A Review of Human Evaluation for Style Transfer

Eleftheria Briakou[1] **Sweta Agrawal**[1] **Ke Zhang**[2] **Joel Tetreault**[2] **Marine Carpuat**[1]

[1]University of Maryland, [2]Dataminr, Inc.

ebriakou@cs.umd.edu, sweagraw@cs.umd.edu, kzhang@dataminr.com,
jtetreault@dataminr.com, marine@cs.umd.edu

Abstract

This paper reviews and summarizes human evaluation practices described in 97 style transfer papers with respect to three main evaluation aspects: style transfer, meaning preservation, and fluency. In principle, evaluations by human raters should be the most reliable. However, in style transfer papers, we find that protocols for human evaluations are often underspecified and not standardized, which hampers the reproducibility of research in this field and progress toward better human and automatic evaluation methods.

1 Introduction

Style Transfer (ST) in NLP refers to a broad spectrum of text generation tasks that aim to rewrite a sentence to change a specific attribute of language use in context while preserving others (e.g., make an informal request formal, Table 1). With the success of deep sequence-to-sequence models and the relative ease of collecting data covering various stylistic attributes, neural ST is a popular generation task with more than 100 papers published in this area over the last 10 years.

Despite the growing interest that ST receives from the NLP community, progress is hampered by the lack of standardized evaluation practices. One practical aspect that contributes to this problem is the conceptualization and formalization of styles in natural language. According to a survey of neural style transfer by Jin et al. (2021), in the context of NLP, ST is used to refer to tasks where styles follow a linguistically motivated dimension of language variation (e.g., formality), and also to tasks where the distinction between style and content is implicitly defined by data (e.g., positive or negative sentiment). Across these tasks, ST quality is usually evaluated across three dimensions: style transfer (has the desired attributed been changed
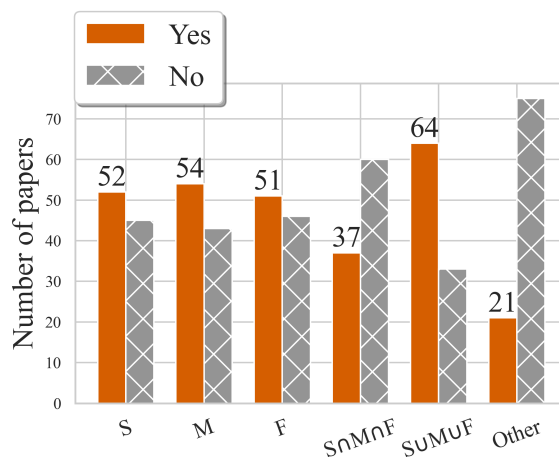

Figure 1: Number of papers employing human evaluations for style transfer (S), meaning preservation (M), fluency (F), all of them (S∪M∪F), at least one of them (S∩M∩F), or another aspect (OTHER).

FORMALITY
Gotta see both sides of the story. *(informal)*
You have to consider both sides of the story. *(formal)*

SENTIMENT
The screen is just the right size. *(positive)*
The screen is too small. *(negative)*

AUTHOR IMITATION
Bring her out to me. *(modern)*
Call her forth to me. *(shakespearean)*

Table 1: Examples of three ST attributes: formality, sentiment and Shakespearean transfer.

as intended?), meaning preservation (are the other attributes preserved?), and fluency (is the output well-formed?) (Pang and Gimpel, 2019; Mir et al., 2019). Given the large spectrum of stylistic attributes studied and the lack of naturally occurring references for the associated ST tasks, prior work emphasizes the limitations of automatic evaluation. As a result, progress in this growing field relies heavily on human evaluations to quantify progress among the three evaluation aspects.

Inspired by recent critiques of human evaluations of Natural Language Generation (NLG) systems (Howcroft et al., 2020; Lee, 2020; Belz et al., 2020, 2021; Shimorina and Belz, 2021), we conduct a structured review of human evaluation for neural style transfer systems as their evaluation is primarily based on human judgments. Concretely, out of the **97 papers we reviewed, 69 of them resort to human evaluation** (Figure 1), where it is treated either as a substitute for automatic metrics or as a more reliable evaluation.

This paper summarizes the findings of the review and raises the following concerns on current human evaluation practices:

1. **Underspecification** We find that many attributes of the human annotation design (e.g., annotation framework, annotators' details) are *underspecified* in paper descriptions, which hampers reproducibility and replicability;

2. **Availability & Reliability** The vast majority of papers do not release the human ratings and do not give details that can help assess their quality (e.g., agreement statistics, quality control), which hurts research on evaluation;

3. **Lack of standardization** The annotation protocols are inconsistent across papers which hampers comparisons across systems (e.g., due to possible bias in annotation frameworks).

The paper is organized as follows. In Section 2, we describe our procedure for analyzing the 97 papers and summarizing their evaluations. In Section 3, we present and analyze our findings. Finally, in Section 4, we conclude with a discussion of where the field of style transfer fares with respect to human evaluation today and outline improvements for future work in this area.

2 Reviewing ST Human Evaluation

Paper Selection We select papers for this study from the list compiled by Jin et al. (2021) who conduct a comprehensive review of ST that covers the task formulation; evaluation metrics; opinion papers and deep-learning based textual ST methods. The paper list contains more than 100 papers and is publicly available (`https://github.com/fuzhenxin/Style-Transfer-in-Text`). We reviewed all papers in this list to determine whether

they conduct either human or automatic evaluation on system outputs for ST, and therefore should be included in our structured review. We did not review papers for text simplification, as it has been studied separately (Alva-Manchego et al., 2020; Sikka et al., 2020) and metrics for automatic evaluation have been widely adopted (Xu et al., 2016). Our final list consists of 97 papers: 86 of them are from top-tier NLP and AI venues: ACL, EACL, EMNLP, NAACL, TACL, IEEE, AAAI, NeurIPS, ICML, and ICLR, and the remaining 11 are pre-prints which have not been peer-reviewed.

Review Structure We review each paper based on a predefined set of criteria (Table 2). The rationale behind their choice is to collect information on the evaluation aspects that are underspecified in NLP in general as well as those specific to the ST task. For this work, we call the former *global criteria*. The latter is called *dimension-specific criteria* and is meant to illustrate issues with how each dimension (i.e., style transfer, meaning preservation, and fluency) is evaluated.

Global criteria can be split into three categories which describe: (1) the ST stylistic attribute, (2) four details about the annotators and their compensation, and (3) four general design choices of the human evaluation that are not tied to a specific evaluation dimension.

For the *dimension-specific criteria* we repurpose the following *operationalisation* attributes introduced by Howcroft et al. (2020): form of response elicitation (direct vs. relative), details on type of collected responses, size/scale of rating instrument, and statistics computed on response values. Finally, we also collect information on the quality criterion for each dimension (i.e., the wording used in the paper to refer to the specific evaluation dimension).

Process The review was conducted by the authors of this survey. We first went through each of the 97 papers and highlighted the sections which included mentions of human evaluation. Next, we developed our criteria by creating a draft based on prior work and issues we had observed in the first step. We then discussed and refined the criteria after testing it on a subset of the papers. Once the criteria were finalized, we split the papers evenly between all the authors. Annotations were spot-checked to resolve uncertainties or concerns that were found in reviewing dimension-specific criteria (e.g., scale of rating instrument is not explicitly

GLOBAL CRITERIA	
task(s)	ST *task(s) covered*
presence of human annotation	*presence of human evaluation*
annotators' details	*details on annotator's background/recruitment process*
annotators' compensation	*annotator's payment for annotating each instance*
quality control	*quality control methods followed to ensure reliability of collected judgments*
annotations' availability	*availability of collected judgments*
evaluated systems	*number of different systems present in human evaluation*
size of evaluated instance set	*number of instances evaluated for each system*
size of annotation set per instance	*number of collected annotations for each annotated instance*
agreement statistics	*presence of inter-annotator agreement statistics*
sampling method	*method for selecting instances for evaluation from the original test sets*

DIMENSION-SPECIFIC CRITERIA	
presence of human evaluation	*whether there exists human evaluation for a specific aspect*
quality criterion name	*quality criterion of evaluated attribute as mentioned in the paper*
direct response elicitation	*presence of direct assessment (i.e., each instance is evaluated on its own right)*
relative judgment type (if applicable)	*type of relative judgment (e.g., pairwise, ranking, best)*
direct rating scale (if applicable)	*list of possible response values*
presence of lineage reference	*whether the evaluation reuses an evaluation framework from prior work*
lineage source (if applicable)	*citation of prior evaluation framework*

Table 2: Descriptions of attributes studied in our structured review.

defined but inferred from the results discussion) and global criteria (e.g., number of systems not specified but inferred from tables). We release the spreadsheet used to conduct the review along with the reviewed PDFs that come with highlights on the human evaluation sections of each paper at `https://github.com/Elbria/ST-human-review`.

3 Findings

Based on our review, we first discuss trends of stylistic attributes as discussed in ST research through the years (§3.1), followed by global criteria of human evaluation (§3.2), and then turn to domain-specific criteria (§3.3).

3.1 Evolution of Stylistic Attributes

Table 3 presents statistics on the different style attributes considered in ST papers since 2011. First, we observe a significant increase in the number of ST papers starting in 2018 (in 2017 there were 8 ST papers; the following year there were 28). We believe this can be attributed to the creation of standardized training and evaluation datasets for various ST tasks. One example is the Yelp dataset, which consists of positive and negative reviews, and is used for unsupervised sentiment transfer (Shen et al., 2017). Another example is

the GYAFC parallel corpus, consisting of informal-formal pairs that are generated using crowdsourced human rewrites (Rao and Tetreault, 2018). Second, we notice that new stylistic attributes are studied through time (21 over the last ten years), with sentiment and formality transfer being the most frequently studied.

3.2 Global Criteria

Annotators Table 4 summarizes statistics about how papers describe the background of their human judges. The majority of works (38%) rely on crowd workers mostly recruited using the Amazon Mechanical Turk crowdsourcing platform. Interestingly, for a substantial number of evaluations (45%), it is unclear who the annotators are and what their background is. In addition, we find that information about how much participants were compensated is missing from all but two papers. Finally, many papers collect 3 independent annotations, although this information is not specified in a significant percentage of evaluations (42%). In short, the ability to replicate a human evaluation from the bulk of current research is extremely challenging, and in many cases impossible, as so much is underspecified.

STYLE	2011	2012	2016	2017	2018	2019	2020	2021	TOTAL
anonymization					1				1
attractiveness				1			1		2
author imitation		1		2	2	1	5		11
debiasing							2		2
social register					1				1
expertise							1		1
formality	1				1	9	10	3	24
gender			1		2	3			6
political slant					2	1	1		4
sentiment				4	14	14	18	3	53
romantic/humorous					2	1	1		4
simile							1		1
excitement							1		1
profanity							1		1
prose				1	1				2
offensive language					1		1		2
multiple						1	1		2
persona			1			1	1		3
poeticness							1		1
politeness					1		1		2
emotion							1		1
TOTAL	**1**	**1**	**2**	**8**	**28**	**31**	**48**	**6**	**125**

Table 3: Number of ST papers per stylistic attribute across years. Some papers evalute multiple style attributes.

CROWD-SOURCING	PAPER'S DESCRIPTION OF ANNOTATORS	COUNT
YES	*"qualification test"*	6
	"number of approved HITs'	2
	"hire Amazon Mechanical Turk workers"	18
NO	*"bachelor or higher degree; independent of the authors'* *"research group", "annotators with linguistic background"* *"well-educated volunteers", "graduate students in* *computational linguistics" "major in linguistics"* *"linguistic background", "authors"*	12
UNCLEAR	*"individuals", "human judges", "human annotators"* *"unbiased human judges", "independent annotators"*	31

Table 4: Annotators' background for human evaluation as described in ST papers.

Annotations' Reliability Only 31% of evaluation methods that rely on crowd-sourcing employ quality control (QC) methods. The most common QC strategies are to require workers to pass a qualification test (Jin et al., 2019; Li et al., 2016; Ma et al., 2020; Pryzant et al., 2020) to hire the top-ranked workers based on pre-computed scores that reflect the number of their past approved tasks (Krishna et al., 2020; Li et al., 2019), to use location restrictions (Krishna et al., 2020), or to perform manual checks on the collected annotations (Rao and Tetreault, 2018; Briakou et al., 2021). Furthermore, only 20% of the papers report inter-annotator agreement statistics, and only 4 papers release the actual annotations to facilitate the reproducibility and further analysis of their results. Without this information, it is difficult to replicate the evaluation and compare different evaluation approaches.

Data Selection Human evaluation is typically performed on a sample of the test set used for automatic evaluation. Most works (62%) sample instances randomly from the entire set, with a few exceptions that employ stratified sampling according to the number of stylistic categories considered (e.g., random sampling from positive and negative classes for a binary definition of style). For 25% of ST papers information on the sampling method is not available. Furthermore, the sample size of instances evaluated per system varies from 50 to 1000, with most of them concentrated around 100.

3.3 Dimension-specific Criteria

Quality Criterion Names Table 5 summarizes the terms used to refer to the three main dimensions of style transfer, meaning preservation, and fluency. As Howcroft et al. (2020) found in the context of NLG evaluation, we see that the names of these dimensions are not standardized for the three ST evaluation dimensions. Each dimension has at least six different ways that past literature has referred to them. We should note that even with the same name, the nature of the evaluation is not necessarily the same across ST tasks: for instance, what constitutes content preservation differs in formality transfer and in sentiment transfer, since the latter arguably changes the semantics of the original text. While fluency is the aspect of evaluation that might be most generalizable across ST tasks, it is referred to in inconsistent ways across papers which could lead to different interpretations by annotators. For instance, the same text could be rated as natural but not grammatical. Overall, the variability in terminology makes it harder to understand exactly what is being evaluated and to compare evaluation methods across papers.

Rating Type Table 6 presents statistics on the rating type (direct vs. relative) per dimension over time. *Direct rating* refers to evaluations where each system output is assessed in isolation for that dimension. *Relative rating* refers to evaluations where two or more system outputs are compared against each other. Rating types were more inconsistently used before 2020, with recent convergences toward direct assessment. Among papers that report rating type, direct assessment is the most frequent approach for all evaluation aspects over the years 2018 to 2021.

Possible Responses Tables 7, 8, and 9 summarize the range of responses elicited for direct and

STYLE

attribute compatibility, formality, politeness level, sentiment, style transfer intensity, attractive captions, attribute change correctness, bias, creativity, highest agency, opposite sentiment, sentiment, sentiment strength, similarity to the target attribute, style correctness, style transfer accuracy, style transfer strength, stylistic similarity, target attribute match, transformed sentiment degree.

MEANING

content preservation, meaning preservation, semantic intent, semantic similarity, closer in meaning to the original sentence, content preservation degree, content retainment, content similarity, relevance, semantic adequacy.

FLUENCY

fluency, grammaticality, naturalness, gibberish language, language quality.

Table 5: Quality criterion names used in ST human evaluation descriptions for the three evaluation dimensions.

	2011	2012	2016	2017	2018	2019	2020	2021	Total
STYLE									
DIRECT	1	1		1	8	10	12	4	40
RELATIVE			1		4	7			12
NONE			2	6	11	11	15		45
MEANING									
DIRECT		1			12	10	18	4	45
RELATIVE				1		4	4		9
NONE	1		2	7	8	11	14		43
FLUENCY									
DIRECT		1		1	10	10	19	4	45
RELATIVE					4	2			6
NONE	1		1	2	8	6	7		46

Table 6: Number of papers using each rating type for the three evaluation dimensions across years.

relative ratings. They cover diverse definitions of scales within each rating type. Across evaluation aspects, the dominant evaluation framework is **direct ratings** on a **5-point scale**. However, while that configuration is what the field tends to focus on, there is clearly a wide array of choices that the field also considers which, once again, makes comparing human evaluations head to head very difficult.

	Rating Scale	(1)	[-2,-1,0,1,2]
		(3)	[-3,-2, -1, 0, 1, 2, 3]
		(1)	[polite, slightly polite, neutral, slightly rude, rude]
		(4)	[positive, negative, neutral]
		(1)	[positive, negative, relaxed, annoyed]
		(1)	[more formal, more informal, neither]
DIRECT (40)		(2)	[0,1,2]
		(2)	[1,2,3]
		(1)	[0,1,2,3,4,5]
		(19)	[1, 2, 3, 4, 5]
		(2)	[1,2,3,4,5,6,7,8,9,10]
		(1)	binary
	Not available	(2)	
RELATIVE (12)	Best selection	(5)	
	Pairwise	(7)	

Table 7: **Style** results. Numbers in parentheses denote paper counts per category. The most popular rating type across each dimension is highlighted.

	Rating Scale	(1)	[-2,-1,0,1,2]
		(6)	[0,1,2]
		(1)	[1,2,3]
		(1)	[1,2,3,4]
DIRECT (45)		(25)	[1, 2, 3, 4, 5]
		(1)	[0,1,2,3,4,5]
		(4)	[1,2,3,4,5,6]
		(3)	[1,2,3,4,5,6,7,8,9,10]
	Not available	(3)	
	Best selection	(3)	
RELATIVE (9)	Pairwise	(3)	
	Ranking	(3)	

Table 8: **Meaning Preservation** results. Numbers in parentheses denote paper counts per category. The most popular rating type across each dimension is highlighted.

	Rating Scale	(1)	["easy to understand", "some grammar errors", "impossible to understand"]
		(1)	["incorrect", "partly correct", "correct"]
		(1)	[0,1]
		(3)	[0,1,2]
		(2)	[1,2,3]
		(4)	[1,2,3,4]
DIRECT (45)		(1)	[0,1,2,3,4]
		(26)	[1, 2, 3, 4, 5]
		(1)	[0,1,2,3,4,5]
		(1)	[1,2,3,4,5,6]
		(2)	[1,2,3,4,5,6,7,8,9,10]
	Not available	(2)	
	Best selection	(1)	
RELATIVE (6)	Pairwise	(4)	
	Ranking	(1)	

Table 9: **Fluency** results. Numbers in parentheses denote paper counts per category. The most popular rating type across each dimension is highlighted.
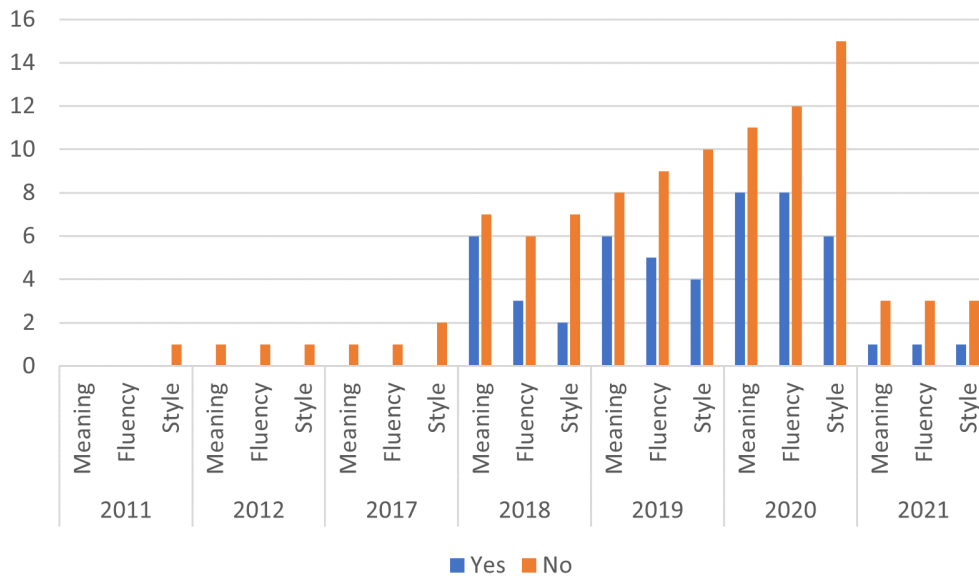
Figure 2: Lineage statistics (i.e., number of papers) for each ST evaluation aspect over years.

Lineage Figure 2 shows how often the human evaluation setup used in each reviewed paper is based on cited prior work, for each dimension over time. Only 19% of papers repurpose or reuse some prior work for the evaluation of style. Most of these papers target ST for formality or sentiment. Even when evaluating fluency or meaning preservation, more than 50% of the papers do not refer to any prior work. This is striking because it suggests that there is currently not a strong effort to replicate prior human evaluations.

For papers that mention lineage, the most common-set up for evaluating meaning preservation (24%) and fluency (28%) is Li et al. (2018). 43% of ST papers that work on sentiment also refer to Li et al. (2018). Some papers follow Agirre et al. (2016) for measuring textual similarity, Heilman et al. (2014) for grammaticality and Pavlick and Tetreault (2016) for formality.

4 Discussion & Recommendations

4.1 Describing Evaluation Protocols

Our structured review shows that human evaluation protocols for ST are mostly underspecified and lack standardization, which fundamentally hinders progress, as it is for other NLG tasks (Howcroft et al., 2020). The following attributes are commonly underspecified:

1. details on the procedures followed for recruiting annotators (i.e., linguistic background of

expert annotators or quality control method employed when recruiting crowd-workers)

2. annotator's compensation to better understand their motivation for participating in the task,

3. inter-annotator agreement statistics,

4. number of annotations per instance (3-5 is the most popular choice of prior work),

5. number of systems evaluated,

6. number of instances annotated (minimum of 100 based on prior works),

7. selection method of the annotated instances (suggestion is same random sampled for all annotated systems).

8. detailed description of evaluated frameworks per evaluation aspect (e.g., rating type, response of elicitation).

Furthermore, we observe that annotated judgments are hardly ever made publicly available and that, when specified, evaluation frameworks are not standardized.

As a result, our first recommendation is simply to include all these details when describing a protocol for human evaluation of ST. We discuss further recommendations next.

4.2 Releasing Annotations

Making human-annotated judgments available would enable the development of better automatic metrics for ST. If all annotations had been released with the papers reviewed, we estimate that more than 10K human judgments per evaluation aspect would be available. Today this would suffice to train and evaluate dedicated evaluation models.

In addition, raw annotations can shed light on the difficulty of the task and nature of the data: they can be aggregated in multiple ways (Oortwijn et al., 2021), or used to account for annotator bias in model training (Beigman and Beigman Klebanov, 2009). Finally, releasing annotated judgments makes it possible to replicate and further analyze the evaluation outcome (Belz et al., 2021).

4.3 Standardizing Evaluation Protocols

Standardizing evaluation protocols is key to establishing fair comparisons across systems (Belz et al., 2020) and to improving evaluation itself.

Our survey sheds light on the most frequently used ST frameworks in prior work. Yet more research is needed to clarify how to evaluate, compare and replicate the protocols. For instance, Mir et al. (2019) point to evidence that relative judgments can be more reliable than absolute judgments (Stewart et al., 2005), as part of their work on designing automatic metrics for ST evaluation. However, research on human evaluation of machine translation shows that this can change depending on the specifics of the annotation task: relative judgments were replaced by direct assessment when Graham et al. (2013) showed that both intra and inter-annotator agreement could be improved by using a continuous rating scale instead of the previously common five or seven-point interval scale (Callison-Burch et al., 2007).

For ST, the lack of detail and clarity in describing evaluation protocols makes it difficult to improve them, as has been pointed out for other NLG tasks by Shimorina and Belz (2021) who propose evaluation datasheets for clear documentation of human evaluations, Lee (2020) and van der Lee et al. (2020) who propose best practices guidelines, and Belz et al. (2020, 2021) who raise concerns regarding reproducibility. This issue is particularly salient for ST tasks where stylistic changes are defined implicitly by data (Jin et al., 2021) and where the instructions given to human judges for style transfer might be the only explicit characterization

of the style dimension targeted. Furthermore, since ST includes rewriting text according to pragmatic aspects of language use, who the human judgments are matters since differences in communication norms and expectations might result in different judgments for the same text.

Standardizing and describing protocols is also key to assessing the alignment of the evaluation with the models and task proposed (Hämäläinen and Alnajjar, 2021), and to understand potential biases and ethical issues that might arise from, e.g., compensation mechanisms (Vaughan, 2018; Schoch et al., 2020; Shmueli et al., 2021).

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Eyal Beigman and Beata Beigman Klebanov. 2009. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 280–287, Suntec, Singapore. Association for Computational Linguistics.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Xformal: A benchmark for multilingual formality style transfer.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Mika Hämäläinen and Khalid Alnajjar. 2021. The great misalignment problem in human evaluation of NLP methods. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 69–74, Online. Association for Computational Linguistics.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180, Baltimore, Maryland. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2021. Deep learning for text style transfer: A survey.

Z. Jin, D. Jin, J. Mueller, N. Matthews, and E. Santus. 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)*, pages 3095–3107. Association for Computational Linguistics.

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

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2020. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, page 101151.

Kiyong Lee. 2020. Annotation-based semantics. In *16th Joint ACL - ISO Workshop on Interoperable Semantic Annotation PROCEEDINGS*, pages 36–48, Marseille. European Language Resources Association.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3304–3313, Hong Kong, China. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Yvette Oortwijn, Thijs Ossenkoppele, and Arianna Betti. 2021. Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 131–141, Online. Association for Computational Linguistics.

Richard Yuanzhe Pang and Kevin Gimpel. 2019. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. In *NGT@EMNLP-IJCNLP*.

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

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

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

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

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6833–6844, Red Hook, NY, USA. Curran Associates Inc.

Anastasia Shimorina and Anya Belz. 2021. The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in nlp. *arXiv preprint arXiv:2103.09710*.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of nlp crowdsourcing. *arXiv preprint arXiv:2104.10097*.

Punardeep Sikka, Manmeet Singh, Allen Pink, and Vijay Mago. 2020. A survey on text simplification. *arXiv preprint arXiv:2008.08612*.

Neil Stewart, Gordon D. A. Brown, and Nick Chater. 2005. Absolute identification by relative judgment. *Psychological Review*, 112(4):881–911.

Jennifer Wortman Vaughan. 2018. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*, 18(193):1–46.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.