SURVEY PAPER

From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey

Enrica Troiano[1,†], Aswathy Velutharambath[1,2] and Roman Klinger[1,*]

[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany and [2]100 Worte Sprachanalyse GmbH, Heilbronn, Germany
*Corresponding author. E-mail: roman.klinger@ims.uni-stuttgart.de

Abstract

Humans are naturally endowed with the ability to write in a particular style. They can, for instance, rephrase a formal letter in an informal way, convey a literal message with the use of figures of speech or edit a novel by mimicking the style of some well-known authors. Automating this form of creativity constitutes the goal of style transfer. As a natural language generation task, style transfer aims at rewriting existing texts, and specifically, it creates paraphrases that exhibit some desired stylistic attributes. From a practical perspective, it envisions beneficial applications, like chatbots that modulate their communicative style to appear empathetic, or systems that automatically simplify technical articles for a non-expert audience.

Several style-aware paraphrasing methods have attempted to tackle style transfer. A handful of surveys give a methodological overview of the field, but they do not support researchers to focus on specific styles. With this paper, we aim at providing a comprehensive discussion of the styles that have received attention in the transfer task. We organize them in a hierarchy, highlighting the challenges for the definition of each of them and pointing out gaps in the current research landscape. The hierarchy comprises two main groups. One encompasses styles that people modulate arbitrarily, along the lines of registers and genres. The other group corresponds to unintentionally expressed styles, due to an author's personal characteristics. Hence, our review shows how these groups relate to one another and where specific styles, including some that have not yet been explored, belong in the hierarchy. Moreover, we summarize the methods employed for different stylistic families, hinting researchers towards those that would be the most fitting for future research.

Keywords: Natural Language Generation; Translation Technology; Style Transfer; Paraphrasing

1. Introduction

Communication comes in a style. Be it in language, visual arts or music, the things that people express have a *content*—what is to be conveyed, and a *style*—how that is done. These two concepts are evident in the Shakespearean verses "*By the pricking of my thumbs, Something wicked this way comes*" (Macbeth, Act 4, Scene 1.), where the content (i.e., the foreseeing of an evil future) is encoded in the slant rhyme with peculiar rhythm and unusual vocabulary choices. Style is thus the form given to a core piece of information, which collocates it into some distinctive communicative categories. For instance, we perceive that the above example is a poem, and specifically, one written in an old variety of English.

[†]Enrica Troiano and Aswathy Velutharambath contributed equally.

The binomial of content and style is interesting from a computational perspective because content can be styled in a controlled manner. By considering these two variables, many studies have dealt with the automatic generation of texts (Gatt and Krahmer 2018), images (Wu, Xu, and Hall 2017), and music (Briot, Hadjeres, and Pachet 2020) that display a number of desired features. Works as such create content from scratch and combine it with style, while a kin line of research transforms styles starting from an already existing piece of content. The rationale is: if style and content are two and separate, one can be modified and the other kept unaltered. This practice is pervasive among humans as well. It can be observed, for instance, any time they give an inventive twist to their utterances and creations (e.g., when conveying a literal gist through a metaphor, or when painting by imitating Van Gogh's singular brush strokes). The field of vision has achieved remarkable success in changing the styles of images (Gatys, Ecker, and Bethge 2016), and following its footsteps, natural language processing (NLP) has risen to the challenge of style transfer in text.

1.1 Style transfer in text: task definition

The goal of textual style transfer is to modify the style of texts while maintaining their initial content (i.e., their main meaning). More precisely, style transfer requires the learning of $p(t' \mid s, t)$: a text t' has to be produced given the input t and a desired stylistic attribute s, where s indicates either the presence or the absence of such an attribute[a] with respect to t. For example, if t is written in a formal language, like the sentence "*Please, let us know of your needs*", then s may represent the opposite (i.e., *in*formality), thus requiring t' to shift towards a more casual tone, such as "*What do you want?*". Therefore, style transfer represents an effort towards conditioned language generation and yet differs from this broader task fundamentally. While the latter creates text and imposes constraints over its stylistic characteristics alone, the style transfer constraints relate to both style, which has to be different between input and output, and content, which has to be similar between the two—for some definition of "similar". In short, a successful style transfer output checks three criteria. It should exhibit a different stylistic attribute than the source text t, it needs to preserve its content, and it has to read as a human production (Mir *et al.* 2019).

1.2 Applications and challenges

Style transfer lends itself well for several applications. For one thing, it supports automatic linguistic creativity, which has a practical entertainment value. Moreover, since it simulates humans' ability to switch between different communicative styles, it can enable dialogue agents to customize their textual responses for the users and to pick the one that is appropriate in the given situation (Gao *et al.* 2019). Systems capable of style transfer could also improve the readability of texts by paraphrasing them in simpler terms (Cao *et al.* 2020) and help in this way non-native speakers (Wang *et al.* 2019b).

The transfer in text has been tackled with multiple styles (e.g., formality and sentiment) and different attributes thereof (e.g., formal vs. informal, sentiment gradations). Nevertheless, advances in these directions are currently hampered by a lack of appropriate data. Learning the task on human-written linguistic variations would be ideal, but writers hardly produce parallel texts with similar content and diverse attributes. If available, resources of this sort might be unusable due to the mismatch between the vocabularies of the source and target sides (Pang, 2019b), and constructing them requires expensive annotation efforts (Gong *et al.* 2019).

The goal of style transfer seems particularly arduous to achieve per se. Most of the time, meaning preservation comes at the cost of only minimal changes in style (Wu *et al.* 2019a), and bold stylistic shifts tend to sacrifice the readability of the output (Helbig, Troiano, and Klinger 2020).

---

[a]We call "attribute" the value (e.g., presence, absence, degree) that a specific style (e.g., formality) can take.

This problem is exacerbated by a lack of standardized evaluation protocols, which makes the adopted methods difficult to compare. In addition, automatic metrics to assess *content preservation* (i.e., if the input semantics is preserved), *transfer accuracy/strength* (i.e., if the intended attribute is achieved through the transfer), and *fluency or naturalness* (i.e., if the generated text appears natural) (Pang and Gimpel 2019; Mir *et al.*, 2019) often misrepresent the actual quality of the output. As a consequence, expensive human-assisted evaluations turn out inevitable (Briakou *et al.* 2021a, 2021b).

1.3 Purpose and scope of this survey

With the spurt of deep learning, style transfer has become a collective enterprise in NLP (Hu *et al.* 2022; Jin *et al.* 2022). Much work has explored techniques that separate style from content and has investigated the efficacy of different systems that share some basic workflow components. Typically, a style transfer pipeline comprises an encoder-decoder architecture inducing the target attribute on a latent representation of the input, either directly (Dai *et al.* 2019) or after the initial attribute has been stripped away (Cheng *et al.* 2020a). Different frameworks have been formulated on top of this architecture, ranging from lexical substitutions (Li *et al.* 2018; Wu *et al.* 2019b) to machine translation (Jin *et al.* 2019; Mishra, Tater, and Sankaranarayanan 2019) and adversarial techniques (Pang and Gimpel 2019; Lai *et al.* 2019). Therefore, the time seems ripe for a survey of the task, and with this paper, we contribute to organizing the existing body of knowledge around it.

The recurring approaches to style transfer make it reasonable to review its methods, but there already exist three surveys that do so (Toshevska and Gievska 2021; Hu et al., 2022; Jin et al., 2022). They take a technical perspective and focus on the *methods* used to transfer styles. Automatic metrics and evaluation practices have been discussed as well in previous publications (Briakou *et al.* 2021a, 2021b). We move to a different and complementary angle which puts focus on the *styles* to be transferred. Our leading motive is a question that is rooted in the field but is rarely faced: *Can all textual styles be changed or transferred?*

Current publications in the field see style transfer by and large from an engineering angle, aiming at acceptable scores for the three style transfer criteria, and comparing their numerical results in a limited fashion: they neglect the peculiarities of the styles that they are transferring. In our view, each style requires robust understanding in itself, as a pre-requisite for the applied transfer models' choice and success. We thus provide a detailed look into both well-established styles, and those that remain under-explored in the literature. Instead of asking *Is that method advantageous for style transfer?*, we are interested in questions like *How well does it perform when dealing with a particular style?* and *Is finding a balance between naturalness, transfer, and content preservation equally difficult for all styles?* In this vein, we propose a hierarchy of styles that showcases how they relate to each other. We not only characterize them separately and by tapping on some insights coming from humanity-related disciplines,[b] but we also illustrate how they have been handled in the context of style transfer, covering the challenges that they pose (e.g., lack of data), their potential applications, and the methods that have been employed for each of them. Further, we observe if such models have been evaluated in different ways (some of which could fit a style more than others), and lastly, we consider how well styles have been transferred with respect to the three style transfer criteria. Our hierarchy incorporates a selection of papers published from 2008 to September 2021 that we found relevant because of their use or development of datasets for the task at hand, for their proposal of methods that later became well-established in the field, or alternatively, for their proposed evaluation measures. A few of these studies tackle Chinese (Su *et al.*

---

[b] Also Jin et al. (2022) compare various styles and their respective definitions, but in a data-driven approach, as features that vary across datasets.

2017; Shang *et al.* 2019), a handful of them deal with multilingual style transfer (Niu, Rao, and Carpuat, 2018; Briakou *et al.* 2021c), but most works address style transfer for English.

The paper is structured as follows. Section 2 summarizes the technical approaches to this task, covering also some recurring evaluation techniques. Our main contribution, organizing styles in a hierarchy, is outlined in Section 3 (with details in Sections 4 and 5). These discussions include descriptions of data, methods, as well as the evaluations employed for their transfer performance. Section 6 concludes this work and indicates possible directions for future research.

1.4 Intended audience

This survey is addressed to the reader seeking an overview of the state of affairs for different styles that undergo transfer. Specifically, we aim for the following.

Readers needing a sharp focus on a specific style. We revise what has been done within the scope of each style, which could hardly be found in works with a more methodological flavor.

Readers preparing for upcoming style transfer studies, interested in the research gaps within the style transfer landscape. On the one hand, this review can help researchers categorize future work among the massive amount produced in this field, indicating similar works to which they can compare their own. This can eventually guide researchers to decide on the appropriate models for their specific case. On the other hand, we suggest possible "new" styles that were not treated yet but which have an affinity to the existing ones.

Readers questioning the relationship between content and style. NLP has fallen short in asking what textual features can be taken as a style and has directly focused on applying transfer procedures—often generating not too satisfying output texts. Without embarking on the ambitious goal of defining the concept of "style", we systematize those present in NLP along some theoretically motivated coordinates.

2. Style transfer methods and evaluation

Our survey focuses on styles and relations among them. To connect the theoretical discussion with the methodological approaches to transfer, we now briefly describe the field from a technical perspective. We point the readers to Jin et al. (2022), Hu et al. (2022) and Toshevska and Gievska (2021) for a comprehensive clustering and review of the existing methods, and to Prabhumoye, Black, and Salakhutdinov (2020) for a high-level overview of the techniques employed in controlled text generation, style transfer included.

Methodological choices typically depend on what data is available. In the ideal scenario, the transfer system can directly observe the linguistic realization of different stylistic attributes on parallel data. However, parallel data cannot be easily found or created for all styles. On the other hand, mono-style corpora that are representative of the attributes of concern might be accessible (e.g., datasets of texts written for children and datasets of scholarly papers), but they might have little content overlap—thus making the learning of content preservation particularly challenging (Romanov *et al.* 2019). Therefore, we group style transfer methods according to these types of corpora, that is, parallel resources [either ready to use (Xu *et al.* 2012; Rao and Tetreault 2018, i.a.) or created via data augmentation strategies (Zhang, Ge, and Sun 2020b, i.a.), and mono-style datasets (Shen *et al.* 2017; Li *et al.*, 2018; John *et al.* 2019, i.a.)]. As illustrated in Figure 1, which adapts the taxonomy of methods presented in Hu *et al.* (2022), the two groups are further divided into subcategories with respect to the training techniques adopted to learn the task.

Throughout the paper, such methods are reported to organize the literature in Tables 1, 3, 5, 7 and 9, which inform the reader about the approach that each study has taken for a given style, the approaches that have not yet been leveraged for it (i.e., no author is reported in a cell of a table), and those that have been indiscriminately applied for multiple styles (e.g., the same authors appear more than once in a table, or appear in many of them).
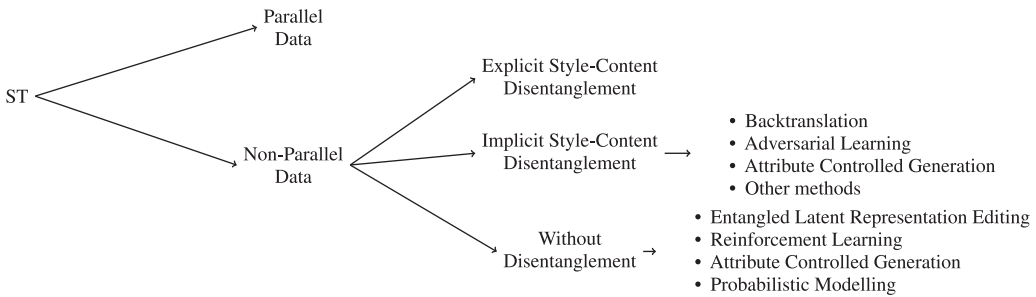
Figure 1. Methods discussed in previous style transfer surveys, adapted from Hu et al. (2022). In contrast, our contribution is the inspection of styles depicted in Figure 2.

2.1 Parallel data

A parallel corpus for transfer would contain texts with a particular stylistic attribute on one side (e.g., formal texts) and paraphrases with a different attribute on the other (e.g., informal texts). When such datasets exist, style transfer can be approached as a translation problem that maps one attribute into the other. Using a corpus of Shakespearean texts and their modern English equivalents, Xu *et al.* (2012) demonstrated the feasibility of style-conditioned paraphrasing with phrase-based machine translation. Later, neural models started to be trained to capture fine stylistic differences between the source and the target sentences, one instance at a time. Jhamtani *et al.* (2017), for example, improved the transfer performance on the Shakespearean dataset by training a sequence-to-sequence architecture with a pointer network that copies some words from the input. Rao and Tetreault (2018) corroborated that machine translation techniques are a strong baseline for style transfer on the Grammarly's Yahoo Answers Formality Corpus, a parallel corpus for formality transfer which turned out to drive the majority of the style transfer research on parallel data (leveraged by Niu *et al.* 2018; Wang *et al.* 2019b; Xu, Ge, and Wei 2019b, among others).

Sequence-to-sequence models achieved remarkable results in conjunction with different style controlling strategies, like multi-task learning (Niu *et al.* 2018; Xu *et al.* 2019b), rule harnessing (Wang *et al.* 2019b), post-editing with grammatical error correction (Ge *et al.* 2019), and latent space sharing with matching losses (Wang *et al.* 2020). Parallel resources, however, are scarce or limited in size. This has triggered a number of attempts to synthesize parallel examples. Zhang *et al.* (2020b) and Jin *et al.* (2019) exemplify this effort. While the former augmented data with translation techniques (i.e., backtranslation and backtranslation with a style discriminator) and a multi-task transfer framework, Jin *et al.* (2019) derived a pseudo-parallel corpus from mono-style corpora in an iterative procedure, by aligning sentences which are semantically similar, training a translation model to learn the transfer, and using such translations to refine the alignments in return.

2.2 Non-parallel data

The paucity of parallel resources also encouraged transfer strategies to develop on mono-style corpora (i.e., non-parallel corpora of texts that display one or more attributes of a specific style). This research line mainly approached the task intending to disentangle style and content, either by focusing the paraphrasing edits on the style-bearing portions of the input texts, or by reducing the presence of stylistic information into the texts' latent representations. On the other hand, a few studies claimed that such disentanglement can be avoided. Therefore, methods working with non-parallel data can be divided into those which do style transfer with an explicit or implicit style-to-content separation and those which operate with no separation.

2.2.1 Explicit style-content disentanglement

Some styles have specific markers in text: expressions like "*could you please*" or "*kindly*" are more typical of a formal text than an informal one. This observation motivated a spurt of studies to alter texts at the level of explicit markers—which are replaced in the generated sentences by the markers of a different attribute. The first step of many such studies is to find a comprehensive inventory of style-bearing words. Strategies devised with this goal include frequency statistics-based methods (Li et al., 2018; Madaan *et al.* 2020), lexica (Wen *et al.* 2020), attention scores of a style classifier (Xu *et al.* 2018; Sudhakar, Upadhyay, and Maheswaran 2019; Helbig *et al.* 2020; Reid and Zhong 2021), or combinations of them (Wu *et al.* 2019b; Lee 2020). As an alternative, Malmi *et al.* (2020) identified spans of text on which masked language models (Devlin *et al.* 2019), trained on source and target domains, disagree in terms of likelihood: these would be the portions of a sentence responsible for its style, and their removal would produce a style-agnostic representation for the input.

Candidate expressions are then retrieved to replace the source markers with expressions of the target attribute. Distance metrics used to this end are (weighted) word overlap (Li *et al.*, 2018), Euclidean distance (Li *et al.*, 2018), and cosine similarity between sentence representations like content embeddings (Li *et al.*, 2018), weighted TF-IDF vectors, and averaged GloVe vectors over all tokens (Sudhakar *et al.* 2019). Some studies resorted instead to WordNet-based retrievals (Helbig *et al.* 2020).

In the last step, (mostly) neural models combine the retrieved tokens with the style-devoid representation of the input, thus obtaining an output with the intended attribute. There are also approaches that skip this step and directly train a generator to produce sentences in the target attribute based on a template (Lee 2020, i.a.). Similar techniques for explicit keyword replacements are relatively easy to train and are more explainable than many other methods, like adversarial ones (Madaan *et al.*, 2020).

2.2.2 Implicit style-content disentanglement

Approaches for explicit disentanglement cannot be extended to all styles because many of them are too complex and nuanced to be reduced to keyword-level markers. Methods for implicit disentanglement overcome this issue. Their idea is to strip the input style away by operating on the latent representations (rather than at the text level). This usually involves an encoder–decoder architecture. The encoder produces the latent representation of the input and the decoder, which generates text, is guided by training losses controlling for the style and content of the output.

Adversarial learning.   Implicit disentanglement has been instantiated by adversarial learning in several ways. To ensure that the representation found by the encoder is devoid of any style-related information, Fu *et al.* (2018) trained a style classifier adversarially, making it unable to recognize the input attribute, while Lin *et al.* (2020) applied adversarial techniques to decompose the latent representation into a style code and a content code, demonstrating the feasibility of a one-to-many framework (i.e., one input, many variants). John et al. (2019) inferred embeddings for both content and style from the data, with the help of adversarial loss terms that deterred the content space and the style space from containing information about one another, and with a generator that reconstructed input sentences after the words carrying style were manually removed. Note that, since John et al. (2019) approximated content with words that do not bear sentiment information, they could also fit under the group of Explicit Style-Content Disentanglement. We include them here because the authors themselves noted that ignoring sentiment words can boost the transfer, but is not essential.

Backtranslation.   A whole wave of research banked on the observation that backtranslation washes out some stylistic traits of texts (Rabinovich *et al.* 2017) and followed the work of

Prabhumoye *et al.* (2018b). There, input sentences were translated into a pivot language and back as a way to manipulate their attributes: the target values were imposed in the backward direction, namely, when decoding the latent representation of the (pivot language) text, thus generating styled paraphrases of the input (in the source language).

Attribute controlled generation. Attribute control proved to be handy to produce style-less representations of the content while learning a code for the stylistic attribute. This emerges, for instance, in Hu *et al.* (2017), who leveraged a variational auto-encoder and some style discriminators to isolate the latent representation and the style codes, which were then fed into a decoder. While the discriminators elicited the disentanglement, the constraint that the representation of source and target sentence should remain close to each other favored content preservation.

Other methods. An alternative path to disentanglement stems from information theory. Cheng *et al.* (2020a) defined an objective based on the concepts of mutual information and variation of information as ways to measure the dependency between two random variables (i.e., style and content). On the one hand, the authors minimized the mutual information upper bound between content and style to reduce their interdependency; on the other, they maximized the mutual information between latent embeddings and input sentences, ensuring that sufficient textual information was preserved.

2.2.3 Without disentanglement

By abandoning the disentanglement venture, some studies argued that separating the style of a text from its content is not only difficult to achieve—given the fuzzy boundary between the two, but also superfluous (Lample *et al.* 2019). This observation became the core of a wave of research that can be categorized as follows.

Entangled latent representation editing. Some works edited the latent representations of the input texts learned by an auto-encoder. A common practice in this direction is to jointly train a style classifier and iteratively update the auto-encoder latent representation by maximizing the confidence on the classification of the target attribute (Mueller, Gifford, and Jaakkola 2017; Liu *et al.* 2020a). Another approach trained a multi-task learning model on a summarization and an auto-encoding task, and it employed layer normalization and a style-guided encoder attention using the transformer architecture (Wang, Hua, and Wan 2019a).

Attribute controlled generation. Proven successful by disentanglement-based studies, methods for learning attribute codes were also applied without the content-vs.-style separation. Lample et al. (2019), for instance, employed a denoising auto-encoder together with backtranslation and an averaged attribute embedding vector, which controlled for the presence of the target attribute during generation. Instead of averaging the one-hot encoding for individual attribute values, Smith *et al.* (2019) used supervised distributed embeddings to leverage similarities between different attributes and perform zero-shot transfer.

Reinforcement learning. Multiple training loss terms have been defined in style transfer to endow the output texts with the three desiderata of content preservation, transfer accuracy, and text naturalness—often referred to as "fluency". The dependency on differentiable objectives can be bypassed with reinforcement learning, which uses carefully designed training rewards (Luo *et al.* 2019a, i.a.). Generally, rewards that cope with the presence of the target attribute are based on some style classifiers or discriminators, those pertaining to naturalness rely on language models; and those related to content preservation use Bleu or similar metrics that compare an output text against some reference.

Gong et al. (2019) worked in a generator-evaluator setup. There, the generator's output was probed by an evaluator module, whose feedback helped improve the output attribute, semantics, and fluency. Two building blocks can also be found in Luo *et al.* (2019b). They approached style transfer as a dual task (i.e., source-to-target and target-to-source mappings) in which, to warm-up the reinforcement learning training, a model was initially trained on a pseudo-parallel corpus. Wu *et al.* (2019a), instead, explored a sequence operation method called Point-Then-Operate, with a high-level agent dictating the text position where the operations should be done and a low-level agent performing them. Their policy-based training algorithm employed extrinsic and intrinsic rewards, as well as a self-supervised loss to model the three transfer desiderata. The model turned out relatively interpretable thanks to these explicitly defined operation steps. Tuning their number, in addition, allowed to control the trade-off between the presence of the initial content and of the target attribute.

An exception among reinforcement learning studies is the cycled reinforcement learning of Xu et al. (2018), which falls within the disentangling picture.

Probabilistic modelling. Despite being a common practice in unsupervised learning, the definition of task-specific losses can lead to training instability. These objectives are empirically determined among a vast number of possible alternatives. To overcome the issue, He *et al.* (2020) formulated a probabilistic generative strategy that follows objectives defined by some principles of probabilistic inference, and which makes clear assumptions about the data. This approach allowed them to reason transparently about their system design and to outperform many works choosing ad-hoc training objectives.

2.3 Evaluation

The methods presented above are usually assessed with metrics that quantify content preservation, transfer accuracy/intensity, and generation of natural-sounding paraphrases. A detailed discussion of the evaluation methods can be found in Mir et al. (2019), Pang (2019a), Briakou *et al.* (2021a) and Briakou *et al.* (2021b), with the latter focusing on human evaluation settings. As they appear in most style transfer publications, we briefly introduce them here and will refer back to them throughout the paper.

Content preservation, i.e., the degree to which an output retains the content of the input, is usually gauged with measures that originated in machine translation. They compute the overlap between the words of the generation system and some reference texts, under the assumption that the two should share much lexical material. Among them are Bleu (Papineni *et al.* 2002) and Meteor (Banerjee and Lavie 2005), often complemented with Rouge (Lin 2004), initially a measure for automatic summaries. Transfer accuracy, i.e., the efficacy of the models in varying stylistic attributes, is usually scored by classifiers: trained on a dataset characterized by the style in question, a classifier can tell if an output text has the target attribute or not. Applied on a large scale, this second criterion can be quantified as the percentage of texts that exhibit the desired attribute. Last comes the naturalness or fluency of the variants that have been changed in style. This is typically estimated with the perplexity of language models, indicating the degree to which a sequence of words in a paraphrase is predictable—hence grammatical.

Focusing on automatic content preservation, Tikhonov *et al.* (2019) advocated that Bleu should be used with some caution in style transfer. They argued that the entanglement between semantics and style in natural language is reflected in the entanglement between the Bleu score measured between input and output and the transfer accuracy. Indeed, they provided evidence that such measures can be easily manipulated: the outputs that a classifier in the generative architecture indicates as having the incorrect attribute could be replaced with sentences which are most similar to the input in their surface form—thus boosting both the reported accuracy and Bleu. Human-written reformulations are necessary in their view for upcoming experiments, as current style

transfer architectures become more sophisticated, and therefore, accuracy and Bleu might be too naive metrics to estimate their performance. Going in a similar direction, the extensive meta-analysis of Briakou *et al.* (2021a) discusses the pitfalls of automatic methods and the need for standardized evaluation practices (including human evaluation) to boost advance in this field.

3. Style hierarchy

Style transfer relies on a conceptual distinction between meaning and form (e.g., De Saussure 1959), but what is this form? It is a dimension of sociolinguistic variation that manifests in syntactic and lexical patterns, that can be correlated with independent variables and that, according to Bell (1984), we shift in order to fit an audience. Bell's characterization emphasizes the intentionality of language variation, accounting only for the styles ingrained in texts out of purpose. Yet, many others emerge as a fingerprint of the authors' identities, for instance from specific markers of people's personality and internal states (Brennan, Afroz, and Greenstadt 2012). This already suggests that different styles have diverse characteristics. However, their peculiar challenges have received little attention in the literature. As a remedy for the lacuna, we bring style transfer closer to the linguistic and sociological theories on the phenomenon it targets. We propose a hierarchy of styles in which we place the relevant body of NLP research.

A recent study by Kang and Hovy (2021) actually groups styles into a handful of categories (*personal*, *interpersonal*, *figurative* and *affective*) based on some social goals achieved through communication. Their work did not investigate specific styles. It rather intended to fertilize research towards a cross-style direction, by combining existing corpora into an overarching collection of 15 styles.[c] By contrast, our hierarchy concentrates on the peculiarities of styles separately, while indicating the methods that have been used and those that have been dismissed for each of them.

To unify the above-mentioned theoretical insights, we make a first, coarse separation between accidental and voluntary styles, structuring them into the *unintended* and *intended* families.[d] The former group copes with the self. It corresponds to the personal characteristics of the authors, which we split into factors that define *between-persons* and *within-person* language variations. Namely, there are stable traits defining systematic differences between writers and short-term internal changes within an individual subject which, in response to situations, do not persist over time (Beckmann and Wood 2017). We call them *persona* and *dynamic states* respectively. The other category of styles is *intended*, as it covers deliberate linguistic choices with which authors adapt to their communicative purpose or environment. Style transfer publications that fall within this group echo what is known as "palimpsest" in literary theories, i.e., the subversion of a text into a pastiche or a parody to imitate an author, degrade a text, or amplify its content (Genette 1997). Among these are styles used to express how one feels about the topic of discussion: a speaker/writer can have a positive sentiment *on* a certain matter, be angry or sad *at* it, be sarcastic *about* it, etc. Of this type are styles *targeted* towards a topic, while others, the *non-targeted* subset, are more independent of it. Some (*circumstantial registers*) are rather dependent on the context in which they are deployed, and they convey a general attitude of the writers, a tone in which they talk or a social posture—an example being formality, that speakers increase if they perceive their interlocutor as socially superior (Vanecek and Dressler 1975). Other styles are socially coded. They can be thought of as *conventional* writing styles tailored to the ideal addressee of the message rather than an actual one, and are typically employed in mass communication, such as scientific, literary, and technical productions.

---

[c]Their data, part of which is annotated with 15 styles, is available at https://github.com/dykang/xslue

[d]We do not include works on the transfer of syntax, such as Bao *et al.* (2019), Balasubramanian *et al.* (2021) and Lyu *et al.* (2021).

These categories subsume a number of individual styles. For instance, *persona* branches out into *personality traits*, *gender and age*, and *background*, which in turn encompasses *country and ethnicity, education, and culture*. Note that the leaves in our hierarchy are the major styles that have been addressed so far by automatic systems, but many others can be identified and explored in future work. We include some in our discussions. Furthermore, we acknowledge that a few styles pertain to both the *unintended* and *intended* branches. Our motivation to insert them under one rather than the other is due to the type of data on which the transfer was made (e.g., *emotion state*) or to how the problem was phrased by the corresponding studies (e.g., *literature*).

The remainder of this paper follows the structure of our hierarchy. We provide a top-down discussion of the nodes, starting from the high-level ones, which are presented from a theoretical perspective, and proceeding towards the leaves of the branches, which is where the concrete style transfer works are examined in relation to the data, the methods and the evaluation procedures that they used.

4. Unintended styles

Writers leave traces of their personal data. Information like one's mental disposition, biological, and social status are revealed by stylometric cues present in a text. These cues might be produced unknowingly, and because of that, they could help to combat plagiarism, foster forensics, and support humanities. On the other hand, accessing knowledge about writers could breach people's privacy and exacerbate demographic discrimination. Hence, while classification-based studies leveraged such latent information to profile people's age and gender (Rosenthal and McKeown 2011; Nguyen *et al.* 2013; Sarawgi, Gajulapalli, and Choi 2011; Fink, Kopecky, and Morawski 2012), geolocation, and personality (Eisenstein *et al.* 2010; Verhoeven, Daelemans, and Plank 2016; Plank and Hovy 2015), the attempt to defeat authorship recognition moved research towards the transfer of such unintended styles—i.e., age, gender, etc.

Arguably the first work to address this problem is that of Brennan *et al.* (2012), who tried to confound stylometric analyses by backtranslating existing texts with available translation services, such as Google Translate and Bing Translator. Their preliminary results did not prove successful, as the writer's identity remained recognizable through the translation passages from source to targets and back, but follow-up research provided evidence that automatic profilers can be effectively fooled (Kacmarcik and Gamon 2006; Emmery, Manjavacas Arevalo, and Chrupała 2018; Shetty, Schiele, and Fritz 2018; Bo *et al.* 2021, i.a.).

Successive style transfer studies narrowed down the considered authors' traits. They tackled stable features that are a proxy for the writers' biography, which we subsume under the category of *persona*, or more *dynamic states* that characterize writers at a specific place and time. It should be noticed that such works rely on a tacit assumption about writers' authenticity: writers express themselves spontaneously and do not attempt to mask their own traits (Brennan *et al.* 2012).

We illustrate the methods used to transfer *unintended* styles in Table 1.

4.1 Persona

Persona includes biographic attributes coping with personality and people's social identity. Individuals construct themselves "as girls or boys, women or men—but also as, e.g., Asian American" (Eckert and McConnell-Ginet 1999), that is, they often form an idea of the self as belonging to a group with a shared enterprise or interest (Tajfel 1974). The interaction within such a group also affects their linguistic habits (Lave and Wenger 1991) as they develop a similar way of talking. In this sense, linguistic style is a key component of one's identity (Giles and Johnson 1987). It manifests some traits unique to a specific person or community (Mendoza-Denton and Iwai (1993) provide insights on the topic with respect to the Asian-American English speech).

Table 1. Style transfer methods and the *unintended* styles of *persona*

| | Parallel | Non-parallel | | |
		Exp. Disent.	Imp. Disent.	No Disent.
Gender	Kang et al. (2019)	Reddy and Knight (2016); Sudhakar (2019); Madaan et al. (2020)	Prabhumoye et al. (2018a, 2018b); Nangi et al. (2021)	Preoţiuc-Pietro et al. (2016); Lample et al. (2019); Liu et al. (2020b)
Age	Kang et al. (2019)	–	–	Lample et al. (2019)
Personality traits.	Bujnowski et al. (2020)	–	Cheng et al. (2020a)	–
Country	Kang et al. (2019)	–	–	Riley et al. (2021)
English variety	–	–	Logeswaran et al. (2018)	Lee et al. (2019); Krishna et al. (2020)
Ethnicity	Kang et al. (2019)	–	–	Krishna et al. (2020)
Education	Kang et al. (2019)	–	–	–

At least to a degree, *persona* styles are implicit in the way people express themselves. As opposed to the *intended* branch of our hierarchy, they are not communicative strategies consciously set in place by the writers, but they are spontaneous indicators of other variables. For instance, it has been shown that women tend to use paralinguistic signals more often than men (Carli 1990), that speakers' vocabulary becomes more positively connoted and less self-referenced in older ages (Pennebaker and Stone 2003), and that subcultures express themselves with a specific slang (Bucholtz 2006).

The transfer of *persona* aims to go from one attribute to the other (e.g., young to old for the style of age), and its main challenge is that different styles are closely intertwined. Age and gender, for instance, can imply each other because "the appropriate age for cultural events often differs for males and females" (Eckert 1997), and therefore, one may not be changed without altering the other. Moreover, there is still a number of styles dealing with people's communicative behaviors and skills which are left unexplored. Future studies could focus on those, like the pairs playful vs. aggressive, talkative vs. minimally responsive, cooperative vs. antagonist, dominant vs. subject, attentive vs. careless, charismatic vs. uninteresting, native vs. L2 speaker, curious vs. uninterested, avoidant vs. involved.

4.1.1 Gender and age

Style transfer treats gender and age as biological facts. The transfer usually includes a mapping between discrete labels: from male to female or vice versa, and from young to old or the other way around (see some examples in Table 2). It should be noticed that such labels disregard the fluidity of one's gender experience and performance, which would be better described along a spectrum (Eckert and McConnell-Ginet 2003), and they represent age as a chronological variable rather than a social one depending on peoples' personal experiences (Eckert 1997). This simplification is not made by style transfer specifically, but it is common to many studies focused on authors' traits, due to how the available datasets were constructed—e.g., in gender-centric resources, labels are inferred from the name of the texts' authors (Mislove *et al.* 2011).

The Rt-Gender corpus created by Voigt *et al.* (2018) stands out among such resources. It was built to research how responses towards a specific gender differ from responses directed to another, in opposition to related corpora that collect linguistic differences between genders. This labeled dataset potentially sets the ground for the next steps in style transfer.

Table 2. Examples of style transfer on a subset of *persona* styles. *Personality traits* sentences come from Shuster *et al.* (2019), *gender*-related ones from Sudhakar *et al.* (2019), the *age*-related examples from Preoţiuc-Pietro *et al.* (2016a), and *background*-related examples from Krishna *et al.* (2020). For each pair, the input is above

Gender and age	**Male:** *this is a spot that's making very solid food, with good quality product*
	Female: *this is a cute spot that's making me very happy, with good quality product*
	Young: *hilton worldwide starts its biggest global career event URL #csr*
	Old: *hilton worldwide launches its largest global career event URL #csr*
Personality traits	**Sweet**: *That is a lovely sandwich*
	Dramatic: *This sandwich looks so delicious! My goodness!*
	Money-minded: *I would totally pay $100 for this plate*
	Optimistic: *It will taste positively wonderful*
Background	**1990**: *He was being terrorized into making a statement by the same means as the other so-called "witnesses"*
	1810: *Terror had been employed in the same manner with the other witnesses, to compel him to make a declaration.*
	British: *As the BMA's own study of alternative therapy showed, life is not as simple as that.*
	American: *As the F.D.A.'s own study of alternative therapy showed, life is not as simple as that.*

Data.   Works on *gender* style transfer typically follow the choice of data by Reddy and Knight (2016), who used tweets posted in the US in 2013 and some reviews from the Yelp[e] dataset, and inferred gender information from the users' names.

For this style, there also exists Pastel,[f] a corpus annotated with attributes of both *unintended* and *intended* styles. That is the result of the crowdsourcing effort conducted by Kang, Gangal, and Hovy (2019), in which ≈41K parallel sentences were collected in a multimodal setting, and which were annotated with the gender, age, country, political view, education, ethnicity, and time of writing of their authors.

The need to collect attribute-specific rewrites further motivated Xu, Xu, and Qu (2019a) to create Alter. As a publicly available tool[g], Alter was developed to overcome one major pitfall of crowdsourcing when it comes to generating gold standards: human annotators might fail to associate textual patterns to a gender label, at least when dealing with short pieces of text. Alter facilitates their rewriting tasks (specifically, to generate texts which are not associated with a particular gender) by providing them with immediate feedback.

Methods.   Though not focused on transfer, Preoţiuc-Pietro, Xu, and Ungar (2016a) were the first to show that automatic paraphrases can exhibit the style of writers of different ages and genders, by manipulating the lexical choices made by a text generator. A phrase-based translation model learned that certain sequences of words are more typically used by certain age/gender groups and, together with a language model of the target demographics, it used such information to translate tweets from one group to the other. Their translations turned out to perform lexical substitution, a strategy that was more directly addressed by others. Reddy and Knight (2016), for instance, performed substitution in order to defeat a gender classifier. They did so with the guidance of three metrics: one measured the association between words and the target gender label, thus indicating the words to replace to fool the classifier as well as possible substitutes; another quantified the

---

[e]https://www.yelp.com/dataset
[f]https://github.com/dykang/PASTEL
[g]https://github.com/xuqiongkai/ALTER

semantic and syntactic similarity between the words to be changed and such substitutes; and the last measured the suitability of the latter in context.

A pitfall of such heuristics, noticed by the authors themselves, is that style and content-bearing words are equal candidates for the edit. Some neural methods bypassed the issue with a similar three-step procedure. That is the case of Sudhakar *et al.* (2019), who proposed a variation of the pipeline in Li et al. (2018). There, (1) only style-bearing words are deleted upon the decision of a Bert-based transformer, where an attention head encodes the stylistic importance of each token in a sentence. Next, (2) candidate substitutes are retrieved: sentences from a target-style corpus are extracted to minimize the distance between the content words of the input and theirs. Lastly, (3) the final output is generated with a decoder-only transformer based on Gpt, having learned a representation of both the content source words and the retrieved attribute words. It should be noted that this method was not designed to transfer genre-related attributes specifically (it achieves different results when dealing with other styles). Also, Madaan et al. (2020) addressed *gender* as an ancillary task. They used a similar methodology (further discussed in Section 5.3 under *Politeness*) that first identifies style at the word level, and then changes such words in the output.

Prabhumoye *et al.* (2018b), instead, separated content and style at the level of the latent input representation, by employing backtranslation as both a paraphrasing and an implicit disentangling technique. Since machine translation systems are optimized for adequacy and fluency, using them in a backtranslation framework can produce paraphrases that are likely to satisfy at least two style transfer desiderata (content preservation and naturalness). To change the input attribute and comply with the third criterion, the authors hinged on the assumption that machine translation reduces the stylistic properties of the input sentence and produces an output in which they are less distinguishable. With this rationale, a sentence in the source language was translated into a pivot language; encoding the latter in the backtranslation step then served to produce a style-devoid representation, and the final decoding step conditioned towards a specific gender attribute returned a stylized paraphrase.

Modelling content and style-related personal attributes separately are in direct conflict with the finding by Kang *et al.* (2019), who pointed out that features used for classifying styles are of both types. As opposed to the studies mentioned above, this work transferred multiple *persona* styles in conjunction (e.g., *education* and *gender*) and did so with a sequence-to-sequence model trained on a parallel dataset. Similarly, the proposal of Liu *et al.* (2020b) did not involve any content-to-style separation. With the aim of making style transfer controllable and interpretable, they devised a method based on a variational auto-encoder that performs the task in different steps. It revises the input texts in a continuous space using both gradient information and style predictors, finding an output with the target attribute in such a space.

Evaluation.   While Reddy and Knight (2016) carried out a small preliminary analysis, others assessed the quality of the outputs with (at least some of) the three criteria, both with automatic and human-based studies. For instance, Sudhakar *et al.* (2019) evaluated the success of the transfer with a classifier and quantified fluency in terms of perplexity. For meaning preservation, other than Bleu (Kang *et al.* 2019; Sudhakar *et al.* 2019), *gender* transfer was evaluated automatically with metrics based on n-gram overlaps (e.g., Meteor) and embedding-based similarities between output and reference sentences [e.g., Embedding Average similarity and Vector Extrema of Liu *et al.* (2016), as found in Kang *et al.* (2019)].

Sudhakar *et al.* (2019) also explored Gleu as a metric that better correlates with human judgments. Initially a measure for error correction, Gleu fits the task of style transfer because it is capable of penalizing portions of texts changed inappropriately while rewarding those successfully changed or maintained. As for human evaluation, the authors asked their raters to judge the final output only with respect to fluency and meaning preservation, considering the transfer of *gender* a too challenging dimension to rate. Their judges also evaluated texts devoid of style-related attributes.

4.1.2 Personality traits

The category of *personality traits* contains variables describing characteristics of people that are stable over time, sometimes based on biological facts (Cattell 1946). Studied at first in the field of psychology, personality traits have also been approached in NLP (Plank and Hovy 2015; Rangel *et al.* 2015, i.a.), as they seem to correlate with specific linguistic features—e.g., depressed writers are more prone to using first-person pronouns and words with negative valence (Rude, Gortner, and Pennebaker 2004). This has motivated research to both recognize the authors' traits from their texts (Celli *et al.* 2014) and to infuse them within newly generated text (Mairesse and Walker 2011).

Computational works typically leverage well-established schemas, like the (highly debated) Myers-Briggs Type Indicators (Myers and Myers 2010) and the established Big Five traits (John, Naumann, and Soto 2008). These turn out particularly useful because they qualify people in terms of a handful of dimensions, either binary (introvert-extrovert, intuitive-sensing, thinking-feeling, judging-perceiving) or not (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism).

Accordingly, a style transfer framework would change the attribute value along such dimensions. Some human-produced examples are the switch from the sweet to dramatic type of personality and the transfer money-minded to optimistic in Table 2 (note that not all attributes addressed in style transfer are equally accepted in psychology). More precisely, each dimension represents a different personality-related style, and this makes traits particularly difficult to transfer: the same author can be defined by a certain amount of all traits, while many other styles only have one dimension (e.g., the dimension of polarity for sentiment), with the two extreme attributes being mutually exclusive (i.e., a sentence is either positively polarized or has a negative valence).

The ability to transfer *personality traits* brings clear advantages. For instance, the idea that different profiles associate to different consumer behaviors (Foxall and Goldsmith 1988; Gohary and Hanzaee 2014) may be exploited to automatically tailor products on the needs of buyers; personification algorithms could also improve health care services, such that chatbots communicate sensitive information in a more human-like manner, with a defined personality, fitting that of the patients; further, they can be leveraged in the creation of virtual characters.

Data.   So far, this task explored the collection of image captions crowdsourced by Shuster et al. (2019), who asked annotators to produce a comment for a given image which would evoke a given personality trait. Their dataset Personality-Captions[h] contains 241,858 instances and spans across 215 personality types (e.g., sweet, arrogant, sentimental, argumentative, charming). Note that these variables do not exactly correspond to personality traits established in psychology. As an alternative, one could exploit the corpus made available by Oraby *et al.* (2018), synthesized with a statistical generator. It spans 88k meaning representations of utterances in the restaurant domain and matched reference outputs which display the Big Five personality traits of extraversion, agreeableness, disagreeableness, conscientiousness, and unconsciousness.[i]

Methods.   Cheng *et al.* (2020a) provided evidence that the disentanglement between the content of a text and the authors' personality (where personalities are categorical variables) can take place. Observing that such a disentanglement is in fact arduous to obtain, they proposed a framework based on information theory. Specifically, they quantified the style-content dependence via mutual information, i.e., a metric indicating how dependent two random variables are, in this case measuring the degree to which the learned representations are entangled. Hence, they defined the objective of minimizing the mutual information upper bound (to represent style and content into

---

[h]http://parl.ai/projects/personality_captions
[i]https://nlds.soe.ucsc.edu/stylistic-variation-nlg

two independent spaces) while maximizing their mutual information with respect to the input (to make the two types of embeddings maximally representative of the original text).

Without complying with any psychological models, Bujnowski *et al.* (2020) addressed a task that could belong to this node in our hierarchy. Neutral sentences were transferred into "cute" ones, i.e., excited, positive, and slangy. For that, they trained a multilingual transformer on two parallel datasets, one containing paired mono-style paraphrases and the other containing stylized rewritings, for it to simultaneously learn to paraphrase and apply the transfer.

Evaluation.   Other than typical measures for style (i.e., style classifiers' accuracy) and content (Bleu), Cheng *et al.* (2020a) considered generation quality, i.e., corpus-level Bleu between the generated sentence and the testing data, as well as the geometric mean of these three for an overall evaluation of their system.

4.1.3 Background

Our last *unintended* style of *persona* is the background of writers. Vocabulary choices, grammatical and spelling mistakes, and eventual mixtures of dialect and standard language expose how literate the language user is (Bloomfield 1927); dialect itself, or vernacular varieties, marked by traits like copula presence/absence, verb (un)inflection, use of tense (Green 1998; Martin and Wolfram 1998) can give away the geographical or ethnic provenance of the users (Pennacchiotti and Popescu 2011). Further, because these grammatical markers are prone to changing along with word meanings, language carries evidence about the historical time at which it is uttered (Aitchison 1981).

In this research, streamline are style transfer works leveraging the idea that there is a "style of the time" (Hughes *et al.* 2012): they performed diachronic linguistic variations, thus taking timespans as a transfer dimension (e.g., Krishna, Wieting, and Iyyer (2020) transferred among the 1810–1830, 1890–1910, 1990–2010 attributes). Others applied changes between English varieties, for instance switching from British to American English (Lee *et al.* 2019), as well as varieties linked to ethnicity, like English Tweets to African American English Tweets and vice versa (Krishna *et al.* 2020), or did the transfer between education levels (Kang *et al.* 2019).

The following are example outputs of these tasks, from Krishna *et al.* (2020): "*He was being terrorized into making a statement by the same means as the other so-called "witnesses".*" (1990) → "*Terror had been employed in the same manner with the other witnesses, to compel him to make a declaration.*" (1810); "*As the BMA's own study of alternative therapy showed, life is not as simple as that.*" (British) → "*As the F.D.A.'s own study of alternative therapy showed, life is not as simple as that.*" (American).

Such variations could be applied in real-world scenarios in order to adjust the level of literacy of texts, making them accessible for all readers or better resonating with the culture of a specific audience. Future research could proceed into more diverse background-related styles, such as those which are not shared by all writers at a given time or in a specific culture, but which pertain to the private life of subsets of them. For instance, considering hobbies as a regular activity that shapes how people talk, at least for some types of content, one could rephrase the same message in different ways to better fit the communication with, say, an enthusiast of plants, or rather with an addressee who is into book collecting.

Data.   Sources that have been used for English varieties are the New York Times and the British National Corpus for English (Lee et al., 2019). Krishna *et al.* (2020) employed the corpus of Blodgett, Green, and O'Connor (2016) containing African American Tweets, and included this dialectal information in their own dataset[j]; as for the diachronic variations that they considered,

---

[j]http://style.cs.umass.edu

texts came from the Corpus of Historical American English (Davies 2012). Also the Pastel corpus compiled by Kang *et al.* (2019) contains ethnic information, which covers some fine-grained labels, like Hispanic/Latino, Middle Eastern, Caucasian, and Pacific Islander. Their resource includes data about the education of the annotators involved in the data creation process, from unschooled individuals to PhD holders.

Methods.   Logeswaran, Lee, and Bengio (2018) followed the line of thought that addresses content preservation and attribute transfer with separate losses. They employed an adversarial term to discourage style preservation, and an auto-reconstruction and a backtranslation term to produce content-compatible outputs. Noticing that the auto-reconstruction and backtranslation losses supported the models in copying much of the input, they overcame the issue by interpolating the latent representations of the input and of the generated sentences.

Other methods used for this style are not based on disentanglement techniques (e.g., Kang *et al.* 2019). Among those is the proposal of Lee et al. (2019), who worked under the assumption that the source attribute is a noisy version of the target one, and in that sense, style transfer is a backtranslation task: their models translated from a "clean" input text to their noisy counterpart, and then denoised it towards the target. Krishna *et al.* (2020) fine-tuned pretrained language models on automatically generated paraphrases. They created a pseudo-parallel corpus of stylized-to-neutral pairs and trained different paraphrasing models in an "inverse" way, that is, each of them learns to recover a stylistic attribute by reconstructing the input from the artificially-created and style-devoid paraphrases. Hence, at testing time, different paraphrasers transferred different attributes (given a target attribute, the model trained to reconstruct it was applied).

Evaluation.   Krishna *et al.* (2020) proposed some variations on the typical measures for evaluation, hinging on an extensive survey of evaluation practices. As for content preservation, they moved away from n-gram overlap measures like Bleu which both disfavors diversity in the output and does not highlight style-relevant words over the others. Instead, they automatically assessed content with the subword embedding-based model by Wieting and Gimpel (2018). With respect to fluency, they noticed that perplexity might misrepresent the quality of texts because it can turn out low for sentences simply containing common words. To bypass this problem, they exploited the accuracy of a RoBerta classifier trained on a corpus that contains sentences judged for their grammatical acceptability. Moreover, they jointly optimized automatic metrics by combining accuracy, fluency and similarity at the sentence level, before averaging them at the corpus level.

4.2 Dynamic states

In the group of *dynamic* styles, we arrange a few states in which writers find themselves in particular contexts. Rather than proxies for stable behaviors or past experiences, they are short-lived qualities, which sometimes arise just in response to a cue. Many facts influencing language slip into this category and represent an opportunity for future exploration. Some of them are: the activity performed while communicating (e.g., moving vs. standing); motivational factors that contribute to how people say the things they say (e.g., hunger, satisfaction); positive and negative moods, as they, respectively, induce more abstract, high-level expressions littered with adjectives, and a more analytic style, focused on detailed information that abounds with concrete verbs (Beukeboom and Semin 2006); the type of communication medium, known to translate into how language is used—for instance, virtual exchanges are fragmentary, have specialized typography, and lack linearity (Ferris 2002).

Another ignored but promising avenue is the transfer of authenticity. Authenticity is a dynamic state transversing all the styles we discussed so far, and at the same time defining a style on its own. In the broader sense, it is related to an idea of truth (Newman 2019), as it regards those qualities of texts which allow to identify their author correctly: this is the type of authenticity underlying the

Table 3. Style transfer methods distributed across *unintended dynamic* styles of our hierarchy

	Parallel	Non-parallel		
		Exp. Disent.	Imp. Disent.	No Disent.
Writing time	Kang 2019	–	–	–
Subjective bias	Pryzant 2020	–	–	–

other *unintended* leaves, i.e., the assumption that writers are spontaneous and do not mask nor alter their personal styles. Besides, a puzzling direction could be that of "values" or "expressive authenticity" (Newman 2019). Writers may be more or less genuinely committed to the content they convey. Authenticity in the sense of sincerity would be the correspondence between people's internal states and their external expressions, with a lack of authenticity resulting in a lie. The binomial authentic-deceptive fits style transfer: all content things being equal, what gives a lie away is its linguistic style (Newman *et al.* 2003). Therefore, an authenticity-aware style transfer tool could help understand deceptive communication or directly unveil it. Yet, the transfer between authenticity attributes appears puzzling because successful liars are those who shape their content in a style that seems convincing and trustworthy (Friedman and Tucker 1990).

Below are the dynamic states that, to the best of our knowledge, are the only ones present in the style transfer literature (they are visualized in Table 3, with some corresponding examples in Table 4).

4.2.1 Writing time

An instance of *dynamic states*-related styles in the literature is the time at which writers produce an utterance. Information revolving around the writing time of texts was collected by Kang *et al.* (2019) and is contained in their Pastel corpus. The authors considered daily time spans such as Night and Afternoon, that represent the stylistic attributes to transfer in text. These attributes were tackled with the methods discussed above, under *persona* and *background* (the success of their transfer was evaluated with the same techniques).

4.2.2 Subjective bias

Talking of subjectivity in language evokes the idea that words do not mirror an external reality, but reflect it as is seen by the speakers (Wierzbicka 1988). In this sense, language has the power to expose personal bias. NLP has risen to a collective endeavor to mitigate the prejudices expressed by humans and reflected in the computational representations of their texts (Bolukbasi *et al.* 2016; Zhao *et al.* 2018a). For its part, style transfer has surged to the challenge of debiasing language by directly operating on the texts themselves.

Although bias comes in many forms (e.g., stereotypes harmful to specific people or groups of people), only one clear-cut definition has been assumed for conditional text rewriting: bias as a form of inappropriate subjectivity, emerging when personal assessment should be obfuscated as much as possible. That is the case with encyclopedias and textbooks whose authors are required to suppress their own worldviews. An author's personal framing, however, is not always communicated openly. This is exemplified by the sentence "*John McCain exposed as an unprincipled politician*", reported in the only style transfer work on this topic (Pryzant *et al.* 2020). Here, the bias would emerge from the word "*exposed*", a factive verb presupposing the truth of its object. The goal of style transfer is to move the text towards a more neutral rendering, like one containing the verb "*described*".

Bias (and the choice of terms that reinforce it) can operate beyond the conscious level (Chopik and Giasson 2017). Further, circumventing one's skewed viewpoints seems to take an expert

Table 4. Examples of *dynamic states*, namely *writing time* [from Kang *et al.* (2019)] and *subjective bias* [taken from Pryzant *et al.* (2020)]. Note that the former is transferred in combination with other styles (i.e., *background*)

Writing time	**Morning:** *the flowers were in full bloom*
	Afternoon: *Tulips are one of the magnificent varieties of flowers*
Subjective Bias	**Biased**: *John McCain exposed as an unprincipled politician*
	De-biased: *John McCain described as an unprincipled politician*

effort—as suggested by the analysis of Pryzant et al. (2020) on their own corpus, senior Wikipedia revisors are more likely to neutralize texts than less experienced peers. Therefore, we collocate the style *subjective bias* under the *unintended* group, and specifically, as a division of *dynamic states* because prior judgments are open to reconsideration.

Data.   Pryzant et al. (2020) released a corpus[k] of aligned sentences, where each pair consists of a biased version and its neutralized equivalent. The texts are Wikipedia revisions justified by a neutral point of view tag, comprising 180k pre and post revision pairs.

Methods.   With the goal of generating a text that is neutralized, but otherwise similar in meaning to an input, Pryzant et al. (2020) introduced two algorithms. One, more open to being interpreted, has two components: a neural sequence tagger that estimates the probability that a word in a sentence is subjectively biased, and a machine translation-based step dedicated to editing while being informed by probabilities about subjectivity. The alternative approach directly performs the edit, with Bert as an encoder and with an attentional Lstm as a decoder leveraging a copy and coverage mechanisms.

Evaluation.   The models' accuracy was equated to the proportion of texts that reproduced the changes of editors. In the human-based evaluation, the success of models was measured with the help of English-speaking crowdworkers who passed preliminary tests proving their ability to identify subjective bias.

5. Intended styles

The second branch of the hierarchy stems from the observation that some linguistic variations are intentional. By *intended* we refer to styles that people modify contextually to the audience they address, their relationship, their social status, and the purpose of their communication. Due to a complex interaction between individuals, society, and contingent situations (Brown and Fraser 1979), it is not uncommon for speakers to change their language as they change their role in everyday life, alternating between non-occupational roles (stranger, friend), professional positions (doctor, teacher), and kinship-related parts (mother, sibling). Such variations occur as much in speech conversations, as they do in texts (Biber 2012).

We split this group of styles into the *targeted* and *non-targeted* subcategories. The *non-targeted* ones, which are the non-evaluative (or non-aspect-based) styles, further develop into the circumstantial and conventional nodes. While all *non-targeted* leaves can be associated with an idea of linguistic variation, many of them are specifically closer to what theoretical work calls "registers" and "genres". Understanding the characteristics of these two concepts would shed light on the linguistic level at which the transfer of *non-targeted* features of text should operate; yet, there is no agreement on the difference between genres and registers, and a precise indication of what

---

[k]https://github.com/rpryzantneutralizing-bias

differentiates them from style is missing as well (Biber 1995). In our discussion, we follow Lee (2001): by genre we mean novels, poems, technical manuals, and all such categories that group texts based on criteria like intended audience or purpose of production; whereas registers are linguistic varieties solicited by an interpersonal context, each of which is functional to immediate use. Therefore, we place the culturally recognized categories to which we can assign texts among the *conventional genres*, and we collocate linguistic patterns that arise in specific situations among the *circumstantial registers*. Note that these two classes of styles are not mutually exclusive: a formal register can be instantiated in an academic prose as well as in a sonnet.

5.1 Targeted

The presence of writers in language becomes particularly evident when they assess a topic of discourse. They applaud, disapprove, and convey values. Communications of this type, which pervade social media, have provided fertile ground for the growth and success of opinion mining in NLP. Opinion mining is concerned with the computational processing of stances and emotions targeted towards entities, events, and their properties (Hu and Liu 2006). The same sort of information is the bulk of study for the *targeted* group in our hierarchy. It is "targeted" because it reflects the relational nature of language, often directed *towards* an object (Brentano 1874): people state their stances or feelings *about* things or *with respect to* properties. Hence, under this group are styles that pertain to the language of evaluations, like *sarcasm* and *emotions*.

The tasks of mining opinions and transferring them are kin in that they use similar texts and observe similar phenomena. Yet, they differ in a crucial respect. Each of them looks for information at different levels of granularity. The former task not only recognizes sentiment and opinions, but also extracts more structured information such as the holder of the sentiment, the target, and the aspects of the target of an opinion (Liu and Zhang 2012). Instead, style transfer only changes the subjective attitudes of writers.

Dealing with evaluations makes the transfer of *targeted* styles particularly troublesome. To appreciate what is at stake here, let us take an example that explicitly mentions an emotion, "*I'm happy for you*". A style transfer task might generate a paraphrase that expresses another state, for instance sadness, and might do so by changing the emotion word into, e.g., "*sad*". Would such a modification change the stylistic attribute and preserve the meaning of the input? This question urges attention: to date, it is unclear whether this research line can aim at satisfying the three transfer criteria, and therefore, whether it addresses style transfer at all. Works in the field have not provided an answer, nor have other studies in NLP offered key insights. As a matter of fact, some of the styles at hand are cognitive concepts whose realization in text is yet to be fully understood (are they content or style, or both?). The problem arises not only with input texts containing explicit markers of style (e.g., "*happy*" for emotions). Even when attitudes are expressed less directly in a sentence (e.g., "*I managed to pass the exam*"), the issue of shifting its stylistic attribute (and only its stylistic attribute) remains. Current studies solely suggest that the transfer is effortless for some texts but not for others, and that it can occur through various strategies—not necessarily by swapping emotion words (Helbig *et al.* 2020).

An exhaustive overview of the relevant style transfer literature is available in Table 5. Examples of the tasks can be found in Table 6.

5.1.1 Emotion state

Language carries a great deal of information about the writers' emotions. These mental states have sparked research based on classification (Abdul-Mageed and Ungar 2017; Felbo *et al.* 2017; Schuff *et al.* 2017, i.a.) and generation (Zhou and Wang 2018; Huang *et al.* 2018; Song *et al.* 2019, i.a.), but they have found little space in the study of transfer. Indeed, the multifaceted ways in which emotions are realized in language—e.g., explicit mentions ("*I am happy*"), implicit pointers

Table 5. Literature on *intended, targeted* styles divided by method

| | Parallel | Non-parallel | | |
		Exp. Disent.	Imp. Disent.	No Disent.
Emotions	Chakrabarty et al. (2021)	Helbig et al. (2020)	Li et al. (2020b); Nangi et al. (2021)	Dryjański et al. (2018); Lample et al. (2019); Smith et al. (2019); Troiano et al. (2020;) Riley et al. (2021)
Sentiment	Jin et al. (2019); Cavalin et al. (2020)	Guerini et al. (2008); Whitehead and Cavedon (2010); Li et al. (2018); Xu et al. (2018); John et al. (2019); Sudhakar et al. (2019); Wu et al. (2019a); Wu et al. (2019b); Lee (2020); Madaan et al. (2020); Wen et al. (2020); Malmi et al. (2020); Reid & Zhong (2021); Lee et al. (2021)	Hu et al. (2017); Shen et al. (2017); Fu et al. (2018); Liao et al. (2018); Logeswaran et al. (2018); Prabhumoye et al. (2018a); Prabhumoye et al. (2018b); Singh and Palod (2018); Tian, Hu, and Yu (2018); Yang et al. (2018); Yang et al. (2019); Zhang et al. (2018b); Zhao et al. (2018b); Fang et al. (2019); Kruengkrai (2019); Leeftink & Spanakis (2019); Lai et al. (2019); Li et al. (2019); Tikhonov et al. (2019); Yamshchikov et al. (2019); Cheng et al. (2020a); Li et al. (2020b); Lin et al. (2020); Nangi et al. (2021)	Mueller et al. (2017); Guu et al. (2018); Dai et al. (2019); Gong et al. (2019); Lample et al. (2019); Luo et al. (2019a); Luo et al. (2019b); Pang and Gimpel (2019); Wang et al. (2019a); Xu et al. (2019b); Chawla & Yang (2020); Gong et al. (2020); He et al. (2020); Li et al. (2020a); Liu et al. (2020b); Mai et al. (2020); Li et al. (2021); Jafaritazehjani et al. (2020); Reif (2022); Riley et al. (2021)
Sarcasm	Peled and Reichart (2017)	–	Chakrabarty et al. (2020a); Mishra et al. (2019)	–
Political slant	Kang et al. (2019)	Sudhakar et al. (2019); Madaan et al. (2020)	Chen et al. (2018); Prabhumoye et al. (2018a, 2018b); Tian *et al.* (2018); Nangi et al. (2021)	–

Table 6. Examples of some *intended (targeted)* styles, namely, *emotion state*, *sentiment,* and *sarcasm* coming from Helbig *et al.* (2020), Li *et al.* (2018) and Mishra *et al.* (2019), respectively

Emotion State	**Anger:** *This soul-crushing drudgery plagues him*
	Joy: *This fulfilling job motivates him*
Sentiment	**Positive**: *great food but horrible staff and very very rude workers!*
	Negative: *great food, awesome staff, very personable and very efficient atmosphere!*
Sarcasm	**Non-sarcastic**: *Hate when the bus is late*
	Sarcastic: *Love when the bus is late*

("*I was on cloud nine*"), descriptions of salient events ("*Cool, I passed the exam!*")—place this phenomenon at the turn between *what* is said and *how* that is done (Casel, Heindl, and Klinger 2021). As emphasized by the works on emotion transfer, it is still debatable whether emotions can be changed without distorting the semantic content of a text (Helbig *et al.* 2020; Troiano, Klinger, and Padó 2020).

Assuming that emotions can be considered a style, their transfer requires rewriting a source text such that the output conveys the same message and a new emotional nuance. Source and target attribute labels can be borrowed from various traditions in psychology. Past research in emotion analysis has used diverse schemas, which describe emotions in multi-dimensional spaces (Buechel and Hahn 2017; Preoţiuc-Pietro *et al.* 2016b) or in terms of some underlying cognitive components (Hofmann *et al.* 2020; Troiano *et al.* 2022; Stranisci *et al.* 2022). On the other hand, style transfer has only leveraged discrete psychological models and has mapped between emotion names. Given a source sentence like "*I was going to knock down a pedestrian with my car*", that the writer associates to a fearful circumstance, a joyful counterpart could be "*I wanted to overturn a pedestrian with my car*" (Troiano *et al.* 2020). There are also publications that do not follow any established emotion schema. That is the case of Lample *et al.* (2019), who performed the transfer between two discrete writer's feelings, i.e., relaxed and annoyed, and Smith *et al.* (2019), who preferred a richer set of labels that mix different affective states and emotions. They put them under the umbrella term of "sentiment", despite including more fine-grained labels than polarity, such as the states of being annoyed, ecstatic, and frustrated.

Chakrabarty, Hidey, and Muresan (2021) are an exception in this panorama. Rather than focusing on the mental states per se, they considered the *appeal* to emotions, as an argumentative strategy that makes texts persuasive to an audience. These authors leveraged the association between emotions and arguments, and rewrote the latter to obtain more trustworthy variants (e.g., without appealing to fear), thus paraphrasing sentences like "*At this dire moment, we all need to amplify our voices in defense of free speech.*" as "*At this crucial moment, we all need to amplify our voices in support of free speech.*".

It should be noted that discrete labels account for only part of humans' emotion episodes. Other aspects are the strength of such experiences, that is, their intensity (Sonnemans and Frijda 1994), and the degree of arousal and dominance that they induce in the concerned individuals (Mehrabian 1996). Style transfer could be done in the future based on such models, for instance by controlling not only what emotion is transferred but also to what degree, similar to other generation studies that condition both the emotion and the emotional strength of texts (Ghosh *et al.* 2017; Goswamy *et al.* 2020, i.a.). This might make the task of changing the emotion connotation more feasible (e.g., the transfer might be possible between different emotions but only for specific levels of intensity).

Since emotions pervade communication, there is an unbounded number of applications where the related branch of style transfer could be put to use—from clinical to political contexts. As an example, style transfer tools might support the production of arguments by infusing a specific emotion in them, thus enhancing their persuasive power; vice versa, they could be employed to strip emotions away from existing arguments in order to isolate their factual core. In the domain of education, they could give an emotional slant to learning materials, to stimulate the learning process (Zull 2006). Augmenting emotions or making them explicit might also facilitate textual understanding for individuals who struggle to interpret the expression of affective states, like people on the autism spectrum, or suffering from alexithymia (Poquérusse *et al.* 2018). In commerce, they could be used to rewrite trailers of books, movies, or the presentation of any other product, with a higher emotional impact. Lastly, any chatbot capable of emotion transfer may adjust the affective connotation for the same semantic gist depending on its users.

We recognize that placing *emotion state* in the *intended* set of styles is a questionable choice. There are some features of this mental fact that stir it towards the *unintended* side: people might not necessarily be aware that emotions seep out of their written productions, neither do they purposefully experience them [emotions are reactions to salient events (Scherer 2005)]. However, publications on emotion transfer used data that humans consciously produced around emotion-bearing events and impressions. Therefore, we include them in the present category.

Data.   There exists a comparably large set of emotion corpora from various domains (Bostan and Klinger 2018), but only a small subset has interested style transfer. Among them are Tec, the corpus of Tweets from Mohammad (2012), Isear, a collection of descriptions of events that elicited emotional responses in their experiencers (Scherer and Wallbott 1994), and the EmpatheticDialogues dataset[l] from Rashkin *et al.* (2019), found in Smith et al. (2019), which encompasses a wide range of mental states. A corpus that is not dedicated to emotions but contains them as personality-related labels is the Personality-Caption dataset (Shuster et al., 2019), leveraged by Li *et al.* (2020b).

Concerning emotions and arguments, Chakrabarty *et al.* (2021) collected 301k textual instances from the subreddit *Change My View*, a forum for persuasive discussions. They created a parallel corpus with the help of a masked language model and a resource that labels nouns and adjectives with their connotations, including the label *Emotion Association* (Allaway and McKeown 2021). The authors matched the words in the arguments they gathered to the entries in such an external dictionary. They masked those which are associated with fear, trust, anticipation and joy, and constrained the replacements proposed by the language model to have a different emotional association than the original one.

A number of other emotion-related datasets could be adopted in the future, which cover different textual domains and follow varied psychological theories. Examples are the 10k English sentences of Buechel and Hahn (2017) labeled with dimensional emotion information in the Valence-Arousal-Dominance schema, the emotion-bearing dialogs of Li *et al.* (2017), and the literary texts made available by Kim, Padó, and Klinger (2017) annotated both with discrete emotions and the communication channels that express them (e.g., description of facial expressions or body movements).

Methods.   Being an under-explored task, emotion style transfer was tackled by Helbig *et al.* (2020) with a pipeline transparent for investigation. Subsequent components (1) identify textual portions to be changed, (2) find appropriate new words to perform the lexical substitution, and (3) from the resulting alternatives, pick one depending on its fluency, content preservation and presence of a target attribute. Each step was instantiated with many strategies, like (1) a rule-based identification of words vs. a selection mechanism informed by the attention scores of an emotion classifier, (2) retrieving new words from WordNet vs. leveraging the similarity between input embeddings and those of possible substitutes, (3) re-ranking the outputs with different weights for the three transfer criteria. The approach of Dryjański *et al.* (2018) used a neural network to perform phrase insertion, but it is similar to that of Helbig *et al.* (2020) in the idea that specific portions of texts should be targeted for the change.

A filtering step based on re-ranking was also explored in Troiano *et al.* (2020), where style transfer is defined as a backtranslation post-processing. The authors leveraged the idea that neural machine translation systems maximize both the output fluency and its faithfulness to the input (thus guaranteeing content preservation and naturalness), and focused on their ability to generate multiple and lexically diverse outputs as a way to promote emotion variability. Hence, with the help of an emotion classifier, they re-ranked backtranslations with respect to their association with the target emotion and to perform the transfer, they selected the text that best fulfilled such a requirement. Similarly, Chakrabarty *et al.* (2021) generated multiple styled rewritings, picking the one with the same meaning as the input—in their case, the one with the highest entailment relation to the original text. Their model was a fine-tuned Bart which learned to generate texts on their parallel data (with the artificially created text being the input and the original argument representing the target). Generation was further controlled by inserting a special separator token as a delimiter for the words that the model needed to edit during fine-tuning.

---

[l]https://github.com/facebookresearch/EmpatheticDialogues

Though not directly formulated in emotion-related terms, an effort of emotion style transfer can be found in Nangi *et al.* (2021). Therefore, the produced paraphrases display a different degree of excitement than the original texts, mirroring the notion of arousal in the continuous models of emotions. This paper aimed at gaining control over the strength of the transfer by integrating counterfactual logic in a generative model. With a series of losses to promote disentanglement, their variational auto-encoder was trained to find two separate embeddings for style and content. Counterfactuals came into play in the form of a generation loss which guided the model to find a new representation for the input attribute, specifically, a representation that can push the prediction made by a style classifier (given the style embeddings) towards the target attribute.

Evaluation. In a small-scale human evaluation, Helbig *et al.* (2020) defined a best-worst scaling task: two annotators chose the best paraphrase for a given sentence, picking among four alternatives generated from different pipeline configurations.

Consistent with the idea of making arguments more trustworthy, Chakrabarty *et al.* (2021) conducted a human evaluation in which workers on Amazon Mechanical Turk rated arguments with respect to the presence of fear, while simultaneously taking into consideration the preservation of meaning (i.e., a trustworthy text would have been penalized if it altered the input meaning).

5.1.2 Sentiment

Sentiment in NLP refers to the expression of a subjective and polarized opinion (Liu 2012). A few works aimed at creating paraphrases that preserve the sentiment but not the content of the input texts [e.g., "*It is sunny outside! Ugh, that means I must wear sunscreen.*" → "*It is rainy outside! Ugh, that means I must bring an umbrella.*", as illustrated in Feng, Li, and Hoey (2019)]. Going in the opposite direction, style transfer rephrases an input text to alter its polarity, which is either positive ("*I was extremely excited in reading this book*"), negative ("*The book was awful*"), neutral ("*I've read the book*"), or is characterized by some polarity gradation ("*That's a quite nice book*").

What a successful transfer of *sentiment* should look like is difficult to establish. The issue becomes clear by considering examples of a transfer input and output, such as "*this restaurant has awesome pizza*" and "*this restaurant has awful pizza*". On the one hand, these sentences are (intuitively) stylistically the same—which casts doubt on the status of sentiment as a style. On the other, they showcase that changing the polarity of a text also affects its semantics. We stand by the view of Tikhonov and Yamshchikov (2018), who denied that sentiment can be taken as a linguistic dimension unrelated to content. They highlighted that if sentiment is not independent of a text's semantics, but rather its function, then the transfer attempt is contradictory (as content changes, so does the "sentiment style"). Consistent with this is an observation of Guu *et al.* (2018), who proposed a generation system able to control for the attribute of a prototype text with a series of edits. With their model having to distort the meaning of the prototype as little as possible, they noticed that an edit like "*my son hated the delicious pizza*" for the prototype "*my son enjoyed the delicious pizza*" would miss the goal of content preservation. To overcome this problem, Prabhumoye *et al.* (2018b) relaxed the condition of keeping the content untouched in favor of maintaining *intent*, or the purpose for which a text was produced (e.g., to move a critique).

Nevertheless, transferring *sentiment* represents today a hallmark for most of the state-of-the-art style transfer methods, due to polarity being represented in many and relatively large datasets, together with its possible industrial applications. A case in point can be found in Gatti *et al.* (2012), who created an application that subverts the messages conveyed by posters by exaggerating their sentiment, both positively and negatively. Moreover, sentiment is relatively easy to recognize: given its polar nature, it has distinctive linguistic markers, and it is often sufficient to perform changes at this lexical level for the transfer to be considered achieved (Fu *et al.* 2019). We hence include *sentiment* in our hierarchy, and we refer to it as a style for convenience, to report on the massive amount of works that did so.

Data.   A fair share of sentiment-polarized datasets consists of mono-style resources. Commonly used are Yelp reviews[m], Amazon reviews[n], and Imdb reviews[o]. Arguing that superior performance is observed for any sequence-to-sequence task with parallel data, Cavalin *et al.* (2020) employed a semantic similarity measure to derive parallel data from (non-parallel) Amazon and Yelp reviews. Also, Jin et al. (2019) and Kruengkrai (2019) derived a pseudo-parallel corpus from mono-style data by aligning semantically similar sentences from the sides of the source and target attributes. For a subset of the Yelp reviews, they collected human-generated styled variations.[p]

Methods.   Many approaches that attempted to obtain a sentiment neutralized latent representation of the content (e.g., Hu et al., 2017) employed methods like adversarial training (Shen et al., 2017; Fu et al., 2018; Zhao *et al.* 2018b; Fang *et al.* 2019; Lin et al., 2020) and fed this latent representation into a decoder to generate content with the desired polarity. Reinforcement learning-based methods have been adopted for sentiment transfer as well, to bypass the dependency on differentiable learning objectives like loss terms (Gong et al., 2019; Luo *et al.* 2019a,b). In the cycled reinforcement learning approach of Xu et al. (2018), a "neutralization" module removed sentiment from the semantic content of a sentence, and an "emotionalization" module introduced the style with the desired attribute in the newly generated text. A policy gradient-based method rewarded the neutralization step using the quality of the generated text from the emotionalization phase.[q]

Explicit disentanglement by identifying and changing style markers has been claimed effective in sentiment style transfer (Guerini, Strapparava, and Stock 2008; Whitehead and Cavedon 2010; Li et al., 2018; Xu et al., 2018; Sudhakar *et al.* 2019; Leeftink and Spanakis 2019; Wu *et al.* 2019b; Lee 2020; Madaan et al., 2020; Malmi *et al.* 2020), because such markers are less subtle compared to those of other styles (e.g., *personality traits*). Strategies designed to this end use frequency statistics-based methods (Li et al., 2018; Madaan et al., 2020), sentiment lexica (Wen et al., 2020), techniques based on the attention scores of a style classifier (Xu et al., 2018; Zhang *et al.* 2018b; Sudhakar *et al.* 2019; Yang *et al.* 2019; Reid and Zhong 2021) or a combination of them (Wu *et al.* 2019b). The sentiment-devoid content is then used as a template to generate text with the target sentiment. Wu *et al.* (2019a) achieved this with the contribution of two agents: one that iteratively proposes where the re-wordings should occur in the text, and another that performs such local changes. In Reid and Zhong (2021), concurrent edits across multiple spans were made possible by generating a template with the Levenshtein edit operations (e.g., insert, replace, delete) which guided the transformation of the input text towards the desired attribute.

As stated by Yamshchikov *et al.* (2019), the fact that content and style are hard to separate at the lexical level does not undermine the possibility that they can be separated in their latent representations—with the quality of such disentanglement depending on the used architecture. The machine translation framework of Prabhumoye *et al.* (2018b), already described in relation to *genre* style transfer (see Section 4.1), aimed at producing a style-devoid representation in the encoding step of the backtranslation. Compared to them, John et al. (2019) pushed the disentanglement even further, by dividing such representation in two separate components, that is, a space of sentiment and a space for the content of the sentence (where the content is defined with bag-of-words, style-neutral features). For a given input, an auto-encoder represented the content (but not the style), which was then fed to the decoder, concatenated with an embedding of the desired output attribute. This is similar to Liao *et al.* (2018), who used two encoders to model content and target attribute (a value of the rating of sentences/reviews representing polarity). Claiming that the

---

[m]https://www.kaggle.com/yelp-dataset/yelp-dataset
[n]https://cseweb.ucsd.edu//\~jmcauley/datasets.html#amazon_reviews
[o]https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
[p]https://github.com/zhijing-jin/IMaT
[q]Xu et al. (2018) use the terms "sentiment" and "emotion" interchangeably (their emotionalization module transfers in fact sentiment). Psychology, on the other hand, separates emotions from other affective states (Scherer 2005).

conditioning structure is essential for the performance of a style transfer model, Lai *et al.* (2019) refrained from treating the target attribute simply as part of the initial vector fed to the decoder. Instead, they concatenated the style vector with the output of a Gated Recurrent Unit (Chung *et al.* 2015) cell at each time step. Style information was implicitly obfuscated at the token level by Lee *et al.* (2021) under the assumption that the alternative option of explicit removal of tokens would result in an information loss. They opted for an adversarial strategy, which reversed the attention scores of a style discriminator to obtain a style-devoid content representation, and they applied conditional layer normalization on this representation to adapt it to the target attribute distribution.

In opposition to typical disentanglement-based studies, Yang *et al.* (2018) noticed that classifiers that guide the decoding step towards the desired attribute can be insufficient (their error signal is sometimes too weak to train the generator), and that their presence in adversarial setups as discriminators can lead to unstable optimization. To solve this problem, the authors moved to language models as a different type of discriminator which overcomes the need for adversarial training: a language model trained on the target sentiment data would not only assign low probabilities to outputs that do not contain the desired sentiment, but it would also allow outcome introspection (which word is responsible for such low probability?). In a similar vein, Li *et al.* (2020b) proposed to gradually incorporate the style-conditional supervision signals in the successive training iterations, as long as the output quality does not degenerate. While these studies focused on the semantics of the input and the generated sentences, Gong, Song, and Bhat (2020) advocated the need for including the representation of their syntactic information in the transfer process. They encoded a sentence by considering dependency trees (to capture word relations) and structured semantic information (i.e., semantic roles) with the help of a Graph Neural Network (Marcheggiani and Titov 2017), providing evidence that they can help a model identify the core information to be preserved.

Many limitations of disentanglement were pointed out in other sentiment-based style transfer studies (e.g., using fix-sized vectors for the latent representations might fail to retain the rich semantic information characterizing long texts), with some of them casting doubt on the feasibility of the style-to-content separation (e.g., Jafaritazehjani *et al.* 2020). As an alternative to the manipulation of latent representations, Dai et al. (2019) added a style embedding as an input to their transformer encoder, while Li *et al.* (2020a) directly proposed a novel architecture composed of two generators and no discriminator. They performed style transfer with a sentence noisification approach: after introducing noise to an input text, they found a number of variations and used them to learn the transfer by having the model reconstruct the original input attribute. The novel method proposed by Li, Sun, and Wang (2021), which did not resort to disentanglement, used a generative adversarial network and a style classifier to regularize the distribution of latent representations from an auto-encoder. Instead, in the generative framework that Guu et al. (2018) presented, a sequence of revisions was produced for some prototype sentences. First, they extracted a prototype from a corpus, next, they sampled an edit vector encoding the edit to be performed: both were fed into the neural editor to produce 1k sequences, and the sequence with the highest likelihood to contain the target attribute was selected.

According to Li *et al.* (2019), a further problem that researchers should consider is that leveraging data from various domains might result in poor transfer performances. A model learned on movie reviews might not be appropriate to transfer polarity on restaurant reviews. Hence, they presented a domain adaptive approach which modifies sentiment in a domain-aware manner. Others focused on how to leverage pretrained text-to-text models. For instance, Mai *et al.* (2020) formulated a "plug and play" approach that allows to employ pretrained auto-encoders, and in which the transfer is learned within the latent space of the auto-encoder itself (i.e., embedding-to-embedding). For few-shot style transfer, Riley *et al.* (2021) leveraged the presumably strong textual representations inherent to T5 (Raffel *et al.* 2020). Their encoder-decoder model was trained to reconstruct a corrupted input. Generation was conditioned on a fixed-width style vector (similar

to Lample et al. (2019)) extracted from the preceding sentence, assuming that style is a feature which spans over large context windows. At inference time, the stylistic vector was inferred from a set of style transfer exemplar pairs. Interestingly, they demonstrated that a single model trained on generic web data can transfer multiple styles, including dialect, emotiveness, formality, and politeness.

Evaluation.   As reported in the analysis of evaluation practices in (sentiment) style transfer by Mir et al. (2019), content preservation is typically evaluated in an automatic fashion with metrics devised for machine translation, like Bleu, language models' perplexity over the generated texts serves as a score for fluency, and sentiment classifiers quantify the transfer strength (i.e., transfer accuracy would be the percentage of output sentences that are classified as belonging to the target attribute). To overcome the limitations of these metrics, they suggested some alternative approaches. In their view, transfer strength is quantified by the Earth Mover's Distance: observing the cost of turning the style distribution of the input into that of the output (Rubner, Tomasi, and Guibas 1998) would acknowledge the transfer even if the output did not properly display the target attribute, but leaned toward it more than the input. With respect to content preservation, the authors experimented with two different settings, i.e., one in which the style-related words coming from a style lexicon were removed and one in which they were masked. Hence, they computed the Word Mover Distance to quantify the distance between the input and output word embeddings (Kusner *et al.* 2015). Lastly, naturalness was assessed via adversarial evaluation, with classifiers having to distinguish the input texts written by humans from the output of the generation system.

Mir et al. (2019) also proposed some best practices with respect to human evaluation, with the main idea that annotators should be asked to perform pairwise comparisons: by rating the stylistic difference between input and output, by comparing the two after masking their style markers, and by choosing which of them is the most natural.

Yamshchikov et al. (2019) leveraged human productions to propose some measures for the decomposition of textual information into content and styles (they corroborated the idea that better decomposition leads to better Bleu scores between output and human paraphrases). Yet another strategy was put forward by Pang and Gimpel (2019). They quantified content preservation as the average of the cosine similarities over all input/output sentence pairs, and observed perplexity using a language model trained on concatenated source and target attribute datasets. Moreover, they introduced a strategy to adapt to the task at hand which summarizes different metrics into a single score.

5.1.3 Sarcasm

Sarcasm represents a form of verbal irony (Kreuz and Glucksberg 1989). Alba-Juez and Attardo (2014) held that the usage of irony covers a spectrum of evaluative purposes: to criticize (negative evaluation), to praise (positive evaluation), or to express a neutral stance. Sarcasm falls within the scope of negative evaluations because it emerges as "a sharp and often satirical or ironic utterance designed to cut or give pain".[r] While some studies hesitated in drawing an exact distinction between irony and sarcasm (Utsumi 2000, i.a.), others did so and considered it as a figure of speech with a specific target and a negative connotation (Clift 1999; Alba-Juez and Attardo 2014, i.a.).

Being a figurative device, sarcasm is also characterized by a contradiction between the literal and intended meaning of a statement. It requires an understanding of the context in which an expression is uttered, or a mutually shared assumption between the involved parties, for the right interpretation to be grasped (Camp 2012). For example, the exclamation "*What a clever idea!*" following a dull statement would be sarcastic, as the intended meaning (i.e., the idea is unclever) conveys an unfavorable assessment, while the utterance "*I now realize what a bad actor you are!*"

---

[r]Merriam-Webster. (n.d.). Sarcasm. In Merriam-Webster.com dictionary. Retrieved October 15, 2021, from https://www.merriam-webster.com/dictionary/sarcasm

(after the actor got an award) would be ironic but devoid of any sarcastic effect. By insisting on the view of sarcasm in terms of meaning inversion, Camp (2012) actually identified distinct sub-classes of sarcasm—depending on the illocutionary force of the text, its evaluative attitude and its propositional content.

Most computational studies dedicated to such a phenomenon revolve around classification. These works investigated the role of lexical features, punctuation, emojis, sentence length, and sentiment, as potential markers of sarcastic texts, and focused predominantly on social media communication (González-Ibáñez, Muresan, and Wacholder 2011; Barbieri, Saggion, and Ronzano 2014; Sulis *et al.* 2016; Ling and Klinger 2016, i.a.). There are also a few studies on sarcasm generation in style transfer. Even though they do not explicitly formulate it as a transfer problem, they essentially use an attribute-mapping principle, where a literal input is translated into a sarcastic one or vice versa. Peled and Reichart (2017) called this task "sarcasm interpretation", which consists in interpreting and spelling out the actual intention of a sarcastic statement.

Data. A parallel sarcasm corpus, arguably the first of its kind, was introduced by Peled and Reichart (2017). These authors crawled tweets with the hashtag "#sarcasm" and used crowdsourcing to generate non-sarcastic alternatives. The resulting dataset includes 3k sarcastic tweets and five non-sarcastic variants for each of them.

Methods. Driven by the idea that sarcastic statements have strong polarized connotations, Peled and Reichart (2017) presented a machine translation-based algorithm targeting textual sentiment to "interpret" sarcasm and turn a sarcastic expression into a literal one. Mishra *et al.* (2019) also leveraged the relation between sarcasm and sentiment, and managed to introduce the figurative-to-literal incongruity using an unsupervised approach with four steps: the first neutralizes the input statement that expresses a negative opinion, by removing the sentiment information with a classifier and a self-attention based filtering—e.g., "*Hate when the bus is late*" → "*the bus is late*"; next, positive sentiment is injected into the neutralized sentence with a sequence-to-sequence model trained on the neutralized and positive sentence pairs—e.g., "*the bus is late*" → "*love when the bus is late*"; the third step retrieves a negative-situation phrase fitting the input from their own collection of facts (e.g., *canceled at short notice*, *getting yelled at by people*) using an information retrieval system, with the input acting as a query (e.g., "*waiting for bus*"); and as a last step, the sarcastic statement is synthesized from the positive keywords and negative situation phrases, with a reinforcement reward.

Chakrabarty *et al.* (2020a) worked with similar assumptions. Their system first reversed the valence of the input sentence by lexical antonym replacement or negation removal—e.g., "*zero visibility in fog makes driving difficult*" → "*zero visibility in fog makes driving easy*". Next, it generated common sense knowledge using Comet (Bosselut *et al.* 2019), a pretrained language model fine-tuned on the ConceptNet knowledge graph (Speer, Chin, and Havasi 2017), by supplying keywords from the input and leveraging the *causes* relation - e.g., (*zero*, *visibility*, *fog*, *driving*, *difficult*) → *accident*. Lastly, this knowledge served to retrieve candidate sentences, which were corrected for grammatical consistency and ranked on a *contradiction* score, similar to a natural language inference problem.

Evaluation. Standard measures useful to quantify the lexical closeness between a candidate and a reference (Bleu, Rouge, Pinc (Chen and Dolan 2011)) were reported for automatic evaluations (Peled and Reichart 2017; Chakrabarty *et al.* 2020a). In addition, Mishra *et al.* (2019) presented a metric, the "percentage of length increment", based on the assumption that sarcasm requires more context than its literal counterpart.

As for the human evaluations, Peled and Reichart (2017) collected ratings on the fluency and the adequacy of an interpretation, Mishra *et al.* (2019) on the fluency and the relatedness to an input, and Chakrabarty *et al.* (2020a) on the creativity, level of sarcasm, humor, and grammaticality. Mishra *et al.* (2019) also had annotators label the sentiment of the transfer outputs.

5.1.4 Political slant

Countless studies have been conducted on the relationship between politics and language (Orwell 1962; Shapiro 1986; Habermas 2006; Spencer-Bennett 2018, i.a.). In the public sphere, verbal communication is strategic for political manoeuvres. It creates meanings around problems and events to favor specific courses of action. The idea that language puts things into a compelling narrative for particular ideologies is one that Foucault (1966) developed further. He went as far as claiming that it is the language that constructs its users—and not the users constructing language, as the twentieth-century linguistics purported [e.g., the Sapir–Whorf hypothesis in Hoijer (1954)]. Indeed, every public debate inaugurates the use of some statements or expressions: to accept one or the other is to embrace an ideology, to present oneself as liberal or conservative, as an activist or a separator, as a victim of the authority or a supporter (Edelman 1985). These roles are the political categories useful for style transfer.

NLP provides a parsimonious solution to address such a style (e.g., it transfers broad attributes like "democratic" and "republican"). However, it simplifies the complexity of political language and the theories revolving around it. The role of the activist, of the authority, etc., not only guides people in opting for certain linguistic variations but it imposes constraints upon *what* they say: a police chief, for instance, is called to praise order over anarchy (Edelman 1985). This picture suggests that content and political slant style are inextricably bound together. Style transfer takes a different perspective and only taps on the communicative attitudes of different political groups. A style transfer result would look like the following: "*as a hoosier, i thank you, rep. visclosky.*" (democratic) → "*as a hoosier, i'm praying for you sir*" (republican). That is, moving from one attribute to the other does not necessarily imply distorting an expressed political opinion, but generating one that keeps the *intent* of the original text (in this case, to thank the senator) while changing the cues about the speaker's political affiliation (Prabhumoye *et al.* 2018b). An exception to this perspective is the work by Chen *et al.* (2018), who treated political slant as a biased opinion to be altered (hence, we include this style among those which are arguably closer to content, marked with an asterisk in Figure 2).

Linguistics-oriented studies that investigated the rhetorical devices of political communication (Rank 1980; Beard 2000; Reisigl 2008; Charteris-Black 2018) remain neglected in style transfer. Yet, they provide fruitful insights. Among others is the idea that debates, arguments, and propaganda are filled with stylistic inventiveness to marshal support and resonate with a large audience [e.g., political messages can be disguised under some words that evoke objectivity—like synonyms of "essential" or "true" (Edelman 1985)]. Future style transfer studies could rewrite the language of promises as ordinary language, devoid of sensationalisms, and rhetoric intents, to observe if the same message is conveyed, whether its persuasive strength changes, and ultimately, to help people establish if certain political claims are valid or are just embellished deceptions.

Data.   Ideated to study responses to gender, the corpus of Voigt et al. (2018) has also supported research in political slant transfer. Rt-Gender is a rich multi-genre dataset, with one subset including Facebook posts from the members of the House and Senate in the United States and their top-level responses. The posts include a label indicating if the Congressperson is affiliated with the Republican or the Democratic party. Posts and responses are publicly available,[s] but all information that could identify the users was removed for privacy.

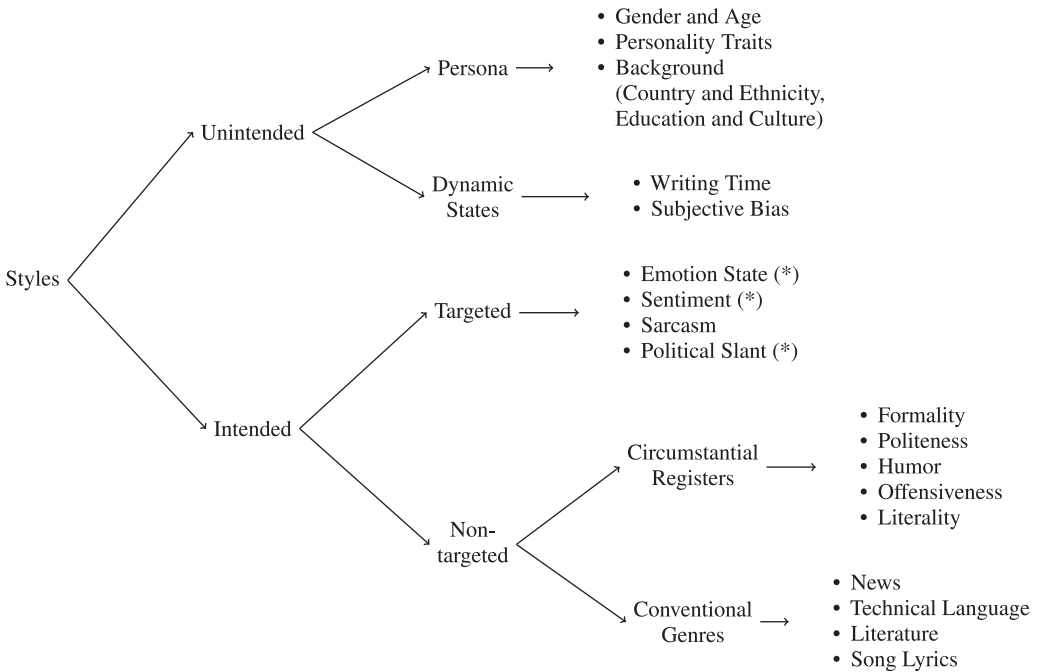[s]https://cs.cmu.edu/~tsvetko/rtgender/

Figure 2. The hierarchy of styles guiding our discussion. Each branch defines different challenges for style transfer and illustrates how styles relate to one another. Asterisks (∗) mark the nodes on the fence between content and style, since altering their attributes brings substantial content loss—they are included in the hierarchy nevertheless, because they have been leveraged for the transfer goal.

The RtGender creators claimed that the dataset is controlled for content by nature, because the members of the Congress discuss similar topics. This represents an advantage for style transfer. According to Prabhumoye *et al.* (2018b), what reveals political slant are both topic and sentiment, markedly different for the two affiliations, like in the examples "*defund them all, especially when it comes to the illegal immigrants*" and "*we need more strong voices like yours fighting for gun control*" uttered by a republican and a democratic, respectively. Researchers interested in deepening such observation could make use of the dataset released by Mohammad *et al.* (2015), as it includes electoral tweets annotated for sentiment, emotion, purpose of the communication (e.g., to agree, disagree, support), and information related to some rhetorical traits (e.g., whether it is sarcastic, humorous, or exaggerated).

To address political opinions more specifically, Chen et al. (2018) collected 2196 pairs of news article headlines found on the platform *all-sides.com*, each of which is either left-oriented or right-oriented, depending on the newspapers and portals where they were published.

Methods. As for the stance flipping task addressed by Chen et al. (2018), the authors started from the observation that not all news headlines are biased enough for a model to learn the task. Hence, they trained a generative model on the body of their articles, whose sentences are not semantically paired. Hence, they reproduced the cross-alignment setting proposed by Shen et al. (2017) to transfer sentiment in the absence of parallel data, training two encoders and two decoders (one for each transfer direction).

No other method has been implemented exclusively for this task. The ones that have been applied are the backtranslation frameworks of Prabhumoye *et al.* (2018b) and Prabhumoye *et al.* (2018a) used for *sentiment* and *gender* style transfer, which include a separate decoder for each

attribute (republican vs. democratic), and the tag-and-generate pipeline proposed by Madaan et al. (2020) in the context of *politeness* (discussed in the next section).

Evaluation.   Prabhumoye *et al.* (2018b) set up a comparison task. Eleven annotators compared the models' outputs with an input sentence. In line with the definition of the task, they had to choose the paraphrase that maintained the intent of the source sentence, while changing its slant. The annotators also had the option to express no preference for any output. Their results showed that most of the time people did not select any of the outputs, suggesting that state-of-the-art systems still have a long way to go.

Chen et al. (2018) framed the human evaluation task as one in which annotators judged the degree to which two headlines have opposite bias. Prabhumoye *et al.* (2018a), instead, refrained from measuring the presence of the target attributes in their human evaluation setting because judgments on political slants can be biased by personal worldviews.

5.2 Non-targeted: circumstantial registers

Registers are functional variations of a language (Halliday 1989). Like the styles subsumed under the *targeted* group, registers have specific lexico-grammatical patterns—e.g., the distribution of pronouns and nouns differs between a casual conversation and an official report (Biber and Conrad 2009). Unlike the *targeted* styles, they are not oriented towards an object, but are general linguistic routines that mirror some behavioral conventions. For example, in high-context cultures the discourse becomes more courteous when addressing an older interlocutor or someone perceived as higher in the social hierarchy. This is a hint of the complexity of this family of styles: as noticed by Hudson (1993), "one man's dialect is another man's register".

We show an overview of the *intended, non-targeted* styles regarding *circumstantial registers* in Table 7 (examples in Table 8). These types of styles have also witnessed the definition of a new framework for style transfer: according to Cheng *et al.* (2020b), a reasonable way of changing the characteristic attributes of a sentence is to take into account the context in which the sentence occurs, and to produce a stylized paraphrase that is coherent with it. The task of contextual style transfer would reproduce more faithfully what happens in real communications, where texts are never uttered out of context (e.g., sentences combine into paragraphs).

The readers might notice that some of these styles could also belong in the *targeted* category. As an example, humor can serve to express an evaluative stance, similar to *sarcasm*. However, such styles are socially-motivated, and we consider them *registers* in that sense.

5.2.1 Formality

The sentences "*His work was impressive and worthy of appreciation*" and "*His work was damn good*" show how texts can vary with respect to formality, an important dimension of linguistic variation (Heylighen and Dewaele 1999) that characterizes the register of a communication act. A formal text is explicit, accurate, and often required to minimize misunderstandings, for instance in academic works and legal documents. On the other hand, an informal text has a spontaneous and phatic nature. Being more relaxed, it can include colloquial/slang terms, ellipses, contractions (Heylighen and Dewaele 1999; Graesser *et al.* 2014; Li *et al.* 2016) and, on social media, also emojis, acronyms, consecutive punctuation (". . .", "!!!").

The concept of (in)formality encompasses multiple features, like seriousness–triviality, shared knowledge and familiarity (Irvine 1979; Brown and Fraser 1979), but style transfer usually adopts the more straightforward dichotomy of formal vs. informal, often treated as endpoints of a continuum (Heylighen and Dewaele 1999; Graesser et al., 2014).

Table 7. Literature on *intended, non-targeted* styles corresponding to *circumstantial registers* in our hierarchy, divided by method

	Parallel	Non-parallel		
		Exp. Disent.	Imp. Disent.	No Disent.
Formality	Niu et al. (2018); Rao and Tetreault (2018); Czeresnia Etinger and Black (2019); Ge et al. (2019); Jin et al. (2019); Xu et al. (2019b); Wang et al. (2019b); Chawla and Yang (2020); Cheng et al. (2020b); Wang et al. (2020); Zhang et al. (2020b); Briakou et al. (2021c); Lai et al. (2021); Yao and Yu (2021)	–	Li et al. (2019); Nangi et al. (2021)	Gong et al. (2019); Luo et al. (2019b); Shang et al. (2019); He et al. (2020); Yang and Klein (2021); Riley et al. (2021); Reif et al. 2022
Politeness	–	Madaan et al. (2020); Reid & Zhong (2021)	–	Riley et al. (2021)
Humor	Weller et al. (2020)	Li et al. (2018); Sudhakar et al. (2019); Madaan et al. (2020)	Li et al. (2020b)	Zhu, Yu, and Wan (2019); Wang et al. (2019a); Xu et al. (2019b)
Offensiveness	Cheng et al. (2020b)	Su et al. (2017); Tran et al. (2020)	–	dos Santos et al. (2018)
Literality	Chakrabarty et al. (2020b)	–	–	–

Table 8. Examples of style transfer on different *circumstantial registers—formality, politeness, humor, figurative language* and *offensiveness*—taken from Rao and Tetreault (2018); Madaan *et al.* (2020); Weller *et al.* (2020); Chakrabarty *et al.* (2020b); dos Santos *et al.* (2018), respectively

Formality	**Informal:** *I'd say it is punk though*
	Formal: *However, I do believe it to be punk*
Politeness	**Impolite**: *Send me the data*
	Polite: *Could you please send me the data*
Humor	**Non-humorous**: *Meet the wealthy donors pouring millions into the 2018 elections*
	Humorous: *Meet the wealthy sadists pouring millions into the 2018 elections*
Figurative/Simile	**Literal**: *You just started staring off into space and smiling dangerously*
	Figurative: *You just started staring off into space and smiling like a lunatic*
Offensive	**Offensive**: *what a f**king circus this is*
	Non-offensive: *what a big circus this is*

Data. Research on *formality* transfer has been largely supported by the Grammarly's Yahoo Answers Formality Corpus (Gyafc).[t] Introduced by Rao and Tetreault (2018), Gyafc contains around 110K formal/informal sentence pairs, where the informal side was generated via crowdsourcing. Next, the corpus curated by Briakou *et al.* (2021c), Xformal,[u] extended formality data

---

[t]https://github.com/raosudha89/GYAFC-corpus
[u]https://github.com/Elbria/xformal-Fostyletransfer

to multiple languages. Like Gyafc, Xformal was built by extracting texts in the topic "family & relationship" from an existing corpus of Yahoo answers. Such texts, which are in Brazilian Portuguese, Italian, and French, were characterized by an informal style. Crowdworkers on the platform Amazon Mechanical Turk[v] provided multiple formal rewrites for each of them.

Depending on a single dataset might hinder the generalization capability over unseen domains. Hence, by taking Gyafc as ground truth, a few works based on data augmentation methods have created and made available more style transfer instances. The formality classifier of Xu *et al.* (2019b) was trained on Gyafc and made predictions on unlabeled texts; such predictions were filtered for a threshold confidence score of 99.5%. Czeresnia Etinger and Black (2019) augmented data with the assumption that POS tags are representative of style-independent semantics. After training a classifier on Gyafc, they applied it on a style-unlabelled corpus and created formal–informal sentence pairs, by aligning sentences that become equal as soon as their respective style markers are replaced with the corresponding POS tags.

Zhang et al. (2020b) augmented approximately 4.9M sentence pairs with three techniques: backtranslation, formality discrimination, and multi-task transfer. Backtranslation employed a sequence-to-sequence model trained on parallel data in the formal to informal direction. It was then used to generate 1.6M informal sentences, given formal ones coming from the "entertainment & music" and "family & relationships" domains on Yahoo Answers L6.[w] Also the formality discrimination method exploited the observation that machine-translated informal texts can be rendered more formal: a number of informal English sentences from Yahoo Answers L6 were translated to different pivot languages and then back, followed by a discriminator with a predefined threshold that further filtered the augmented data, giving a total of 1.5M pairs. While these two strategies used the newly generated texts to augment data, the multi-task transfer method relied on sentence pairs annotated from previous tasks. For that, style transfer was formulated as a problem of Grammatical Error Correction under the assumption that informal sentences are prone to containing grammatical errors, character repetitions, spelling mistakes, unexpected capitalization, and so on. Accordingly, to improve the transfer of formality, they used the training data points for the Grammatical Error Correction task as augmented texts, namely, the Gec data (Mizumoto *et al.* 2011; Tajiri, Komachi, and Matsumoto 2012) and the Nucle corpus (Dahlmeier, Ng, and Wu 2013).

Different from such resources, the Enron-Context corpus released by Cheng *et al.* (2020b) contains paragraph-level data. It includes emails randomly sampled from the Enron dataset (Klimt and Yang 2004), in which sentences identified as informal by human annotators were rewritten in a more formal manner.

Methods.   The availability of a relatively large parallel dataset has made *formality* transfer a go-to task. Rao and Tetreault (2018) spurred extensive research, benchmarking the performance of phrase-based and neural machine translation for this style. Following their work, Ge et al. (2019) performed style transfer on the Gyafc corpus as a problem of grammatical error correction.

Others have moved the challenge of formality transfer into a multi-lingual setting: Niu *et al.* (2018) opted for a multi-task learning approach to jointly perform monolingual transfer and multilingual formality-sensitive machine translation; Briakou *et al.* (2021c) leveraged machine translation for inter-language style transfer, learned both in a supervised and unsupervised manner. The translation model of Yang and Klein (2021) conditioned the output translation towards formality with the help of future discriminators. These consisted in some style predictors operating on an incomplete text sequence, which inform as to whether the desired attribute will hold for the complete text sequence, and can thus help adjust the generators' original probabilities.

Many solutions were motivated by the need for massive amounts of parallel data to prevent overfitting in machine translation models. Among them are data augmentation attempts, like those by Czeresnia Etinger and Black (2019) and Zhang *et al.* (2020b). The latter employed augmented texts to pretrain models, but acknowledging that such texts are less than perfect, the models were subsequently fine-tuned on the original natural data. Xu *et al.* (2019b) augmented data with a formality classifier. They trained a transformer model on a parallel corpus with each instance prefixed with a token to indicate the direction of transfer, such that a single model could go from formal to informal and vice versa.

This was also achieved by Wang et al. (2020), a work belonging to the line of research that leverages pretrained language models. A sequence-to-sequence model with a single encoder captured the style-independent semantic representations with auxiliary matching losses, and two decoders were dedicated to each target attribute, jointly trained for bi-directional transfer. In Chawla and Yang (2020), a pretrained language model-based discriminator helped to maximize the likelihood of the target attribute being in the output, and a mutual information maximization loss between input and output supported diversity in generation. Lai, Toral, and Nissim (2021) worked on the parallel texts from Gyafc to fine-tune large pretrained language models, Gpt-2 (Radford *et al.* 2019) and Bart (Lewis *et al.* 2020) and augmented them with rewarding strategies based on style discriminators (targeting the transfer of the attributes) and Bleu (targeting content preservation). They argued that pretrained models contribute to better content preservation, even with limited training data.

Wang *et al.* (2019b) transformed informal sentences into formal ones in a rule-based fashion, with some transfer rules incorporated in their language model. The encoder was presented with an input as a concatenation of the original informal sentence and its formal revision to mitigate the consequent problem of noisy parallel data. Yao and Yu (2021) explored a similar architecture. The encoder's input was created by concatenating the original sentence and additional information, comprising a list of all matched rules and the corresponding text alternatives, arranged as tuples. Keeping all rules in the input allowed the model to identify which ones to use dynamically.

Other approaches in *formality* transfer that circumvented the use of parallel corpora were reinforcement learning (Xu *et al.* 2019b) and probabilistic modelling (He et al., 2020). The work by Cheng *et al.* (2020b) stands out in this panorama, in that it alters the formality of sentences while simultaneously considering the topic coherence to the text surrounding them. The context-aware model they proposed employs one decoder that translates the joint features from two separate encoders (which represent the main sentence and its contextual paragraph, respectively).

Evaluation.   Outside NLP, researchers have used measurements based on diagnostic linguistic features to quantify the formality of text. A popular measure is the F-score (formality score) which is sensitive to the frequencies of different word classes in text, ranging from articles and pronouns to adjectives and interjections (Heylighen and Dewaele 1999). There also exists a composite score that measures formality: Defined by Graesser et al. (2014), it is based on five principal component dimensions of Coh-Metrix,[x] and it takes into account syntax, discourse, and goals of communication (e.g., syntactic simplicity, referential cohesion, word concreteness, narrativity).

Style transfer studies have never opted for these measures. Indeed, while Rao and Tetreault (2018) raised the issue that the evaluation of style transfer (both human and automatic) is in need for best practices, *formality* transfer has insisted on evaluating the transfer accuracy with a style classifier, in line with other styles.

---

[x]http://www.cohmetrix.com/

5.2.2 Politeness

Linguistic politeness reflects the evaluation of a social context. Guided by a person's experience of social interactions (Meier 1995; Holtgraves 2001) and socio-cultural environment, politeness can uphold interpersonal relationships. Its markers (e.g., "*please*") affect how the speaker is perceived: as a considerate individual or, on the contrary, as discourteous (Meier 1995). Most studies in style transfer focus on the broad attributes of "polite" and its opposite, "impolite". However, according to some theories, the latter should be explicitly distinguished from rudeness, which is always intentional—impoliteness can instead occur accidentally (Segarra 2007; Terkourafi 2008).

Politeness transfer would change a formulation like "*You are wrong*" into "*I think you might be mistaken*". To date, this style appears in a limited number of publications, despite its link to formality as well as its potential to assist automatic writing (e.g., to help non-native speakers produce polite responses, as they might ignore some nuances in the target language).

Data.   The transfer task in Madaan et al. (2020) is restricted to action-derivatives (e.g., "*Let's stay in touch*") which are rewritten as polite requests (e.g., "*Can you call me when you get back?*"). As these constructs are frequent in official communication, the authors built a politeness dataset starting from a collection of emails exchanged within the Enron corporation, contained in the Enron corpus (Klimt and Yang 2004). With the application of some filtering heuristics, 1.39 million sentences were gathered, annotated, and filtered with a politeness score assigned by a classifier. This dataset is open source[y] and includes both the texts and the politeness scores.

Politeness labels are also present in the resource of Danescu-Niculescu-Mizil *et al.* (2013). Included in the collection of styled corpora from Kang and Hovy (2021), it encompasses 10k requests produced in the context of Wikipedia edits and other administrative functions, as well as Stack Exchange, where requests are related to a variety of topics. Their work focused on the politeness markers of requests, characterized by strategies that minimize imposition through indirect phrases (e.g., "*Could you please . . .*") or apologies (e.g., "*I'm sorry, but . . .*").

Method.   The task was introduced by Madaan et al. (2020). Observing the complex, socio-cultural nature of politeness, these authors limited their study to the use of formal language among North American English speakers. They defined impoliteness as a lack of politeness markers and adopted a tag-and-generate approach. The linguistic realizations of the potential marker positions were tagged in the source sentence and the target attribute markers were then generated in such positions.

Reid and Zhong (2021), who tested their method on the same dataset, introduced an unsupervised explicit disentanglement procedure. First, it transformed input texts into style-agnostic templates thanks to the attention scores of a style classifier; then, it filled the tagged positions in the templates using fine-tuned pretrained language models. Unlike other infilling methods for style transfer (Wang *et al.* 2019b; Malmi *et al.* 2020), theirs allowed concurrent edits over multiple textual spans.

Evaluation.   For the automatic evaluation of transfer accuracy, Madaan et al. (2020) calculated the percentage of generated sentences on which a classifier recognized the target attribute. For human evaluation, their annotators judged the match with the target attribute on a 5-point scale.

5.2.3 Humor

Most theories on linguistic humor agree that this phenomenon arises from an incongruity (Morreall 1983; Gruner 1997; Rutter 1997, i.a.). Just like sarcasm, which assumes the existence of two incompatible interpretations for the same text, humor is given by the resolution of such interpretations (Raskin 1979; Attardo and Raskin 1991; Ritchie 1999). In order to understand

---

[y]https://github.com/tag-and-generate/politeness-dataset

a joke, the receiver needs to identify the punchline (i.e., an incongruity) and then to resolve it by grasping its relationship with the main context of utterance. In communication, humor can serve as a tool to relieve tension or lighten the mood, encourage solidarity, further interactions within groups, and introduce new perspectives (Meyer 2006). On the other hand, humor can cause communication failures, if not perceived as intended.

A significant gap exists between computational studies of the style *humor*[z] and the theories underlying this concept, which remains also overlooked in style transfer. This style has an extremely subjective nature and, unlike others, it is not characterized by a defined pair of opposite attributes. In fact, only a few researchers considered the labels "non-humorous" and "humorous" (Weller, Fulda, and Seppi 2020), while the majority of them did the transfer between the attributes "humorous", "factual", and "romantic" (Li et al., 2018; Sudhakar *et al.* 2019; Wang *et al.* 2019a). This indicates a possible future line of research in which factuality and romantic intimacy could stand as styles by themselves.

Data. Weller *et al.* (2020) used the Humicroedit dataset (Hossain, Krumm, and Gamon 2019), a resource where crowdworkers made single word edits to render a regular news headline more humorous (e.g., "*Meet the wealthy donors pouring millions into the 2018 elections*" → "*Meet the wealthy sadists pouring millions into the 2018 elections*"). Humicroedit contains around 15k edited headlines. A similar corpus was presented by West and Horvitz (2019). It was curated using an online game by asking participants to edit a humorous headline and make it sound serious. Not evaluated to date, this dataset could be useful for future research.

Additional data can be found in the Captions corpus (Gan *et al.* 2017), which provides humorous captions describing images. Romantic and factual labels are also present as attributes opposite to humorous. Instead, researchers who prefer to treat "non-humorous" as such opposite could make use of the Short Text Corpus for Humor Detection[aa] and the Short Jokes Dataset[ab] indicated by Kang and Hovy (2021). These authors also provided a small sample of texts (2k instances) which allow to consider personal romanticism as a style on its own, with the two attributes "romantic" and "non-romantic".

Method. Weller et al. (2020) did an exploratory investigation of the usability of the Humicroedit humor-based corpus for style transfer purposes. A transformer-based sequence-to-sequence model was trained for humor generation and a random POS tag replacement was taken as a baseline.

As humor is not the main focus of the other works mentioned above, we refer the reader to their respective discussions, under *formality* and *sentiment*.

Evaluation. Weller *et al.* (2020) conducted a human-based evaluation regarding the fluency and the level of humor of texts, which were rated on a 5-point scale. The authors reported that the manually edited sentences were considered more humorous than the machine-generated ones, which in turn were better than random replacements. This positively asserted the potential for the humor generation task, highlighting at the same time the subjectivity of the phenomenon in question. A similar conclusion was drawn by Amin and Burghardt (2020). Focusing on the broader task of humor generation, they analyzed possible evaluation approaches: human ratings on a Likert scale for humorousness, human ratings on a Likert scale for the likeness that a humorous text was written by a human—the soft Turing test as in Yu, Tan, and Wan (2018)—and "humorous frequency" as the proportion of funny instances out of a set of generated texts. All of them failed to present a criterion to evaluate humor in text objectively.

---

[z]Amin and Burghardt (2020) presents a comprehensive overview of research in computational humor generation
[aa]https://github.com/CrowdTruth/Short-Text-Corpus-For-Humor-Detection
[ab]https://github.com/amoudgl/short-jokes-dataset

5.2.4 Offensiveness

Under the expression "offensive language", we place facts related to abusive language and harmful/hateful speech (Nobata *et al.* 2016; Schmidt and Wiegand 2017; Davidson *et al.* 2017a). Offensiveness is the negative extremity in the formality and politeness spectrum, and it is usually resorted to with the intention of attracting attention, offending[ac] or intimidating, and to express anger, frustration and resentment (Sue *et al.* 2007; Popuşoi, Havârneanu, and Havârneanu 2018). Extensive research has stemmed from this phenomenon, typically observed in the current social media-communicating world, where any type of information can be publicly discussed. While offensive behaviour detection (Razavi *et al.* 2010; Davidson *et al.* 2017b; Founta *et al.* 2019, e.g.) has aimed at identifying and prohibiting offensive material that exists online, style transfer studies like Su et al. (2017) and dos Santos, Melnyk, and Padhi (2018) reformulated offensive texts (e.g., "*That is f**king disgusting*") in more gentle terms (e.g., "*That is repulsive*"), or removed profanities (Tran, Zhang, and Soleymani 2020).

Whether a text is derogatory or hurtful does not solely depend on the presence of abusive words. Waseem *et al.* (2017) brought up a typology of abusive language detection tasks which clarifies that language can be belittling even without explicit slurs or an explicit target person (or group of persons) to whom it is directed. Rhetorical questions and comparisons are only two examples of how toxicity can emerge without swear words (van Aken *et al.* 2018), but harm can find its way into language with many more and more complex strategies—e.g., jokes and sarcasm (Wiegand, Ruppenhofer, and Eder 2021). While these insights encourage researchers to make informed decisions as to the most appropriate features to consider, depending on the type of offensiveness in question, works in style transfer do not necessarily consider all such factors.

In the future, studies related to this group of styles could address the challenge of making texts not only less toxic but also more inclusive of minorities.

Data.   To overcome the lack of parallel data, dos Santos *et al.* (2018) opted to create a non-parallel resource, and did so by employing the offensive language and hate speech classifier from Davidson *et al.* (2017b). The final dataset contains approximately 2M and 7M sentences from Twitter and Reddit, respectively, with the majority of instances being non-offensive. Also Cheng *et al.* (2020b) created a parallel dataset of offensive and non-offensive texts (the latter were crowdsourced by asking annotators to produce two non-offensive alternatives for a given offensive input).

As for dictionary-based approaches, several open-access sources are available. For instance, Tran *et al.* (2020) compiled a vocabulary of offensive terms by crawling a list of more than 1k English expressions made available by Luis von Ahn's research group,[ad] and an online platform that contains an ever-growing inventory of profanities.[ae]

Method.   dos Santos *et al.* (2018) employed an encoder–decoder model with an attention mechanism. They ensured output quality with a cycle consistency loss and the help of a collaborative classifier providing signal about the effectiveness of the transfer. Interestingly, they noted that their model was unable to handle implicit offensive content (e.g., ordinarily inoffensive words used offensively), indicating that offensiveness cannot always be addressed at a lexical level by changing a few words.

Still, other researchers focused on the editing of offensive lexical items. For paraphrasing profane texts in Chinese, Su et al. (2017) manually devised a rule-based system, equipped with an extensive set of profanity detection and paraphrasing strategies (the rules were language-specific, hence the system is not extendable to other languages). Similarly, Tran *et al.* (2020) developed a

---

[ac]It should be noted that some studies (e.g., Waseem and Hovy 2016; Davidson *et al.* 2017b) refrain from equating "hate speech" to language with offensive intentions, while others treat both as the same category to be detected (Plaza-del Arco *et al.* 2021; Grimminger and Klinger 2021).

[ad]https://www.cs.cmu.edu/~biglou/resources/bad-words.txt

[ae]https://www.noswearing.com/

transparent modular pipeline around the idea that a text is offensive if it contains profanity. The pipeline had different modules. First comes the retrieval module: it extracts ten part-of-speech (POS) tag sequences from a dataset of non-offensive texts, which are similar to the POS sequence found in an offensive sentence. Next is the generation module, which creates non-offensive sentences by matching the words from the input into possible positions in the generated POS sequences, and then filling the unmatched positions with a pretrained language model. An edit step further corrects word order. The selected output was the one with the best fluency, meaning preservation and transfer—which in this case corresponds to the absence of profanities.

Evaluation. In addition to the regular metrics for content preservation and fluency, dos Santos *et al.* (2018) reported the classification accuracy using the classifier from Davidson *et al.* (2017b).

5.2.5 Literality

Figurative language can be considered a style because it embellishes things that could be said plainly (e.g., the statement "*He is a couch potato*" creatively conveys that a person is inactive). It includes (but is not limited to) metaphors, similes, idioms, and oxymorons, each of which has distinctive features and requires different levels of cognitive processing. Expressions of this type have non-standard meanings, which are somewhat derivative of their literal ones (Paul 1970). This makes the distinction between figurative and literal styles blurred. Instead of dichotomies, they represent different sites on a continuum (Gibbs Jr. and Colston 2006).

Computational studies on figurative language have favored metaphors (Niculae and Yaneva 2013), but the only form of figurative expression that has entered the style transfer literature is the simile, "a figure of speech comparing two essentially unlike things and often introduced by *like* or *as*" (Paul 1970). Similes are figurative precisely because the items they compare are essentially dissimilar from one another (Bredin 1998), unlike direct comparisons. Thus, "*She is like her mother*" is not a simile, while "*Her smile is like sunshine*" is.

Chakrabarty, Muresan, and Peng (2020b) were the first to frame *simile* generation as a style transfer task. Their goal was to replace the literal expression (usually an adjective or an adverb) at the end of a sentence with a figurative substitute (e.g., "*You just started staring off into space and smiling dangerously*" → "*You just started staring off into space and smiling like a lunatic*").

Data. A parallel dataset for similes with approximately 87k sentences was created by Chakrabarty *et al.* (2020b). It was built in an automatic manner, crawling self-labelled simile expressions from Reddit via the comparative phrase *like a* (e.g., "*The boy was like an ox*"). The authors employed Comet (Bosselut et al., 2019), a pretrained language model fine-tuned on the ConceptNet (Speer *et al.* 2017) knowledge graph, to replace the logical object of the comparison (here, "*an ox*") with its shared property (here, "*being strong*") in order to generate the parallel sentence (e.g., "*The boy was strong*").

Method. Chakrabarty *et al.* (2020b) exploited a simplified lexical structure followed by a simile, with clearly defined roles for the lexical elements. In the example "*Her smile is like sunshine*", the author intended to describe the topic, *her smile*, by comparing it to a logical object, *sunshine*, via a shared property, i.e., their brightness. The parallel dataset they curated with distant supervision served to fine-tune Bart (Lewis et al., 2020), a pretrained language model that is a combination of bidirectional and auto-regressive transformers. They also conducted experiments with baseline models based on conditional generation, metaphor masking, and retrieval using Comet (Bosselut et al., 2019). Hence, they demonstrated that incorporating structured common sense knowledge through Comet is effective and can be employed in related creative text generation tasks. The fine-tuned Bart model successfully generated novel sentences and generalized over unseen properties.

Table 9. Literature on *intended, non-targeted* styles corresponding to *conventional genres*, divided by method

| | Parallel | Non-parallel | | |
		Exp. Disent.	Imp. Disent.	No Disent.
News	–	–	Romanov et al. (2019); Fu et al. (2018)	Gatti et al. (2015); Gatti et al. (2016); Lee et al. (2019); Zhang et al. (2018a); Chen et al. (2021)
Tech.	Cao et al. (2020)	–	–	–
Literature	Xu et al. (2012); Jhamtani et al. (2017); Carlson et al. (2018); Bujnowski et al. (2020)	Gero et al. (2019)	Romanov et al. (2019)	Mueller et al. (2017); Pang and Gimpel (2019); Shang et al. (2019); Krishna et al. (2020); He et al. (2020)
Lyrics	–	–	–	Lee et al. (2019); Krishna et al. (2020)

Evaluation.   For automatic evaluation, Chakrabarty *et al.* (2020b) reported Bleu after removing the common prefix in the generated and reference sentences. Moreover, they leveraged BertScore (Zhang *et al.* 2020a), a measure indicating the similarity between candidate and reference sentences that uses contextual embeddings, for the contextual vectors of the logical object of the comparison phrases. Human evaluation aimed at comparing the literal utterances against six generated outputs, rated on a scale of 1-to-5 with respect to creativity, overall quality, relevance of the comparison object in portraying the shared property, and relevance of the suggested comparison object in the given topic context.

5.3 Non-targeted: conventional genres

Established textual varieties, like poems, newspaper articles and academic productions flow into the *conventional* category (see an overview in Table 9). This family of styles includes institutionalized types of communication, which are encoded within one (or many) culture(s) (Biber 1995). Hence, they follow some systematic norms, and for this reason they are different from *circumstantial* styles, in which linguistic choices are due to social and contingent situations.

Different genres (henceforth, styles) are recognizable by some markers that can be more or less explicit (e.g., *the objective of this paper is*. . . vs. *once upon a time*. . .) (Coutinho and Miranda 2009). Scientific articles, for instance, put constraints on one's vocabulary choices and syntactic structures, as opposed to literary genres, which allow for freer linguistic constructions (e.g., including evaluative adjectives, metaphors, etc.) (Biber 1995). Their transfer includes objectives like the versification of a prose, the satirization of a novel, or the simplification of technical manuals. Tasks with such kinds of styles are appealing for end users—turning poems into paraphrases has the potential to support education and transforming existing news headlines to produce catchier ones can be useful for advertisement. They also bear a potential value from a theoretical perspective: style transfer can foster academic attempts to describe what genre is, because manipulating markers offers different conditions of investigation, and this might help explain how readers decide about the membership of a text into a certain category.

5.3.1 Forums/newspapers

While the transfer of *newspaper*-based attributes has taken a number of forms, early attempts involved the concept of "blending". Blending consists in rephrasing and incorporating a piece of

Table 10. Examples of style transfer outputs on different *conventional genres* of text—*forums & newspapers, literature, technical language,* and *song lyrics*—taken from Chen *et al.* (2021); Xu et al. (2012); Cao et al. (2020); Lee et al. (2019)

Newspapers	*Economic Frame*: *"It's time for Congress to take action," says a spokesman for the bill's sponsors, who want a flexible spending limit*
	Legality Frame: *"Illegal aliens' is a growing problem in the country," says a spokesman for the measure's sponsors*
Literature	**Early Modern English**: *I will bite thee by the ear for that jest*
	Contemporary English: *I'll bite you by the ear for that joke*
Technical Language	**Expert**: *Many cause dyspnea, pleuritic chest pain, or both*
	Layman: *The most common symptoms, . . ., are shortness of breath and chest pain*
Song Lyrics	**Hip-hop**: *Yo, where the hell you been?*
	Pop: *Yo, where the hell are you?*

text with a secondary (arbitrary) idea, to produce an utterance that evokes not only the original meaning but also the newly juxtaposed one. For instance, a given expression (a slogan like "*Make love not war*", or a cliché, a song, a movie title) can be blended with the daily news (e.g., the headline "*Women propose sex strike for peace*"), such that the result will contain a reference to both (e.g., "*Make peace not war*" (Gatti *et al.* 2015)). These initial works did not explicitly formulate the task as style transfer, but as one where the stylistic attributes used to communicate the news of the day are rendered more similar to a well-known expression.

Without tapping on notions related to creativity, Lee et al. (2019) addressed the problem of transferring the stylistic features of forums to news (e.g., "*i guess you need to refer to bnet website then*" → "*I guess you need to refer to the bnet website then*"), which in their view amounts to a task of formality transfer and Fu et al. (2018) ventured the goal of scientific paper to newspaper title transfer ("*an efficient and integrated algorithm for video enhancement in challenging lighting conditions*" → "*an efficient and integrated algorithm, for video enhancement in challenging power worldwide*"). The transfer was also made between the stylistic attributes of different newspapers. Zhang, Ding, and Soricut (2018a) showed that publishers can be taken proxies for style (e.g., the New York Times[af] has a different stylistic cipher from the Associated Press[ag]) as they tend to use different wording patterns.

Taking a different approach, a line of research addressed the problem of "reframing" news. This type of conditioned paraphrasing consists in changing the perspective from which a topic is conveyed (Chen et al., 2021), for the audience to focus on some of its aspects and prefer a particular interpretation. There, the stylistic attributes of *newspapers* are the frames that are evoked by a piece of text (e.g., economics-, legality-related frames). These can prompt two texts to have the same denotation/reference but different connotations, which is the case for "*undocumented workers*" and "*illegal aliens*". This task is similar to the argument rewriting discussed with respect to *emotional state*, it is close to *sentiment* (as it connects to rewriting with a more positive or negative presentation of the topic) and it touches upon the notion of contextual style transfer (discussed under *formality*) because it needs to ensure that an output sentence is coherent with the surrounding context. Some examples are in Table 10.

Data. A useful newspaper dataset for style transfer was created by De Mattei *et al.* (2020), even though their work regarded style-aware generation rather than transfer. They collected news that are lexically similar from two newspapers, a subset of which are topic-aligned. Gatti *et al.* (2016)

[af]https://www.nytimes.com
[ag]https://apnews.com/

used the news of the day, extracted from the RSS feed of the New York Times and BBC News, and Lee et al. (2019) resorted to articles from the New York Times and comments from Reddit.

Another dataset dedicated to news articles is the Gigaword corpus[ah] (Parker *et al.* 2011). This resource was acquired over several years by the Linguistic Data Consortium, and it spans seven international sources of English newswire (i.e., Agence France-Presse, Associated Press Worldstream, Central News Agency of Taiwan, Los Angeles Times/Washington Post Newswire Service, New York Times, Xinhua News Agency, and Washington Post/Bloomberg Newswire Service). Fu et al. (2018) focused instead on news titles. They built a dataset[ai] of 108,503 titles belonging to the science and technology categories and which come from the UC Irvine Machine Learning Repository (Dua and Graff 2017). As an attribute opposite to "news", their corpus contains scientific-oriented language, specifically paper titles crawled from academic websites.

The reframing study of Chen et al. (2021) made use of the corpus published by Card *et al.* (2015). Encompassing more than 35k news articles about death penalty, gun control, immigration, same-sex marriage and tobacco, the corpus is annotated with 15 framing dimensions (e.g., economics, morality, politics) developed by Boydstun *et al.* (2014).

Methods.   Gatti et al. (2015) performed lexical substitution by extracting keywords from the news and inserting them in well-known expressions coming from slogans, movie titles, song titles, and clichés: after pairing the two data based on a similarity measure, they used a dependency metrics to find the probability for the words in the slogan of being replaced with the same part-of-speech keywords from the news.

More recent neural attempts aimed at transferring news titles to scientific paper titles. This was done by Romanov et al. (2019), who fit in the picture of disentanglement based on adversarial methods. They had an encoder produce a continuous style vector and a meaning vector for a given input. Compared to other adversarial approaches, these authors employed two complementary forces. One was a discriminator that penalized the encoder if the meaning embeddings still carried information about style; the other was a motivator, and it pushed the encoder to produce representations that facilitate the correct attribute classification—encouraging, rather than penalizing, was proven to make the separation between the two types of embeddings bolder.

Moving on to news reframing, Chen et al. (2021) characterized the problem in the following terms: given three consecutive sentences and a target frame, the middle sentence can be masked, and a new one generated to fill in such blank, which contains the target frame and links the preceding and follow up sentences coherently. The authors trained one generation model for each frame, and experimented with three strategies. Namely, fine-tuning a sequence-to-sequence model on a specific frame, including knowledge about named entities to promote topic coherence, and adding examples in the training data (the sentence to be generated has a different frame compared to the surrounding ones).

Evaluation.   De Mattei et al. (2020) put forward the idea that *news* styles are more difficult to judge than others (e.g., sentiment), and that humans are not as reliable judges of said styles as machines. They proposed a framework for the automatic evaluation of style-aware generation that seems handy for style transfer as well. Their automatic classifier had to distinguish the newspaper style of lexically aligned headlines: such an alignment pushed the classifier to make decisions based on stylistic information rather than content-related one.

With respect to human evaluation, Gatti et al. (2015) asked people if an output headline was grammatically correct and if it could work as a headline for a given article, while Chen et al. (2021) conducted an extensive study in which they presented crowdworkers with multiple reframings for

---

[ah]https://catalog.ldc.upenn.edu/LDC2011T07
[ai]https://github.com/fuzhenxin/textstyletransferdata

an input text, which had to be evaluated for their contextual coherence, topical congruence, and presence of a given frame.

5.3.2 Technical language

The curse of knowledge, an expression introduced by Camerer, Loewenstein, and Weber (1989), is a cognitive bias that arises in communication, for instance between professionals in a certain field and less expert people. It can be observed when a well-informed agent assumes understanding from less informed ones, thus hampering a successful exchange of ideas. Style transfer methods can be applied to such situations to simplify language and mitigate the lack of shared knowledge between the two parties.

The task of automatic rewriting to make texts more easily readable (while securing their relevant information) has sparked wide attention in NLP (Wubben, van den Bosch, and Krahmer 2012; Zhang and Lapata 2017; Zhao *et al.* 2018c), but only one work follows the paradigm of style transfer. With a focus on scientific (or technical) texts, Cao et al. (2020) performed expertise style transfer suggesting reformulations of sentences like "*Many cause dyspnea, pleuritic chest pain, or both.*" as "*The most common symptoms, regardless of the type of fluid in the pleural space or its cause, are shortness of breath and chest pain.*". Their goal was to demonstrate how paraphrasing medical jargon can promote better understanding. Hence, for this task, the stylistic attribute of a text is given by the level of domain knowledge that the text involves.

Data. An obvious prerequisite for style transfer in a specialized genre is the availability of domain-specific data. Cao et al. (2020) introduced an expert-annotated parallel corpus[aj] in the medical domain. It was derived from human-written medical references tailored for consumers vs. healthcare professionals who, in their view, are set apart by two major knowledge gaps: one related to technical terminology ("*dyspnea*" → "*shortness of breath*") and one related to the understanding of empirical evidence (e.g., "*About 1/1,000*" → "*quite small*").

Methods. The major contribution of Cao et al. (2020) was the dataset itself, that they evaluated with five state-of-the-art models from prior style transfer (Hu et al., 2017; Li et al., 2018; Dai et al., 2019) and text simplification studies (Shardlow and Nawaz 2019; Surya *et al.* 2019).

Evaluation. The adopted evaluation methods in Cao et al. (2020) were transfer accuracy based on a classifier's performance, fluency based on the perplexity of a fine-tuned Bert model, and content preservation computed in terms of Bleu. In their human evaluation study, laypeople rated content preservation in the model-generated output on a 1-to-5 scale, given both the input and human-produced gold references. The metrics Sari (Xu *et al.* 2016) was also used to evaluate language simplicity, as it compares the n-grams in the generated output with the input and human references, taking into account the words that were added, deleted and retained by the model. The authors concluded that for transfers regarding this style, there exists a substantial difference between the quality of machine-produced and human-produced texts.

5.3.3 Literature

Literature-centered styles have sparked many formulations of style transfer. Most of them tackle the problem of making an old text sound more modern, but ultimately, this type of task shifts the attributes of several styles simultaneously. Even those works that present themselves as mapping text between diachronically different language varieties, in fact, transfer between textual structures (e.g., from sonnets to plain sentences), including differences at various levels of granularity:

[aj]https://srhthu.github.io/expertise-style-transfer/

in the register, in the vocabulary choices, in the senses of words, and in the syntactical constructions (Jhamtani et al., 2017). This also occurs in some studies that focus on author imitation—i.e., rewriting sentences as if that was done by a well-known author, to mimic their stylistic touch (He et al., 2020).[ak]

In this light, *literature* in style transfer seems related to a notion of idiostyle (i.e., a space of linguistic idiosyncrasies specific to writers), which makes it kin to the *background* node of *persona* in our hierarchy. Nevertheless, we dedicate a separate discussion to it as an *intended* style because the writers' artistic speech might reflect the (unintentionally expressed) style of the time but does not coincide with it—within certain time spans, it is actually the idiostyle of established writers that creates a linguo-typological variant of literary texts (Sydorenko 2018). Moreover, such idiostyles need to be (intentionally) adapted to the genre of the writers' literary productions, as these are intended to have an audience.

There exist many examples of this stream of research. Shang et al. (2019) paraphrased old Chinese poems, Bujnowski et al. (2020) and Carlson, Riddell, and Rockmore (2018) switched between the prose attributes of various versions of the Bible ("*Then Samuel gave him an account of everything, keeping nothing back*" → "*And Samuel told all things, and did not hold back*"); Xu et al. (2012), Jhamtani et al. (2017), and He et al. (2020) accounted for the features of Shakespearean plays, transferring Early Modern to contemporary English ("*I will bite thee by the ear for that jest*" → "*I'll bite you by the ear for that joke*") or vice versa ("*Send thy man away*" → "*Send your man away*"). A similar goal was addressed by Pang and Gimpel (2019) but with Dickens' literature, while Krishna *et al.* (2020) performed style transfer with different styles and attributes, transforming tweets into Shakespearean-like texts, Shakespearean texts into Joyce-sounding writings,[al] Joyce-authored texts into Bible-styled ones, and Bible verses into poems. These works hence exemplify that there are transfer works in which the shift does not occur along one conceptual dimension (e.g., presence vs. absence of Shakespeare's style), but rather go from a style to another (e.g., from Shakespeare to Joyce). Therefore, to view style as a non-categorical variable seems a good option for this task. As delineated in Romanov et al. (2019), this would not only account for the reality of language in which the attributes of different genres[am] overlap, but if applied to the literature of specific authors, it would allow to understand how each author relates to the others in a continuous stylistic space.

Gero *et al.* (2019) offered yet another perspective, which radically rethinks the relation of style to content. They delineated a well-defined notion of style in literature, starting from an early quantitative study by Mendenhall (1887), which revealed that writers present some systematic features in their vocabulary, word length, word frequencies and compositions. To Gero et al. (2019), this means that words that are most frequently used (i.e., non-content words) are actually those most indicative of one's literary style. They thus showed that non-content words allow a classifier to determine style, and they leveraged those to transfer between gothic novels, philosophy books, and pulp science fiction, hereafter sci-fi.

Data.   Carlson *et al.* (2018) contributed to fixing the lack of parallel data for style transfer. They collected a high-quality parallel corpus without the involvement of any automatic alignment effort. Their resource contains 34 versions of the Bible produced by professionals and which are

---

[ak]A challenge of this family of styles is given by the name of the characters present in a story, which differs from author to author –an interesting study in this direction was made by Stamatatos (2017).

[al]Note that we did not mention transfer works that shift style from one author to the other by including multiple authors (e.g., Syed *et al.* 2020; Singh *et al.* 2021). As opposed to the Shakespeare-Joyce example given above, which paraphrases texts conditioned on a diachronical dimensions and with respect to their poem or poetry nature, these works take style as persistent characteristics of specific individuals. Hence, they cannot be generalized and subsumed under any specific style category in our hierarchy.

[am]By "genre" we mean what Romanov et al. (2019) call "'register".

naturally aligned, given the structure of such texts, i.e., in chapters and verses. Each version corresponds to an English stylistic value (e.g., archaic, simple, American). They made the dataset available for the texts that were already public.

Pang and Gimpel (2019) limited themselves to two variants of English, with the old one taken from Dickens' works in Project Gutenberg[an] and the modern version from the Toronto Books Corpus. Focusing on Chinese, Shang et al. (2019) constructed a parallel corpus containing old and modern versions of poems. Xu et al. (2012) made a sentence-aligned corpus of Shakespearean plays and their modern translations freely available. Krishna et al. (2020) built a non-parallel English corpus containing 15 M sentences, which contain 11 styles, including the Bible, Shakespeare, James Joyce. Lastly, the philosophy texts, sci-fi and gothic novels of Gero et al. (2019) also come from mono-style sources. They were extracted from Project Gutenberg and the Pulp Magazine Archive,[ao] respectively.

Methods.   The first attempt at dealing with literature styles explored statistical machine translation (Xu et al., 2012); on top of that, Carlson *et al.* (2018) went for sequence-to-sequence translation models, trained for each target attribute. A sequence-to-sequence network was also leveraged by Jhamtani et al. (2017). They added both a pointer that facilitates the copy of input words, and a dictionary of shakespearean-to-modern word pairs which allows to retrofit pre-trained word embeddings, thus accounting for novel words or words that have changed in meaning.

On the unsupervised side, Pang and Gimpel (2019) experimented with models that include losses corresponding to the three criteria, and that could be used both for model tuning and selection. Among such losses, many of which had been already explored (Shen et al., 2017, i.a.), they tried to favor content preservation with a reconstruction loss, a cyclic consistency loss (similar to the former, but with the transfer happening twice, i.e., from source to target and back), and a paraphrase loss obtained with sentence–paraphrase pairs coming from a parallel dataset.

Author mimicking was addressed with the probabilistic approach of He et al. (2020); similarly aiming at minimizing the manually defined objectives (e.g., content-to-style separation), the semi-supervised method of Shang et al. (2019) employed an encoder-decoder that learns to represent a style within a specific latent space, and a projection function that maps the latent representations of one attribute onto the other. The two steps leveraged non-parallel and parallel data respectively. Instead, Krishna *et al.* (2020) adopted their inverse paraphrasing approach already introduced with the *background* styles.

Style and content were handled separately by Gero et al. (2019). In line with their POS-based characterization of style, they defined some low-level linguistic features (e.g., frequency of pronouns, prepositions) as the style of a text, and they performed style transfer by inputting an encoder-decoder with only the content words, which allowed the generation to maintain them while adjusting the features of the target attribute. By contrast, Mueller et al. (2017) refrained from defining editing features or rules. Claiming that revisions of combinatorial structures are unlikely to be found by simple search procedures, they addressed the Shakespearization of language as a problem of finding improved rewrites of a text.

Evaluation.   To measure the quality of paraphrases, Carlson *et al.* (2018), Jhamtani et al. (2017) and Xu et al. (2012) accompanied Bleu, a measure that fundamentally favors textual similarity at the word level, with Pinc, which instead rewards the diversity of the output from the source text thanks to the number of n-grams in the candidate output that do not appear in the source. To measure the transfer strength criterion, Xu et al. (2012) used a language model to compute the posterior probability that a sentence was generated from a model of the target language.

[an]https://www.gutenberg.org
[ao]https://archive.org/details/scifi-corpus

Pang and Gimpel ([2019](#)) introduced a way to measure the success of transfer by aggregating the metrics: an adjusted geometric mean between the accuracy, content preservation and perplexity, which penalizes perplexity scores that are too low, often achieved with short phrases but not meaningful sentences. For human evaluation, their annotators decided which of two generated sentences they preferred with respect to the three transfer criteria. The sentences were taken from different model variants, to observe the correlation between human judgments and each system.

5.3.4 Song lyrics

"*Yo, where the hell you been?*" → "*Yo, where the hell are you?*" is an example of transfer from Lee et al. ([2019](#)), who shifted the genre of lyrics between Hip Hop and Pop songs. A similar attempt was made by Krishna *et al.* ([2020](#)). Their work did not directly alter lyrics attributes (i.e., the music category to which lyrics would belong), but it mapped such texts to a completely different style. As a result, for instance, they made lyrics gain the style of tweets produced by African American English writers (e.g., given the input "*It's a good thing you don't have bus fare*", an output would be "*It's a goof thing u aint gettin no ticket*").

Data.   This task leveraged non-parallel lyrics resources from MetroLyrics[ap] in which more than 500k songs are associated to specific music genres.

Methods.   Lee et al. ([2019](#)) treated the problem as a denoising one, with the same model used to transfer the *background* of *persona* described in Section [2.2.2](#). The non-parallel source data were noised with a model trained on clean-noisy sentence pairs extracted from a language learner forum; the newly synthesized texts were then re-ranked according to their proximity to the target attribute and to the meaning of the source inputs; lastly, a denoising model was trained to find the probability of a clean text (i.e., the target), given the noisy one (i.e., the source).

Evaluation.   Unlike other studies, Lee et al. ([2019](#)) defined the transfer strength criterion as the ratio between the probability of the output belonging to the target domain and the probability of observing it in the source domain.

6. Discussion and conclusion

Style transfer seems to have a bright future ahead owing to its myriad of applications, from online communication (e.g., as an assistant) to studies within NLP (e.g., for data augmentation), and its potential to reveal facts about language. "Operating at all linguistic levels (e.g., lexicology, syntax, text linguistics, and intonation) [. . .] style may be regarded as a choice of linguistic means; as deviation from a norm; as recurrence of linguistic forms; and as comparison." (Mukherjee [2005](#)). Language is creative, it is situated, and has to do with our communicative competence: its users can give new meanings to old words (Black [1968](#)), produce utterances within a particular time and place (Bamman, Dyer, and Smith [2014](#)), and determine if they are appropriate in specific contexts (Hymes [1966](#)). Hence, the variety of realizations in which the same message can be shaped stems from many distinct factors. On the one hand are variations related to personal differences between speakers (e.g., a person's class, gender, social environment) and on the other are those occurring within the speech acts of a single speaker (Labov [1966](#)). We unified these insights into a hierarchy of styles, as a way to relate them to one another.

Our discussion started from the frameworks typically used to learn the task. We summarized the method-oriented survey of Hu et al. ([2022](#)), and showed that many publications consider transfer as a problem of translation between attributes, others assume that style lurks in certain

---

[ap]http://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics

portions of texts and transform it with localized textual changes, or leverage special training functions to reach the three output desiderata. Tables 1, 3, 5, 7, and 9 give an overview of the studies we detailed by style and method, and they further include some recent pre-prints that we did not explicitly mention in the main text. Are current methods sufficient to tackle the complexity of a style of interest? The tables show that not all methods have been evaluated for all styles. The reader is left with the decision of whether this is a signal for promising research gaps, or instead points at an important caveat of style transfer. Namely, some approaches might be acceptable to alter, e.g., sentiment, like retrieval-based frameworks, but they might miss the mark for styles in which paraphrases can be expected to be bolder, non-limited to lexical changes (Yamshchikov et al., 2019). In this sense, our style-oriented survey was also meant to encourage new technical development.

More importantly, we pushed style transfer to question the styles it addresses, while acknowledging that many others (and more varied attributes than binary ones) could be explored. Our analysis revealed that some are under-explored and inherently difficult to transfer. An example is *humor*, a multifaceted phenomenon with tremendous variation depending on the culture and the social settings in which it is deployed. Further, many styles are intertwined. For instance, we put *background* with other stable traits as an inter-speaker difference (i.e., under *persona*), but this choice does not account for speakers shifting their general speech patterns over time (similar to a *dynamic state*), as a result of moving to a different dialect region or interacting with different social groups. On a higher level in the hierarchy, style contaminations are possible between *intended* styles, and between them and *unintended* subsets, e.g., one can write a poem while being romantic, and a certain cultural background can emerge while being more or less polite. This is also reflected in the varied ways in which the publications themselves formulate the transfer problem. A case in point is *literature*, which fits multiple positions in the hierarchy, as it is addressed by some as a diachronic variation (Romanov et al., 2019) and by others as author mimicking (He et al., 2020).

The interconnection between the *unintended* and *intended* branches of the hierarchy exemplifies that styles are a multidimensional concept and cannot always be told apart from one another. Informative in this regard are a number of studies that did not revolve around transfer, such as those by Riloff *et al.* (2013), Mohammad, Shutova, and Turney (2016) and Felt and Riloff (2020) concerned with the link between affective states (e.g., *emotion state*) and figurative language (i.e., *literality*). At the same time, only some combinations of stylistic attributes might be acceptable. As pointed out in an investigation of style inter-dependence (Kang and Hovy 2021), the presence of impoliteness and positive sentiment in the same text might be paradoxical.

A more serious theoretical understanding of style could inform future computational research. For one thing, it could cast doubt on the possibility of addressing style transfer with any feature of text that can be shifted along some dimensions and that appears to tie in with some extra-propositional content of texts—a trend that currently dominates the field. If anything, evaluation approaches can be refined for said styles. The outputs of state-of-the-art systems reveal indeed that the available evaluation metrics are inadequate, but the problem might be upstream. Namely, the three criteria quantified by such metrics arguably generalize across styles. Is a successful system for the transfer of sentiment supposed to maintain meaning as much as a politeness-conditioned system? Precisely because different styles have different linguistic realizations, expecting that the systems addressing them (often, the very same system) perform similarly seems somewhat unreasonable. Transfer, meaning, and grammaticality may be variously reached for each style, making it more urgent to ask "*to what extent can a method changing the polarity of a text retain its semantics?*" than measuring if it did. In other words, an investigation of transfer with respect to individual styles can redefine the task at hand and reconsider the attainable goals.

Readers might have noticed that we indistinctly called "style" both linguistic variations (e.g., formality) and aspects that underlie them (gender correlates with, but is not, style). We also disregarded if the selected articles actually deal with a feature of language that corresponds to *how*

things are said: all the styles that the body of research presents as such were included in our hierarchy. In fact, this field lacks a stable definition of style—unsurprisingly, since no consensus exists on it.

Neither did we take the challenge to define "style" ourselves. We gave a loose characterization of it, adapting one that is established among linguists (Bell 1984). That is, style correlates to external factors, of which gender and personality are an instance. Still, the example outputs we provided convey the following: to assume that a text can be paraphrased with any attribute corresponds to taking style and content as independent variables. In style transfer, the binomial is thought of in terms groups of "semantic equivalence" subsuming textual instances that differ with respect to their stylistic attribute. However, this view has an evident consequence for the field: if shaping a meaning into specific attributes seems unfeasible (e.g., the transfer of sentiment comes at the expense of losing content, contradicting the independence assumption), then such attributes cannot define a goal for style transfer. Content is information predictive of a future (e.g., what word comes next?), while style is additional information prior to generation and tapping on some personal states of the writers. It is grounded in reality, in the human experience (e.g., gender, ethnicity), and ultimately, in the reasons that push speakers to communicate and that current machines (struggling to transfer) do not have.

Competing interests. We do not declare any competing interests.

References

Abdul-Mageed M. and Ungar L. (2017). EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver: Association for Computational Linguistics, pp. 718–728.

Aitchison J. (1981). *Language Change: Progress or Decay?* London: Fontana.

Alba-Juez L. and Attardo S. (2014). The evaluative palette of verbal irony. *Evaluation in Context* **242**, 93–116.

Allaway E. and McKeown K. (2021). A unified feature representation for lexical connotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online. Association for Computational Linguistics, pp. 2145–2163.

Amin M. and Burghardt M. (2020). A survey on approaches to computational humor generation. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Online. International Committee on Computational Linguistics, pp. 29–41.

Attardo S. and Raskin V. (1991). Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research* **4**(3-4), 293–348.

Balasubramanian V., Kobyzev I., Bahuleyan H., Shapiro I. and Vechtomova O. (2021). Polarized-VAE: Proximity based disentangled representation learning for text generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online. Association for Computational Linguistics, pp. 416–423.

Bamman D., Dyer C. and Smith N. A. (2014). Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, MD: Association for Computational Linguistics, pp. 828–834.

Banerjee S. and Lavie A. (2005). Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72.

Bao Y., Zhou H., Huang S., Li L., Mou L., Vechtomova O., Dai X.-y and Chen J. (2019). Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 6008–6019.

Barbieri F., Saggion H. and Ronzano F. (2014). Modelling sarcasm in Twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Baltimore, MD: Association for Computational Linguistics, pp. 50–58.

Beard A. (2000). *The Language of Politics*. Intertext (London): Routledge.

Beckmann N. and Wood R. E. (2017). Dynamic personality science. Integrating between-person stability and within-person change. *Frontiers in Psychology* **8**, 1486.

Bell A. (1984). Language style as audience design. *Language in Society* **13**(2), 145–204.

Beukeboom C. J. and Semin G. R. (2006). How mood turns on language. *Journal of Experimental Social Psychology* **42**(5), 553–566.

Biber D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* **8**(1), 9–37.

Biber D. and Conrad S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Black M. (1968). *The Labyrinth of Language*. New York: Mentor.

Blodgett S. L., Green L. and O'Connor B. (2016). Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, TX: Association for Computational Linguistics, pp. 1119–1130.

Bloomfield L. (1927). Literate and illiterate speech. *American Speech* **2**(10), 432–439.

Bo H., Ding S. H. H., Fung B. C. M. and Iqbal F. (2021). ER-AE: Differentially private text generation for authorship anonymization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online. Association for Computational Linguistics, pp. 3997–4007.

Bolukbasi T., Chang K.-W., Zou J. Y., Saligrama V. and Kalai A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems* **29**, 4349–4357.

Bosselut A., Rashkin H., Sap M., Malaviya C., Celikyilmaz A. and Choi Y. (2019). COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 4762–4779.

Bostan L.-A.-M. and Klinger R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, NM: Association for Computational Linguistics, pp. 2104–2119.

Boydstun A. E., Card D., Gross J., Resnick P. and Smith N. A. (2014). Tracking the development of media frames within and across policy issues. In *American Political Science Association 2014 Annual Meeting Paper*.

Bredin H. (1998). Comparisons and similes. *Lingua* **105**(1), 67–78.

Brennan M., Afroz S. and Greenstadt R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)* **15**(3), 1–22.

Brentano F. (1874). *Psychology from an Empirical Standpoint*. London: Routledge and Kegan Paul.

Briakou E., Agrawal S., Tetreault J. and Carpuat M. (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*. Online and Punta Cana: Association for Computational Linguistics, pp. 1321–1336.

Briakou E., Agrawal S., Zhang K., Tetreault J. and Carpuat M. (2021b). A review of human evaluation for style transfer. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Online. Association for Computational Linguistics, pp. 58–67.

Briakou E., Lu D., Zhang K. and Tetreault J. (2021c). 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*. Online. Association for Computational Linguistics, pp. 3199–3216.

Briot J.-P., Hadjeres G. and Pachet F. (2020). *Deep Learning Techniques for Music Generation*. Cham: Springer.

Brown P. and Fraser C. (1979). Speech as a marker of situation. In *Social Markers in Speech*. Cambridge: Cambridge University Press, pp. 33–62.

Bucholtz M. (2006). Word up: Social meanings of slang in california youth culture. *A Cultural Approach to Interpersonal Communication: Essential Readings* **243**, 267.

Buechel S. and Hahn U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia: Association for Computational Linguistics, pp. 578–585.

Bujnowski P., Ryzhova K., Choi H., Witkowska K., Piersa J., Krumholc T. and Beksa K. (2020). An empirical study on multi-task learning for text style transfer and paraphrase generation. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*. Online. International Committee on Computational Linguistics, pp. 50–63.

Camerer C., Loewenstein G. and Weber M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy* **97**(5), 1232–1254.

Camp E. (2012). Sarcasm, pretense, and the semantics/pragmatics distinction*. *Noûs* **46**(4), 587–634.

Cao Y., Shui R., Pan L., Kan M.-Y., Liu Z. and Chua T.-S. (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*. Online. Association for Computational Linguistics, pp. 1061–1071.

Card D., **Boydstun A. E.**, **Gross J. H.**, **Resnik P. and Smith N. A.** (2015). The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing: Association for Computational Linguistics, pp. 438–444.

Carli L. L. (1990). Gender, language, and influence. *Journal of Personality and Social Psychology* **59**(5), 941–951.

Carlson K., **Riddell A. and Rockmore D.** (2018). Evaluating prose style transfer with the bible. *Royal Society Open Science* **5**(10), 171920.

Casel F., **Heindl A. and Klinger R.** (2021). Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*. Düsseldorf: KONVENS 2021 Organizers, pp. 49–61.

Cattell R. B. (1946). Personality structure and measurement. I. The operational determination of trait unities. *British Journal of Psychology* **36**(2), 88.

Cavalin P., **Vasconcelos M.**, **Grave M.**, **Pinhanez C. and Alves Ribeiro V. H.** (2020). From disjoint sets to parallel data to train Seq2Seq models for sentiment transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online. Association for Computational Linguistics, pp. 689–698.

Celli F., **Lepri B.**, **Biel J.-I.**, **Gatica-Perez D.**, **Riccardi G. and Pianesi F.** (2014). The workshop on computational personality recognition 2014. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1245–1246.

Chakrabarty T., **Ghosh D.**, **Muresan S. and Peng N.** (2020a). Rô3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, pp. 7976–7986.

Chakrabarty T., **Hidey C. and Muresan S.** (2021). ENTRUST: Argument reframing with language models and entailment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online. Association for Computational Linguistics, pp. 4958–4971.

Chakrabarty T., **Muresan S. and Peng N.** (2020b). Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online. Association for Computational Linguistics, pp. 6455–6469.

Charteris-Black J. (2018). *Analysing Political Speeches*. Basingstoke: Macmillan International Higher Education.

Chawla K. and Yang D. (2020). Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online. Association for Computational Linguistics, pp. 2340–2354.

Chen D. and Dolan W. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, OR: Association for Computational Linguistics, pp. 190–200.

Chen W.-F., **Al Khatib K.**, **Stein B. and Wachsmuth H.** (2021). Controlled neural sentence-level reframing of news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana: Association for Computational Linguistics, pp. 2683–2693.

Chen W.-F., **Wachsmuth H.**, **Al Khatib K. and Stein B.** (2018). Learning to flip the bias of news headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pp. 79–88.

Cheng P., **Min M. R.**, **Shen D.**, **Malon C.**, **Zhang Y.**, **Li Y. and Carin L.** (2020a). Improving disentangled text representation learning with information-theoretic guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, pp. 7530–7541.

Cheng Y., **Gan Z.**, **Zhang Y.**, **Elachqar O.**, **Li D. and Liu J.** (2020b). Contextual text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online. Association for Computational Linguistics, pp. 2915–2924.

Chopik W. J. and Giasson H. L. (2017). Age differences in explicit and implicit age attitudes across the life span. *The Gerontologist* **57**(suppl_2), S169–S177.

Chung J., **Gulcehre C.**, **Cho K. and Bengio Y.** (2015). Gated feedback recurrent neural networks. In *International Conference on Machine Learning*. PMLR, pp. 2067–2075.

Clift R. (1999). Irony in conversation. *Language in Society* **28**(4), 523–553.

Coutinho M. A. and Miranda F. (2009). To describe genres: Problems and strategies. In *Genre in a changing world*. Fort Collins, CO: The WAC Clearinghouse and Parlor Press, pp. 35–55.

Czeresnia Etinger I. and Black A. W. (2019). Formality style transfer for noisy, user-generated conversations: Extracting labeled, parallel data from unlabeled corpora. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Hong Kong: Association for Computational Linguistics, pp. 11–16.

Dahlmeier D., **Ng H. T. and Wu S. M.** (2013). Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta: Association for Computational Linguistics, pp. 22–31.

Dai N., **Liang J.**, **Qiu X. and Huang X.** (2019). Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 5997–6007.

Danescu-Niculescu-Mizil C., **Sudhof M.**, **Jurafsky D.**, **Leskovec J. and Potts C.** (2013). A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 250–225.

Davidson T., **Warmsley D.**, **Macy M. and Weber I.** (2017a). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17*, pp. 512–515.

Davidson T., **Warmsley D.**, **Macy M. and Weber I.** (2017b). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17*, pp. 512–515.

Davies M. (2012). Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora* **7**(2), 121–157.

De Mattei L., **Cafagna M.**, **Dell'Orletta F. and Nissim M.** (2020). Invisible to people but not to machines: Evaluation of style-aware HeadlineGeneration in absence of reliable human judgment. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, pp. 6709–6717.

De Saussure F. (1959). *Course in General Linguistics*. New York: Philosophical Library.

Devlin J., **Chang M.-W.**, **Lee K. and Toutanova K.** (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)*. Minneapolis, MN: Association for Computational Linguistics, pp. 4171–4186.

dos Santos C., **Melnyk I. and Padhi I.** (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)*. Melbourne: Association for Computational Linguistics, pp. 189–194.

Dryjański T., **Bujnowski P.**, **Choi H.**, **Podlaska K.**, **Michalski K.**, **Beksa K. and Kubik P.** (2018). Affective natural language generation by phrase insertion. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 4876–4882.

Dua D. and Graff C. (2017). UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. Available at https://archive.ics.uci.edu/ml/citation_policy.html

Eckert P. (1997). Age as a sociolinguistic variable. In *The Handbook of Sociolinguistics*. Malden, MA: Wiley, pp. 151–167.

Eckert P. and McConnell-Ginet S. (1999). New generalizations and explanations in language and gender research. *Language in Society* **28**(2), 185–201.

Eckert P. and McConnell-Ginet S. (2003). *Language and Gender*. Cambridge: Cambridge University Press.

Edelman M. (1985). Political language and political reality. *PS* **18**(1), 10–19.

Eisenstein J., **O'Connor B.**, **Smith N. A. and Xing E. P.** (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, pp. 1277–1287.

Emmery C., **Manjavacas Arevalo E. and Chrupała G.** (2018). Style obfuscation by invariance. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, NM: Association for Computational Linguistics, pp. 984–996.

Fang L., **Li C.**, **Gao J.**, **Dong W. and Chen C.** (2019). Implicit deep latent variable models for text generation. 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)*. Hong Kong: Association for Computational Linguistics, pp. 3946–3956.

Felbo B., **Mislove A.**, **Søgaard A.**, **Rahwan I. and Lehmann S.** (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen: Association for Computational Linguistics, pp. 1615–1625.

Felt C. and Riloff E. (2020). Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*. Online. Association for Computational Linguistics, pp. 136–145.

Feng S. Y., **Li A. W. and Hoey J.** (2019). Keep calm and switch on! Preserving sentiment and fluency in semantic text exchange. 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)*. Hong Kong: Association for Computational Linguistics, pp. 2701–2711.

Ferris S. P. (2002). Writing electronically: The effects of computers on traditional writing. *Journal of Electronic Publishing* **8**(1). DOI 10.3998/3336451.0008.104.

Fink C., **Kopecky J. and Morawski M.** (2012). Inferring gender from the content of tweets: A region specific example. *Proceedings of the International AAAI Conference on Web and Social Media* **6**(1), 459–462.

Foucault M. (1966). *Les Mots et les choses*. Paris: Editions Gallimard.

Founta A. M., **Chatzakou D.**, **Kourtellis N.**, **Blackburn J.**, **Vakali A. and Leontiadis I.** (2019). A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science, WebSci '19*. New York, NY: Association for Computing Machinery, pp. 105–114.

Foxall G. R. and Goldsmith R. E. (1988). Personality and consumer research: Another look. *Journal of the Market Research Society* **30**(2), 111–125.

Friedman H. S. and Tucker J. S. (1990). Language and deception. In *Handbook of Language and Social Psychology*. Oxford: Oxford University Press, pp. 257–270.

Fu Y., Zhou H., Chen J. and Li L. (2019). Rethinking text attribute transfer: A lexical analysis. In *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo: Association for Computational Linguistics, pp. 24–33.

Fu Z., Tan X., Peng N., Zhao D. and Yan R. (2018). Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence,* Vol. 32. New Orleans: AAAI, pp. 663–670. DOI 10.1609/aaai.v32i1.11330.

Gan C., Gan Z., He X., Gao J. and Deng L. (2017). Stylenet: Generating attractive visual captions with styles. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 955–964.

Gao X., Zhang Y., Lee S., Galley M., Brockett C., Gao J. and Dolan B. (2019). Structuring latent spaces for stylized response generation. 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)*. Hong Kong: Association for Computational Linguistics, pp. 1814–1823.

Gatt A. and Krahmer E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* **61**, 65–170.

Gatti L., Guerini M., Callaway C. B., Stock O. and Strapparava C. (2012). Creatively subverting messages in posters. In *ICCC*, pp. 175–179.

Gatti L., Özbal G., Guerini M., Stock O. and Strapparava C. (2015). Slogans are not forever: Adapting linguistic expressions to the news. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Gatti L., Ozbal G., Guerini M., Stock O. and Strapparava C. (2016). Heady-lines: A creative generator of newspaper headlines. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces*, pp. 79–83.

Gatys L. A., Ecker A. S. and Bethge M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423.

Ge T., Zhang X., Wei F. and Zhou M. (2019). Automatic grammatical error correction for sequence-to-sequence text generation: An empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 6059–6064.

Genette G. (1997). *Palimpsests: Literature in the Second Degree*, vol. 8. Lincoln: University of Nebraska Press.

Gero K., Kedzie C., Reeve J. and Chilton L. (2019). Low level linguistic controls for style transfer and content preservation. In *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo: Association for Computational Linguistics, pp. 208–218.

Ghosh S., Chollet M., Laksana E., Morency L.-P. and Scherer S. (2017). Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver: Association for Computational Linguistics, pp. 634–642.

Gibbs Jr.., **R. W. and Colston H. L.** (2006). Figurative language. In *Handbook of Psycholinguistics*. Amsterdam: Elsevier, pp. 835–861.

Giles H. and Johnson P. (1987). Ethnolinguistic identity theory: A social psychological approach to language maintenance. *International Journal of the Sociology of Language* **68**, 69–99.

Gohary A. and Hanzaee K. H. (2014). Personality traits as predictors of shopping motivations and behaviors: A canonical correlation analysis. *Arab Economic and Business Journal* **9**(2), 166–174.

Gong H., Bhat S., Wu L., Xiong J. and Hwu W.-M. (2019). Reinforcement learning based text style transfer without parallel training corpus. 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)*. Minneapolis, MN: Association for Computational Linguistics, pp. 3168–3180.

Gong H., Song L. and Bhat S. (2020). Rich syntactic and semantic information helps unsupervised text style transfer. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 113–119.

González-Ibáñez R., Muresan S. and Wacholder N. (2011). Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, OR: Association for Computational Linguistics, pp. 581–586.

Goswamy T., Singh I., Barkati A. and Modi A. (2020). Adapting a language model for controlled affective text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona: International Committee on Computational Linguistics, pp. 2787–2801.

Graesser A. C., McNamara D. S., Cai Z., Conley M., Li H. and Pennebaker J. (2014). Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal* **115**(2), 210–229.

Green L. (1998). Aspect and predicate phrases in african-american vernacular english. In *African-American English: Structure, History, and Use*. New York: Routledge, pp. 37–68.

Grimminger L. and Klinger R. (2021). Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Online. Association for Computational Linguistics, pp. 171–180.

Gruner C. R. (1997). *The Game of Humor: A Comprehensive Theory of Why We Laugh*. New Brunswick, NJ: Transaction Publishers.

Guerini M., **Strapparava C. and Stock O.** (2008). Valentino: A tool for valence shifting of natural language texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech: European Language Resources Association (ELRA).

Guu K., **Hashimoto T. B.**, **Oren Y. and Liang P.** (2018). Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics* **6**, 437–450.

Habermas J. (2006). Political communication in media society: Does democracy still enjoy an epistemic dimension? the impact of normative theory on empirical research1. *Communication Theory* **16**(4), 411–426.

Halliday M. A. K. (1989). *Spoken and Written Language. Language Education*. Oxford: Oxford University Press.

He J., **Wang X.**, **Neubig G. and Berg-Kirkpatrick T.** (2020). A probabilistic formulation of unsupervised text style transfer. In *Proceedings of International Conference on Learning Representations*.

Helbig D., **Troiano E. and Klinger R.** (2020). Challenges in emotion style transfer: An exploration with a lexical substitution pipeline. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. Online. Association for Computational Linguistics, pp. 41–50.

Heylighen F. and Dewaele J.-M. (1999). Formality of language: Definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4.

Hofmann J., **Troiano E.**, **Sassenberg K. and Klinger R.** (2020). Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona: International Committee on Computational Linguistics, pp. 125–138.

Hoijer H. E. (1954). *Language in Culture; Conference on the Interrelations of Language and Other Aspects of Culture*. Chicago: University of Chicago Press.

Holtgraves T. M. (2001). *Language as Social Action: Social Psychology and Language Use*. New York: Psychology Press.

Hossain N., **Krumm J. and Gamon M.** (2019). "President vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. 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)*. Minneapolis, MN: Association for Computational Linguistics, pp. 133–142.

Hu M. and Liu B. (2006). Opinion extraction and summarization on the web. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2, AAAI'06*. AAAI Press, pp. 1621–1624.

Hu Z., **Lee R. K.-W.**, **Aggarwal C. C. and Zhang A.** (2022). Text style transfer: a review and experimental evaluation. *SIGKDD Explorations Newsletter* **24**(1), 14–45.

Hu Z., **Yang Z.**, **Liang X.**, **Salakhutdinov R. and Xing E. P.** (2017). Toward controlled generation of text. *Proceedings of Machine Learning Research* **70**, 1587–1596.

Huang C., **Zaïane O.**, **Trabelsi A. and Dziri N.** (2018). Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, LA: Association for Computational Linguistics, pp. 49–54.

Hudson R. A. (1993). *Sociolinguistics*. Cambridge: Cambridge University Press.

Hughes J. M., **Foti N. J.**, **Krakauer D. C. and Rockmore D. N.** (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of The National Academy of Sciences of The United States of America* **109**(20), 7682–7686.

Hymes D. (1966). Two types of linguistic relativity. In *Sociolinguistics*. De Hague: Mouton, pp. 114–158.

Irvine J. T. (1979). Formality and informality in communicative events. *American Anthropologist* **81**(4), 773–790.

Jafaritazehjani S., **Lecorvé G.**, **Lolive D. and Kelleher J.** (2020). Style versus content: A distinction without a (learnable) difference? In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona: International Committee on Computational Linguistics, pp. 2169–2180.

Jhamtani H., **Gangal V.**, **Hovy E. and Nyberg E.** (2017). Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*. Copenhagen: Association for Computational Linguistics, pp. 10–19.

Jin D., **Jin Z.**, **Hu Z.**, **Vechtomova O. and Mihalcea R.** (2022). Deep learning for text style transfer: A survey. *Computational Linguistics* **48**(1), 155–205.

Jin Z., **Jin D.**, **Mueller J.**, **Matthews N. and Santus E.** (2019). IMaT: Unsupervised text attribute transfer via iterative matching and translation. 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)*. Hong Kong: Association for Computational Linguistics, pp. 3097–3109.

John O. P., **Naumann L. P. and Soto C. J.** (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of Personality Theory and Research*. New York: Guilford Press, pp. 114–158.

John V., **Mou L.**, **Bahuleyan H. and Vechtomova O.** (2019). Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 424–434.

Kacmarcik G. and Gamon M. (2006). Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL, 2006 Main Conference Poster Sessions*. Sydney: Association for Computational Linguistics, pp. 444–451.

Kang D., **Gangal V. and Hovy E.** (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)*. Hong Kong: Association for Computational Linguistics, pp. 1696–1706.

Kang D. and Hovy E. (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)*. Online. Association for Computational Linguistics, pp. 2376–2387.

Kim E., **Padó S. and Klinger R.** (2017). Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Vancouver: Association for Computational Linguistics, pp. 17–26.

Klimt B. and Yang Y. (2004). The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*. Berlin: Springer, pp. 217–226.

Kreuz R. J. and Glucksberg S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General* **118**(4), 374–386.

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

Kruengkrai C. (2019). Learning to flip the sentiment of reviews from non-parallel corpora. 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)*. Hong Kong: Association for Computational Linguistics, pp. 6311–6316.

Kusner M., **Sun Y.**, **Kolkin N. and Weinberger K.** (2015). From word embeddings to document distances. In *International Conference on Machine Learning*. PMLR, pp. 957–966.

Labov W. (1966). *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Lai C.-T., **Hong Y.-T.**, **Chen H.-Y.**, **Lu C.-J. and Lin S.-D.** (2019). Multiple text style transfer by using word-level conditional generative adversarial network with two-phase training. 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)*. Hong Kong: Association for Computational Linguistics, pp. 3579–3584.

Lai H., **Toral A. and Nissim M.** (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)*. Online. Association for Computational Linguistics, pp. 484–494.

Lample G., **Subramanian S.**, **Smith E. M.**, **Denoyer L.**, **Ranzato M. and Boureau Y.** (2019). Multiple-attribute text rewriting, OpenReview.net. In *7th International Conference on Learning Representations, ICLR 2019*. New Orleans, LA, May 6–9, 2019.

Lave J. and Wenger E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Lee D., **Tian Z.**, **Xue L. and Zhang N. L.** (2021). Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. 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)*. Online. Association for Computational Linguistics, pp. 93–102.

Lee D. Y. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language Learning & Technology* **5**(3), 37–72.

Lee J. (2020). Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer. In *Proceedings of the 13th International Conference on Natural Language Generation*. Dublin: Association for Computational Linguistics, pp. 195–204.

Lee J., **Xie Z.**, **Wang C.**, **Drach M.**, **Jurafsky D. and Ng A.** (2019). Neural text style transfer via denoising and reranking. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. Minneapolis, MN: Association for Computational Linguistics, pp. 74–81.

Leeftink W. and Spanakis G. (2019). Towards controlled transformation of sentiment in sentences. In *International Conference on Agents and Artificial Intelligence*.

Lewis M., **Liu Y.**, **Goyal N.**, **Ghazvininejad M.**, **Mohamed A.**, **Levy O.**, **Stoyanov V. and Zettlemoyer L.** (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, pp. 7871–7880.

Li D., **Zhang Y.**, **Gan Z.**, **Cheng Y.**, **Brockett C.**, **Dolan B. and Sun M.-T.** (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)*. Hong Kong: Association for Computational Linguistics, pp. 3304–3313.

Li H., **Graesser A. C.**, **Conley M.**, **Cai Z.**, **Pavlik P. I. and Pennebaker J. W.** (2016). A new measure of text formality: An analysis of discourse of mao zedong. *Discourse Processes* **53**(3), 205–232.

Li J., **Jia R.**, **He H. and Liang P.** (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)*. New Orleans, LA: Association for Computational Linguistics, pp. 1865–1874.

Li X., **Chen G.**, **Lin C. and Li R.** (2020a). DGST: A dual-generator network for text style transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online. Association for Computational Linguistics, pp. 7131–7136.

Li X., **Sun S. and Wang Y.** (2021). Text style transfer: Leveraging a style classifier on entangled latent representations. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. Online. Association for Computational Linguistics, pp. 72–82.

Li Y., **Li C.**, **Zhang Y.**, **Li X.**, **Zheng G.**, **Carin L. and Gao J.** (2020b). Complementary auxiliary classifiers for label-conditional text generation. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(05), 8303–8310.

Li Y., **Su H.**, **Shen X.**, **Li W.**, **Cao Z. and Niu S.** (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei: Asian Federation of Natural Language Processing, pp. 986–995.

Liao Y., **Bing L.**, **Li P.**, **Shi S.**, **Lam W. and Zhang T.** (2018). QuaSE: Sequence editing under quantifiable guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, pp. 3855–3864.

Lin C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 74–81.

Lin K., **Liu M.-Y.**, **Sun M.-T. and Kautz J.** (2020). Learning to generate multiple style transfer outputs for an input sentence. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*. Online. Association for Computational Linguistics, pp. 10–23.

Ling J. and Klinger R. (2016). An empirical, quantitative analysis of the differences between sarcasm and irony. In Sack H., Rizzo G., Steinmetz N., Mladenić D., Auer S. and Lange C., (eds), *The Semantic Web*. Cham: Springer International Publishing, pp. 203–216.

Liu B. (2012). *Sentiment Analysis and Opinion Mining*. Vermont: Morgan & Claypool.

Liu B. and Zhang L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*. Boston, MA: Springer, pp. 415–463.

Liu C.-W., **Lowe R.**, **Serban I.**, **Noseworthy M.**, **Charlin L. and Pineau J.** (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, TX: Association for Computational Linguistics, pp. 2122–2132.

Liu D., **Fu J.**, **Zhang Y.**, **Pal C. and Lv J.** (2020a). Revision in continuous space: Unsupervised text style transfer without adversarial learning. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(05), 8376–8383.

Liu D., **Fu J.**, **Zhang Y.**, **Pal C. and Lv J.** (2020b). Revision in continuous space: Unsupervised text style transfer without adversarial learning. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(05), 8376–8383.

Logeswaran L., **Lee H. and Bengio S.** (2018). Content preserving text generation with attribute controls. In Bengio S., Wallach H., Larochelle H., Grauman K., Cesa-Bianchi N. and Garnett R., (eds), *Advances in Neural Information Processing Systems*, vol. 31. Red Hook, NY: Curran Associates, Inc.

Luo F., **Li P.**, **Yang P.**, **Zhou J.**, **Tan Y.**, **Chang B.**, **Sui Z. and Sun X.** (2019a). Towards fine-grained text sentiment transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 2013–2022.

Luo F., **Li P.**, **Zhou J.**, **Yang P.**, **Chang B.**, **Sui Z. and Sun X.** (2019b). A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*.

Lyu Y., **Liang P. P.**, **Pham H.**, **Hovy E.**, **Póczos B.**, **Salakhutdinov R. and Morency L.-P.** (2021). StylePTB: A compositional benchmark for fine-grained controllable text style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online. Association for Computational Linguistics, pp. 2116–2138.

Madaan A., **Setlur A.**, **Parekh T.**, **Poczos B.**, **Neubig G.**, **Yang Y.**, **Salakhutdinov R.**, **Black A. W. and Prabhumoye S.** (2020). Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, pp. 1869–1881.

Mai F., **Pappas N.**, **Montero I.**, **Smith N. A. and Henderson J.** (2020). Plug and play autoencoders for conditional text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online. Association for Computational Linguistics, pp. 6076–6092.

Mairesse F. and Walker M. A. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics* **37**(3), 455–488.

Malmi E., **Severyn A. and Rothe S.** (2020). Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online. Association for Computational Linguistics, pp. 8671–8680.

Marcheggiani D. and Titov I. (2017). Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen: Association for Computational Linguistics, pp. 1506–1515.

Martin S. and Wolfram W. (1998). The sentence in African-American vernacular English. In *African American English: Structure, History, and Use*. New York: Routledge, pp. 11–36.

Mehrabian A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* **14**(4), 261–292.

Meier A. (1995). Defining politeness: Universality in appropriateness. *Language Sciences* **17**(4), 345–356.

Mendenhall T. C. (1887). The characteristic curves of composition. *Science* **9**(214), 237–249.

Mendoza-Denton N. and Iwai M. (1993). They speak more caucasian: Generational differences in the speech of Japanese-Americans. In *Proceedings of the First Annual Symposium About Language and Society*, pp. 58–67.

Meyer J. C. (2006). Humor as a double-edged sword: Four functions of humor in communication. *Communication Theory* **10**(3), 310–331.

Mir R., **Felbo B.**, **Obradovich N. and Rahwan I.** (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)*. Minneapolis, MN: Association for Computational Linguistics, pp. 495–504.

Mishra A., **Tater T. and Sankaranarayanan K.** (2019). A modular architecture for unsupervised sarcasm generation. 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)*. Hong Kong: Association for Computational Linguistics, pp. 6144–6154.

Mislove A., **Lehmann S.**, **Ahn Y.-Y.**, **Onnela J.-P. and Rosenquist J.** (2011). Understanding the demographics of twitter users. *Proceedings of the International AAAI Conference on Web and Social Media* **5**(1), 554–557.

Mizumoto T., **Komachi M.**, **Nagata M. and Matsumoto Y.** (2011). Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai: Asian Federation of Natural Language Processing, pp. 147–155.

Mohammad S. (2012). #emotional tweets. In *∗SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal: Association for Computational Linguistics, pp. 246–255.

Mohammad S., **Shutova E. and Turney P.** (2016). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Berlin: Association for Computational Linguistics, pp. 23–33.

Mohammad S. M., **Zhu X.**, **Kiritchenko S. and Martin J.** (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management* **51**(4), 480–499.

Morreall J. (1983). *Taking Laughter Seriously*. Albany: Suny Press.

Mueller J., **Gifford D. and Jaakkola T.** (2017). Sequence to better sequence: Continuous revision of combinatorial structures. In Precup D. and Teh Y. W., (eds), *Proceedings of the 34th International Conference on Machine Learning*, vol. 70. PMLR, pp. 2536–2544.

Mukherjee J. (2005). Stylistics. In *Encyclopedia of Linguistics*. New York: Taylor & Francis, pp. 1043–1045.

Myers I. B. and Myers P. B. (2010). *Gifts Differing: Understanding Personality Type*. London: Nicholas Brealey Publishing.

Nangi S. R., **Chhaya N.**, **Khosla S.**, **Kaushik N. and Nyati H.** (2021). Counterfactuals to control latent disentangled text representations for 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)*. Online. Association for Computational Linguistics, pp. 40–48.

Newman G. E. (2019). The psychology of authenticity. *Review of General Psychology* **23**(1), 8–18.

Newman M. L., **Pennebaker J. W.**, **Berry D. S. and Richards J. M.** (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* **29**(5), 665–675.

Nguyen D., **Gravel R.**, **Trieschnigg D. and Meder T.** (2013). "How old do you think i am?" A study of language and age in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* **7**(1), 439–448.

Niculae V. and Yaneva V. (2013). Computational considerations of comparisons and similes. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*. Sofia: Association for Computational Linguistics, pp. 89–95.

Niu X., **Rao S. and Carpuat M.** (2018). Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1008–1021.

Nobata C., **Tetreault J.**, **Thomas A.**, **Mehdad Y. and Chang Y.** (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web, WWW '16*. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee, pp. 145–153.

Oraby S., **Reed L.**, **Tandon S.**, **S. T. S.**, **Lukin S. and Walker M.** (2018). Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Melbourne: Association for Computational Linguistics, pp. 180–190.

Orwell G. (1962). Politics and the English Language. In *Inside the Whale and Other Essays*. Harmondsworth: Penguin Books, pp. 143–157.

Pang R. Y. (2019a). The daunting task of real-world textual style transfer auto-evaluation. *arXiv preprint, arXiv: 1910.03747*.

Pang R. Y. (2019b). Towards actual (not operational) textual style transfer auto-evaluation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Hong Kong: Association for Computational Linguistics, pp. 444–445.

Pang R. Y. and Gimpel K. (2019). Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Hong Kong: Association for Computational Linguistics, pp. 138–147.

Papineni K., **Roukos S.**, **Ward T. and Zhu W.-J.** (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318.

Parker R., **Graff D.**, **Kong J.**, **Chen K. and Maeda K.** (2011). *English Gigaword*, 5th edn. Philadelphia, PA: Linguistic Data Consortium.

Paul A. M. (1970). Figurative language. *Philosophy & Rhetoric* **3**(4), 225–248.

Peled L. and Reichart R. (2017). Sarcasm SIGN: Interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver: Association for Computational Linguistics, pp. 1690–1700.

Pennacchiotti M. and Popescu A.-M. (2011). A machine learning approach to twitter user classification. *Proceedings of the International AAAI Conference on Web and Social Media* **5**(1), 281–288.

Pennebaker J. W. and Stone L. D. (2003). Words of wisdom: Language use over the life span. *Journal of personality and social psychology* **85**(2), 291–301.

Plank B. and Hovy D. (2015). Personality traits on Twitter—or—How to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Lisboa: Association for Computational Linguistics, pp. 92–98.

Plaza-del Arco F. M., **Halat S.**, **Padó S. and Klinger R.** (2021). Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. In *Forum for Information Retrieval Evaluation, Virtual Event/India*

Popuşoi S. A., **Havârneanu G. M. and Havârneanu C. E.** (2018). "Get the f#∗k out of my way!" Exploring the cathartic effect of swear words in coping with driving anger. *Transportation Research Part F: Traffic Psychology and Behaviour* **56**, 215–226.

Poquérusse J., **Pastore L.**, **Dellantonio S. and Esposito G.** (2018). Alexithymia and autism spectrum disorder: A complex relationship. *Frontiers in psychology* **9**, 1196.

Prabhumoye S., **Black A. W. and Salakhutdinov R.** (2020). Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona: International Committee on Computational Linguistics, pp. 1–14.

Prabhumoye S., **Tsvetkov Y.**, **Black A. W. and Salakhutdinov R.** (2018a). Style Transfer Through Multilingual and Feedback-Based Back-Translation. *arXiv: 1809.06284*.

Prabhumoye S., **Tsvetkov Y.**, **Salakhutdinov R. and Black A. W.** (2018b). Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne: Association for Computational Linguistics, pp. 866–876.

Preoţiuc-Pietro D., **Xu W. and Ungar L.** (2016a). Discovering user attribute stylistic differences via paraphrasing. *Proceedings of the AAAI Conference on Artificial Intelligence* **30**(1). DOI 10.1609/aaai.v30i1.10393.

Preoţiuc-Pietro D., **Schwartz H. A.**, **Park G.**, **Eichstaedt J.**, **Kern M.**, **Ungar L. and Shulman E.** (2016b). Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. San Diego, CA: Association for Computational Linguistics, pp. 9–15.

Pryzant R., **Martinez R. D.**, **Dass N.**, **Kurohashi S.**, **Jurafsky D. and Yang D.** (2020). Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 480–489.

Rabinovich E., **Patel R. N.**, **Mirkin S.**, **Specia L. and Wintner S.** (2017). Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia: Association for Computational Linguistics, pp. 1074–1084.

Radford A., **Wu J.**, **Child R.**, **Luan D.**, **Amodei D.**, **Sutskever I. and** et al. (2019). Language models are unsupervised multitask learners. *Technical Report. OpenAI blog*.

Raffel C., **Shazeer N.**, **Roberts A.**, **Lee K.**, **Narang S.**, **Matena M.**, **Zhou Y.**, **Li W. and Liu P. J.** (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**(140), 1–67.

Rangel F., **Rosso P.**, **Potthast M.**, **Stein B. and Daelemans W.** (2015). Overview of the 3rd author profiling task at PAN 2015. In *Conference and Labs of the Evaluation Forum (CLEF), 2015*.

Rank H. (1980). Analyzing political rhetoric. *The English Journal* **69**(9), 38–43.

Rao S. and Tetreault J. (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)*. New Orleans, LA: Association for Computational Linguistics, pp. 129–140.

Rashkin H., **Smith E. M.**, **Li M. and Boureau Y.-L.** (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 5370–5381.

Raskin V. (1979). Semantic mechanisms of humor. In *Proceedings of the Fifth Annual Meeting of the Berkeley Linguistics Society*, vol. 5, pp. 325–335.

Razavi A. H., **Inkpen D.**, **Uritsky S. and Matwin S.** (2010). Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*. Berlin/Heidelberg: Springer, pp. 16–27.

Reddy S. and Knight K. (2016). Obfuscating gender in social media writing. In *Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*, pp. 17–26.

Reid M. and Zhong V. (2021). LEWIS: levenshtein editing for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online. Association for Computational Linguistics, pp. 3932–3944.

Reif E., **Ippolito D.**, **Yuan A.**, **Coenen A.**, **Callison-Burch C. and Wei J.** (2022). A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin: Association for Computational Linguistics, pp. 837–848.

Reisigl M. (2008). 11. Rhetoric of Political Speeches. In *Handbook of Communication in the Public Sphere*. New York: De Gruyter Mouton, pp. 243–270.

Riley P., **Constant N.**, **Guo M.**, **Kumar G.**, **Uthus D. and Parekh Z.** (2021). TextSETTR: Few-shot text style extraction and tunable targeted restyling. 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)*. Online. Association for Computational Linguistics, pp. 3786–3800.

Riloff E., **Qadir A.**, **Surve P.**, **De Silva L.**, **Gilbert N. and Huang R.** (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, WA: Association for Computational Linguistics, pp. 704–714.

Ritchie G. (1999). Developing the incongruity-resolution theory. In *Proceedings of the AISB Symposium on Creative Language: Stories and Humour*, pp. 78–85.

Romanov A., **Rumshisky A.**, **Rogers A. and Donahue D.** (2019). Adversarial decomposition of text representation. 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)*. Minneapolis, MN: Association for Computational Linguistics, pp. 815–825.

Rosenthal S. and McKeown K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, OR: Association for Computational Linguistics, pp. 763–772.

Rubner Y., **Tomasi C. and Guibas L. J.** (1998). A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE, pp. 59–66.

Rude S., **Gortner E.-M. and Pennebaker J.** (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* **18**(8), 1121–1133.

Rutter J. (1997). *Stand-up as Interaction: Performance and Audience in Comedy Venues*. University of Salford (United Kingdom).

Sarawgi R., **Gajulapalli K. and Choi Y.** (2011). Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pp. 78–86.

Scherer K. R. (2005). What are emotions? And how can they be measured? *Social Science Information* **44**(4), 695–729.

Scherer K. R. and Wallbott H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology* **66**(2), 310–328.

Schmidt A. and Wiegand M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia: Association for Computational Linguistics, pp. 1–10.

Schuff H., **Barnes J.**, **Mohme J.**, **Padó S. and Klinger R.** (2017). Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen: Association for Computational Linguistics, pp. 13–23.

Segarra C. (2007). *How to Become a True Professional*. Maitland, FL: Xulon Press, Incorporated.

Shang M., **Li P.**, **Fu Z.**, **Bing L.**, **Zhao D.**, **Shi S. and Yan R.** (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)*. Hong Kong: Association for Computational Linguistics, pp. 4937–4946.

Shapiro M. J. (1986). Language and politics. *Annual review of applied linguistics* **7**, 74–85.

Shardlow M. and Nawaz R. (2019). Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 380–389.

Shen T., **Lei T.**, **Barzilay R. and Jaakkola T.** (2017). Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pp. 6830–6841.

Shetty R., **Schiele B. and Fritz M.** (2018). A4nt: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, pp. 1633–1650.

Shuster K., **Humeau S.**, **Hu H.**, **Bordes A. and Weston J.** (2019). Engaging image captioning via personality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12516–12526.

Singh A. and Palod R. (2018). Sentiment transfer using seq2seq adversarial autoencoders. *arXiv preprint, arXiv: 1804.04003*.

Singh H., **Verma G.**, **Garimella A. and Srinivasan B. V.** (2021). DRAG: Director-generator language modelling framework for non-parallel author stylized rewriting. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online. Association for Computational Linguistics, pp. 863–873.

Smith E. M., **Gonzalez-Rico D.**, **Dinan E. and Boureau Y.** (2019). Zero-shot fine-grained style transfer: Leveraging distributed continuous style representations to transfer to unseen styles. *CoRR*, abs/1911.03914.

Song Z., **Zheng X.**, **Liu L.**, **Xu M. and Huang X.** (2019). Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 3685–3695.

Sonnemans J. and Frijda N. H. (1994). The structure of subjective emotional intensity. *Cognition & Emotion* **8**(4), 329–350.

Speer R., **Chin J. and Havasi C.** (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17*. AAAI Press, pp. 4444–4451.

Spencer-Bennett J. (2018). *Moral talk: Stance and Evaluation in Political Discourse*. New York: Routledge.

Stamatatos E. (2017). Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia: Association for Computational Linguistics, pp. 1138–1149.

Stranisci M. A., **Frenda S.**, **Ceccaldi E.**, **Basile V.**, **Damiano R. and Patti V.** (2022). APPReddit: A corpus of reddit posts annotated for appraisal. In *Proceedings of The 13th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association.

Su H.-P., **Huang Z.-J.**, **Chang H.-T. and Lin C.-J.** (2017). Rephrasing profanity in Chinese text. In *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC: Association for Computational Linguistics, pp. 18–24.

Sudhakar A., **Upadhyay B. and Maheswaran A.** (2019). "Transforming" delete, retrieve, generate approach for controlled 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)*. Hong Kong: Association for Computational Linguistics, pp. 3269–3279.

Sue D. W., **Capodilupo C. M.**, **Torino G. C.**, **Bucceri J. M.**, **Holder A.**, **Nadal K. L. and Esquilin M.** (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist* **62**(4), 271–286.

Sulis E., **Irazú Hernández Farías D.**, **Rosso P.**, **Patti V. and Ruffo G.** (2016). Figurative messages and affect in twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems* **108**, 132–143. New Avenues in Knowledge Bases for Natural Language Processing.

Surya S., **Mishra A.**, **Laha A.**, **Jain P. and Sankaranarayanan K.** (2019). Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 2058–2068.

Sydorenko I. (2018). The notion of idiostyle in linguistic studies of literary texts. *Advanced Linguistics*, 15–20. DOI 10.20535/2617-5339.2018.1.148751.

Syed B., **Verma G.**, **Srinivasan B. V.**, **Natarajan A. and Varma V.** (2020). Adapting language models for non-parallel author-stylized rewriting. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 9008–9015.

Tajfel H. (1974). Social identity and intergroup behaviour. *Social Science Information* **13**(2), 65–93.

Tajiri T., **Komachi M. and Matsumoto Y.** (2012). Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island: Association for Computational Linguistics, pp. 198–202.

Terkourafi M. (2008). *Toward a Unified Theory of Politeness, Impoliteness, and Rudeness*. New York: De Gruyter Mouton, pp. 45–76.

Tian Y., **Hu Z. and Yu Z.** (2018). Structured content preservation for unsupervised text style transfer. *arXiv preprint, arXiv: 1810.06526*.

Tikhonov A., **Shibaev V.**, **Nagaev A.**, **Nugmanova A. and Yamshchikov I. P.** (2019). Style transfer for texts: Retrain, report errors, compare with rewrites. 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)*. Hong Kong: Association for Computational Linguistics, pp. 3936–3945.

906 E. Troiano *et al.*

Tikhonov A. and Yamshchikov I. P. (2018). What is wrong with style transfer for texts? *arXiv preprint, arXiv: 1808.04365*.

Toshevska M. and Gievska S. (2021). A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*, 1–1. DOI 10.1109/TAI.2021.3115992.

Tran M., **Zhang Y. and Soleymani M.** (2020). Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona: International Committee on Computational Linguistics, pp. 2107–2114.

Troiano E., **Klinger R. and Padó S.** (2020). Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona: International Committee on Computational Linguistics, pp. 4340–4354.

Troiano E., **Oberländer L.**, **Wegge M. and Klinger R.** (2022). x-enVENT: A corpus of event descriptions with experiencer-specific emotion and appraisal annotations. In *Proceedings of The 13th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association.

Utsumi A. (2000). Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics* **32**(12), 1777–1806.

van Aken B., **Risch J.**, **Krestel R. and Löser A.** (2018). Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels: Association for Computational Linguistics, pp. 33–42.

Vanecek E. and Dressler W. (1975). Bericht uber psycholinguistische experimente zur sprechvariation. *Weiner Linguistische Gazette* **9**, 17–38.

Verhoeven B., **Daelemans W. and Plank B.** (2016). TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association (ELRA), pp. 1632–1637.

Voigt R., **Jurgens D.**, **Prabhakaran V.**, **Jurafsky D. and Tsvetkov Y.** (2018). RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA).

Wang K., **Hua H. and Wan X.** (2019a). Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Advances in Neural Information Processing Systems, 32*.

Wang Y., **Wu Y.**, **Mou L.**, **Li Z. and Chao W.** (2019b). 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)*. Hong Kong: Association for Computational Linguistics, pp. 3573–3578.

Wang Y., **Wu Y.**, **Mou L.**, **Li Z. and Chao W.** (2020). Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona: International Committee on Computational Linguistics, pp. 2236–2249.

Waseem Z., **Davidson T.**, **Warmsley D. and Weber I.** (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC: Association for Computational Linguistics, pp. 78–84.

Waseem Z. and Hovy D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. San Diego, CA: Association for Computational Linguistics, pp. 88–93.

Weller O., **Fulda N. and Seppi K.** (2020). Can humor prediction datasets be used for humor generation? Humorous headline generation via style transfer. In *Proceedings of the Second Workshop on Figurative Language Processing*. Online. Association for Computational Linguistics, pp. 186–191.

Wen Z., **Cao J.**, **Yang R. and Wang S.** (2020). Decode with template: Content preserving sentiment transfer. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, pp. 4671–4679.

West R. and Horvitz E. (2019). Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances". *Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 7265–7272.

Whitehead S. and Cavedon L. (2010). Generating shifting sentiment for a conversational agent. In *Proceedings of the NAACL HLT, 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles, CA: Association for Computational Linguistics, pp. 89–97.

Wiegand M., **Ruppenhofer J. and Eder E.** (2021). Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online. Association for Computational Linguistics, pp. 576–587.

Wierzbicka A. (1988). *The Semantics of Grammar*, vol. 18. Amsterdam: John Benjamins Publishing.

Wieting J. and Gimpel K. (2018). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne: Association for Computational Linguistics, pp. 451–462.

Wu C., **Ren X.**, **Luo F. and Sun X.** (2019a). A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence: Association for Computational Linguistics, pp. 4873–4883.

Wu X., **Xu K. and Hall P.** (2017). A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Science and Technology* **22**(6), 660–674.

Wu X., **Zhang T.**, **Zang L.**, **Han J. and Hu S.** (2019b). Mask and infill: Applying masked language model for sentiment transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, (IJCAI-19)*, pp. 5271–5277.

Wubben S., **van den Bosch A. and Krahmer E.** (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island: Association for Computational Linguistics, pp. 1015–1024.

Xu J., **Sun X.**, **Zeng Q.**, **Zhang X.**, **Ren X.**, **Wang H. and Li W.** (2018). Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne: Association for Computational Linguistics, pp. 979–988.

Xu Q., **Xu C. and Qu L.** (2019a). ALTER: Auxiliary text rewriting tool for natural language generation. 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): System Demonstrations*. Hong Kong: Association for Computational Linguistics, pp. 13–18.

Xu R., **Ge T. and Wei F.** (2019b). Formality style transfer with hybrid textual annotations. *arXiv preprint, arXiv: 1903.06353*.

Xu W., **Napoles C.**, **Pavlick E.**, **Chen Q. and Callison-Burch C.** (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics* **4**(4), 401–415.

Xu W., **Ritter A.**, **Dolan B.**, **Grishman R. and Cherry C.** (2012). Paraphrasing for style. In *Proceedings of COLING 2012*. Mumbai: The COLING 2012 Organizing Committee, pp. 2899–2914.

Yamshchikov I. P., **Shibaev V.**, **Nagaev A.**, **Jost J. and Tikhonov A.** (2019). Decomposing textual information for style transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Hong Kong: Association for Computational Linguistics, pp. 128–137.

Yang K. and Klein D. (2021). FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online. Association for Computational Linguistics, pp. 3511–3535.

Yang P., **Lin J.**, **Xu J.**, **Xie J.**, **Su Q. and Sun X.** (2019). Specificity-driven cascading approach for unsupervised sentiment modification. 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)*. Hong Kong: Association for Computational Linguistics, pp. 5508–5517.

Yang Z., **Hu Z.**, **Dyer C.**, **Xing E. P. and Berg-Kirkpatrick T.** (2018). Unsupervised text style transfer using language models as discriminators. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7298–7309.

Yao Z. and Yu H. (2021). Improving formality style transfer with context-aware rule injection. 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)*. Online. Association for Computational Linguistics, pp. 1561–1570.

Yu Z., **Tan J. and Wan X.** (2018). A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne: Association for Computational Linguistics, pp. 1650–1660.

Zhang T., **Kishore V.**, **Wu F.**, **Weinberger K. Q. and Artzi Y.** (2020a). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhang X. and Lapata M. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen: Association for Computational Linguistics, pp. 584–594.

Zhang Y., **Ding N. and Soricut R.** (2018a). SHAPED: Shared-private encoder-decoder for text style adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, LA: Association for Computational Linguistics, pp. 1528–1538.

Zhang Y., **Ge T. and Sun X.** (2020b). Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, pp. 3221–3228.

Zhang Y., **Xu J.**, **Yang P. and Sun X.** (2018b). Learning sentiment memories for sentiment modification without parallel data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, pp. 1103–1108.

Zhao J., **Zhou Y.**, **Li Z.**, **Wang W. and Chang K.-W.** (2018a). Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, pp. 4847–4853.

Zhao J. J., **Kim Y.**, **Zhang K.**, **Rush A. M. and LeCun Y.** (2018b). Adversarially regularized autoencoders. In Dy J. G. and Krause A., (eds), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018. Proceedings of Machine Learning Research*, vol. 80. PMLR, pp. 5897–5906.

Zhao S., **Meng R.**, **He D.**, **Saptono A. and Parmanto B.** (2018c). Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, pp. 3164–3173.

Zhou X. and Wang W. Y. (2018). MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne: Association for Computational Linguistics, pp. 1128–1137.

Zhu M., **Yu Z. and Wan X.** (2019). A neural approach to irony generation. *arXiv preprint, arXiv: 1909.06200*.

Zull J. E. (2006). Key aspects of how the brain learns. *New Directions for Adult and Continuing Education* **110**, 1–10.