective reference. The corpora with *emails* and *legal* texts had been pseudonymized already, no information on authors is available. For less-privacy-sensitive text sources, such as *subtitles*, *political* speeches, *news* and *Wikipedia*, we report all information shared in the original corpus, e.g., URLs. The sentences from *fiction* and *science*, which we collected ourselves, are cited appropriately in order to acknowledge intellectual property rights. People involved in creating our dataset were compensated at least following minimum wage requirements.

Acknowledgments

This work was partially funded by the Faculty of Humanities of the University of Klagenfurt. Further, we especially thank Udo Hahn for valuable input and discussions.

References

Fadi Abu Sheikha and Diana Z. Inkpen. 2010. Automatic classification of documents by formality. In *International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–5. IEEE.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Adrien Barbaresi and Kay-Michael Würzner. 2014. For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *Proceedings of NLP4CMC workshop (KONVENS 2014)*, pages 2–10. Hildesheim University Press.

Andreas Blombach, Natalie Dykes, Philipp Heinrich, Besim Kabashi, and Thomas Proisl. 2020. A corpus of German Reddit exchanges (GeRedE). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6310–6316, Marseille, France. European Language Resources Association.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021a. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021b. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Julian Brooke and Graeme Hirst. 2014. Supervised ranking of co-occurrence profiles for acquisition of continuous lexical attributes. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2172–2183, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Coling 2010: Posters*, pages 90–98, Beijing, China. Coling 2010 Organizing Committee.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217, Online. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's how you say it: Identifying appropriate register for chatbot language design. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 102–109, Kyoto, Japan. Association for Computing Machinery.

Yu Cheng, Zhe Gan, Yizhe Zhang, Oussama Elachqar, Dianqi Li, and Jingjing Liu. 2020. Contextual text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2915–2924, Online. Association for Computational Linguistics.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 76–86, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2020. CodE Alltag 2.0 — a pseudonymized German-language email corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4466–4477, Marseille, France. European Language Resources Association.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2021. Acquiring a formality-informed lexical resource for style analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2028–2041, Online. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring language style in chatbots to increase perceived product value and user engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 301–305, Glasgow, United Kingdom. Association for Computing Machinery.

Manaal Faruqui and Sebastian Padó. 2012. Towards a model of formal and informal address in English. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 623–633, Avignon, France. Association for Computational Linguistics.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Alexander Geyken, Adrien Barbaresi, Jörg Didakowski, Bryan Jurish, Frank Wiegand, and Lothar Lemnitzer. 2017. Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik*, 45(2):327–344.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: Definition, measurement and behavioral determinants. Technical report, Center "Leo Apostel", Free University of Brussels, Brussels.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California. Association for Computational Linguistics.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Bryan Klimt and Yiming Yang. 2004. The Enron corpus: A new dataset for email classification research. In *Machine Learning: Proceedings of the 15th European Conference on Machine Learning (ECML 2004)*, 3201, pages 217–226, Pisa, Italy. Springer.

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

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7439–7468, Dublin, Ireland. Association for Computational Linguistics.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

William Labov. 1972. *Sociolinguistic Patterns*. University of Philadelphia Press, Philadelphia, Pennsylvania, USA.

Shibamouli Lahiri. 2015. SQINKY! A corpus of sentence-level formality, informativeness, and implicature. ArXiv, abs/1506.02306.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. In *Computational Linguistics and Intelligent Text Processing. Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 446–457, Tokyo, Japan. Springer.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *Semantic Systems. The Power of AI and Knowledge Graphs (SEMANTiCS 2019)*, pages 272–287, Karlsruhe, Germany. Springer.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press, Cambridge, United Kingdom.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS 2017)*, volume 30, pages 4765–4774. Curran Associates, Inc.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Maria Nadejde, 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 *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8568–8575.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and HB. *Sawtooth Software, Inc.*

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224, Denver, Colorado. Association for Computational Linguistics.

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

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the Enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, Portland, Oregon. Association for Computational Linguistics.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Dariush Saberi, John Lee, and Jonathan James Webster. 2020. Automatic assistance for academic word usage. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2163–2168, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of German online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2017, page

1241–1244, Shinjuku, Tokyo, Japan. Association for Computing Machinery.

Tatjana Scheffler. 2014. A German Twitter snapshot. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2284–2289, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4937–4946, Hong Kong, China. Association for Computational Linguistics.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3573–3578, Hong Kong, China. Association for Computational Linguistics.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online. Association for Computational Linguistics.

A Appendix

A.1 Models for Formality Scoring

Fine-tuned Transformer Models. For fine-tuning transformers, we used the recommended and default parameter settings of the *FLAIR* framework (Akbik et al., 2019) (version 0.10):

- learning rate = 5.0e-5
- maximal epochs = 10
- optimizer = AdamW
- scheduler = linear scheduler with warmup
- warmup fraction = 0.1
- mini batch size = 4

Table 3 shows the results for fine-tuning different transformers on our dataset in a 10-fold cross-validation setting. We experimented with the German transformer models GBERT-base, GBERT-large, GELECTRA-base, GELECTRA-large (all from Chan et al. (2020)), and WECHSEL-RoBERTa-base-german (Minixhofer et al., 2022). The large models possess a high fluctuation in performance. Therefore, we chose the best-performing (and less expensive) GBERT-base model for our experiments on German data.

Model	Spearman's ρ
GBERT-base	0.919 ($\pm$0.009)
GELECTRA-base	0.918 ($\pm$0.011)
GBERT-large	0.109* ($\pm$0.274)
GELECTRA-large	0.322* ($\pm$0.426)
WECHSEL-RoBERTa-base	0.912* ($\pm$0.009)

Table 3: Results for different transformer models on our dataset (10-fold cross-validation); significant differences (at least $p < 0.05$) are marked with '*'.

Table 4 displays the performances of transformer models used in a cross-dataset setting on the original data. We report results for fine-tuning regression models on our dataset and *PT16* in a 10-fold cross-validation setting. For the formality classifier we fine-tuned ourselves, the GBERT model fine-tuned on *FP12*, we achieved perfect accuracy on the original test split of this dataset.

Dataset	Model	Spearman's ρ
ours (de)	XLM-RoBERTa	0.893 ($\pm$0.010)
ours (en)	BERT	0.891 ($\pm$0.010)
PT16 (de)	GBERT	0.762 ($\pm$0.011)
PT16 (en)	XLM-RoBERTa	0.776 ($\pm$0.016)
PT16 (en)	BERT	0.820 ($\pm$0.010)

Table 4: Results for transformer-based regression models used in a cross-dataset setting on the original dataset (10-fold cross-validation).

Feature-based Models. For the feature-based models, we used *spaCy* (3.3) (Honnibal et al., 2020) and its language model *de_core_news_sm* for basic NLP processing routines. We utilized the *benepar* library (Kitaev and Klein, 2018; Kitaev et al., 2019) (version 0.2) for constituency parsing and scored the formality of a word given its word embedding as proposed by Eder et al. (2021). Emotional features are based on the lexicon by Buechel et al. (2020), whereas sentiment was determined with the German *TextBlob* module (0.4.3).[10] We used *scikit-learn.org* (1.0.2) for the ridge regression implementation with the default parameters. We compared two sets of features adapted from Pavlick and Tetreault (2016) and Chhaya et al. (2018). In Table 5, we list the concrete features we employed per setting.

[10] https://textblob-de.readthedocs.io/

∼ **Pavlick and Tetreault (2016)**	∼ **Chhaya et al. (2018)**
• average token length • average sentence length in tokens • Flesch Reading Ease • proportion of hedge phrases • proportion of first person pronouns • proportion of third person pronouns • proportion of upper case words • proportion of lower case words • proportion of title case words • proportion of punctuation • proportion of emoticons and emojis • proportion of contractions • one-hot features for named entity types (e.g., *person*, *location*) • average word formality score • sentiment	
• average sentence length in characters • one-hot features for token uni-, bi- and trigrams • relative frequencies of POS tags • average height of constituency trees • relative frequencies of constituency productions • one-hot features for combinations of dependency relation, POS tag of governor and POS tag of subordinate • GBERT embeddings	
	• average word values for the emotions: valence, arousal, dominance, joy, anger, sadness, fear and disgust

Table 5: Linguistic features used for formality scoring.

Number of Parameters. Table 6 shows the number of parameters for the feature-based architectures and the transformer models.

Model	**Parameters**
∼ Chhaya et al. (2018)	26
∼ Pavlick and Tetreault (2016)	106K
GBERT-base	110M
BERT-base	110M
XLM-RoBERTa-base	125M
GELECTRA-base	110M
GBERT-large	335M
GELECTRA-large	335M
WECHSEL-RoBERTa-base	125M

Table 6: Number of parameters per model.

A.2 Annotation

We restricted the pool of crowdworkers to German native speakers from Germany, Austria, and Switzerland who were older than 18 years. No further information on the demographics of the annotators is accessible. The crowdworkers were compensated following the minimum wage defined by the German government (€ 9.60 per hour at the time of annotation). *Clickworker*, the crowdsourcing platform we used, does not provide separate qualification tests. Rather it ensures the qualification of the crowdworkers by their own filtering methods (e.g., project-independent online tests/training or evaluation of the work results). The German annotation guidelines can be found in the project repository alongside the dataset.

A.3 Computing Details

We carried out our experiments on a NVIDIA RTX A40 GPU with 48GB RAM. We estimate a total computational budget of 72 GPU hours. Fine-tuning GBERT-base, BERT-base, or XLM-RoBERTa-base on our dataset took under 15 minutes per model. Fine-tuning these models on *PT16* required about 45 minutes per model. Fine-tuning GBERT on *FP12* took about two hours, and fine-tuning models on German or English documents needed under five minutes. Training ten ridge regression models for 10-fold cross-validation was completed in under two minutes for the feature set based on Chhaya et al. (2018) and in under 15 minutes for the feature set based on Pavlick and Tetreault (2016).