Contextualizing Variation in Text Style Transfer Datasets

Stephanie Schoch Wanyu Du Yangfeng Ji
Department of Computer Science
University of Virginia
Charlottesville, VA 22904
{sns2gr,wd5jq,yangfeng}@virginia.edu

Abstract

Text style transfer involves rewriting the content of a source sentence in a target style. Despite there being a number of style tasks with available data, there has been limited systematic discussion of how text style datasets relate to each other. This understanding, however, is likely to have implications for selecting multiple data sources for model training. While it is prudent to consider inherent stylistic properties when determining these relationships, we also must consider how a style is realized in a particular dataset. In this paper, we conduct several empirical analyses of existing text style datasets. Based on our results, we propose a categorization of stylistic and dataset properties to consider when utilizing or comparing text style datasets.

1 Introduction

The general task of text style transfer involves rewriting source content in a target style. Currently, there are a number of text style transfer tasks with available data, such as formality (Rao and Tetreault, 2018), bias (Pryzant et al., 2020), sentiment (He and McAuley, 2016), humor or romance (Gan et al., 2017), offensiveness, (Nogueira dos Santos et al., 2018), authorship or time period (Xu et al., 2012), and personal attributes (Kang et al., 2019). While these specific tasks are often modeled in isolation, the general task definition remains consistent. As such, a natural question arises of what the relationship is between the stylistic variation of specific tasks.

Stylistic variation can arise from a number of factors such as communicative intent, topic, and speaker-receiver dynamics (Biber and Conrad, 2019), yet within the task of text style transfer, our view of a style is constrained to the context of each specific dataset. Therefore, understanding the tasks as well as the relationships between different tasks requires considering the stylistic properties and potential contextual and social factors (Hovy and Yang, 2021; Hovy, 2018) underpinning them, as well as the dataset characteristics (Bender and Friedman, 2018) and intersection of influences giving rise to the realization of style within a dataset.

From an application standpoint, considering these influences can provide a more comprehensive understanding of important textual features. There is a body of work already looking at how to identify generic features to increase target task performance (Li et al., 2019) or to compute similarity of textual features to select data for transfer learning (Ruder and Plank, 2017). In the context of text style transfer, these approaches first require understanding what features should be shared across tasks. For example, Zhang et al. (2020) leveraged the stylistic features shared between grammatical error correction data and formality to increase model performance on formality transfer datasets.

In addition to textual features such as stylistic properties, existing work also suggests that context of dataset creation should be taken into account when identifying compatible data or assessing possible out-of-distribution generalizability. For example, the similarity between how sentiment information is reflected in different domains affects adaptation performance (Li et al., 2019), and many models can achieve high performance on natural language inference tasks through task-limiting annotation artifacts (Gururangan et al., 2018; Poliak et al., 2018). In other words, factors such as data source and annotation method can create underlying textual features that can impact performance and limit generalizability. Thus, in combination, these existing works on leveraging inherent stylistic similarities (Zhang et al., 2020) or similar style-representations in different dataset domains (Li et al., 2019), as well as identifying task-limiting dataset properties (Gururangan et al., 2018; Poliak

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 226–239,
Aberdeen, Scotland, UK, 20-24 September 2021.

Dataset	Stylistic Task	Domain	Annotation	Size		
				Train	Dev	Test
Flickr	Romantic→Humorous	Image Captions	Manual	6k	500	500
Shakespeare	Shakespeare→Modern	Literature, SparkNotes	Automatic	18.4k	1.2k	1.5k
GYAFC-FR	Informal→Formal	Yahoo Answers (Online)	Manual	52k	2.8k	1.3k
GYAFC-EM	Informal→Formal	Yahoo Answers (Online)	Manual	52.6k	2.9k	1.4k
Biased-word	Subjective→Neutral	Wikipedia (Online)	Automatic	53.8k	700	1k
Fluency	Disfluent→Fluent	Telephone Conversations	Manual	173.7k	10.1k	7.9k

Table 1: An overview of the datasets used for exploratory analyses. Task describes the source-target direction used in our experiments and domain and annotation show general categorizations. Size provides statistics of the data splits, with standard, pre-existing data splits used when available.

et al., 2018) indicate that analysis of both stylistic properties and dataset characteristics, as well as the potential interdependencies between them, is warranted.

In this paper, we consider two primary categories of textual variation within the context of text style transfer: **stylistic characteristics** and **dataset characteristics**. We perform a series of empirical analyses to demonstrate the visible influence of both style and dataset characteristics on the performance of text style transfer models. Then, we present a categorization of style and dataset properties for consideration when utilizing or comparing style transfer datasets. Finally, we discuss the downstream applications for contextualizing variation in text style datasets, including multi-task learning, data selection, and generalizability. Our work and suggestions fall within the context of and align with recent work on incorporating social factors in natural language processing systems (Hovy and Yang, 2021) and characterizing datasets (Bender and Friedman, 2018).

2 Empirical Analyses

As an exploratory step, we question whether we can distinguish differences arising from style or dataset properties when comparing empirical results across datasets. We identify a set of aligned English datasets used for supervised text style transfer that exhibit differences ranging from style, annotation method, and domain. We further restrict our selection to datasets in which a single stylistic attribute is transferred between classes. Specifically, we look at **GYAFC-EM** & **GYAFC-FR** (Rao and Tetreault, 2018), **Shakespeare** (Xu et al., 2012), **Biased-word (Bias)** (Pryzant et al., 2020), **Fluency** (Wang et al., 2020; Godfrey et al., 1992), and **Flickr** (Gan et al., 2017). We provide dataset overviews in Table 1, with detailed dataset descriptions provided in Appendix A. We perform a preliminary qualitative analysis to get an initial impression of the data differences.

First Impression of Data: Of the six datasets, four were manually annotated and two were automatically annotated. For manually annotated datasets, GYAFC-EM and GYAFC-FR utilized crowdsourced rewrites, Flickr utilized crowdsourced sentences with only visual context shared between annotators, and Fluency utilized expert annotations of the target attribute. Both automatically annotated datasets (Bias, Shakespeare) were created through identification of existing data sources. While each style task is unique (other than two domains of GYAFC for formality), in terms of style we observe that Shakespeare has a significantly different temporal context than all other datasets, and Fluency involves a stylistic attribute that, ideally, the sentence pairs in all other datasets should possess.[1]

Beyond our qualitative observations, we perform an exploratory multi-task learning experiment, described in the following subsection.

2.1 Multi-Task Learning

As a toy experiment, we ask the question "*What would our results look like if we naively train on all style transfer tasks, with no considerations beyond the fact that the tasks share a general task defi-*

[1]Fluency is frequently a criteria used in text style transfer evaluation (Mir et al., 2019; Briakou et al., 2021; Prabhumoye et al., 2018).

nition?[2] We essentially ignore all considerations for style or dataset properties. Our expectation is that negative transfer will occur due to the lack of consideration for factors such as domain (Pan and Yang, 2009; Li et al., 2019)[3], but we are interested in whether all tasks share similar performance patterns or if performance on any tasks diverge from the overall set. If the latter, is there any intuitive explanation for the divergences?

We further expect that the degree of negative transfer will be impacted by the degree of difference of stylistic or data properties, relative to the full set of pre-training datasets. Specifically, we anticipate some level of alignment with our initial impression of the data: the alternate temporal context of Shakespeare may increase degree of negative transfer, yet the inherent stylistic connection with Fluency may lessen the degree of negative transfer.

Experimental Setup We utilize two experimental settings: GPT-2 directly fine-tuned on each dataset, and GPT-2 with multi-task pre-training on all datasets followed by fine-tuning on each target dataset. For both settings, we initialize GPT-2 with the pre-trained parameters from Radford et al. (2019). For our multi-task experimental setup, we follow prior works (Liu et al., 2015, 2019; Raffel et al., 2020) to perform multi-task learning for the baseline GPT-2 model (Wang et al., 2019): we initialize GPT-2 with the pre-trained parameters from Radford et al. (2019), then we jointly pre-train on all style tasks in a supervised manner and fine-tune on each individual style transfer task. [4]

For multi-task learning, we construct our pre-training dataset by randomly shuffling the training examples from all datasets. During pre-training, each training example from each individual task is seen at least once per epoch. All of the training examples in the largest dataset are seen exactly once per epoch, while all training examples for the smallest dataset are seen multiple times per epoch (proportional to the ratio between the training set size of the largest-scale task and the smallest-scale task). For the fine-tuning step, we leverage the multi-task pre-trained model and further fine-tune on each individual supervised task, saving the model with the lowest validation set loss as our final model for evaluation.

Dataset	Task	BLEU-og	BLEU-mt	%og
Shakespeare	shake2mod	24.47	11.33	0.463
Fluency	dis2fl	96.59	96.69	1.001
Flickr	rom2fun	8.14	7.18	0.882
GYAFC-EM	inf2fr	69.96	65.16	0.931
GYAFC-FR	inf2fr	75.16	74.72	0.994
Biased	subj2neut	93.73	93.41	0.996

Table 2: Experiments conducted using GPT-2, where BLEU-og represents directly fine-tuning the original GPT-2 on the target task, BLEU-mt represents multi-task pre-training using all datasets and fine-tuning on the target task, and %og represents the relative performance of multi-task pre-training in comparison to the performance of the original GPT-2 (computed by dividing BLEU-mt by BLEU-og).

Results We report BLEU (Papineni et al., 2002) in Table 2 as a measure of content preservation.[5] We compare the performance between directly fine-tuning the original GPT-2 on the target task (BLEU-og) and firstly multi-task pretraining the original GPT-2 then fine-tuning it on the target task (BLEU-mt).

Negative transfer is identified as a performance drop in BLEU-mt, i.e. %og < 1.00. Since the style transfer datasets in use are diverse across domain and stylistic properties, we expect negative transfer to occur in the multi-task learning setting. However, we are specifically looking at the overall performance pattern as an initial step in determining what properties may underlie such differences, which should be accounted for in a taxonomy.

While most tasks perform within a 12% margin below the original GPT-2 performance, we observe two divergences: with multi-task learning, the Shakespeare-to-modern task performed at less than 50% of the original GPT-2 performance, and the disfluent-to-fluent task experienced a slight performance increase. Performance on Fluency exceeded our initial expectation that the degree of negative transfer would simply be lower compared to other datasets, but overall the divergences with Shakespeare and Fluency match our expectations based on our initial impression of the data style differences. Specifically, we attribute the performance drop on the Shakespeare dataset to limited suitability for combining the data sources likely due to the stylistic attribute pertaining to different temporal

[2]The general task definition is rewriting the source content of a text in a target style (see section 1)

[3]Negative transfer occurs when transferred knowledge negatively impacts target performance (Pan and Yang, 2009).

[4]GPT-2 models were each trained on a single NVIDIA GTX 1080 Ti GPU.

[5]We use the BLEU implementation from Koehn et al. (2007).

context, and we attribute the Fluency dataset performance increase to high suitability for combining the data sources likely due to its stylistic attribute pertaining to a textual criteria that is assumed to be inherent to the other data.

With regard to dataset differences, we note the potential impact of dataset size on performance: to maintain consistency of the model architecture, we utilize the same model configuration with GPT-2 across datasets and experimental settings. In the case of performance on the Flickr dataset (see Table 1), it is possible that such a model configuration may overfit on the dataset. However, this alone fails to account for our observations of performance pattern divergences.

Beyond overall pattern, we observe an unexpectedly wide range of BLEU scores across datasets, which we expect could be attributable to differences in either dataset creation or style. There may be stylistic differences in how style information is encoded that impact content preservation. For example, some styles may have more words that encode both style and content information which may increase the difficulty of content retention (Cao et al., 2020), yet other styles may be characterized by stylistic attributes encoded in only a few key words or phrases (Fu et al., 2019). However, these differences may also be attributable to dataset creation. We expect that if the attribute-encoding words are constrained to a few words or phrases as a property of a style itself, then a dataset's style classes should be highly distinguishable using lexical features; in other words, the decision boundary when classifying styles should stay at the lexical level (Fu et al., 2019).

To test these hypotheses and help explain the range of BLEU scores, we perform two complementary experiments. First, we compute sentence similarity metrics averaged over each dataset to 1) identify if there is a relationship between BLEU scores and baseline sentence pair similarities, and 2) identify datasets with high similarity across class boundaries that constrain stylistic attributes to a few words or phrases. Second, we perform classification and ablation studies using a set of linguistic features defined on each dataset. For datasets with high sentence similarities, if a style can be well-represented by a few style-encoding words or phrases, then we expect high classification performance using only lexical features. Conversely, if a style cannot be isolated to a few words and phrases, we expect low classification performance using lexical features alone, in which case a high sentence similarity is likely attributable to dataset properties rather than inherent style properties.

Dataset	JS ↑	LD ↓	LD-norm ↓	F1-Score ↑
Shakespeare	0.0845	14.79	0.9029	0.0583
Fluency	0.9941	0.366	0.0271	0.9751
Flickr	0.2257	11.92	0.7728	0.3623
GYAFC-EM	0.4471	7.924	0.5616	0.4207
GYAFC-FR	0.4565	7.723	0.5375	0.4500
Biased	0.9137	2.529	0.0763	0.9689

Table 3: Jaccard Similarity (JS), Levenshtein Distance (LD), normalized Levenshtein Distance (LD-norm), and F1-Score. Sentence similarity measures quantify the distance between target and source for the training sets with arrows indicating direction for more similar sentences.

2.2 Similarity Metrics

We calculate token-based Jaccard Similarity, token-based Levenshtein distance, and F_1-score between the source and target training sets. We also report Levenshtein distance normalized by sentence length, $LD_{norm}(\boldsymbol{s}, \boldsymbol{t}) = \left(\frac{LD(\boldsymbol{s},\boldsymbol{t})}{\max |\boldsymbol{s}|,|\boldsymbol{t}|}\right)$ where $LD(\boldsymbol{s}, \boldsymbol{t})$ is the Levenshtein distance, $\boldsymbol{s}, \boldsymbol{t}$ refer to sentences in a sentence pair, and $|\cdot|$ refers to the number of tokens in a sentence. Scores are reported in Table 3.[6]

We see some relationships between similarities in Table 3 and GPT-2 performances in Table 2 in that the datasets with the lowest BLEU scores (Shakespeare and Flickr) have the lowest baseline similarities, and the datasets with the highest BLEU scores (Fluency and Bias) have the highest baseline similarities. We therefore can identify the Fluency and Bias datasets as being of particular relevance for the lingusitic features analysis. Specifically, our hypothesis is that if the Bias and Fluency styles can truly be isolated to few words as the sentence similarities would suggest, then the classification performance should be high using only lexical features. In contrast, if the dataset properties influence variation through constrained stylistic representation, then we expect low classification accuracy using lexical features.

[6]We do not distinguish between source and target direction as these metrics are symmetric in our setting (see Appendix B).

Group	Features
Lexical Complexity	Average word length, average syllable count (with & without stopwords)
Readability	# complex words (≥ 3 syllables)*, Flesch Reading Ease Score, Flesch-Kincaid Grade Level
Lexical Diversity	Unique unigrams & bigrams, with punctuation removed*
POS tags	Universal POS tag distribution*, Penn Treebank POS tag distribution*
Sentence length	Sentence length (words & total tokens)
Phrases	# noun phrases*, # verb phrases*, average length of noun phrases*, average length of verb phrases*, # dependent clauses*, average length of dependent clauses*
Subjectivity	# 1st, 2nd, & 3rd person pronouns*, Subjectivity & Sentiment polarity according to TextBlob sentiment module
Bag-of-Words	Bag-of-words feature representation

Table 4: Linguistic feature groups: lexical (top), syntactic (gray in middle), and other (bottom). Features features denoted with an asterisk (*) are normalized by sentence length.

2.3 Linguistic Features Analysis

We define linguistic features to refer to properties characterizing textual variation primarily at the lexical or syntactic level, where the "other" category in Table 4 indicates features that may capture slight semantic variation (subjectivity) or reflect overall lexical tendencies (bag-of-words). Features are adopted from prior works (Pavlick and Tetreault, 2016; Abu-Jbara et al., 2011; Roemmele et al., 2017) and listed in Table 4, with further description in Appendix C.

We train logistic regression classifiers with $\ell 1$-regularization and feature scaling on the full feature set for each text style dataset. Next, we train and subsequently test classifiers with all features ablated except the specified subset, and identify important features as those with minimal relative performance drop compared to full-feature classification accuracy. Results are shown in Table 5. We further quantify the magnitude of variation by computing the Jensen-Shannon (JS) divergence for each feature, and indicate the cells corresponding to features with divergences ≥ 0.075 in Table 5 in bold.[7]

Datasets with the lowest BLEU scores (Flickr and Shakespeare) have more distributed salient class features across linguistic levels, further reflected in a higher number of features with large divergence magnitudes (≥ 0.075). For the high BLEU and sentence similarity datasets of interest (Bias, Fluency), the inverse of this is true. For Bias and Fluency we see consistently low classification

[7]Table 6 in Appendix D shows a JS-divergence heatmap.

	Flick	Shake	GY-FR	GY-EM	Bias	Flu.
FF	75.6	76.1	80.7	80.9	63.5	55.3
LexC	**51.7**	62.2	**65.6**	64.4	52.6	50.7
Read	**55.7**	52.1	**62.1**	63.3	52.0	51.0
LexD	52.4	49.6	51.2	52.0	**50.4**	**54.4**
UPOS	**59.4**	59.3	62.3	**60.8**	54.4	51.6
XPOS	62.3	59.7	65.1	66.1	55.0	51.7
SenL	**51.8**	**56.7**	56.2	51.7	50.3	51.0
Phr	**54.2**	58.2	53.6	53.4	52.9	51.8
Sub	**60.5**	**60.4**	51.7	52.9	57.0	50.4
BoW	74.2	72.4	71.5	71.7	62.2	50.3

Table 5: Classification accuracy using linguistic feature groupings described in Table 4, with Full Features (FF) indicating the entire suite of features. Classification accuracy for features with Jensen-Shannon divergences ≥ 0.075 are in bold.

performance across ablations, including the lexical feature ablations. These results support our hypotheses and further suggest that neither stylistic differences nor dataset characteristics alone can be used to relate text style datasets. Rather, both influences as well as their interactions require consideration.

In the following section, we propose a taxonomy of style and dataset property categories that can contribute to variation in text style transfer datasets. Additionally, we note that when introducing these properties, we view style *as the targeted stylistic property within the context of a text style dataset*.

3 Variation From Style and Data Properties

Our empirical analyses demonstrate the visible influence of both style and dataset properties on how a style is represented in a given dataset. In addition to brief mentions of influences of dataset creation in section 1, we can identify an intuitive reason for these dual influences. While linguistic approaches exist to analyze textual variation (Halliday and Matthiessen, 2013; Holmes and Wilson, 2017; Biber, 2012), we suggest that the processes of linguistic-based stylistic analysis and text style transfer typically occur in inverse directions: linguistic analysis may work from human-written text and then analyze stylistic variation, whereas text style transfer may work from pre-existing ideas of targeted stylistic variation and then create datasets of human-written text that meet stylistic expectations. In other words, to create a text style transfer dataset or train a text style transfer model, the researcher should have a notion of the desired style against which to judge the resulting artifact. Intuitively, this process can lead to process-attributable

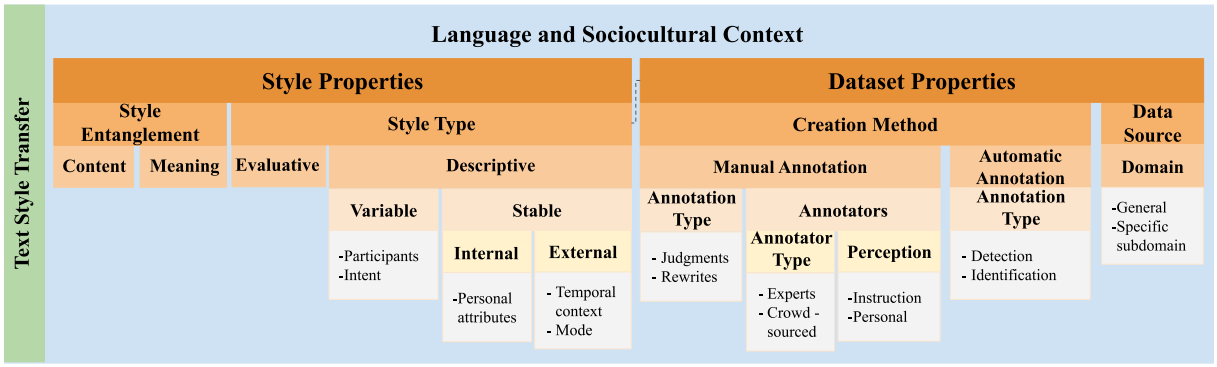

Figure 1: Framework overview visualizing style and dataset properties discussed throughout section 3. Boxes with bullet points indicate example considerations within each category. We contextualize both style and dataset properties within language and sociocultural context as all language is implicitly reflective of these influences (Hovy and Yang, 2021).

variation secondary to and alongside the intended stylistic variation.

Based on our results and observations, we consider ***stylistic properties*** as properties influencing textual variation that are inherent to a particular style and ***dataset properties*** as factors influencing textual variation due to how a particular dataset was created. We detail style and dataset properties in the following subsections and visualize the major distinctions in Figure 1.

3.1 Stylistic Properties

We group stylistic properties under two broad categories: *style entanglement* and *style type*.

3.1.1 Style Entanglement

Although some recent approaches to style transfer model style and content words separately (Li et al., 2018), or try to disentangle style and content representations (John et al., 2019; Kazemi et al., 2019), this approach may be less effective when used to transfer styles in which a higher ratio of words embed both style and content information. We can consider this ratio of dual-embedding a property inherent to the style. Specifically, we can consider how entangled the style and the content or semantic meaning is, where *content entanglement* refers to whether changes to the style result in additions or reductions in the total content details, and *meaning entanglement* refers to whether changes to the style can retain the content details but alter the semantic meaning. As an example of this distinction, sentiment transfer, which has been regarded previously as transfer between negative and positive style (Shen et al., 2017; Prabhumoye et al., 2018) alters semantic meaning while retaining most content, yet transferring between styles such as expert-to-layman can retain meaning but lead to content detail reductions due to the difficulty of preserving content from professional sentences (Cao et al., 2020).

3.1.2 Style Type

Style can refer to the *individuating* sense or *evaluative* sense of a text (Crystal and Davy, 1969). We refer to ***evaluative styles*** as styles distinguished by general properties that address overall textual quality corresponding with rules of usage and composition, effectiveness of expression (Strunk and White, 1999) or based on overall quality evaluation and judgments (Williams and Bizup, 2017). Stylistic variation occurs solely along evaluative lines, independent of situational context or language choice. From our empirical experiments, we can consider the Fluency dataset representative of a dataset in which the transferred stylistic attribute refers to an evaluative sense of style.

We consider ***descriptive styles*** as distinguished by stylistic properties that characterize textual variation through influences such as the underlying communicative intent, the situational or social factors influencing language choice, and the attributes of the producer of the text. We can further differentiate descriptive styles by the stability or variability of the targeted stylistic property.

Stability of Targeted Style Properties On one end of the spectrum variable stylistic properties (high variance, low stability) are characterized by dynamically shifting language to convey information a certain way, which may be reflective of factors such as the underlying intent in producing the text or the social dynamics of a situation. For exam-

ple, politeness can shift based on social dynamics such as social distance and relative power between participants (Brown et al., 1987) independently of the directness of communication, such as formality [8] in email (Peterson et al., 2011). From our empirical experiments, we consider Flickr, GYAFC, and Bias as reflective of variable targeted properties.

At the other end of the spectrum, more stable targeted stylistic properties (low variance, high stability) remain more consistent across social situations and arise from relatively stable internal or external context. These may reflect internal context such as the personal attributes of the producer of text (Kang et al., 2019), or external context such as the temporal context at time of text production or stylistic properties inherent to the mode of distribution. Example datasets include the PASTEL dataset (Kang et al., 2019) annotated for personal attributes such as gender and age group, and the Shakespeare dataset (Xu et al., 2012) which can be considered reflective of authorship (Xu, 2017) or temporal context.[9]

3.2 Dataset Properties

While in the previous section we discussed properties inherent to specific styles, in this section we discuss properties of datasets to which textual variation is attributable. We identify the broad categories of properties due to *creation method* and *data source*. In this context, creation method refers to the general method of creating sentence pairs (automatic or manual annotation, as well as any properties arising from utilizing a specific method, such as influences of annotator background or perceptions) and data source refers to characteristics (such as domain) from where the source data was collected. We provide more detailed discussion in the following subsections.

3.2.1 Creation Method

Generally speaking, datasets can be created via manual annotation, such as through judgments or rewrites, or via automatic annotation, such as through filtering data that has a target attribute (i.e., detection with a classifier). With particular attention on manual annotation, in addition to potential generalizability-limiting data properties arising from artifacts of the *annotation method* and *annotation type* ((Geva et al., 2019), also, see section 1), the *annotators* themselves can influence stylistic variation. For example, model performance has been improved by incorporating annotator identifiers as features (Geva et al., 2019) and by augmenting machine translation models with distinct translator styles identifiable in the training data (Wang et al., 2021). In the case of Wang et al. (2021), using annotator styles resulted in BLEU score variations of up to $+4.5$ points.

Underlying these influences, annotator properties that may give rise to textual variation could include the background of the annotator such as experts or crowd-sourced workers, and the perception the annotators have of the style task. Similar to human evaluation of outputs, perception may arise due to personal understanding or the wording of instructions presented.[10].

Data Source - Domain: Differences in domain can be reflected in entirely different word meanings and contexts of use (Li et al., 2019), as well as different manners of encoding attribute information such as sentiment (Blitzer et al., 2007; Li et al., 2019). In addition to differences of a single style between domains, the domains themselves have different levels of stylistic diversity (Kang and Hovy, 2021). Further, while the properties characterizing a style may be inherent to how a style is realized *within* a domain, there is a distinction in how the style is reflected *between* domains that necessitates domain being considered as a dataset property influencing variation in text style datasets.

4 Interplay Between Style and Data Properties

Bender and Friedman (2018) proposed data statements for documenting dataset contextual factors such as language variety, speaker demographics, annotator demographics, speech situation, and text characteristics (e.g. genre, topic). The style and dataset properties we discuss as potentially contributing to variation in text style transfer datasets show some alignment with those proposed for data statements as such factors contribute to linguistic variation in a general sense. However, our categorization specifically operates within the context

[8] Formality is closely related to politeness (Kang and Hovy, 2021)

[9] Regarding distribution mode, Abu-Jbara et al. (2011) suggested a set of linguistic features differentiating written and audio styles.

[10] Schoch et al. (2020) discuss potential influences of framing effects of questions or instructions on results in human evaluation of outputs, and we suggest similar effects could influence dataset properties resulting from annotation of inputs.

of text style transfer datasets for which there are unique considerations and important distinctions between sources of variation and downstream implications or applications.

In the previous subsections, we discussed style properties and dataset properties to which variation in text style transfer datasets can be attributed. In this section, we discuss the interdependence of style and data properties in text style transfer datasets in terms of context-dependence of and interactions between sources of variation.

Style and Data Property Interactions While we previously considered the potential impact of both style and dataset characteristics independently, these characteristics may have underlying interactions and influences on one another. Specifically, certain types of stylistic properties may be more or less amenable to certain dataset creation methods or sources, and vice versa.

With regard to the stability of stylistic properties, dataset properties such as annotation method may be indirectly influenced when transferring across relatively stable stylistic properties. For example, machine translation models have been found to exhibit stylistic bias through reflecting demographically-biased training data (Hovy et al., 2020). While this demonstrates that the demographics of annotators can serve as an important dataset characteristic, it also demonstrates the potential to transfer across relatively stable stylistic properties, such as personal attributes (Kang et al., 2019). However, as the stylistic properties are inherent to the annotator, there may be constraints on dataset creation through manual data annotation, such as potential limitations and additional considerations for using methods such as human judgments. This underscores additional considerations for and potential challenges of selecting data from two styles that may have underlying influences on how datasets are constructed.

Context-Dependence of Variation Relatedly, contextual considerations come into play with respect to the the Shakespeare to Modern English style transfer task, a dataset also reflective of transfer across stable, contextual boundaries. The Shakespeare to Modern English transfer task can be considered as transferring across temporal context, or as the characteristic style of a single author (Xu, 2017). In this case, while an influence of sociocultural context is apparent when considering the original data sources, the targeted stylistic variation occurs across such context boundaries. Thus, source of variation for textual features arising from external context lies with whether the intent is present for a dataset to represent a transfer across context boundaries, rather than an artifact reflecting specifics of dataset creation. This is illustrated in Figure 1 as a dashed line connecting style type to dataset properties.

With further regard to dataset creation, it is important to acknowledge that while we consider many properties arising from social influences as dynamic and variable influences giving rise to particular *styles*, a dataset will indirectly and inadvertently reflect such social context during creation to some degree. As such, we also must consider social factors **not** related to the actual targeted style, but rather arising from the dataset creation process. As an example of this consideration, we can't simply say two sentiment datasets from the same general domain (such as restaurant reviews) are equivalent if one was constructed with reviewers who had anonymity (in a sense mitigating some of the direct social pressure or influence) and the other was constructed with reviewers who were not anonymous and were thus subject to increased social pressure. By understanding both data and style differences and their interactions within a particular context, these potential differences or hidden influences can be more easily identified. In summary, the interactions between style and data properties are complex. While we have suggested interactions between context and sources of influence, there are likely correlations that exist based on sources of variation which future work can investigate.

5 Influences and Applications

In the previous sections, we demonstrated visible influences of style and dataset properties on performance, categorized a set of style and dataset properties for consideration, and discussed the potential interactions between sources of variation. We conclude by discussing several applications of understanding the sources of variation in text style transfer datasets. Specifically, we look at multi-task learning, domain adaptation, and generalizability.

Multi-Task Learning and Domain Adaptation Multi-task learning aims to jointly train a model with auxiliary tasks to complement learning of the target task. When determining which auxiliary objectives to incorporate, multi-task learning for

various NLP tasks has been shown to benefit from knowledge about both *dataset characteristics* and *stylistic properties*. For example, multi-task learning performance gains for NLP tasks such as POS tagging and text classification are predictable from dataset characteristics (Kerinec et al., 2018; Bingel and Søgaard, 2017). With regard to stylistic properties, within the context of multi-task learning for style transfer Zhang et al. (2020) achieved performance gains by leveraging an intuitive stylistic connection between formality data and grammatical error correction data.[11].

While multi-task learning can be viewed as a form of parallel transfer learning, we can view domain adaptation as a form of sequential transfer learning and look at similar applications of contextualizing stylistic variation. Li et al. (2019) found that leveraging generic style and content information outperformed generic content information alone for domain adaptation, however, the closeness of sentiment information (target attribute) in the source and target domains impacted performance. In other words how the style was reflected in the particular dataset (i.e., a dataset characteristic) was related to the benefit provided by the adaptation. Based on the combined evidence in this section, we can thus support applying analysis of both style and dataset properties for transfer learning data selection, including multi-task learning and domain adaptation, in text style transfer. We suggest that the taxonomy presented in this paper can assist exploration of systematic data selection methods in these and related application areas.

Generalizability One of the underlying motivations for pursuing multi-task learning and domain adaptation is the issue of generalizability. In the context of style transfer, we can consider generalizing a model for one style across different data distributions with the same stylistic attribute, or across similar domains yet different stylistic attributes. In either case, how the model learns to represent the generic style or content information is vital for successful transfer. As we've demonstrated throughout prior sections, considering both style and dataset properties can aid in identifying sources from which possible issues may arise in terms of along which dimensions stylistic attributes may significantly differ, or which artifacts or influences of dataset creation may influence generalizability secondary to any stylistic considerations. Considerations to this end may prove beneficial both in the dataset creation process as well as when considering how a model may perform beyond a specific dataset.

[11]Other styles, such as impoliteness and offense, are also highly dependent on each other (Kang and Hovy, 2021)

6 Conclusion

In this paper, we conducted a set of exploratory analyses to assess the visibility or influence of both style and dataset characteristics on text style transfer. Based on these observations, we proposed a categorization of stylistic and dataset properties that can contribute to variation in text style transfer datasets and described the applications in which these properties may be influential, limiting, or leveragable.

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions. We also thank Diyi Yang and Jingfeng Yang for a series of helpful discussions.

References

Amjad Abu-Jbara, Barbara Rosario, and Kent Lyons. 2011. Towards style transformation from written-style to audio-style. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 248–253, Portland, Oregon, USA. Association for Computational Linguistics.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus linguistics and linguistic theory*, 8(1):9–37.

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

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel Tetreault, and Marine Carpuat. 2021. A review of human evaluation for style transfer. *arXiv preprint arXiv:2106.04747*.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

David Crystal and Derek Davy. 1969. *Investigating English Style*. Indiana University Press, Bloomington & London.

Yao Fu, Hao Zhou, Jiaze Chen, and Lei Li. 2019. Rethinking text attribute transfer: A lexical analysis. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 24–33, Tokyo, Japan. Association for Computational Linguistics.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Janet Holmes and Nick Wilson. 2017. *An introduction to sociolinguistics*. Routledge.

Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 42–49, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark. Association for Computational Linguistics.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, Ph.D) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. 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 1696–1706, Hong Kong, China. Association for Computational Linguistics.

Dongyeop Kang and Eduard Hovy. 2021. Style is NOT a single variable: Case studies for cross-stylistic language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2376–2387, Online. Association for Computational Linguistics.

Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi. 2019. Style and content disentanglement in generative adversarial networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 848–856. IEEE.

Emma Kerinec, Chloé Braud, and Anders Søgaard. 2018. When does deep multi-task learning work for loosely related document classification tasks? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 1–8, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text 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 3304–3313, Hong Kong, China. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and 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 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921, Denver, Colorado. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. 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 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Melissa Roemmele, Andrew S Gordon, and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses. In *SIGKDD 2017 Workshop on Machine Learning for Creativity*, pages 13–17.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "This is a problem, don't you agree?" Framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6833–6844.

Anders Søgaard, Barbara Plank, and Dirk Hovy. 2014. Selection bias, label bias, and bias in ground truth. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 11–13, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

William Strunk and E. B. White. 1999. *The elements of style*, 4th ed edition. Allyn and Bacon, Boston.

Shaolei Wang, Wangxiang Che, Qi Liu, Pengda Qin, Ting Liu, and William Yang Wang. 2020. Multi-task self-supervised learning for disfluency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9193–9200.

Yue Wang, Cuong Hoang, and Marcello Federico. 2021. Towards modeling the style of translators in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1193–1199.

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.

Joseph M. Williams and Joseph Bizup. 2017. *Style: lessons in clarity and grace*, twelfth edition edition. Pearson, Boston.

Wei Xu. 2017. From shakespeare to Twitter: What are language styles all about? In *Proceedings of the Workshop on Stylistic Variation*, pages 1–9, Copenhagen, Denmark. Association for Computational Linguistics.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee.

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 Dataset Details

We selected English text style datasets with a single transferred stylistic attribute between two classes. Of importance for inclusions were datasets that exhibited different creation methods: both automatically annotated and human annotated. Where available, we used the original (or pre-existing, as with the case of the Shakespeare dataset) train/val/test data splits. Links to each dataset are provided through the respective citations.

Fluency Contains aligned sentence pairs labeled as fluent or disfluent, from the English Switchboard (SWBD) Corpus (Godfrey et al., 1992; Wang et al., 2020). Train/val/test split: 173.7k/10.1k/7.9k

GYAFC-EM & GYAFC-FR Contain aligned sentence pairs labeled as informal or formal, from the *Entertainment & Music* and *Family & Relationships* domains, respectively, of the question answering forum Yahoo Answers (Rao and Tetreault, 2018). GYAFC-EM & GYAFC-FR datasets can be requested at https://github.com/raosudha89/GYAFC-corpus. GYAFC-EM Train/val/test split: 52.6k/2.9k/1.4k; GYAFC-FR Train/val/test split: 52k/2.8k/1.3k

Biased-Word Contains aligned sentence pairs labeled as subjective or neutral, crawled from 423,823 Wikipedia editor neutrilization revisions between 2004 and 2019 (Pryzant et al., 2020). Train/val/test split: 53.8k/700/1k

Flickr Contains sentence pairs captioning an image, labeled as romantic or humorous (Gan et al., 2017). We created a 6k/500/500 Train/val/test split since only the original 7k training instances are available.

Shakespeare Contains sentence pairs labeled as Shakespeare or modern English (Xu et al., 2012). Sentences are crawled from 17 Shakespeare plays from Sparknotes [12], which provides the modern counterparts. Following Jhamtani et al. (2017), we use 15 plays for training, with *Twelfth Night* used for validation, and *Romeo and Juliet* used for testing.

B Similarity Metrics

In Table 3 we do not distinguish between source and target direction due to the symmetry of metrics in our setting. We provide further justification below:

Jaccard similarity can be defined as

$$\frac{\mathcal{V}_{\{\boldsymbol{s}^{(k)}\}} \cap \mathcal{V}_{\{\boldsymbol{t}^{(k)}\}}}{\mathcal{V}_{\{\boldsymbol{s}^{(k)}\}} \cup \mathcal{V}_{\{\boldsymbol{t}^{(k)}\}}} \tag{1}$$

where $\mathcal{V}_{\{\boldsymbol{s}^{(k)}\}}$ denotes the set of vocabulary words existing in a source sentence $\{\boldsymbol{s}^{(k)}\}$ and $\mathcal{V}_{\{\boldsymbol{t}^{(k)}\}}$ denotes the set of vocabulary words existing in a target sentence $\{\boldsymbol{t}^{(k)}\}$. By the commutative property, $\mathcal{V}_{\{\boldsymbol{s}^{(k)}\}} \cap \mathcal{V}_{\{\boldsymbol{t}^{(k)}\}} = \mathcal{V}_{\{\boldsymbol{t}^{(k)}\}} \cap \mathcal{V}_{\{\boldsymbol{s}^{(k)}\}}$ and $\mathcal{V}_{\{\boldsymbol{s}^{(k)}\}} \cup \mathcal{V}_{\{\boldsymbol{t}^{(k)}\}} = \mathcal{V}_{\{\boldsymbol{t}^{(k)}\}} \cup \mathcal{V}_{\{\boldsymbol{s}^{(k)}\}}$, making Jaccard similarity symmetric. Word-based Levenshtein distance is defined as the minimum number of edit operations to convert $\{\boldsymbol{s}^{(k)}\}$ to $\{\boldsymbol{t}^{(k)}\}$ through insertions, deletions, and substitutions. Substitutions are symmetric by definition, and insert and delete operations to convert $\{\boldsymbol{s}^{(k)}\}$ to $\{\boldsymbol{t}^{(k)}\}$ are simply reversed when converting $\{\boldsymbol{t}^{(k)}\}$ to $\{\boldsymbol{s}^{(k)}\}$. In $LD_{norm}(s,t)$, we normalize by $\max|s|,|t|$, which is invariant to order. Finally,

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{2}$$

where $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$.

In our setting, $TP = w \in s \cap t$, $FP = w \in \boldsymbol{s} \backslash \boldsymbol{t}$, and $FN = w \in \boldsymbol{t} \backslash \boldsymbol{s}$. By these definitions, *FP* and *FN* are reversed when source and target are reversed, and therefore by definition, F_1 is symmetric when comparing source and target sentence pairs.[13]

C Linguistic Features

Lexical Complexity Lexical complexity refers to the complexity of words based on the length or number of syllables. We use average word length in characters (Pavlick and Tetreault, 2016) and average number of syllables, with and without stopwords.

Lexical Diversity Size of vocabulary has been used as a feature for style categorization in prior work (Abu-Jbara et al., 2011). We chose to include unigrams and bigrams to reflect diversity of vocabulary as well as diversity of expression.

[12] https://www.sparknotes.com/

[13] Acronyms refer to "True Positives" (TP), "False Positives" (FP), and "False Negatives" (FN). We consider target as ground truth and copy source over as a "generated" target. We essentially consider positives as words that are generated and negatives as words that are not generated, with truth values corresponding to whether or not a word *should* have been generated.

POS Tags POS tags have been used extensively in the stylistic analysis of text, including formality (Pavlick and Tetreault, 2016) and written-style vs. audio-style (Abu-Jbara et al., 2011). Granularity of POS tags has stylistic implications, such as implications for different specific punctuation types (Strunk and White, 1999), so we include Universal and Treebank POS tags for course-grained and fine-grained stylistic information, respectively. [14]

Both Universal and Treebank POS tags are processed using Stanza (Qi et al., 2020), which correspond with the Universal Dependencies (McDonald et al., 2013) POS tags and the Penn Treebank (Marcus et al., 1993) English POS tagset.

Sentence Length Sentence length has stylistic implications (Strunk and White, 1999) and has been used as a feature to classify various styles, such as written-style and audio style (Abu-Jbara et al., 2011) and formality (Pavlick and Tetreault, 2016). We include sentence length in words and sentence length in tokens to account for punctuation differences.

Phrases Measures of phrases and clauses have been used for stylistic analysis in terms of syntactic complexity (Abu-Jbara et al., 2011). We include measures of noun phrases, verb phrases, and dependent clauses.

Readability We adopt the readability measures Flesch-Kincaid Grade Level score (Pavlick and Tetreault, 2016) and ratio of complex words (Abu-Jbara et al., 2011) from prior studies.

Subjectivity We adopted several measures of subjectivity from Pavlick and Tetreault (2016) and adapted the measure ratio of pronouns (Abu-Jbara et al., 2011) by measuring the individual type counts of 1st, 2nd, and 3rd person pronouns.

Bag-of-Words We include the bag-of-words feature to account for cross-class vocabulary differences.

D Jensen-Shannon Divergence

While we indicate large Jensen-Shannon Divergences in Table 5, we include the full range of Jensen-Shannon Divergence results in Table 6 in a numerical format as well.

	Flick	Shake	GY-FR	GY-EM	Bias	Flu.
FF	0.022	0.019	0.019	0.027	0.003	0.004
LexC	**0.086**	0.039	**0.132**	0.054	0.047	0.004
Read	**0.081**	0.040	**0.079**	0.056	0.050	0.013
LexD	0.067	0.049	0.041	0.050	0.031	**0.108**
UPOS	**0.088**	0.052	0.066	**0.075**	0.034	0.011
XPOS	0.063	0.042	0.052	0.056	0.026	0.008
SenL	**0.137**	**0.090**	0.070	0.062	0.013	0.017
Phr	**0.105**	0.056	0.064	0.065	0.030	0.024
Sub	**0.107**	**0.075**	0.054	0.057	0.064	0.016
BoW	0.018	0.015	0.013	0.011	0.002	0.002

Table 6: Jensen-Shannon divergence between source and target on each test set using feature groupings in Table 4. Scores ≥ 0.075 are made bold.

[14] Although we used state-of-the-art tools to extract features such as part-of-speech tags, we do note the possibility of tool performance differences across datasets (Søgaard et al., 2014). However, as we utilize the same tool for both the classification and ablation study as well as the divergence scores, we expect the impact of tool performance within a dataset to have minimal impact on resulting conclusions.