say the same but differently

Anna Maria Wegmann

Say the Same but Differently

Computational Approaches to Stylistic Variation and Paraphrasing

Anna Maria Wegmann

Wie het zelfde anders zegt, zegt iets anders.

Paul Scholten

(English: He/she who says the same thing differently, says something else.
German: Wer dasselbe anders sagt, sagt etwas anderes.)

SIKS Dissertation Series No. 2025-36
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN: 978-90-393-7935-6

Printed by: Ipskamp Printing

Cover by: Mirte Ebel

Say the Same but Differently

Computational Approaches to Stylistic Variation and Paraphrasing

Zeg hetzelfde maar dan anders

Computationele methoden voor stijlvariatie en parafraseren

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht

op gezag van de rector magnificus, prof.dr.ir. W. Hazeleger,

ingevolge het besluit van het college voor promoties

in het openbaar te verdedigen op

maandag 13 oktober 2025 des middags te 2.15 uur

door

Anna Maria Wegmann

geboren op 01 juli 1994
te Kirchheimbolanden, Duitsland

Promotoren:

Prof. dr. C.J. van Deemter

Copromotoren:

Dr. D.P. Nguyen

Beoordelingscommissie:

Prof. dr. A.P.J. van den Bosch

Dr. S. Degaetano-Ortlieb

Prof. dr. J. Grieve

Prof. dr. M. Poesio

Prof. dr. H. Zinsmeister

This work was supported by the *Digital Society - The Informed Citizen* research program, which was financed by the Dutch Research Council (NWO) under grant number 410.19.007 and supported by the stakeholder EMMA.

Contents

Part I

Opening Remarks

Contents

1

Introduction

Today, if someone says "I'm going to Las Vegas!", we might immediately assume that they are going to Las Vegas to gamble. Three hundred years ago, when a man in the Netherlands said, "I'm going to Utrecht!", a similar assumption would have been that he is going to Utrecht to have sex with other men. In these utterances, there is no explicit mention of homosexuality or gambling—the "surface-level" meanings of the two sentences are just about visiting the named city. Still, for the second utterance, who the speaker is (i.e., a man), the context (i.e., three hundred years ago in the Netherlands) and knowledge of the world (i.e., the Utrecht sodomy trials in the 18th century) tell a supplementary story. In short, understanding language is more complex than it might seem.

At the heart of this dissertation lies *language variation*, that is, **different ways of saying the same thing**. Human language is full of it. For instance, both words "Utrechter" and "Utrechtenaar" refer to an inhabitant of the city of Utrecht. Out of the two, "Utrechtenaar" is the historically more accurate term (van Lieshout et al., 2024)—certainly three hundred years ago—while "Utrechter" is the more common term today. Historically, "Utrechtenaar" also has the connotation of referring to gay men (van Lieshout et al., 2024). Today, when someone uses "Utrechtenaar" over "Utrechter" to refer to themselves, we might know more about them—for example, that they are more likely part of the local queer community. We can say that the use of "Utrechtenaar" carries additional social meaning.

Language models today are often not able to represent, let alone understand such nuances. Consider two Dutch sentences "Ik ben een Utrechter." and "Ik ben een Utrechtenaar.". The translation tool DeepL translates both of these sentences to "I am a Utrecht resident.". This translation is perfectly reasonable as both "Utrechter" and "Utrechtenaar" refer to an inhabitant of the city of Utrecht. We can say that, in this instance, it is good that DeepL is "robust to language variation". However, there are situations where it might be helpful for translation systems to be "sensitive to language variation". Imagine a news report in Dutch where two people refer to themselves as an

"Utrechter" and as an "Utrechtenaar", respectively. Translating both as "Utrecht resident" might obscure subtle distinctions in background and social identity—possibly leading to confusion or loss of narrative nuance.

I

This is problematic because natural language processing (NLP) systems are increasingly used throughout society, with the latest surge being caused by the release of CHATGPT in November 2022. One might use CHATGPT to write a dissertation, DEEPL to translate a dissertation summary from English to Dutch and GOOGLE SEARCH to find relevant academic publications. NLP systems can automatically caption YouTube videos, translate Dutch rental agreements into English, summarize complex legal texts in simpler language, search thousands of medical texts for conditions with similar symptoms and support numerous other real-world applications.

In this dissertation, I focus on **accounting for language variation in NLP models**. There are two core problems that I address: (1) creating NLP models that are *robust* to language variation and (2) creating NLP models that are *sensitive* to language variation. Before I specify my research question further, let me give a short overview of NLP in recent years and then unpack the problem of robustness and sensitivity.

A Short History of NLP

If you enter "writing a successful dissertation" to find relevant websites using GOOGLE SEARCH, Google starts by converting your search terms into a numerical form that a machine can process. One of the main challenges in NLP over the last decades has been how to represent the meaning of a word or phrase in machine-readable form—using real numbers, arranged in vectors and matrices. In this century, one of the most successful ideas applied to NLP is the distributional hypothesis. It has its origins in philosophy and linguistics and states that words are similar in meaning if they appear in similar contexts (Harris, 1954; Firth, 1957; Wittgenstein, 1953; Brunila and LaViolette, 2022). The distributional hypothesis informs training tasks for *encoder* (e.g., Google's BERT in Devlin et al., 2019 and Meta's ROBERTA in Liu et al., 2019) and *decoder* models (e.g., OpenAI's GPT-2 in Radford et al., 2019 and CHATGPT in OpenAI, 2022 and DEEPSEEK in Guo et al., 2025). Encoder models learn to represent text as real numbered vectors—they encode text in machine-readable form. These real numbered vectors are often called *embeddings*. Decoder models learn probability distributions for texts. They generate text based on the learned probability distributions—they decode text from real-numbered vectors and matrices.

Text embeddings and encoder models were at the forefront of a lot of the advancements in NLP in the last decade (Mikolov et al., 2013; Devlin et al., 2019). The real-numbered vectors led to leaps in solving previously difficult classification tasks like natural language inference (i.e., does a hypothesis text follow from a premise text?) or paraphrase classification (i.e., do two texts have the same meaning?), which are instrumental to practical applications like information retrieval (e.g., GOOGLE SEARCH[1]). With scaling up training to huge amounts of text data and billion parameter model sizes as well as post-training advances (i.e., altering models after initial training), de-

[1] https://blog.google/products/search/search-language-understanding-bert/

coder models have largely taken over functions previously taken on by encoder models, at the latest with the release of ChatGPT in November 2022. Today, decoder models are often able to solve a variety of problems "zero-shot" using in-context learning (i.e., *prompting*): We can now directly ask a language model for an answer to a problem without having to train it for that specific problem first. For example, we can ask ChatGPT if "Lucy had studied so hard." follows from "Anna did a lot worse than her good friend Lucy on the test because she had studied so hard." and it will usually generate the correct answer.[2] And, perhaps more impressively, we can now ask large language models trained only on text data to do tasks thought almost impossible for them in 2021 like coding or playing chess.[3]

Overall, the challenge of representing text meaning in machine readable form is foundational in NLP, as language models rely on representing the meaning of texts as vectors and matrices to process and further manipulate natural language.

Model Robustness to Language Variation

Up until now, I employed a rather vague notion of "surface-level" meaning. Let me get a bit more specific. NLP has typically been interested in representing what I call the *referential meaning*[4] or semantic content[5] of words and texts. Coming back to the initial example, the referential meaning of the word "Utrechter" is what the word "Utrechter" refers to, that is an actual inhabitant of the city of Utrecht in the Netherlands. Different forms can have the same referential meaning: "Utrechtenaar" refers to the same concept, an inhabitant of Utrecht. Being able to map words with the same referential meaning to a similar computational representation can be and has been very useful for typical *semantic* NLP tasks like paraphrase classification or natural language inference. One should not have to use the exact term "Utrechter" when searching for documents on the inhabitants of the city of Utrecht with Google Search. Efforts in NLP to treat different linguistic forms with similar referential meanings similarly are what I call making NLP models ***robust to language variation***.

Language models are not necessarily robust to language variation. NLP tools might perform differently—usually less well—for people who use less common language varieties (Grieve et al., 2025). Someone writing in African American English can expect lower performance of NLP systems than someone writing in "Standard" American English (Blodgett et al., 2016; Ziems et al., 2023) and asking ChatGPT the same question in Bulgarian costs more U.S. dollars than using Standard American English (Petrov et al., 2023). Furthermore, different surface forms expressing the same refer-

[2] Example taken from WNLI (Wang et al., 2018).

[3] The impressive thing about language models playing chess is not anymore that a machine is good at playing chess. This is not surprising since 1997 when IBM's Deep Blue system beat grandmaster Garry Kasparov. The surprising and impressive part is that now, language models learn to play chess as a "side product". ChatGPT is not a specific expert system designed for game playing or chess. It has not even been taught the rules of chess. It only learned to manipulate text.

[4] I am using "referential meaning" in a sense common to literature in sociolinguistics, e.g., Campbell-Kibler (2011). It is somewhat connected to de Saussure (1916)'s separation of signifier (word) and signified (described idea or thing). However, note that the term "referential meaning" is not generally used in linguistics and might invoke different understandings in semiotics. See more in Section 2.2.

[5] Throughout the dissertation I use referential meaning and (semantic) content interchangeably.

ential meaning can influence the performance of NLP models. For example, spelling variations have been shown to impact classification accuracy (Han and Baldwin, 2011; Eisenstein, 2013; Barteld et al., 2016, 2018), while paraphrasing a prompt can similarly lead to both increases and decreases in model performance (Mizrahi et al., 2024; Wahle et al., 2024).

Initial efforts in NLP to improve robustness were often centered around text normalization, that is transforming text into a more standardized form (Han and Baldwin, 2011; Liu et al., 2011; Barteld et al., 2018; van der Goot et al., 2021) and selecting "good quality", standardized language for training only (Gururangan et al., 2022). In our example that could mean to alter "Utrechtenaar" to something closer to "Utrechter" in a text normalization step before handing it to the Google Search pipeline or CHATGPT. However, the ideology and existence of a standard form of language has been questioned in other fields (Bell, 2014; Craft et al., 2020) and increasingly so in NLP (Eisenstein, 2013; Gururangan et al., 2022). Another common approach is to include different forms in training while encouraging models to treat forms with the same referential meaning similarly (Eisenstein, 2013; Barteld et al., 2018; Piktus et al., 2019; Liu et al., 2020; Nguyen et al., 2021). For example, this could mean including examples of "Utrechtenaar" next to examples of "Utrechter" in the training dataset and manipulating model representations so that the meaning vector of "Utrechtenaar" is closer to that of "Utrechter".

Model Sensitivity to Language Variation

In contrast to only considering the referential meaning for "Utrechtenaar" and "Utrechter", one could also argue that the choice of an author to say "Utrechtenaar" over "Utrechter" carries social meaning beyond the referential. Utrecht became widely associated with homosexual men in 1730 when a wave of sodomy trials swept through the Netherlands starting with the execution of 18 men in Utrecht (van Lieshout et al., 2024). Soon the word "Utrechtenaar" was used in a derogatory way in everyday language throughout the country and became synonymous with being homosexual. In 1947, an Utrecht newspaper introduced the term "Utrechter" as an alternative to "Utrechtenaar" to escape the second meaning of the term. Up until today, with the negative association to "Utrechtenaar" fading, the term "Utrechter" is more commonly used for inhabitants of Utrecht. However, the dialectal variant "Utrechtenaor" was consistently common in some local speech communities and, recently, there have been efforts to reclaim "Utrechtenaar" by the local queer community (van Lieshout et al., 2024). All that said, we can probably say we have more information about a speaker that uses "Utrechtenaar" compared to "Utrechter". They are more likely to be an Utrecht local and belong to the queer community. In other words, when we represent "Utrechtenaar" and "Utrechter" with almost the same vector, we might lose social information and connotation about a considered text.

"Utrechtenaar" is a lexical variant of "Utrechter". This is only one aspect of the more general concept of ***language variation***. Language variation is present and pervasive in all aspects of language. Language variation has been studied by comparing different ways of saying the same thing (e.g., in Labov, 1972). Language variation can, for

example, be lexical (e.g., "Utrechter" vs. "Utrechtenaar"), syntactic (e.g., "I gave the dissertation to Dong" vs. "I gave Dong the dissertation") or orthographic (e.g., "writing" vs. "writin"). Like with "Utrechtenaar", such variation is often not random, but systematically linked to regional, social, and contextual factors (Nguyen et al., 2016; Eckert, 2008). Language variation carries *social meaning*, i.e., information about the social background and identity of the language user (Nguyen et al., 2021).

Being able to represent different language varieties, can be useful for form- or style-based NLP tasks like authorship verification (i.e., have two texts been written by the same author?), or style transfer (i.e., change the style of a given text without changing the referential meaning). A machine translation system should not just translate the referential meaning of a formal cover letter, but also keep the original format and use a style that resembles the formal style in the original text. I call efforts in NLP to represent different varieties—whether in dataset curation, training, representation or evaluation approaches—making NLP more ***sensitive to language variation***.

Language models are not necessarily sensitive to language variation. In NLP, researchers have typically trained language models without considering the full range of variation in English or other languages (Bender and Friedman, 2018; Gururangan et al., 2022; Grieve et al., 2025). As a result, models typically overrepresent more common varieties and underrepresent less common ones. Even if language models are sensitive to language variation, they might still display undesirable behaviors (Grieve et al., 2025): They might be covertly racist by assigning less-prestigious jobs to speakers of African American English (Hofmann et al., 2024) or generate condescending answers and stereotype users that use non-standard language varieties like Nigerian English (Fleisig et al., 2024). Such stereotyping might occur because speakers of these varieties are represented in training datasets through speech of others that reflects stereotypes (Grieve et al., 2025). Intentionally and fairly representing different language varieties and language forms has typically been neglected in NLP (Hovy, 2018; Nguyen et al., 2021; Hovy and Yang, 2021). However, explicitly adding social information in the form of sociodemographic variables (e.g., age, gender, social class) can help the performance of NLP models (Volkova et al., 2013; Wu et al., 2021; Hovy and Yang, 2021). Further, including information about a used language variety and social information about the user can enable accessibility efforts (e.g., simplifying a text for a child or summarizing a text for a non-expert) or further improve the quality of the NLP system by adapting the generated text to the user (Dudy et al., 2021; Stoop and van den Bosch, 2014).[6] Additionally, ensuring that language models accurately represent a wider range of human language can be a goal in itself (Blodgett et al., 2020).

In summary, both how to make NLP models more sensitive to language variation and how to make them more robust to it are important considerations in NLP that need to be explored more. Making NLP models more robust to language variation is helpful for semantic tasks like natural language inference and enables invariance across different surface form and social groups (Lucy et al., 2024). Making NLP models more sensitive to language variation can help form- and style-based tasks like authorship

[6] Such personalization efforts can lead to potential harms specifically related to privacy. See Section 7.3 for a discussion.

verification and enables adaptation to social groups (Lucy et al., 2024). There might not exist a one-size-fits-all solution. For example, making models more robust to language variation with normalization efforts might come at the expense of sensitivity to that variation. Overall, I argue that we should consider language variation more in NLP systems.

1.1. Research Questions

The overarching motivation behind this dissertation is to *develop NLP methods that account for language variation.* With accounting for language variation, I specifically refer to (1) making models sensitive to form- and style-based differences (2) making models robust to form- and style-based differences. I present four research questions that I address in this dissertation. All of them address a small fraction of the overarching motivation in that each research question focuses on a specific research context (e.g., tokenizers as a building block of LLMs). I group the research questions into three parts that focus on studying (1) tasks requiring sensitivity to language variation and tasks requiring robustness to language variation, (2) the former, and (3) the latter.

Study A: Tokenization for Variation-Robust and Variation-Sensitive Tasks

The first part of this dissertation focuses on tasks requiring robustness to language variation (e.g., for semantic tasks like natural language inference, labels do not depend on whether a text uses British or American spelling) and tasks requiring sensitivity to language variation (e.g., for form-based tasks like authorship verification, labels depend on whether a text uses British or American spelling). Tasks that require robustness to language variation might need a different approach from LLMs than tasks requiring sensitivity to language variation—down to the fundamental building blocks of LLMs. I investigate one such fundamental building block of LLMs: tokenizers.

> **Research Question 1 (RQ1)**: How do different key algorithmic decisions for tokenizers influence the performance on downstream tasks: Tasks requiring robustness to language variation and tasks requiring sensitivity to language variation?

Tokenizers break up input strings and determine the actual tokens that are fed into language models. As discussed previously, language variation is systematically linked to regional, social and contextual factors (Coupland, 2007; Eckert, 2012; Nguyen et al., 2016) and some language varieties are more common than others. Tokenizers behave differently for linguistic forms that are less common (Matthews et al., 2024; Ovalle et al., 2024), which in turn influences the features a language model can work with. In Chapter 3, I introduce the distinction between variation-robust and variation sensitive-tasks, highlighting how tokenization might affect them differently. Then, I investigate whether tokenizers in fact differently impact downstream performance on

tasks that are robust to language variation and tasks that are sensitive to language variation.

Study B: Representing Linguistic Style

In the second part of this dissertation, I focus on developing text representations that are sensitive to one aspect of language variation: linguistic style. Specifically, I investigate how to evaluate text representations on their sensitivity to linguistic style (RQ2a) and how to develop neural text representations that are sensitive to linguistic style (RQ2b).

> **Research Question 2a (RQ2a)**: How can we evaluate whether text representations are sensitive to changes in linguistic style?

Linguistic style has been extensively studied in (socio-)linguistics (e.g., Labov, 1972; Bell, 1984; Eckert, 2008) and also received some attention in NLP (Danescu-Niculescu-Mizil et al., 2012; Neal et al., 2017; Gatt and Krahmer, 2018; Jin et al., 2022). In NLP, there are several general evaluation benchmarks for different linguistic phenomena (e.g., Wang et al., 2018, 2019) but less emphasis has been put on linguistic style. In Chapter 4, I develop STEL, the first framework to evaluate text representations in their abilities to represent style.

> **Research Question 2b (RQ2b)**: How can we build neural representations of linguistic style that are disentangled from referential meaning?

Recently, training objectives based on the authorship verification task have been used to train neural vector representations of text that are sensitive to author style (Boenninghoff et al., 2019b; Hay et al., 2020; Zhu and Jurgens, 2021). In Chapter 5, I experiment with using the same task to train neural text representations that are sensitive to linguistic style. The assumption is that two texts written by the same author are more likely to be written in the same style than two texts written by different authors. However, representations trained on the authorship verification task might suffer from being sensitive to not just style, but also to referential meaning. This is because style and referential meaning are often correlated (Gero et al., 2019; Bischoff et al., 2020). An author using primarily British spelling might more often discuss British politics or British landmarks. As a result, a model solving the authorship verification task might succeed in part by representing the referential meaning of texts. In Chapter 5, I introduce a variation of the authorship verification training task by increasing the likelihood that utterances from different authors are about similar topics. I evaluate the newly trained neural text representations on a variation of STEL that pitches style against referential meaning.

Study C: Paraphrasing Across Speakers

In the last part of this dissertation, I consider tasks that require robustness to language variation. Specifically, I focus on detecting paraphrases in dialog. The task of paraphrase classification across speakers, requires robustness to shifts in viewpoint (e.g., Speaker 1: "That book is mine." becomes Speaker 2: "That book is yours."), contextual information (e.g., "several years" might be "a while" for the topic of dis-

cussion) and speaker varieties (e.g., African American English vs. Standard American English).

I

> **Research Question 3 (RQ3)**: How can we detect paraphrases across speakers in dialog?

Repeating or paraphrasing what the previous speaker said has time and time again been found to be important in human-to-human or human-to-computer dialogs: It encourages elaboration (Rogers, 1951; Miller and Rollnick, 2013; Hill, 1992; Shah et al., 2022), de-escalation (Vecchi et al., 2005, 2019), and can increase the perceived response quality of dialog systems (Weizenbaum, 1966; Dieter et al., 2019). In Chapter 6, I investigate paraphrases across turns in dialog. Dialog is a setting that is uniquely sensitive to context (Grice, 1957, 1975; Davis, 2003) and might make matching the same referential meanings especially difficult. I provide an operationalization of context-dependent paraphrases in dialog, develop a training for crowd workers to classify paraphrases in dialog and introduce a dataset with utterance pairs from NPR and CNN news interviews annotated for context-dependent paraphrases. I compare training encoder models and in-context learning with decoder models to automatically detect paraphrases across speaker turns in dialog.

1.2. Main Contributions

With this dissertation, I make theoretical, empirical and artifact contributions.[7]

Theoretical Contributions

- I introduce and motivate the separation of NLP tasks into variation-sensitive and variation-robust tasks (Chapter 3). This separation emphasizes the different requirements models might have to fulfill to deal with language variation in different scenarios. I motivate and explain why tokenizers might play a crucial role in both variation-sensitive and variation-robust tasks.
- I provide an evaluation framework called `STEL` to evaluate computational text representations on how well they represent linguistic style (Chapter 4). I make the framework modular to cater to different definitions of style.
- I motivate the use of "hard-negatives" for learning stylistic text representations (Chapter 5) to make style representations independent from referential meaning. I also motivate a variation of `STEL` to test a model's ability to prioritize style over referential meaning in its representations.
- I introduce a definition of context-dependent paraphrases and motivate the importance of detecting paraphrases in context—particularly across speakers in dialog (Chapter 6).

[7] I use "I" to introduce and conclude my dissertation. However, this work was made possible only through the support of many others. See the research chapters for the respective collaborators and the acknowledgments for my broader support system.

Empirical Contributions

- I am the first to investigate tokenizers while accounting for language variation (Chapter 3). I show that the best algorithmic choices for tokenizers vary on tasks requiring robustness to and sensitivity to language variation. I show that the under-researched "pre-tokenization" step of tokenizers has the biggest influence on performance. Moreover, I find that common evaluation measures of tokenizers work less well to predict downstream performance on tasks that are sensitive to language variation.
- I am the first to systematically evaluate computational text representations in their sensitivity to linguistic style (Chapter 4). I find that neural text representations are better than previous feature-engineered approaches in representing style across several dimensions.
- I train neural text representations of style using an authorship verification training task that pitches referential meaning against style (Chapter 5). I am the first to evaluate neural representations on their ability to represent linguistic style independent of referential meaning.
- I am the first to operationalize, annotate and automatically detect context-dependent paraphrases across turns in dialog (Chapter 6). I reach promising results with both decoder and encoder models. When identifying text spans that constitute paraphrase pairs, encoder models profit from not being able to hallucinate quotes.

Artifact Contributions

- I provide a compilation of existing and original tasks to evaluate models on tasks requiring sensitivity and robustness to language variation (Chapter 3).
- I provide data for four style dimensions demonstrating the introduced `STEL` evaluation framework to evaluate text representations on their sensitivity to linguistic style (Chapter 4).[8]
- I release stylistic text representations, or *style embeddings* (Chapter 5), that have been appreciated in the NLP community for their independence from referential meaning.[9]
- I provide a dataset of context-dependent paraphrases in news interviews[10], and a trained encoder model to classify paraphrases in dialog (Chapter 6).[11] Additionally, I iteratively developed and share instructions for a hands-on annotator training in detecting paraphrases in dialog.[12] While annotation procedures for NLP tasks are typically not mentioned as individual contributions, I still consider

[8] `https://github.com/nlpsoc/STEL`
[9] `https://huggingface.co/AnnaWegmann/Style-Embedding`
[10] `https://huggingface.co/datasets/AnnaWegmann/Paraphrases-in-Interviews`
[11] `https://huggingface.co/AnnaWegmann/Highlight-Paraphrases-in-Dialog-ALL`
[12] `https://annawegmann.github.io/Paraphrases.html`

this hands-on and iteratively tested training of annotators a valuable contribution as descriptions of annotation instructions and procedures are hard to find and increasingly relevant when considering human variation.

1.3. Outline

The remainder of this dissertation continues with a background section, four research chapters and a conclusion. Note that this is a paper-based dissertation and consists of a collection of four conference papers, that are tied together and put into perspective with the help of an introduction, a background section and a conclusion.

Part I: Background Part I concludes by providing background on the two research areas that are at the core of my dissertation: NLP methods and linguistic style. First, I introduce the foundations of the historically successful transformer encoder models and argue why, even after the broad success of pure decoder models like CHATGPT, encoder models are still relevant to this day. Second, I provide an overview of different style definitions and style operationalizations in sociolinguistics.

Part II: Tokenization for Variation-Robust and Variation-Sensitive Tasks Chapter 3 introduces the separation of NLP tasks into tasks requiring robustness and sensitivity to language variation. I motivate the relevance to one foundational building block of LLMs: tokenizers. I show that the best tokenizer varies for tasks requiring robustness and sensitivity to language variation.

Part III: Representing Linguistic Style Chapters 4 and 5 focus on a task requiring sensitivity to language variation: representing the linguistic style of a text. Specifically, Chapter 4 introduces an evaluation framework to evaluate in how far text representations represent linguistic style. Chapter 5 uses contrastive learning and demonstrates that hard negatives are crucial to train content-independent style representations.

Part IV: Paraphrasing Across Speakers Chapter 6 focuses on a task requiring robustness to language variation: detecting paraphrases across speaker turns in dialog. I provide annotation procedures that deal with disagreements among annotators, provide a dataset and experiment with decoder and encoder models to detect paraphrases computationally. For paraphrase span identification, encoder models profit from not being able to hallucinate quotes.

Part V: Conclusion Chapter 7 summarizes the main findings of my dissertation and discusses future research directions when it comes to accounting for language variation in NLP.

2

Background

2.1. Natural Language Processing

Natural language processing (NLP) is a research field combining computer science, linguistics and artificial intelligence that studies algorithms that process and generate natural language. Here, natural language stands for human languages like Dutch, Chinese or Tamil that developed "naturally" over time, as opposed to artificial languages like programming languages or mathematics. NLP models take natural language as input or return natural language as output.[1] The recently surging large language models (LLMs) like CHATGPT[2] do both, they receive natural language as input, process and manipulate it, and return natural language as output. Real-world applications that make use of NLP models include information retrieval systems like GOOGLE SEARCH, machine translation systems like DEEPL and dialog systems like CHATGPT. By now, NLP systems are increasingly and commonly used for many different applications and by many different social groups across society.

Note that I use *language models* as an umbrella term for both encoder models and decoder models. Historically, only models that can generate text were considered language models. In this thesis, I use a broad understanding of the term language model including encoder models. Now, the term language model is usually used for neural NLP models that encode or generate texts.

Text Representations

Vector representations of texts are fundamental to the field of NLP. Generally, text representations transform texts into real numbered vectors (called text embeddings)

[1] NLP models can also take other modalities (e.g., images) as input or output. However, every NLP model processes natural language of some form. In this dissertation, I only consider text data.

[2] CHATGPT is technically an application build on top of an LLM like GPT-4. I am using CHATGPT as a stand in for OpenAI's underlying LLMs as it is the more broadly known term outside of NLP.

that represent the object of interest — usually the meaning of a text. I focus on modern neural approaches in this dissertation. However, classic approaches like TF-IDF (Spärck Jones, 1972) or word n-grams were (and sometimes still are) successfully used for some time (Jurafsky and Martin, 2025).

Word Embeddings

The neural representation model that transformed the field of NLP in 2013, is the word embedding algorithm WORD2VEC (Mikolov et al., 2013). It used the continuous bag-of-words training task.[3] For continuous bag-of-words, the word that is currently learned is masked out and the neural model has to learn to predict it from the 10 surrounding words.[4] It is called bag-of-words as the model cannot make use of the order or structure of other words surrounding the masked out word; all 20 context words are just "put in a bag" and considered without their structure. WORD2VEC provides 300-dimensional vector representations of words. It represents words that appear in similar contexts close to each other. This idea can be traced back to the distributional hypothesis (i.e., similar words appear in similar contexts) introduced in linguistics and philosophy (Harris, 1954; Firth, 1957; Wittgenstein, 1953; Brunila and LaViolette, 2022). WORD2VEC represents words in a dense, continuous vector space capable of capturing more nuanced semantic relationships than previous approaches. Soon other word embedding models like GLOVE (Pennington et al., 2014) and FASTTEXT (Bojanowski et al., 2017) followed. Still, word embeddings had a problem: They return single, static vectors for a word, no matter what context it appears in. However, a word like "run" can take on different meanings depending on what context it appears in. Consider the difference in meaning between "run" in "run a company" and "run a marathon".

Transformers

In the 2010s, there were other efforts in NLP to use neural networks to represent and manipulate not just single words but sequences of words to retain structural information and word dependencies. In 2017, this led to the break-through *transformer* architecture (Vaswani et al., 2017), see also Figure 2.1. Transformers remain the dominant architecture for language models to date. One of the key innovations of transformers is the principle of *self-attention*. For a given input word, self-attention computes a weighted average of all words in the input sequence, where the weights reflect the "relevance" of each word to the one being considered. For example, consider the sentence "This is my dissertation.", "This" might receive higher relevance for "dissertation" than "is" or "my". Self-attention can be applied in parallel which is a huge advantage compared to previously common architectures like RNNs that process everything sequentially. Further, it led to contextualized text representations: Vector representations of the word "run" in a given input sequence are now different for input sequences "run a company" and "run a marathon".

[3] Also the skip-gram training task. It's similar to bag-of-words, so I skip-that.

[4] This as well as other numbers like the 300 dimensions following later are hyperparameters that can take on other values. I use specific numbers for illustrative purposes.

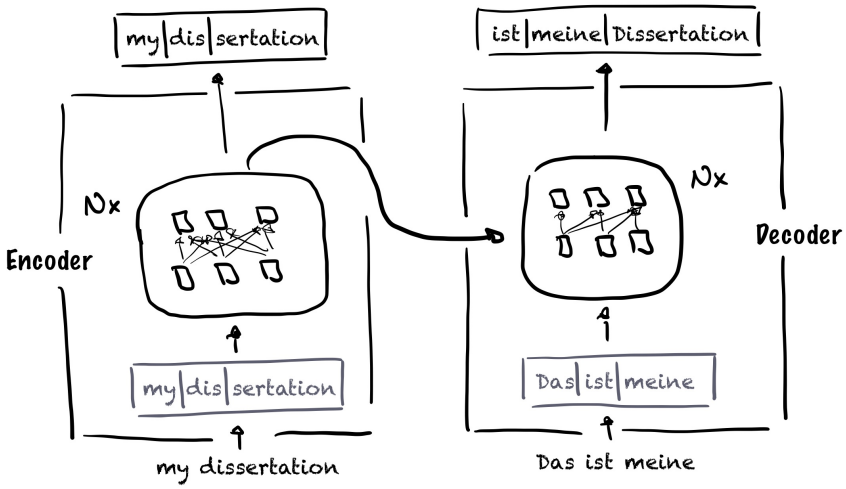


Figure 2.1: Simplified Transformer Architecture. I display a simplified schematic of the encoder-decoder transformer architecture (Vaswani et al., 2017). The original schematic includes sequences of nonlinearities and feed forward neural networks, positional encodings and output probabilities. I display the tokenizer separately in **grey** as I zoom in on tokenizers in Research Chapter 3. The tokenizer cuts the input text into individual tokens. This is signified by the single blocks in the figure. The transformer is an architecture that can include an encoder and a decoder (as displayed), just an encoder or just a decoder. The attention mechanism is bidirectional in the encoder models, and unidirectional in the decoder model. Encoder models are typically trained using a masked language modeling task. Decoder models are typically trained using a next word prediction task. In the figure, a machine translation model is displayed. "Dissertation" is predicted as the next word in a German translation of "dissertation".

The original transformer architecture consists of an *encoder* and a *decoder*. Encoder models are typically trained using *masked language modeling*, where the model predicts missing words in a sequence (e.g., "I defended my ___ successfully."). By learning to perform masked language modeling, encoders learn representations of texts. Decoder models, by contrast, are usually trained using *next word prediction*, where the model predicts the next word in a sequence of words (e.g., "I defended my ___"). By learning to perform next word prediction, decoder models learn to generate texts. One of the first successful models making use of the transformer architecture was the encoder model BERT (Devlin et al., 2019), which is still used for classification tasks. The first successful decoder models included GPT-2 (Radford et al., 2019). There are also models, such as T5 (Raffel et al., 2020), that combine both encoder and decoder architectures as suggested in Vaswani et al. (2017). While encoder models where more heavily studied with the release of BERT, decoder models have increasingly become the dominant type of language model—especially following the release of CHATGPT in November 2022.

Before an input can be processed by either an encoder or decoder model, it must first be converted into a sequence of tokens in a tokenization step. Generally, the *tokenizer* is a separate algorithm that splits raw input text into smaller units—tokens—that are then fed to the language model. Then, each token is mapped to a unique token ID, which is in turn used to extract learned vector representations (cf. the small blocks in Figure 2.1). Tokens might be whole word (e.g., "the"), superwords (e.g., "of the"), subwords (e.g., "th"), or even individual characters (e.g., "t"). For example in Figure 2.1, the input phrase "my dissertation" might be tokenized into {"my", " dis", "sertation"}. Tokenization typically consist of several stages including normalization (e.g., lowercasing all text as done for BERT UNCASED in Devlin et al., 2019), pre-tokenization (e.g., whitespace splitting, cf. Mielke et al., 2021) and segmentation. The segmentation step is arguably the most important. Here, the tokenizer segments a text based on a learned vocabulary and segmentation rules. The most common tokenizer algorithm in recent language models is Byte-Pair Encoding or BPE (Sennrich et al., 2016). Among others, it is used by LLAMA 3 (Grattafiori et al., 2024), MIXTRAL (Jiang et al., 2024), DEEPSEEK-R1 (Guo et al., 2025) and GPT-4 (Achiam et al., 2023). Other algorithms are discussed in Mielke et al. (2021). Tokenization is particularly important when models encounter previously unseen words (Mielke et al., 2021), when different types of tasks are considered (cf. Chapter 3) and for computation, as splitting a text into more tokens requires more calculations and reduces the amount of text that fits within the model's context window (Mielke et al., 2021).

The Importance of Encoder Models and Text Representations

I describe why up until today, good text representations are relevant in NLP and that encoder models are important. I argue for both as I am creating text representations for linguistic style in Chapter 5 and evaluate and train encoder models in all research chapters.

The relevance of text representations In the previous section, I established that text representations have been a major area of interest in NLP over the past decades. However, one could question their continued relevance with the following argument: Why do we still need separate representations of text—like precomputed embeddings that are provided as additional input to models—when generative models can already (or will soon be able to) solve all the problems we needed text representations for in the first place? For example, instead of using text representations to extract relevant documents for a GOOGLE SEARCH, we could directly ask a generative model to extract and return relevant documents across all documents provided to the model. However, the same example of GOOGLE SEARCH reveals a setting where text representations remain relevant: Retrieval Augmented Generation or RAG (Ram et al., 2023). RAG is a popular method in information retrieval that retrieves an answer to the user's question by comparing the embedding vector of the question with the embedding vectors of documents stored in a relevant database. This approach allows access to a much larger document pool since generative models are restricted in the number of tokens they can process at once. Further, RAG can complement a generative model by leveraging a curated up-to-date and domain-specific database, thereby increasing accuracy and redu-

cing hallucinations (Gao et al., 2024)—instances where the language model generates incorrect or fabricated information. Beyond RAG, there are several other applications that benefit from high-quality text representations such as measuring the semantic change of words over time (Shoemark et al., 2019; Wegmann et al., 2020; Schlechtweg et al., 2020), or clustering similar documents (Grootendorst, 2022).

The relevance of encoder models In the introduction, I established that encoder models drove a lot of the progress in NLP in the last decade. However, decoder models have since become the more heavily studied language model architecture. Despite this global shift in research focus, encoder models remain widely used in practice (Warner et al., 2024). This continued use is probably a result of state-of-the-art encoder models being typically much smaller than state-of-the-art decoder models, with encoder models using millions and decoder models juggling billions of parameters. As a result, encoder models are more efficient to train and can be applied more quickly than decoder models. Moreover, encoder models continue to perform well for discriminative tasks like classification where they can outperform or perform similarly to much larger decoder models (Warner et al., 2024). Overall, in applications where quick, cheap and accurate replies are essential, encoder models can be preferable over prompting big decoder models for every step.

Sociolinguists generally think of styles as different ways of saying the same thing. In every field that studies style seriously, however, this is not so

Penelope Eckert

I

2.2. Linguistic Style

In this dissertation, language variation is a central concept. Language variation has been studied across many different research fields. For concreteness, I borrow terminology and concepts from *sociolinguistics*—often from the U.S. American perspective[5]—as it is one area where researchers have extensively investigated different forms of language variation. Broadly, sociolinguistics studies how language interacts with society, with a focus on linking language variation and change to social meaning (Meyerhoff, 2006; Bell, 2014; Holmes and Wilson, 2017). While my research chapters are also motivated by sociolinguistic goals (e.g., enabling the study of linguistic accommodation with the help of style embeddings in Chapter 5), this dissertation primarily focuses on representing linguistic style and language variation and does not yet describe varieties in specific populations or more broadly connect language variation to social meaning (cf. future work in Chapter 7.3).

Language variation and style In this dissertation, two related concepts are discussed: language variation and linguistic style. Out of the two, language variation is the broader concept compared to linguistic style, practically encompassing all forms in which language can vary. Variation in linguistic style is only one, narrower form of variation in language (Bell, 1984; Eckert and Rickford, 2001; Coupland, 2007). I focus on explaining the literature on linguistic style further as it is the main focus of Part II. I discussed language variation in Section 1 in Chapter 1. Note that linguistic style and language variation are terms that are fundamentally intertwined. Although this section centers on linguistic style, it also engages with key considerations related to other types of language variation central to sociolinguistics.

[5] When discussing sociolinguistics in this dissertation, I usually refer to the U.S. American tradition of sociolinguistics. There are rich traditions in other parts of the world as well. See an overview of sociolinguistics around the world in Ball et al. (2023).

2.2.1. Style in General

Style has received a lot of attention in fields like sociolinguistics, stylometry, forensic linguistics, but also in natural language processing. With the term researchers usually aim to the describe the form of a text[6] (i.e., how something is said) more so than its referential meaning (i.e., what is said). Describing a style can be understood as studying what makes a phrasing distinctive (Irvine, 2001; Crystal, 2011) according to whatever lens we currently choose to look through. However, it is a highly ambiguous and elusive term that has previously been conceptualized in many different, sometimes conflicting ways.[7] It is impossible to give a complete overview here. Instead, I provide an overview of the various style operationalizations that have influenced this dissertation (Sections 2.2.2 and 2.2.3). Before doing so, I will briefly make some general remarks.

Referential meaning Referential meaning[8] is a term that is commonly used in sociolinguistics (Labov, 1972; Lavandera, 1978; Campbell-Kibler, 2011; Nguyen et al., 2016, 2021). The term appears to point toward two possible traditions: (1) the Mill-Frege-Russell-Montague compositional semantics tradition in theoretical and formal semantics (Labov, 1972; Weiner and Labov, 1983), which (among other things) formalizes the meaning of sentences through compositional grammars that systematically pair linguistic expressions with logical representations (see, e.g., van Benthem and ter Meulen, 2011), in a way that centers around the notion of reference, or (2) de Saussure (1916)'s structuralist tradition (Hernández-Campoy, 2016), where a linguistic sign consists of the signified (i.e., mental concept, e.g., an inhabitant of the city of Utrecht) and a signifier (i.e., the sound pattern or word, e.g., "Utrechter") matched by arbitrary social conventions, where meaning is further determined by the sign's relationship and contrast with other signs within the linguistic system.

The referential meaning of a sentence is, roughly speaking, the factual message that lies at the core of the sentence (i.e, the "What") , as opposed to the linguistic means through which this message is expressed linguistically (i.e., the "How"). This distinction is especially compelling in (early) variationist sociolinguistics (Labov, 1972; Eckert, 2012) which studies quantitatively how different ways of saying the same thing—known as variants of a *sociolinguistic variable*—are connected to external social factors. Two variants are often assumed to have the same referential meaning if they are the same in a truth-conditional sense (i.e., they are true in exactly the same situations). The "social and/or stylistic significance" of variants that have the same referential meaning might differ considerably (Labov, 1972; Weiner and Labov, 1983).

A precise interpretation of this definition poses practical challenges. It has been argued, for example, that only Labov (1972)'s original object of study—phonological variables—can leave referential meaning untouched whereas all other variables including lexical and syntactic variables will necessarily change the referential mean-

[6] I do not discuss the term style in other modalities. Note, however, that one might also discuss the distinctive style of a painting or an outfit.

[7] Scholars have even argued against using the term style at all, e.g., Crystal (2011).

[8] Other terms that capture a similar notion include "content", "referential content", "semantic content", "semantic meaning", "linguistic meaning" or "denotational meaning" in sociolinguistic work like Eckert and Rickford (2001); Meyerhoff (2006); Eckert (2008, 2012); Bell (2014); Holmes and Wilson (2017) and Ball et al. (2023). Throughout the dissertation I use referential meaning and (semantic) content interchangeably.

ing as well (Lavandera, 1978; Campbell-Kibler, 2011; Sun et al., 2023). Further, Eckert (2008, 2012) argues that a used style systematically connects an utterance to the social world, and style can thus not be fully separated from referential meaning. Generally, one might argue that every two forms necessarily contrast in meaning (Clark, 1992). Some work in sociolinguistics side-steps the problem of meaning equivalence by identifying and studying the contexts in which a set of linguistic forms count as alternants (Campbell-Kibler, 2011; Christensen and Jensen, 2022).
Nonetheless, I believe the notion of referential meaning to be useful because it draws our attention to cases in which form varies but referential meaning remains approximately the same (Weiner and Labov, 1983). In NLP, for example, we often face practical challenges where it is useful to cluster similar referential meanings (e.g., GOOGLE SEARCH) or adapt to different styles (e.g., Machine Translation) which requires a separation of style and referential meaning of some form.

Speech production in psycholinguistics Psycholinguistics is an interdisciplinary research field that studies how language is processed in the human brain. An influential psycholinguistic model is Levelt (1989)'s model of speech production which separates the human cognitive process of speech production into a conceptualizer (that creates an idea or message), formulator (that phrases the message) and articulator (that produces the phonetic speech signal). The formulator can be influenced by contextual factors like the alignment to an interlocutor or the importance of avoiding misunderstandings (Garrod and Anderson, 1987; van Deemter et al., 2012). There are clear parallels between "referential meaning" in sociolinguistics, the notion of a "message" in the psycholinguistics model, and the notion of a "message" in the classic modular natural language generation pipeline (which took Levelt's pipeline as one of its sources of inspiration) (Gatt and Krahmer, 2018), see Section 2.2.4. All these notions are separated from the articulation, formulation, form or linguistic style of an utterance or text.

Style across research fields There are several fields that study linguistic style in some capacity. I generally focus on style operationalizations and definitions from U.S. American *sociolinguistics* including work like Labov (1972) and Eckert (2008). Sociolinguistics generally studies the relationship between language and society with a focus on language change and variation. Other fields that study style include *corpus linguistics* that study language use in text corpora including work like Biber (1988) and Biber and Conrad (2019). Typical applications might include comparing language between different genres like scientific papers or poems. *Forensic linguistics* studies style in the context of law and crime investigation and is typically interested in recognizing a style or *idiolect* unique to investigated individuals (Coulthard, 2004). Practical insights from forensic linguistics also reciprocally influence *stylistics* and *stylometry* that more generally studies linguistic style in language. Specific applications might include investigating the style of literary authors (Holmes, 1985), or attributing disputed literary works (Mosteller and Wallace, 1963; Burrows, 2002; Stamatatos, 2009). Style has also been investigated in the general field of natural language processing to characterize authors (e.g., age or gender in Koppel et al., 2002; Nguyen et al., 2013), detect stylistic inconsistencies (Collins et al., 2004; Stamatatos, 2009) or adapt styles in machine translation

(Rabinovich et al., 2017; Niu et al., 2017, 2018). Note that these fields are not strictly separable. Methods from corpus linguistics can inform sociolinguistics, forensic linguistics uses methods from stylometry and so on. Further, there are several fields that can be connected to linguistic style that I do not specifically discuss here like *discourse analysis*, *digital humanities*, *linguistic anthropology*, or *sociology*.

2.2.2. Style Conceptualizations

In this section, I discuss differences and commonalities in style when it comes to style definitions and conceptualizations.

Whose style? Style is usually discussed in a relative sense, as a distinctive difference between two objects of study (Irvine, 2001). However, the objects of study vary. In (socio-)linguistics, style has most often been discussed as intra-individual variation (Bell, 1984; Irvine, 2001; Labov, 2006; Meyerhoff, 2006): The variation in language use of the same speaker across different situations. Famously, Labov (1972) compared the style of individuals when he manipulated them to pay more or less attention to speech. Sociolinguistics have studied the relationship between the variation within individuals and the variation between individuals or groups on the "social dimension" (Bell, 1984; Irvine, 2001)—for example in that the style of individuals indexes membership of certain social groups (Eckert, 2008) like g-dropping might index a southern region in the US (Campbell-Kibler, 2007). The collective linguistic repertoire of social groups have also been called style (Eckert and Rickford, 2001)—for example, the language use of women (Degaetano-Ortlieb, 2018; Degaetano-Ortlieb et al., 2021)—which deviates from the focus on the individual. Next to styles of individuals or groups, domains, genres and registers have also been objects of style research (Biber and Conrad, 2019; Grieve, 2023). News reports, blogs and conversations might display very different patterns in their linguistic features which might be called the style of a news report, blog or conversation (Biber and Conrad, 2019; Irvine, 2001; Grieve et al., 2011). There are many other objects of study that also have been described stylistically. For example, Trump's Twitter account over the course of a decade (Clarke and Grieve, 2019), the communication within social networks (Dodsworth and Benton, 2020), 200 years of court proceedings (Degaetano-Ortlieb, 2018), different sections of the same text (Degaetano-Ortlieb and Teich, 2017), academic disciplines (Andresen and Zinsmeister, 2017) or the literary texts by specific authors or from specific time periods (Coupland, 2007).

What is the function of style? Style has been described in different ways. Some scholars emphasize the aesthetic aspects of style[9], seeing it as a matter of linguistic variation with no or limited function (Biber and Conrad, 2019). Others argue that style is fundamentally embedded in social meaning, shaping social identity, relationships, and interactions (Campbell-Kibler et al., 2006; Coupland, 2007; Eckert, 2008, 2012). Stylistic choices have been considered to mark and reflect social categories like demographic variables and social identities (Labov, 1972; Eckert, 1989). Labov (1972) found that differences in pronunciation of /r/ were correlated with social class, while Eckert (1989) found that self-identified "burnouts" at a school in Detroit used fewer standard linguistic features than the more college-bound "jocks". Additionally, stylistic choices have been argued to not just reflect social identity or group membership but to be an active part of indexing social identities as a positioning of individuals within the social world (Eckert, 2008, 2012). For example, Bucholtz (1999) found the use of lin-

[9] In contrast to style, Biber studied "registers" more extensively which he defines as varieties of languages associated with a particular situational context (Biber and Conrad, 2019). Biber's registers have communicative function. Other researchers might call Biber's registers styles as well.

guistic features associated with African American Vernacular[10] to index a kind of masculinity (Eckert, 2012). Coupland (2007) emphasizes that style is interactive. Speakers might accommodate to or distance themselves from the style of their interlocutors or audiences (Giles and Powesland, 1975; Bell, 1984), thereby shaping social relationships (Coupland, 2007). For example, Bourhis and Giles (1977) found that integrative Welsh learners (i.e., those learning Welsh out of interest in their cultural heritage) used more accent features to distance themselves from the researcher's RP English (i.e., a "standard" pronunciation of English) when they called Welsh a "a dying language" (Bourhis and Giles, 1977). Similarly, Barrett (2006) found that Spanish functioned as a means of expressing solidarity and resisting the authority of Anglo managers among Spanish-speaking employees. Further, Bell famously found that New Zealand newscasters shifted their pronunciation when talking to audiences considered to have higher or lower status (Bell, 2014). Finally, when considering Biber and Conrad (2019)'s genre and register as style, style might serve further functions like structuring discourse and fulfilling communicative purposes.

When does style vary? Fields like forensic linguistics are interested in *idiolects*, i.e., the distinctive and idiosyncratic stylistic choices of individuals (Coulthard, 2004)—essentially the aspects of an individual's style that remain relatively stable across all situations. Other aspects of style are generally considered variable. Linguistic style is often considered in a relative sense. Irvine (2001) even argues that a particular style cannot be explained in isolation but must be understood in relation to other styles, as styles operate within a system of social differentiation and ideology. Style has usually been considered dependent on the social and situational context (Eckert and Rickford, 2001; Meyerhoff, 2006): Style can vary depending who the speaker is addressing (e.g., Bell, 1984; Giles and Powesland, 1975). Style also shifts depending on the topic of discussion (Bell, 1984; Rickford and McNair-Knox, 1994). For example, Holliday (2021) finds that biracial Black men displayed fewer African American intonational features when discussing police narratives. Additionally, the mode of communication (e.g., writing and speech in Biber, 1988; sign language in Kusters and Lucas, 2022) as well as the *genre* (i.e., conventions to structure complete texts like news reports or academic papers) and *register* (i.e., functional use of features in a given context like technical jargon in a contract) can have an influence (Biber, 2012). Generally, speech conditions (e.g., speech to a crowd, speech in the courtroom in Ervin-Tripp, 2001) can influence stylistic choices. Other situational factors that shape style can include the social identity and group membership of interlocutors (Eckert and Rickford, 2001), the current historical period (Coupland, 2007; Biber and Conrad, 2019), or the speaker's personal goals (Meyerhoff, 2006). However, even with a fixed audience and setting, style might shift dynamically based on a speaker's decision, which in turn might shape the social context (Eckert, 2012; Coupland, 2007).

[10] While narrow stylistic (Section 2.2.3) and dialectal features probably overlap, dialects tend to not be considered style but a different type of language variation more closely tied to the speaker's social background, such as their geographic region (Biber and Conrad, 2019; Grieve et al., 2025). Nonetheless, some researchers have also considered dialects as a kind of social style (Coupland, 2007).

Is style a choice? Labov (1972) was interested in a speaker's "natural" style (also: vernacular; or authentic speech in Eckert, 2003) that emerges when the speaker pays less attention to their speech—offering "direct access to language untainted by the interference of reflection or social agency" (Eckert, 2003). Originally, Labov understood a speaker's vernacular as a reflection of their broad social identity, not an active choice. He compared speech of individuals under different conditions, e.g., by asking the same person for directions twice in NYC stores and comparing the second answer to the first when the speaker arguably paid more attention to their speech. In more recent approaches in sociolinguistics, style has been argued to be more agentive—reflecting a speaker's identity but also performing and constructing it (Eckert, 2012). For example, the development of linguistic practices of trans activists can be tied to their agency in the creation of their own identity (Zimman, 2019). Additionally, speakers might choose a certain style for performative function like getting attention or for persuasion (Ervin-Tripp, 2001). More broadly, speakers might decide on a certain style to position themselves in their community. As mentioned above, they can make use of the *indexical field* (Eckert, 2008), that is, the range of social meanings associated with a linguistic variable. Depending on the perspective of the hearer, the linguistic variable indexes a membership to a social group which by association leads to the indexing of character traits stereotypically associated with members of that group. For example, the use of the "ing" ending over the "in" ending for a verb can be interpreted as the speaker being educated or pretentious (Campbell-Kibler, 2007; Eckert, 2008).

2.2.3. Style Operationalizations

No matter the specific definition, style is usually operationalized through patterns in linguistic features. I give a brief overview of some narrow style features and broad style dimensions.

Narrow style features The considered linguistic features are usually narrow phonological, lexical, syntactic, orthographic or discourse features (Biber and Conrad, 2019; Crystal and Davy, 1969; Stamatatos, 2009). See some examples in Table 2.1. U.S. American sociolinguistic work has often focused on investigating single phonological features quantitatively as well as qualitatively (Labov, 1972; Bell, 1984; Eckert, 2008; Kirk, 2023). Forensic linguistics, stylistics and corpus linguistics has typically investigated broader sets of lexical and grammatical linguistic features (Biber, 1988; Stamatatos, 2009). Influential feature operationalizations include Biber's Multidimensional Analysis or MDA (Biber, 1988; Biber and Conrad, 2019), modern extensions of MDA (e.g., Grieve et al., 2011; Clarke and Grieve, 2017) and LIWC (Pennebaker et al., 2015). LIWC and MDA consist of several dimensions or features like first person pronouns or negation or part-of-speech that can be used to quantitatively investigate patterns in linguistic features. For further analysis, like investigating the communicative function of linguistic features (Biber, 1988) or profiling authors with patterns in linguistic features (Neal et al., 2017), the linguistic features are used in concert with statistical approaches like dimensionality reduction with factor analysis and distance measures like Burrow's Delta and classifiers like SVM.

Broader style dimensions Next to studying collections of narrow linguistic features, researchers have also considered broader style categories or dimensions that are arguably made up of distinctively patterned collections of linguistic features (Heylighen and Dewaele, 1999; Biber, 1988). Such dimensions could be identified based on their communicative function (e.g., involved versus informational function in Biber, 1988) or based on external categories like the time period, the social background of speakers or situational factors (Campbell-Kibler et al., 2006; Grieve et al., 2025). Generally, such broader dimensions come with the drawback of inviting disagreement on what an interesting dimension is and what specific style features might constitute the considered dimension. Formality is one of the most agreed upon broader dimension of style (Hovy, 1987; Heylighen and Dewaele, 1999) but it also does not have consistent operationalizations (Pavlick and Tetreault, 2016; Fang and Cao, 2009; Heylighen and Dewaele, 2002). Informal style might use more contractions, interjections and idioms or might just be recognizable with the adjective or part-of-speech frequencies. In stylistics or stylometry one might consider the style of a specific author or time period as the broader style dimension of interest (Stamatatos, 2009). This allows for an operationalization of style based on a collection of documents using information-theoretic measures (Degaetano-Ortlieb et al., 2018) or language models (Huang and Grieve, 2024). In natural language processing specifically, most work does not try to define style but rather uses it as an umbrella term for general attributes of texts (Jin et al., 2022)—even including attributes that are related to content like sentiment (Reif et al., 2022). Despite the disagreements, broader style dimensions can be useful to study and understand: For example, formality can indicate familiarity between people, and goals in an interaction and might as a result be relevant to consider for dialog systems (Pavlick and Tetreault, 2016; Vanderlyn and Vu, 2025). Next to formality, other common style dimensions include simple/complex style (Flesch, 1948; Hovy, 1987; Pavlick and Nenkova, 2015), abstract/concrete style (Semin and Fiedler, 1988), and restricted/elaborated style (Bernstein, 2003).

Type	Variable	Examples
LEXICAL	word length	e.g., average word length in Biber's MDA, cf. Grieve (2007)
	sentence length	e.g., distribution of average sentence length, cf. Grieve (2007)
	vocabulary richness	operationalizes vocabulary diversity with measures like type-token ratio in Biber's MDA, number of words occurring once, ... cf. Stamatatos (2009)
	function words	e.g., word frequency distributions of statistically determined most common words, typically including words like "the", "be", "to", cf. Stamatatos (2009)
	pronoun use	e.g., word frequency distributions of first, second, ... person pronouns using Biber's MDA features or LIWC dimensions
	hedge words	e.g., "at about", "something like" as hedges in Biber MDA features; "maybe", "perhaps" in tentative dimension in LIWC
	quantifiers	e.g., "each", "all" as quantifier words or "everybody" or "anybody" as quantifier pronouns in Biber's MDA
	...	
LEXICAL	part-of-speech	e.g., noun, verb, adjective, ... in Biber's MDA
	passive voice	e.g., agentless passives in Biber's MDA
	subordination features	e.g., that-relative clause vs. wh-relative clause (e.g., the dog who vs. the dog that) in Biber's MDA
	negation	e.g., "need no water' as negative concord in Eckert (2008), "not" in analytic negation in Biber's MDA, negation words in LIWC
	invariant be	e.g., "He be working." in Rickford and McNair-Knox (1994)
	zero copula	e.g., "She nice" in Rickford and McNair-Knox (1994)
	...	
MORPHOLOGICAL	word endings	e.g., g-dropping (Campbell-Kibler, 2007), gerunds in Biber's MDA
	nominalizations	e.g., ending in -tion, -ment
	verb morphology	e.g., be as main or auxillary verb in Biber's MDA
	...	
ORTHOGRAPHIC	compression	e.g., train a compression model on one text and use it to estimate how similar in style another text is, cf. Stamatatos (2009)
	character types	e.g., hashtags, emojis, exclamation marks in Clarke and Grieve (2017), uppercase character, digits in Stamatatos (2009)
	character n-grams	cf. Stamatatos (2009)
	lengthening	e.g., "cooool" in Nguyen and Grieve (2020)
	number substitutions	e.g., "2day" in Crystal (2008)
	...	
DISCOURSE	contraction use	e.g., Biber's MDA
	discourse particle	e.g., "well", "now" in Biber's MDA
	...	
PHONETIC	postvocalic /r/	e.g., more or less clear pronunciation of /r/ sound after vowel, cf. (Labov, 1972)
	intervocalic /t/	e.g., full or flapped voicing of /t/ between two vowel sounds making "writer" sound like "rider" (Bell, 1984)
	...	

Table 2.1: Overview of narrow style operationalizations used in different fields for English. I display specific linguistic features that have been used to operationalize linguistic style. These have been investigated separately (Campbell-Kibler, 2009) and collectively (Biber, 1988). I categorize the linguistic features into lexical (i.e., word choice), syntactic (i.e., sentence structure), morphological (i.e., word structure and inflection), discourse (i.e., larger structure), orthographic (i.e., spelling and punctuation) and phonetic (i.e., pronunciation and sound patterns) features. Note that the categorization into lexical, syntactic, morphological etc. are mine and might overlap, e.g., g-dropping might also be considered an orthographic or phonological variable and character n-grams might encode different morphemes. The relevant reference for Biber's MDA is Biber (1988) and for LIWC is Pennebaker et al. (2015). This table was inspired by and partially filled with elements from the table of stylometric features in Stamatatos (2009) and Neal et al. (2017). For further references and examples consider also Grieve (2007) and Biber (1988).

2.2.4. Applications of Style in Natural Language Processing

Stylistic NLP tasks Style has been considered explicitly for different tasks in NLP. The most prominent probably being *authorship attribution* (Neal et al., 2017)—the task of assigning an author to a given text based on linguistic features, and *style transfer* (Jin et al., 2022)—the task of changing the style but not the referential meaning of a text (e.g., the informal "Come and sit!" is transferred to the formal "Please consider taking a seat."). Authorship attribution is closely related to practical applications like the attribution of disputed literary works to known authors (e.g., were essays in the Federalist papers written by Alexander Hamilton or James Madison?), the detection of authors of harassing or fraudulent texts (e.g., who wrote the code for malicious software) and the detection of plagiarism (e.g., was this dissertation written by a chatbot) (Stamatatos, 2009; Arabnezhad et al., 2020; Maneriker et al., 2021; Manolache et al., 2022; Crothers et al., 2023; Huang and Grieve, 2024). In turn, style transfer is closely related to personalization (Flek, 2020; Jin et al., 2022), e.g., changing a text's style to simpler language for non-experts. Another common task related to authorship attribution is *author profiling*—the task of recovering author characteristics based on the text they wrote (Rangel et al., 2013; Nguyen et al., 2013). Note, that author profiling can be useful to improve the performance on some NLP tasks (Hovy, 2015). However, identifying an author's gender, age, personality type etc. has increasingly been criticized for privacy concerns (Brennan et al., 2012; Li et al., 2018; Elazar and Goldberg, 2018; Lison et al., 2021). See a discussion in Section 7.3.

NLG Natural language generation (NLG) systems have to fundamentally determine what information to generate (often called the message or content) and how to generate it (the form or style) (Gatt and Krahmer, 2018). In rule-based NLG, this distinction is usually made explicit. For example, in data-to-text pipelines—NLG systems that convert structured data such as formulas in knowledge bases to text—separate planning stages are devoted to content (i.e., deciding what factual information to include) and form (i.e., deciding how to express the content) (Reiter, 2025). In contrast, neural NLG systems often handle content and style implicitly, generating a text end-to-end without explicit planning stages for determining what and how to say something. Nevertheless, or perhaps precisely because of a lack of planning (Reiter, 2025), generating texts in specific styles remains an important challenge also for neural NLG systems (Ficler and Goldberg, 2017; Gatt and Krahmer, 2018). While early work in NLG often focused on factual correctness (Gatt and Krahmer, 2018), much recent work focuses on how to say the same thing in different ways. An example of content-focused generation is Winograd's SHRDLU blocks world system (Winograd, 1972), which allowed users to interact with a simulated "blocks world" using natural language commands like "put the blue pyramid on the block". In contrast, Hovy (1987) aimed to achieve different communicative goals by choosing different styles (e.g., varying the formality) to word the same message. Recent work in NLG on style include efforts to personalize the style in machine translation systems (Rabinovich et al., 2017; Niu et al., 2017, 2018), control readability in summarization (Goyal et al., 2022; Dreyer et al., 2023; Ribeiro et al., 2023) and project personality traits in dialog systems (Mairesse and Walker, 2007; Oraby et al., 2018) or personalize responses for individual users (Flek, 2020).

Other uses Language models can be brittle when it comes to performance across different styles or language varieties (Pan et al., 2022; Mizrahi et al., 2024). Paraphrasing prompts using different stylistic features (Wahle et al., 2024), or enlarging (post-)training datasets with (synthetic) variations have helped increase model performances and make models more robust to stylistic and other language variation (Wang et al., 2023c; Chen et al., 2024).

2.2.5. Linguistic Style in this Dissertation

In this background section, I have established that linguistic style is an ambiguous and elusive term that has previously been conceptualized, operationalized and applied in different ways.

In this dissertation, I use a broad conceptualization of the term linguistic style, which I define in contrast to referential meaning (cf. Chapter 4 and Chapter 5): Linguistic style concerns how information is expressed, rather than what is expressed. I draw on both broad dimensions (cf. Section 2.2.3, e.g., formal vs. informal style) and narrow features of linguistic style (cf. Section 2.2.3, e.g., contraction vs. no contraction use) when evaluating the sensitivity of NLP methods to linguistic style in Chapter 4. To isolate stylistic effects, I vary these dimensions while keeping the referential meaning approximately constant. I mainly consider the linguistic styles of individuals (Chapter 5 and Chapter 3), but also investigate variation in spelling, dialects and across registers (Chapter 3). I assume that linguistic style is more likely to be the same for two utterances written by the same author than for two utterances written by different authors—an assumption that underlies the authorship verification approaches in Chapter 3 and Chapter 5.

Part II

Variation-Robust and Variation-Sensitive Tasks

So far I have discussed the background of two areas of research that are central to my dissertation: NLP methods and linguistic style. I have also explained that I aim to contribute to efforts that account for language variation. In Part II of this dissertation, I systematically introduce NLP tasks that require robustness and NLP tasks that require sensitivity to language variation. I motivate the relevance of language variation to one foundational building block of LLMs: tokenizers. I show that the best tokenizer varies for tasks requiring robustness and sensitivity to language variation. Overall, language variation might be important to consider at all stages of building LLMs.

Contents

3

Tokenization is Sensitive to Language Variation

This chapter is based on Wegmann, A., Nguyen, D. & Jurgens, D. (2025). Tokenization is Sensitive to Language Variation. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 10958—10983). https://doi.org/10.18653/v1/2025.findings-acl.572. See a video of the conference presentation here: https://youtu.be/GnEpTTj4fc8.

Abstract

Variation in language is ubiquitous and often systematically linked to regional, social, and contextual factors. Tokenizers split texts into smaller units and might behave differently for less common linguistic forms. This might affect downstream LLM performance differently on two types of tasks: Tasks where the model should be robust to language variation (e.g., for semantic tasks like NLI, labels do not depend on whether a text uses British or American spelling) and tasks where the model should be sensitive to language variation (e.g., for form-based tasks like authorship verification, labels depend on whether a text uses British or American spelling). We pre-train BERT base models with the popular Byte-Pair Encoding algorithm to investigate how key tokenization design choices impact the performance of downstream models: the corpus used to train the tokenizer, pre-tokenizer and the vocabulary size. We find that the best tokenizer varies on the two task types and that the pre-tokenizer has the biggest overall impact on performance. Further, we introduce a new approach to estimate tokenizer impact on downstream LLM performance, showing substantial improvement over metrics like Rényi efficiency. We encourage

Author contributions: AW developed the idea, prepared the data, implemented the experiments, and wrote the manuscript. DN and DJ provided supervision and feedback throughout the entire research process.

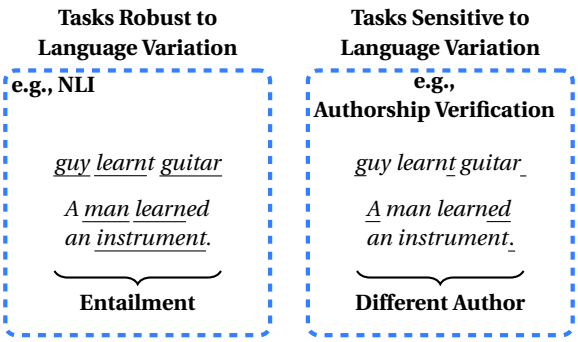


Figure 3.1: Different types of tasks might profit from different tokenizer settings. We investigate whether the same tokenizer performs equally well for tasks that require robustness to language variation (i.e., semantic-focused tasks like NLI) and for tasks that require sensitivity to language variation (i.e., form-focused tasks like authorship verification). Intuitively, one needs to look at semantic signals (e.g., "guitar" and "instrument") for NLI and at form- or style-based signals (e.g., "learned" and "learnt") for Authorship Verification.

II

more work on language variation and its relation to tokenizers and thus LLM performance.

3.1. Introduction

Variation in language is ubiquitous, manifesting in forms such as lexical variation (e.g., vocabulary choice like "lift" vs. "elevator"), spelling variation (e.g., spelling "u" instead of "you"), and syntactic variation (e.g., word order like "I gave her the ball." and "I gave the ball to her."). Such variation is often systematic rather than random, and is linked to regional, social, and contextual factors (Nguyen et al., 2016; Coupland, 2007; Eckert, 2012).

Tokenizers break up input strings and determine the tokens that are fed into language models. Tokenizers build their vocabulary based on a *fitting corpus* and have a given *vocabulary size*. A *pre-tokenizer* can prioritize or prevent the creation of certain tokens (cf. Section 3.3.2). Depending on these settings, tokenizers might behave differently for linguistic forms that are less common (see Section 3.3). For example, spelling variants of common words can be split into more tokens than their standard spelling (Matthews et al., 2024), e.g., "doing" might be one token, while "doin" might be split into "do" and "in". As a result, LLMs might need to learn to recompose such spellings to represent their meaning, suffering in performance compared to representing words as one token. In contrast, smaller tokens could be useful for recognizing form-based patterns (e.g., distinguish the "in" from the "ing" ending). We ask **RQ1**: *Do the same tokenizer settings (i.e., fitting corpus, vocabulary size and pre-tokenizer) perform well for two types of downstream tasks: Tasks whose gold labels are* ***robust to language variation*** *(e.g., tasks that involve analyzing the semantic meaning of texts like NLI) and tasks whose gold labels are* ***sensitive to language variation*** *(e.g., tasks centered around*

	Example	Label
GLUE	How do I hire an ethical hacker? Where can I find ethical hackers?	duplicate
+typo	Wher're cacn I fien ethical hackprs? Hjw fo I hie an ehtical hacker?	duplicate
+dialect	How do I hire me an ethical hacker? Where can I find me ethical hackers?	duplicate

(a) Tasks requiring robustness to language variation.

Dataset	Task
AV	Authorship Verification
PAN	Author Change Classification
CORE	Register Classification
NUCLE	Error Classification
Dialect	Dialect Classification

(b) Tasks requiring sensitivity to language variation.

Table 3.1: Evaluation Tasks. We evaluate tokenizers on tasks that require (a) robustness and (b) sensitivity to language variation. For (a), we use GLUE tasks, adding spelling perturbations using Wang et al. (2021) and dialect perturbations using Ziems et al. (2023). Examples are taken from QQP. For (b) we use a newly compiled selection of tasks containing our own authorship verification dataset, an author change classification dataset (Bevendorff et al., 2024), a register classification dataset (Laippala et al., 2023), a grammatical error correction dataset (Dahlmeier et al., 2013) and a dialect classification dataset generated with Ziems et al. (2023).

II

the form, style or language variety of texts like authorship verification and dialect classification)? For example, for semantic tasks like NLI, models should perform equally well, regardless of whether the text is written in Standard American English or African American English. In contrast, for form-based tasks like authorship verification (i.e., are two texts written by the same author?), models should be able to distinguish British and American spelling. Intuitively a model needs to use different signals (see Figure 3.1) to solve tasks requiring robustness and tasks requiring sensitivity to language variation, and might thus profit from different tokenizer settings. We test this by pretraining BERT base models (Devlin et al., 2019) with different tokenizer settings of the de-facto standard tokenizer: Byte-Pair Encoding or BPE (Sennrich et al., 2016).

The evaluation of tokenizers with fast proxy measures—usually based on how effectively a tokenizer compresses a reference corpus—offers a popular and cheaper alternative to training larger LMs for comparing different tokenizers. However, common proxy measures do not consistently achieve a high correlation with downstream performance (Zouhar et al., 2023; Cognetta et al., 2024; Schmidt et al., 2024). One reason could be that such measures are task-agnostic: They predict the same performance for a given downstream corpus, regardless of task or label differences (e.g., in Figure 3.1, NLI and authorship verification have different labels for the same sentence pair). We therefore investigate **RQ2**: *Can simple task-aware measures based on logistic regression better predict tokenizer downstream performance on variation-robust and variation-sensitive tasks than the common task-agnostic measures Rényi Efficiency and Corpus Token Count?*

Contributions We make the following three contributions: (1) We provide our code and a selection of English tasks (see Table 3.1) to evaluate models on tasks requiring sensitivity and robustness to language variation.[1] (2) We find that the impact of tokenizers on downstream performance varies depending on whether a task requires robust-

[1] https://github.com/nlpsoc/Tokenization-Language-Variation

ness or sensitivity to language variation, and that the pre-tokenizer has the biggest influence on downstream performance across task types. However, aggregated performance differences remain small and highlight the need for future work to investigate different types of language variation individually. (3) Further, we find that logistic regression performance has a higher correlation with BERT downstream performance than metrics like Corpus Token Count or Rényi efficiency (Zouhar et al., 2023). We provide practical suggestions to evaluate and build better tokenizers. With this work, we hope to encourage more work on language variation and its relation to tokenizers to build better, fairer and more robust LLMs from the ground up.

3.2. Related Work

Next to Byte-Pair Encoding (Sennrich et al., 2016), there exist several other subword (and other) tokenizers (Mielke et al., 2021). For example, character and byte-based tokenizers have been argued to be more robust to spelling variations (Mielke et al., 2021; Libovický et al., 2022; Xue et al., 2022). In this study we focus on BPE as it has been the most common tokenization algorithm in recent LLMs (e.g., LLAMA 3, MIXTRAL, DEEPSEEK V3 and GPT-4).

There is no universally agreed-upon standard to evaluate tokenizers. Tokenizers have been evaluated *intrinsically* (i.e., without training LLMs) and *extrinsically* (i.e., considering the performance of larger LMs pre-trained with the considered tokenizer). Common intrinsic methods include: the average number of subwords produced per word and correlated measures like Corpus Token Count (Rust et al., 2021; Scao et al., 2022; Ali et al., 2024; Gallé, 2019; Schmidt et al., 2024), information theoretic measures like Shannon Entropy or Rényi efficiency (Zouhar et al., 2023) and morphological alignment (Gow-Smith et al., 2022; Uzan et al., 2024). Extrinsic measures include LM perplexity (Shliazhko et al., 2024; Zevallos and Bel, 2023; Gowda and May, 2020), cross-entropy loss (Rajaraman et al., 2024), computational training cost (Ali et al., 2024) and downstream task performances (Schmidt et al., 2024; Ali et al., 2024). It is computationally infeasible to train SOTA LLMs end-to-end for each version of the tokenizer one wants to evaluate. Recent work tackled this by training "smaller" generative language models with 350M–2.5B parameters for each tokenizer (Schmidt et al., 2024; Ali et al., 2024). However, even such smaller models still can take several days to train. We measure downstream task performance on models with 110M parameters taking less than 15 GPU hours to train per model.

Tokenizer algorithms and settings can affect a LLM's performance, for example, on tasks including numbers like arithmetic (Thawani et al., 2021; Wallace et al., 2019), on tasks including domain specific vocabulary or jargon like coding or medicine (Gu et al., 2021; Dehaerne et al., 2022; Zan et al., 2023), on different scripts and languages (Petrov et al., 2023; Rust et al., 2021; Limisiewicz et al., 2023; Ahia et al., 2023; Velayuthan and Sarveswaran, 2025) and when translating between languages (Gallé, 2019; Libovický et al., 2022; Zhang et al., 2022). To the best of our knowledge monolingual tokenizers have been underinvestigated in relation to language variation. Monolingual tokenizers and tokenizer settings have recently been investigated on broader selections of NLU

tasks (Schmidt et al., 2024; Ali et al., 2024), presumably with the underlying assumption that for a given language like English, there exists a best tokenizer for most, if not all, tasks. We investigate this assumption for two systematically different types of tasks: tasks requiring robustness and tasks requiring sensitivity to language variation.

3.3. Tokenizer Settings

Tokenizers break up texts into smaller units. These units are fed into the language model as input. We investigate different variations of the most popular tokenization algorithm: Byte-Pair Encoding or BPE (Sennrich et al., 2016). BPE is a subword tokenization algorithm (i.e., it breaks rare words down into subwords, bottoming out in bytes in the worst case[2]). The vocabulary is built iteratively: It starts out with a base vocabulary of distinct bytes and merges them based on token frequency in the fitting corpus until the desired vocabulary size is reached. We vary the algorithm on three parameters: (1) the vocabulary size, (2) the pre-tokenizer and (3) the fitting corpus.

3.3.1. Vocabulary Size

Common vocabulary sizes range from 30k to 64k in monolingual models, to about 128k (LLAMA 3) and 200k (GPT-4O) in recent multilingual models. To the best of our knowledge, previous work on tokenizer vocabulary size mostly tested mid-range vocabulary sizes between 32k and 100k on NLU tasks, finding only small differences in performance between vocabulary sizes (Ali et al., 2024; Schmidt et al., 2024). However, vocabulary size might still play a role when dealing with language variation. When using mid-range vocabulary sizes, less common words or spelling variants might not be represented with one token anymore, influencing a word's position in the embedding space (Matthews et al., 2024). Character-based models like ByT5 show better robustness to spelling variation than subword models with larger vocabulary sizes (Libovický et al., 2022; Xue et al., 2022; Tay et al., 2022). Similarly, tokenizers using low-range (in the extreme case character-level) vocabulary sizes might be more robust to spelling variation, as, for example, swapping of characters will not drastically alter the character-based segmentation of a word. We experiment with the following vocabulary sizes: 500 and 4k (low-range vocabulary size), 32k and 64k (mid-range vocabulary size) and 128k (high-range vocabulary size).

3.3.2. Pre-Tokenizer

Pre-tokenizers split the fitting corpus into word-like units (Mielke et al., 2021), so-called "pre-tokens" (e.g., "I'm fine." → "I", "'m", "fine", ".") which are then further tokenized independently. That is, in building its vocabulary, the BPE algorithm cannot merge tokens that cross pre-token boundaries (e.g., "I'm"). Pre-tokenizers are usually expressed with regular expressions that include hard-coded knowledge about a language or script (Velayuthan and Sarveswaran, 2025), e.g., whitespaces separate words in Latin script and "'m" is a contraction in English. The most common English

[2] adapted from "rare words are broken down into a collection of subword units, bottoming out in characters in the worst case." in Bostrom and Durrett (2020)

pre-tokenizers are more or less elaborate variations of splitting texts based on whitespaces. For example, GPT-2's pre-tokenizer additionally separates different *Unicode Character Categories* (e.g., letters, numbers and punctuation) but leaves single leading whitespaces attached to words (Radford et al., 2019).

Pre-tokenizers influence compression effectiveness (Radford et al., 2019), NLU downstream (Schmidt et al., 2024) and arithmetic performance.[3] However, pre-tokenizers have not been investigated for tasks that require sensitivity to language variation. Pre-tokenizers might play a role for words that merge Unicode Character Categories, for example, when letters are substituted with numbers (e.g., "2day" or "c000l", see Eger et al. (2019) for more). If a pre-tokenizer generally splits off numbers and letters, such words can never become part of the vocabulary. Further, not splitting on whitespace might better represent syntactic variation by including frequent phrases like "of the". Frequent compositional phrases are processed faster by humans (Arnon and Snider, 2010), and building tokenizers that align closer to such cognitive processes (e.g., Yang et al., 2020) might also be beneficial for semantic tasks.

We compare five different pre-tokenizers: (a) not using a pre-tokenizer (NO), (b) isolating whitespaces (WS), (c) leaving single whitespaces attached to words (_WS), (d) LLAMA 3's pre-tokenizer (LLAMA3) and (e) GPT-2's widely used pre-tokenizer (GPT2). LLAMA3 and GPT2 can be understood as extensions of the _WS pre-tokenizer. Among others, the LLAMA3 pre-tokenizer splits off Unicode Character Categories (e.g., punctuation) but leaves one leading non-letter character attached to letters (e.g., 'm). The GPT2 pre-tokenizer is similar to LLAMA3, but separates Unicode Character Categories more. See Appendix A.1.2 for the regular expressions describing the WS, _WS, LLAMA3 and GPT2 pre-tokenizer.

3.3.3. Fitting Corpus

Sub-word tokenizers construct their vocabulary based on a fitting corpus, adding tokens based on the distribution of tokens in that corpus. Lexical, morphological and spelling variation are all intuitively connected to the fitting corpus. If a fitting corpus does not show a variation, it can never be part of a single token in the vocabulary.

We sampled fitting corpora with a size of approximately 1.5 billion words. **PubMed** was randomly sampled from The Pile's PubMed Abstracts (Gao et al., 2020). The PubMed Abstracts consist of 30M abstracts from biomedical articles. **Wikipedia** was randomly sampled from Wikipedia articles from a snapshot from June 1st, 2023 after the plain text of articles was extracted.[4] **Twitter** was sampled from the Decahose Twitter stream[5] throughout the year 2021, with queries on almost every day of the year 2021. We only select English tweets based on Twitter's internal language identification system. We exclude retweets. **Miscellaneous** was sampled from a variety of domains with no overlap with the other fitting corpora. It includes Reddit, literature sources (fanfictions and books), news articles and comments, question answering websites, reviews, mails,

[3] right-to-left pre-tokenization of numbers improved arithmetic for LLAMA 3, see bluesky.
[4] using `https://github.com/LorcanJConnolly/WikiTextExtractor`.
[5] Decahose provided access to 10% of real time tweets sampled by Twitter.

transcripts, blogs, Common Crawl, scientific articles, code and mathematical formulas. See details in Appendix A.1.1.

We expect the fitting corpora to differ in the lexical, syntactic and lexical variation they exhibit. For example, PubMed probably contains less spelling and lexical variation than Twitter. Intuitively, a tokenizer constructed on PubMed should thus be less capable in representing stylistic variation than one constructed on Twitter. However, it remains unclear how important the fitting corpus composition is. Zhang et al. (2022) investigate different compositions of the fitting corpus in the multilingual setting. They find a surprising robustness to language imbalance in the fitting corpus for languages sharing the same script.

3.4. Evaluation Tasks

We compare tokenizers on classification tasks that require robustness to language variation (Section 3.4.1) and tasks that require sensitivity to language variation (Section 3.4.2). Models solving these two types of tasks should need to make use of more semantic and form-based signals respectively and might have different requirements for a tokenizer (cf. Figure 3.1). We select tasks that strike a balance between being sufficiently challenging and staying within the capabilities of our pre-trained BERT models. See an overview in Table 3.1.

3.4.1. Tasks Robust to Language Variation

First, we evaluate tokenizers on tasks where the gold label is robust to language variation. We use GLUE (Wang et al., 2018), a standard NLP benchmark that was also used to evaluate BERT at its introduction Devlin et al. (2019), and is within the capabilities of our pre-trained BERT models. We compare tokenizers on the following four GLUE tasks: SST-2 (sentiment classification), QQP (paraphrase classification), MNLI and QNLI (NLI tasks). For details on our task selection see Appendix A.2.1. Ideally, a robust tokenizer performs consistently across all versions of **GLUE**: the original, primarily written in Standard American English, the spelling-transformed **GLUE+typo** and the dialect-transformed **GLUE+dialect**.

GLUE+typo We use `textflint` (Wang et al., 2021) to introduce simulated typos and spelling errors to our tasks, similar to Libovický et al. (2022). It uses random character swapping and a list of common spelling errors.

GLUE+dialect We use `Multi-VALUE` (Ziems et al., 2023) to introduce simulated dialectal variation to our GLUE tasks. `Multi-VALUE` makes use of 189 dialectal perturbation rules (Ziems et al., 2023) operationalized based off of eWAVE (Kortmann et al., 2020). For each example in the GLUE tasks, we randomly choose between transformations to Appalachian English, Chicano English, Colloquial Singapore English, Indian English and Urban African American English dialects.

3.4.2. Tasks Sensitive to Language Variation

Second, we evaluate tokenizers on tasks that are sensitive to language variation. The gold label for such tasks should be sensitive to stylistic or form-based signals.

AV Models performing **a**uthorship **v**erification (i.e., are two texts were written by the same author?) usually need to be sensitive to different styles and forms used by different authors (Zhu and Jurgens, 2021; Wegmann et al., 2022; Wang et al., 2023a). Past work has found authorship verification to be sensitive to tokenization, with significant gaps between the BERT and RoBERTa tokenizers (Zhu and Jurgens, 2021). Therefore, we curate a new authorship verification dataset of 40.8k train, 2.5k dev and 4.8k test pairs of texts from different domains, similar in distribution to the Miscellaneous corpus (cf. Section 3.3). Labels are balanced in the test set. See details in Appendix A.2.2.

PAN We use the PAN 2024 Multi-Author Writing Style Analysis task to predict whether an author shift occurs between two consecutive paragraphs extracted from Reddit (Bevendorff et al., 2024). Specifically, we use the 'hard' task, where paragraphs are about the same topic. We sample such that labels are balanced. This results in a training set of 18k and a dev set of 4k instances.

NUCLE We use the NUCLE 3.3 corpus (Dahlmeier et al., 2013) for multi-label classification of the errors that were made by English learners in a given sentence. NUCLE was annotated by professional English instructors for 27 error types (e.g., verb tense or article use). It includes 22k unique sentences with errors. We add a sample of 5k error-free sentences from the same dataset. We split it into a train (80%) and dev (10%) set.

CORE We use the **C**orpus of **O**nline **R**egisters of **E**nglish (Laippala et al., 2023) for register classification. Register is one of the most important factors associated with linguistic variation (Biber, 2012). We use 8 main register labels (e.g., spoken or informational description) for multi-class prediction. To increase the occurence of rarer labels, we split long texts and reach a train size of 30k and a dev size of 5k. See Appendix A.2.2 for details.

Dialect We randomly sample 60k instances from GLUE-dialect (Section 3.4.1) and the original GLUE task, to create a dialect classification task for five dialects and Standard American English in the original GLUE texts. We use 50k texts for the train and 5k for the dev set.

3.5. Modeling

For each investigated tokenizer (cf. Section 3.3), we pre-train three BERT models with different seeds. We use encoder instead of decoder models, as encoder models tend

to reach higher performance for classification tasks for low parameter settings. This allows us to train models using fewer GPU hours.

Pretraining BERT models We experiment with pre-training tiny BERT models using a token count T close to 3300M that is exponentially bigger than the 4.6M parameters P, similar to the ratio in the original BERT papers Devlin et al. (2019); Turc et al. (2019). However, we find that for the same compute, using a bigger model size P and less tokens T improves the training loss. Chinchilla's scaling law might also hold for smaller encoder models, specifically optimal parameter count could scale with the token size for a fixed compute $P_{OPT} \approx T^{23/27}$ (Hoffmann et al., 2024). For the remainder of this work, we use the base BERT model architecture with 110M parameters, initialize all weights randomly and pre-train on 750M tokens sampled in sequences of 512 from the Miscellaneous corpus (Section 3.3.3) and use a batch size of 32 and 45k steps. For further details and hyperparameters, see Appendix A.3.

Fine-tuning BERT Unless otherwise specified, we use 3 epochs, a max sequence length of 128, a batch size of 32 and a learning rate of 2e-5 to perform the classification tasks. We evaluate on the dev set for GLUE. For comparability, we use the same setup for tasks requiring sensitivity to language variation. Only for the authorship verification task we use a contrastive training setup, then use the dev set to find an optimal cosine similarity threshold and calculate accuracy on the test set.

3.6. Results

We show the performance of fine-tuned BERT models on tasks that require robustness to language variation in Table 3.2 and tasks that require sensitivity to language variation in Table 3.3. When we investigate a specific setting (e.g., the fitting corpus in first three rows in Table 3.3), we only change that setting and leave the other at their "default settings" to ensure comparability and isolate the effect of each individual setting without exhaustive testing of all possible combinations. We use the following default values for the three settings: the miscellaneous fitting corpus, the GPT2 pre-tokenizer and a vocabulary size of 32k.

Note that the performance differences averaged over three seeds tend to be relatively small, which is consistent with previous work comparing different tokenizer algorithms on downstream tasks (Ali et al., 2024; Schmidt et al., 2024). To ensure significance, we compute the pairwise McNemar, 1947's test for the pre-tokenizer, fitting corpus and vocabulary size settings, see Figure 3.2. For the significance testing, we consider classifications by models with the same settings but different seeds to be stemming from the same rater. We use the Bonferroni correction (Dunn, 1961) for our total of 26 tests.

RQ1: Tokenizer settings perform differently on tasks robust and sensitive to language variation Overall, tasks sensitive to language variation profit more from tokenizers that encode more variation through a larger vocabulary size (§3.6.3).

	Model	orig	+typo	+dialect	AVG
Fitting Corpus	PubMed	80.8 ± 0.0	**69.1** ± 0.2	78.6 ± 0.2	*76.2* ± 0.0
	Wikipedia	80.7 ± 0.3	68.6 ± 0.2	**79.3** ± 0.2	*76.2* ± 0.2
	Twitter	**81.1** ± 0.0	69.1 ± 0.5	78.8 ± 0.1	**76.4** ± 0.2
Pre-Tokenizer	NO	72.1 ± 1.0	61.6 ± 0.1	70.1 ± 0.3	67.9 ± 0.3
	WS	80.8 ± 0.4	68.2 ± 0.3	**79.3** ± 0.3	*76.1* ± 0.3
	_WS	80.8 ± 0.3	**68.9** ± 0.1	79.0 ± 0.2	**76.2** ± 0.1
	LLAMA3	80.9 ± 0.1	68.2 ± 0.2	79.0 ± 0.0	*76.1* ± 0.0
	GPT2	**81.3** ± 0.4	68.2 ± 0.2	79.2 ± 0.4	**76.2** ± 0.3
Vocabulary Size	500	77.2 ± 2.5	**70.3** ± 2.6	75.6 ± 2.0	74.4 ± 2.4
	4k	80.5 ± 0.8	**70.3** ± 0.9	78.6 ± 0.8	**76.4** ± 0.8
	32k	**81.3** ± 0.4	68.2 ± 0.2	**79.2** ± 0.4	76.2 ± 0.3
	64k	80.8 ± 0.4	67.6 ± 0.6	**79.2** ± 0.2	75.9 ± 0.4
	128k	78.7 ± 2.0	64.6 ± 1.9	76.1 ± 2.6	73.1 ± 2.2

Table 3.2: Performance on tasks requiring robustness to language variation. We display BERT performances, averaged on the original four GLUE tasks and their perturbations using spelling mistakes (+typo) and dialectal transformations (+dialect).We provide the mean and standard deviation (±) over three seeds respectively. We boldface the best performances for each column and investigated setting. For the averaging column (AVG), italics indicate tokenizers with no statistically significant difference from the best-performing tokenizer (cf. Figure 3.2).

Note that the best-performing tokenizer settings are not always consistent across the individual variation-robust and variation-sensitive tasks (e.g., vocabulary size for GLUE+typo and GLUE in Table 3.2). We suspect that this is due to differences in the types of language variation present in the specific task datasets (e.g., character-level tokens seem to be more robust to spelling variation). Future work could investigate different types of language variation individually (e.g., spelling vs. lexical variation).

3.6.1. Pre-Tokenizers

Pre-tokenizers have the greatest influence on performance For both task types, the range of performance values is largest for pre-tokenizers, second largest for vocabulary size and smallest for fitting corpus. This is surprising as, to the best of our knowledge, pre-tokenizers have received the least attention in previous work. For both task types, using pre-tokenizers improves performance over using no pre-tokenizer.

Pre-tokenizer performance differs more between individual tasks than between task types. For both task types, _WS is among the best performing pre-tokenizers. Combining leading whitespaces with letters (_WS) generally improved performance over isolating whitespaces (WS). For dialect and spelling-transformed GLUE, CORE, NUCLE and Dialect, the whitespace-based pre-tokenizers _WS and WS perform the best. For AV, PAN and the original GLUE tasks, GPT2 and LLAMA3 perform better. Tasks like grammatical error detection (NUCLE) could be seen as identifying deviations from a standard, and might benefit from pre-tokenizers that include tokens with typical combinations of Unicode Character Categories. One of their main differences of GPT2 and LLAMA3 the whitespace-based pre-tokenizers is that the combination of different Unicode Character Categories (e.g., punctuation and letters) within the same token is

	Model	AV (acc)	PAN (acc)	CORE (acc)	NUCLE (F1)	DIALECT (F1)	AVG
Corpus	PubMed	81.7 ± 0.1	65.2 ± 0.5	55.9 ± 0.6	21.8 ± 1.2	87.9 ± 0.1	62.5 ± 0.4
	Wiki	81.9 ± 0.1	65.5 ± 0.5	55.5 ± 0.6	**23.5** ± 0.2	**88.9** ± 0.5	63.1 ± 0.2
	Twitter	**82.9** ± 0.6	**66.7** ± 0.4	**56.5** ± 0.6	21.4 ± 0.9	88.3 ± 0.2	**63.2** ± 0.2
Pre-Tok.	NO	81.8 ± 0.3	59.9 ± 1.0	51.7 ± 0.2	16.3 ± 0.2	77.3 ± 0.2	57.4 ± 0.2
	WS	75.5 ± 0.9	66.1 ± 0.4	55.1 ± 0.9	**23.0** ± 0.1	**88.8** ± 0.2	61.7 ± 0.3
	_WS	82.5 ± 0.3	66.3 ± 1.6	**56.6** ± 0.7	22.6 ± 0.5	88.4 ± 0.2	**63.3** ± 0.3
	LLAMA3	82.5 ± 0.2	**66.9** ± 0.6	56.5 ± 0.6	21.1 ± 1.2	88.6 ± 0.0	63.1 ± 0.3
	GPT2	**82.6** ± 0.5	66.6 ± 1.2	56.3 ± 1.2	21.8 ± 0.9	88.4 ± 0.6	63.1 ± 0.4
Size	500	78.2 ± 0.9	62.6 ± 0.2	51.1 ± 0.6	13.1 ± 0.9	85.6 ± 0.4	58.1 ± 0.6
	4k	81.9 ± 0.1	**67.8** ± 0.6	55.1 ± 1.0	17.4 ± 2.8	87.9 ± 0.7	62.0 ± 0.6
	32k	82.6 ± 0.5	66.6 ± 1.2	**56.3** ± 1.2	21.8 ± 0.9	**88.4** ± 0.6	**63.1** ± 0.4
	64k	**82.7** ± 0.4	67.2 ± 0.7	54.9 ± 1.5	**22.0** ± 1.1	88.1 ± 0.6	63.0 ± 0.6
	128k	80.1 ± 2.2	62.4 ± 3.4	51.2 ± 2.2	19.0 ± 2.7	81.5 ± 5.1	58.9 ± 3.0

Table 3.3: Performance on tasks requiring sensitivity to language variation. We display BERT performances on our Authorship Verification (AV), the PAN, the CORE, and the multi-Dialect dataset. We provide the mean and standard deviation (±) over three seeds respectively. For the averaging column (AVG), italics indicate tokenizers with no statistically significant difference from the best-performing tokenizer (cf. Figure 3.2).

less often allowed. By preventing combinations of categories like punctuation and letters, the vocabulary of GPT2 can include a broader range of tokens that only consist of letters, e.g., "_queer" in Table 3.4. This could explain the better average performance of GPT2 on the original GLUE task.In contrast, LLAMA3 also allows the mixing of one initial punctuation mark with letters, e.g., including "'all" in Table 3.4, seems to especially help LLAMA3 solve PAN. Future work could investigate different individual tasks and the influence of pre-tokenizers individually.

3.6.2. Fitting Corpus

On tasks requiring sensitivity to language variation, Twitter performs best This aligns with our expectation that the Twitter corpus include more spelling variation and a larger set of language varieties than PubMed and Wikipedia. Interestingly, Wikipedia performs the best on NUCLE and DIALECT. Grammatical error detection (NUCLE) could be seen as identifying deviations from a standard, and might benefit from training on corpora with few grammatical errors—like Wikipedia.

On tasks requiring robustness to language variation, Twitter performs surprisingly well Interestingly, Twitter performs indistinguishably from other fitting corpora on tasks that require robustness to language variation. Originally, we expected a more standardized corpus like Wikipedia to perform better, as it should lead to less spelling variation and thus more "full words" in the vocabulary (e.g., "precursor" in Table 3.4).

3.6.3. Vocabulary Size

A larger vocabulary size might be useful for tasks requiring sensitivity to language variation Performance peaked at 4k for tasks requiring robustness to language vari-

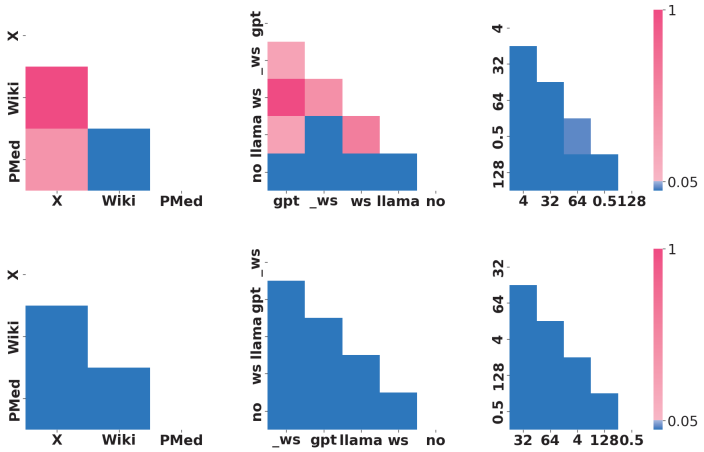


Figure 3.2: Pairwise significance testing of models. We use pairwise McNemar (1947)'s test to test whether there is a significant difference between classifications done by models trained with different tokenizers on the tasks requiring robustness to language variation (first row) and tasks requiring sensitivity to language variation (second row). Tokenizers are sorted by mean performance. Blue colors show statistical significance, while pink colors are above the 0.05 threshold.

ation and at 32k for tasks requiring sensitivity to language variation. It seems that a smaller vocabulary size is sufficient to learn common standard forms and more tokens are needed to include style and form variations. For example, "jumper" is only included with vocabulary size 64k, see Table 3.4. Note that optimal vocabulary sizes might scale with model size (Tao et al., 2024). For both task types, the standard deviation is highest for the smallest (500) and the largest vocabulary size (128k).

Smaller vocabulary is more robust to spelling variation The vocabulary size of 500 consistently performs worse than larger vocabulary sizes. It seems that BERT struggles with long input sequences and learning to compose words for small vocabulary sizes. An exception is the GLUE+typo task, where the tokenizer of size 500 performs best. This is consistent with character-based tokenizers being more robust to spelling variation (cf. Section 3.3.1).

3.7. Pre-evaluating Downstream Tokenizer Impact

A popular alternative to training several LLMs to evaluate tokenizers is using fast *intrinsic* measures on the token distribution on the downstream task corpus. However, common intrinsic measures are not consistently correlated with downstream performance (Rust et al., 2021; Zouhar et al., 2023; Schmidt et al., 2024). Further, they do not make use of task labels and can thus be considered *task agnostic*. Imagine a corpus on which you want to perform both, a task requiring robustness to language variation and a task requiring sensitivity to language variation (cf. Figure 3.1). Task-agnostic in-

Group	Choice	Examples of tokens
Fitting Corpus	Wiki	_queer _precursor _l ma o
	Twitter	_queer _prec urs or _lmao
	PubMed	_qu e er _precursor _l ma o
Pre-Tokenizer	NO	_ y'all_ que er_ b 4
	WS	_ y'all _ que er _ b4
	_WS	_y'all _que er _b 4
	LLAMA3	_y 'all _que er _b 4
	GPT2	_y ' all _queer _b 4
Vocabulary Size	500	_t e h _l g b t q i a + _j um p er
	4k	_te h _l g bt q ia + _j um per
	32k	_te h _l g bt q ia + _j umper
	64k	_teh _lg bt q ia + _jumper
	128k	_teh _lgbt q ia + _jumper

Table 3.4: Examples of Vocabulary Differences. We display examples of sequences tokenized by tokenizers with varying fitting corpora, pre-tokenizers and vocabulary size. We represent whitespaces within tokens as _. The domain of the fitting corpus affects the set of unique words in the vocabulary (e.g., "queer" is part of Twitter but not PubMed). The size determines the number of "rarer" words in the fitting corpus that are added to the vocabulary (e.g., the British variant "jumper" of "sweater" is only added with size 64k). The pre-tokenizer determines what types of words can be part of the vocabulary (e.g., "b4" can not be part of GPT2 and LLAMA3). Note that allowing less Unicode Character Category combinations (e.g., numbers and letters) increases the broadness of the vocabulary (e.g., GPT2 allows for less combinations than other pre-tokenizers and is the only one including "queer" in its vocabulary).

trinsic measures like Rényi efficiency and Corpus Token Count will always predict the same performance for both tasks. However, we found that the type of task can have an influence on what tokenizer settings work better (cf. Section 3.6). We experiment with a task-dependent measure: the performance[6] of logistic regression with task labels as the dependent variable.

Logistic regression We tokenize the task texts. Then, we use the resulting tokens as features for logistic regression. We use a bag-of-tokens approach and do not consider word order. We do not consider frequency but use binary information on token presence. We use Cartesian products of tokens for NLI tasks. Specifically, given two sentences in an NLI setting, we generate all possible combinations of individual tokens from both sentences and use these as features. For PAN and Authorship Verification, we limit the Cartesian product to tokens that appear in both sentences and tokens that only appear in one sentence.[7] For multi-classification and multi-label task we train separate one-vs-rest logistic regression models. We use l1-regularization with C=0.4. Note that one could probably increase logistic regression performance by tuning parameters and features more specifically to the investigated tasks.

[6] We use the same F1 and accuracy performance metrics as for BERT.
[7] PAN and AV had considerably longer texts than the GLUE NLI tasks leading to otherwise too many features.

Measure	Robust	Sensitive
Rényi Efficiency	-.22	-.03
Corpus Token Count	-.45	.37
logistic regression	.85	.84

Table 3.5: Correlation of Intrinsic Measures with BERT Performances. Correlations between the Rényi Efficiency ($\alpha = 2.5$), Corpus Token Count, and logistic regression predictions with the performance of the finetuned BERT models on the tasks robust to language variation and the tasks sensitive to language variation respectively. Logistic regression has the highest correlation values across all tasks. A higher Corpus Token Count is negatively correlated with performance on tasks robust to language variation and positively correlated with tasks sensitive to language variation.

3.7.1. Results

We display the Pearson correlation of Rényi efficiency, Corpus Token Count and logistic regression performances with the fine-tuned BERT performances in Table 3.5. Similar to Schmidt et al. (2024) we find a light negative correlation of LM performances with Rényi efficiency on NLU tasks.

Corpus token count has a negative correlation for tasks requiring robustness and a positive correlation for tasks requiring sensitivity to language variation. Corpus Token Count and correlated measures are often used to assess the compression effectiveness of a tokenizer on a reference corpus or downstream task. A higher value entails more tokens in the corpus and thus less effective compression. More effective compression is commonly believed to be a sign of a better tokenizer, and has often been thought to be correlated with better downstream performance (Rust et al., 2021; Ali et al., 2024; Velayuthan and Sarveswaran, 2025). We show that the correlation is flipped for tasks that are robust and sensitive to language variation highlighting the difference in tokenizer requirements for the two task types.

RQ2: Logistic regression correlates with downstream performance Among the three considered measures, we find the highest correlation between logistic regression performances and BERT performances. Additionally, logistic regression is similarly correlated for both task types and can compare tokenizers of varying vocabulary sizes, which is not possible with Rényi efficiency and Corpus Token Count (for details refer to Appendix A.4). Note that we by no means question the usefulness of measures like Corpus Token Count or Rényi efficiency for assessing different tokenizers and what they do. However, they might be less suited to estimate downstream performance without additional modifications.

3.8. Conclusion

In this work, we investigated tokenizer settings for tasks that require robustness and tasks that require sensitivity to language variation. BPE settings perform differently on

the two task types. We make three practical suggestions for selecting tokenizer settings: (1) Pay the most attention to the pre-tokenizer. It influences how Unicode Character Categories can be combined (e.g., "b4" in Table 3.4), and what words can ultimately be part of the vocabulary (e.g., "_queer" in Table 3.4). (2) Choose a bigger vocabulary size for settings that require sensitivity to language variation. (3) Use a small machine learning classifier—e.g., a logistic regression classifier—to evaluate how different tokenizers affect performance on tasks in your domain. For example, for general-purpose language models, pick a tokenizer that this model consistently predicts to perform well across both variation-robust and variation-sensitive tasks. Tokenizer settings seem to perform differently on tasks that are robust and tasks that are sensitive to language variation. But tokenizers might also be sensitive to what types of language variation (e.g., lexical vs. syntactic) are present. We think it is crucial to investigate different types of language variation individually in future work to ultimately build better, fairer and more robust LLMs.

Limitations

Note that our taxonomy of tasks relies on splitting tasks into semantic-focused tasks (i.e., considering what is said) and form-based tasks (i.e., considering how it is said). However, a strict distinction is challenging since most tasks are best solved using a mixture of typical content and typical form signals. For example, recognizing the lengthening of words (e.g., "cooool") can be helpful for the semantic task of sentiment classification (Brody and Diakopoulos, 2011) and content information (e.g., the topic) can improve authorship verification performance (Wegmann et al., 2022). Still, this distinction clarifies the main signals that our different task types should rely on and enables us to compare tokenizer settings on these two different types of tasks.

We evaluate tokenizers on 110M parameter encoder models as encoder models tend to reach higher performance for classification tasks in low parameter settings. However, we risk not accounting for emergent capabilities of popular larger generative language models. For example, the number of context tokens for recent models are generally much larger than the original 512 tokens of BERT base.

We find that logistic regression correlates with downstream performance on classification tasks. For more complex tasks like regression or generation tasks, logistic regression might not be applicable. However, we still see it as a great benefit that logistic regression can enable us to quickly test tokenizers on a variety of different tasks and make a more informed decision on what tokenizers to test more rigorously.

For tasks requiring robustness to language variation, many more tasks could have been included. For example, future work could include more challenging tasks like challenge NLI datasets, e.g., ANLI (Nie et al., 2020a). Further, future work could investigate different types of tasks such as question answering or arithmetic tasks. For tasks requiring sensitivity to language variation, we originally planned to include more classification tasks that are tied to well-established factors of language variation, e.g., age prediction and geographic location prediction (Nguyen et al., 2016). However, we repeatedly encountered difficulties with cleanly separating semantic from form-based

cues. For age prediction and location prediction specifically, logistic regression models made extensive use of content words. Based on the coefficients of our logistic regression models, we excluded tasks that we expected to mainly require sensitivity to language variation, but where the models primarily relied on content words.

We varied one tokenizer setting at a time while leaving the other two on their "default values" (i.e., miscellaneous fitting corpus, GPT2 pre-tokenizer and vocabulary size of 32k), in order to isolate the effect of each individual setting without exhaustive testing of all possible combinations. Future work could explore interdependencies between tokenizer settings through more comprehensive testing.

The tokenizer settings seem to have different effects when the pre-training corpus and fitting corpus differ. The pre-training corpus might influence in how far a model can leverage the diversity encoded in the tokenizer. See Appendix A.5 for further experiments. Note, however, that recent work suggests to only use vocabulary that appears in the training corpus (Land and Bartolo, 2024)—which is unlikely when different corpora are used for model training and tokenizer fitting.

Ethical Considerations

Our pre-training and fitting corpora are largely based on publicly shared and accessible datasets from popular online forums and web pages (e.g., Wikipedia, Reddit, Common Crawl, ...). Unfortunately, these datasets were mostly collected without explicit consent from users and might lead to (among others) privacy concerns. Individuals might be identifiable from their written texts. However, we hope that the risks of reusing already published datasets are minimal. We collected tweets from Twitter in 2021 using Twitter's official API, but opt to not share them publicly. While we aim to include different language varieties in our datasets, they might not be representative of English language use across different social groups. For example, we expect a skew towards straight, white, American, young and male authors. We caution against using our datasets and tasks for general claims about broad selection of language varieties. We confirm to have read and that we abide by the ACL Code of Ethics. Beside the mentioned ethical considerations, we do not foresee immediate risks of our work.

Acknowledgements

We thank the anonymous ARR reviewers for their constructive comments. We thank the colleagues in the various NLP Groups at Utrecht University and the Blablablab at the University of Michigan and, specifically, Kees van Deemter, Antal van den Bosch, Yupei Du, Qixiang Fang, Melody Sepahpour-Fard, Shane Kaszefski Yaschuk, Elize Herrewijnen, Sanne Hoeken, Hugh Mee Wong, Jian Zhu and Pablo Mosteiro for, among others, feedback on writing and presentation. We thank colleagues at EMNLP 2024 for their insights into tokenization, and, specifically, Vilém Zouhar. We thank Laura Wegmann for feedback on writing. This research is supported by the "Digital Society - The Informed Citizen" research programme, which is (partly) financed by the Dutch Research Council (NWO), project 410.19.007. This research is supported in part by

the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Part III

Building Neural Style Representations

So far, I have defined NLP tasks requiring robustness (e.g., paraphrase classification) and NLP tasks requiring sensitivity to language variation (e.g., authorship verification). I showed that language variation might be important to consider at all stages of building LLMs—including tokenizers. In Part III of this dissertation, I focus on a task requiring sensitivity to language variation: representing the linguistic style of a text. Specifically, in Chapter 4, I introduce a linguistically informed evaluation framework to evaluate in how far text representations represent linguistic style. I find that neural representations outperform vanilla feature-based representations like character 3-grams and function word frequencies in their sensitivity to well-established dimensions of style. In Chapter 5, I use contrastive learning to train style representations. I demonstrate that hard negatives are crucial to train text representations that are independent from referential meaning. Beyond the technical contributions, this part of my dissertation underscores the broader need for more sensitivity to language variation in NLP models.

Contents

4

Does It Capture STEL? A Modular, Similarity-based Linguistic Style Evaluation Framework

In this first chapter of Part III on Building Neural Style Representations, I discuss methods to evaluate text representations in their capacity to represent "linguistic style"—a prerequisite to building better style representations. This chapter is based on Wegmann, A., & Nguyen, D. (2021). Does It Capture STEL? A Modular, Similarity-based Linguistic Style Evaluation Framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7109-7130). `https://doi.org/10.18653/v1/2021.emnlp-main.569`. See a video of the conference presentation here: `https://youtu.be/WPbxyOrDK6w`. Some results were updated compared to the original publication, most notably the RoBERTa results on STEL.

Abstract

Style is an integral part of natural language. However, methods for evaluating a language model's ability to represent style are rare, often application-specific and usually do not control for content. We propose the modular, fine-grained and content-controlled *similarity-based STyle EvaLuation framework* (STEL) to evaluate any model or method that can calculate the similarity of two texts in terms of their linguistic style. We illustrate STEL with two general *dimensions* of style (formal vs. informal style and simple vs. complex style) as well as two specific *features* of style (contrac'tion and numb3r substitution). We find that BERT-based methods outperform vanilla versions of commonly used style representations like character 3-grams, punctuation frequency and LIWC-based approaches. We invite the

Author contributions: AW developed the idea, prepared the data, implemented the experiments, and wrote the manuscript. DN provided supervision and feedback throughout the entire research process.

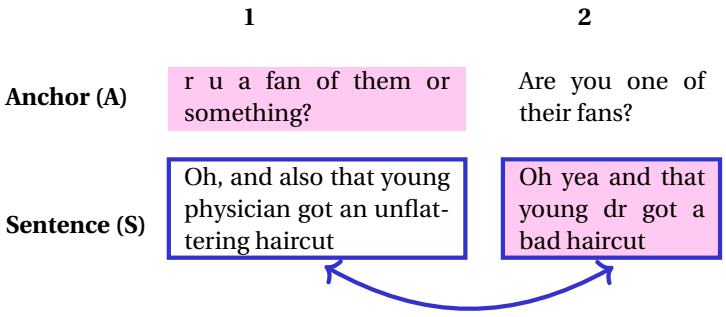


Figure 4.1: STEL Task Instance. Anchor 1 (A1) and anchor 2 (A2) and the alternative sentences 1 (S1) and 2 (S2) are split along the same style dimension (here: formal/informal). The pink text is less formal than the **black** text. The sentences and anchors are paraphrases of each other. The STEL task is to order S1 and S2 to match the order A1-A2 (here: informal, then formal). Here, the correct order is S2-S1, so the reverse of what is displayed.

III

addition of further tasks and task instances to STEL and hope to facilitate the improvement of style-sensitive representations.

4.1. Introduction

Natural language is not only about what is said (i.e., content), but also about how it is said (i.e., *linguistic style*). Linguistic style and social context are often considered to be interrelated (e.g., Coupland, 2007; Bell, 2014). For example, people can accommodate their linguistic style to each other based on social power differences (Danescu-Niculescu-Mizil et al., 2012). Furthermore, linguistic style can affect a reader as it can, e.g., change the persuasiveness of arguments (El Baff et al., 2020) or the success of pitches on crowdsourcing platforms (Parhankangas and Renko, 2017). As a result, style can be important to address in natural language generation (Ficler and Goldberg, 2017), including identity modeling in dialogue systems (Li et al., 2016), style preservation in machine translation (Niu et al., 2017; Rabinovich et al., 2017) and text generation from structured data (van der Sluis and Mellish, 2009). Further, style is relevant for natural language understanding, e.g., in author profiling (Rao et al., 2010), abuse detection (Markov et al., 2021) or understanding conversational interactions (Danescu-Niculescu-Mizil and Lee, 2011).

In NLP, there are several evaluation benchmarks—datasets and tasks used to assess model performance on specific objectives like named entity recognition (e.g., CoNLL in Tjong Kim Sang and De Meulder, 2003) or natural language inference (e.g., GLUE in Wang et al., 2018, 2019)—but less emphasis has been put on linguistic style. Nevertheless, natural language processing literature shows a variety of approaches for the evaluation of *style measuring methods*. With style measuring methods, we refer to models or methods that, either, provide text representations that are sensitive to differences ins style (most often vector representations as, e.g., in Hay et al., 2020), or, can compare two texts on style (e.g., the edit distance between two texts or the predicted value of a cross-encoder that rates the style similarity of sentence pairs). Style measuring

methods have been tested on whether they group texts by the same authors together (Hay et al., 2020; Bevendorff et al., 2020b), whether they can correctly classify the style of a text (Niu and Carpuat, 2017; Kang and Hovy, 2021) and whether human-annotated words of similar style[1] are similarly represented (Akama et al., 2018). However, these evaluation approaches are (i) often application-specific, (ii) often do not test for fine-grained style differences and (iii) usually do not control for content and (iv) are rarely used to compare different style measuring methods.

We propose the **modular, fine-grained and content-controlled *similarity-based STyle EvaLuation framework* (STEL)** to address these shortcomings (i)-(iv): (1) Style is a highly ambiguous and elusive term (e.g., wildly different definitions in Biber and Conrad, 2019; Crystal and Davy, 1969; Labov, 2006; Xu, 2017). We propose a **modular** framework where components can be removed or added to fit an application or specific understanding of style. (2) Variation in style can be subtle. Our proposed evaluation framework can be used to test for **fine-grained** style differences. (3) Style is hard to disentangle from content as the two are often correlated (e.g., Gero et al., 2019; Bischoff et al., 2020). For example, people probably speak more formally in a job interview with a potential new manager than in a bar with friends. At the same time, the distribution of topics that are discussed in these two settings might be different too: In a job interview one will probably talk more about personal qualifications and past professional experiences than in a bar among friends. Thus, language models and methods might pick up on spurious content correlations in a benchmark that does not **control for content** (similar to correlations in NLI in Poliak et al., 2018). (4) We demonstrate STEL on English and apply it to several style measuring methods.

An example STEL task is shown in Figure 4.1. The task is to order sentence 1 (S1) and sentence 2 (S2) to match the style order of anchor 1 (A1) and anchor 2 (A2). A group of STEL task instances can contrast different types of style: Here, we demonstrate two general *dimensions* of style (formal/informal and simple/complex) as well as two specific *features* of style (contraction and number substitution). By design, the style features are easy to identify. Thus, the STEL task instances contrasting different style features are easier to solve than the STEL task instances contrasting different style dimensions. STEL contains 815 task instances per dimension and 100 task instances per feature (see Table 4.1). To be evaluated on STEL, methods do not need to be able to classify styles directly. Instead, any method that can calculate the style *similarity* between two sentences can be evaluated: Style measuring methods that (1) calculate similarity values directly (e.g., edit distance or cross-encoders in Reimers and Gurevych, 2019) and (2) represent a sentence's style as a vector (e.g., Hay et al., 2020; Ding et al., 2019) by using a distance or similarity measure between them (e.g., cosine similarity). STEL components can easily be created from parallel sets of paraphrases which differ along a style dimension (Section 4.3), e.g., sets of paraphrases that vary along the formal/informal dimension as in Rao and Tetreault (2018).

[1] Akama et al. (2018) created a dataset of Japanese word pairs that are rated for their style similarity on a scale of five values based on 15 human annotations. The words were selected for annotation by asking human annotators to judge which words in an utterance would be altered in different situational contexts.

Component Name	#	GT	Example Anchor 1 (A1)	Anchor 2 (A2)	Sentence 1 (S1)	Sentence 2 (S2)
formal/-informal	815	✗	r u a fan of them or something?	Are you one of their fans?	Oh, and also, that young physician got an unflattering haircut.	Oh yea and that young dr got a bad haircut.
simple/-complex	815	✗	These rock formations are made of sandstone with layers of quartz.	These rock formations are featureally composed of sandstone with layers of quartz.	The Odyssey is an ancient Greek epic poem attributed to Homer.	The Odyssey is an old Greek epic poem written by Homer.
number substitution	100	✗	<3 friends forever	<3 friends 4ever	D00d $30 is heaps cheap, that must work out to just a couple of bucks an hour	Dude $30 is heaps cheap, that must work out to just a couple of bucks an hour
contraction	100	✓	In that time, it's become one of the world's most significant financial and cultural capital cities.	In that time, it has become one of the world's most significant financial and cultural capital cities.	Will doesn't refer to any particular desire, but rather to the mechanism for choosing from among one's desires.	Will does not refer to any particular desire, but rather to the mechanism for choosing from among one's desires.

Table 4.1: STEL Examples. We give an example for each component of STEL: Formal/informal and simple/complex for the more complex style dimensions as well as number substitution and contraction for the simpler style features. The task is to order sentence 1 (S1) and sentence 2 (S2) to match the style order of anchor 1 (A1) and anchor 2 (A2). If the original order (i.e., S1-S2) is correct, the ground truth column (GT) shows a ✓.

Contribution With this work, we contribute **(a)** the modular, fine-grained and content-controlled STEL framework (Section 4.3), **(b)** 1830 validated task instances for the considered style components (Section 4.4) and **(c)** baseline results of STEL on 18 style measuring methods (Section 4.5). We find that the transformer-based RoBERTa and BERT base models outperform simple versions of commonly used style measuring approaches like LIWC, punctuation frequency or character 3-grams. We invite the addition of complementary tasks and hope that this framework will facilitate the development of improved style-sensitive models and methods. Our data and code are available on GitHub.[2]

4.2. Related Work

Linguistic style has been analyzed from different perspectives and along different dimensions. A speaker's style can, for example, be influenced by the situation, the speaker's identity or the speaker's choices (Preoţiuc-Pietro et al., 2016; Flekova et al., 2016; Nguyen et al., 2016; Bell, 1984). In NLP, previously commonly analyzed "style" dimensions include formal/informal, simple/complex, abstract/concrete and polite/impolite (Pavlick and Nenkova, 2015; Pavlick and Tetreault, 2016; Paetzold and Specia, 2016; Brooke and Hirst, 2013; Madaan et al., 2020).

[2] https://github.com/nlpsoc/STEL

Linguistic style is usually defined to be distinct from content (or: referential meaning). However, style is often found to be correlated with content (e.g., Gero et al., 2019). NLP researchers have found different ways to take content out of the equation: They avoid the use of content-specific features such as content words (Neal et al., 2017; Stamatatos, 2017), compare the style of a text with its paraphrase (Preoţiuc-Pietro et al., 2016; Niu and Carpuat, 2017) or use texts with low semantic similarity scores or different topic labels but written by the same author to learn the stylistic choices of authors (e.g., Boenninghoff et al., 2019a). Others choose no or only limited control for content (e.g., Zangerle et al., 2020; Kang and Hovy, 2021). There has been considerable work in creating parallel datasets of sentence-level paraphrases with shifting style (Xu et al., 2012, 2016; Rao and Tetreault, 2018; Krishna et al., 2020). The task of generating paraphrases of text fragments with different style properties is sometimes also called *style transfer*.

There is little work on general evaluation benchmarks for style measuring methods. Kang and Hovy (2021) use style classification tasks to compare language models. Only models that classify style into the given 15 "style" dimensions (e.g., formality, sarcasm, ...) can be evaluated. They do not control for content. Other related tasks are the PAN *Authorship Verification* (Kestemont et al., 2020) and *Style Change Detection* (Zangerle et al., 2020) tasks which aim at identifying whether two documents or consecutive paragraphs have been written by the same author. In their current version both tasks do not control for topic. However, Kestemont et al. (2020) controls for domain (here: 'fandom' of the considered 'fanfictions'). The best performing model for Kestemont et al. (2020) was a neural LSTM-based siamese network (Boenninghoff et al., 2020), which is conceptually similar to some variants of sentence BERT (Reimers and Gurevych, 2019). The PAN setup assumes that authors tend to write in a relatively consistent style. Based on similar assumptions, the field of *authorship attribution* aims to determine which author wrote a given document.

Especially in authorship attribution, recurring style features include character n-grams, punctuation, average word length or function word frequency (Neal et al., 2017; Grieve, 2007; Stamatatos, 2009). Other recurring methods for style measurement include LIWC (Pennebaker et al., 2015; Danescu-Niculescu-Mizil et al., 2011; El Baff et al., 2020), and learned vector representations of words and sentences (Akama et al., 2018; Ding et al., 2019; Hay et al., 2020). Niu and Carpuat (2017) suggest that style variations are already represented in commonly used neural embeddings.

Binary and more fine-grained style classification has been employed at the word, text fragment as well as document level (Danescu-Niculescu-Mizil et al., 2013a; Pavlick and Nenkova, 2015; Preoţiuc-Pietro et al., 2016; Pavlick and Tetreault, 2016; Kang et al., 2019). Traditionally, documents considered in authorship attribution were longer than 1,000 words (e.g., Eder, 2013), but recently there has been increased interest in text fragments with fewer than 300 words (e.g., Brocardo et al., 2013; Boenninghoff et al., 2019a).

4.3. Our Style Evaluation Framework

We introduce the modular, fine-grained, and content-controlled similarity-based STyle EvaLuation framework (STEL). STEL tests a (language) model's ability to capture the style of a sentence.

Style definitions Style has received a lot of attention in fields like sociolinguistics, stylometry, forensic linguistics, but also in natural language processing. With the term researchers usually aim to the describe the form of a text (i.e., how something is said) more so than its referential meaning (i.e., what is said).[3] However, beyond this distinction style has been conceptualized in many different, often conflicting, ways. Some have defined style as "aesthetic preferences" of specific authors or time periods without any communicative function (Biber and Conrad, 2019), while others describe it as a socially meaningful clustering of features (Campbell-Kibler et al., 2006) or encompassing all forms of language variation (Crystal and Davy, 1969) shared by a group of people. In NLP, the conceptualizations of style have been even more broad, arguably encompassing any general attribute of a text (Jin et al., 2022), including its sentiment and politeness. For a broader overview of style definitions and operationalizations refer to Background Section 2.2.

Modular operationalization of style We refrain from meddling in the style definition debate and stick with the broad notion of "how vs. what". STEL consists of tasks that contrast pair of sentences with the same referential meaning and thus takes the "what" out of the equation. For the "how", we take inspiration from Campbell-Kibler et al. (2006)'s description of style as a clustering of features. We consider single *features* (i.e., more specific linguistic choices or "features") as well as more general *dimensions* of style (i.e., more complex combinations of style features). By not only using complex style dimensions, but also small scale and simpler features, STEL allows for controlled and **fine-grained** testing. We can easily make sure that only the features and no other aspects change (cf. Table 4.1). Depending on one's goal and understanding of style, some *components* (i.e., dimensions or features) should be excluded and others should be added to this **modular** framework. We exemplify the framework's more complex dimensions with the formal/informal distinction as this previously has been the most agreed upon dimension of style (Heylighen and Dewaele, 1999; Labov, 2006). Additionally, we use the simple/complex dimension which has often been of interest in NLP (Wubben et al., 2012; Al-Thanyyan and Azmi, 2021; Jablotschkin et al., 2024), has been used in connection to linguistic-stylistic choices (Haaften and Leeuwen, 2021; Pavlick and Nenkova, 2015) and has an English parallel corpus available (Xu et al., 2016). We exemplify the framework's simpler style dimensions (i.e., features) with number substitutions (e.g., number → numb3r) and contraction usage (e.g., can not → can't). Contractions are a common part of stylistic or variational analysis (Biber, 1988; Grieve, 2011). See Table 4.1 for examples for each component.

[3] In Natural Language Generation (NLG), this distinction has also been used, usually called tactical vs. strategical NLG (Thompson, 1977; van der Sluis and Mellish, 2009).

Controlling for content It is difficult to clearly separate style from content, e.g., Stamatatos (2017) and Gero et al. (2019). Specific scenarios might correlate with both style and content. For example, in a job interview applicants might use a more formal style and talk more about their profession than in a more informal setting at a bar. Then, a model that generally rates texts about jobs as formal and texts about beverage choices as informal might perform well at formality prediction if one does not control for content. In other words, models that correctly use style features could sometimes be indistinguishable from those that use topical features. To control for content, we use parallel paraphrase datasets (Section 5.3.2), which consist of a set of sentences written in one style and a parallel set of sentences written in another. The paraphrase corpora we use (Xu et al., 2016; Rao and Tetreault, 2018) have annotators rewrite reference sentences with the direction to keep the content the same. Similarly, we rewrite reference sentences by removing number substitutions and contractions for the STEL features (Section 4.4). Note, however, that the parallel datasets might include pairs that do not show complete content equivalence. Insisting on complete equivalence would limit the set to sentences to those that are practically identical at the string level (Dolan and Brockett, 2005; Bhagat and Hovy, 2013), making it impossible to compare more complex dimensions of styles.

Task setup We test a method's style measuring capability with tasks of the setup shown in Figure 4.1. The sentences (S1 and S2) have to be ordered to match the order of the anchor sentences (A1 and A2). Here, 'r u' (A1) and 'Oh yea' (S2) are written in a more informal style than their respective paraphrases A2 ('Are you') and S1 ('Oh, and also'). Thus, the correct order is S2, then S1. We call this setup the *quadruple setup*. Additionally, we explore a second task setup, the *triple setup*, which leaves out anchor 2 (A2). There, the task is to decide which of the two sentences matches the style of anchor 1 (A1) the most. The two different setups are similar to the triple and quadruple training instances in the field of metric learning, e.g., de Vazelhes et al. (2020); Law et al. (2016) and Kaya and Bilge (2019).

4.4. STEL Task Creation

We describe the task instances of STEL: First, we create potential task instances (Section 4.4.1). Second, we describe problems with the created instances (i.e., ambiguity in Section 4.4.2). Third, we filter out the problematic instances via crowdsourcing (Section 4.4.3).

4.4.1. Potential Task Instances

We create potential task instances on the basis of parallel paraphrase datasets written in style 1 and style 2 respectively. For a (style 1, style 2) paraphrase pair (anchors in Table 4.1), we randomly select another paraphrase pair (sentences in Table 4.1). Again randomly, we decide which of the anchor pair is anchor 1 (A1) and which is anchor 2 (A2) and fix that ordering for all future considerations. We do the same for the sentence pair. The answer to the STEL task (cf. Figure 4.1) is labeled as ✓or S1-S2 if A1 was

taken from the same style set as S1, e.g., both from style 1. Otherwise the order is reversed.

Formal/informal dimension We use the test and tune split of the Entertainment_Music GYAFC subcorpus (Rao and Tetreault, 2018) as the parallel paraphrase dataset. It consists of a set of informally phrased sentences and a parallel set of crowd-sourced formal paraphrases. We create 918 potential STEL formal/informal task instances.

Simple/complex dimension We use the test and tune split from Xu et al. (2016). It consists of English Wikipedia sentences and 8 crowd-sourced simplifications per sentence. For each Wikipedia sentence, we randomly draw the parallel paraphrase out of the 8 simplifications. We discard sentences that are too close to the original via the character edit distance of 3 or lower. From this parallel paraphrase dataset, we create 1195 potential STEL simple/complex task instances.

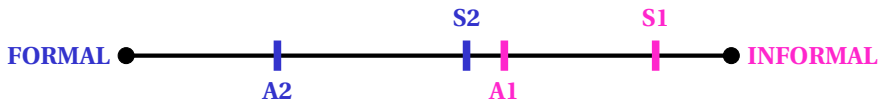
Contraction feature We create the parallel contraction dataset from the December 2018 abstract dump of English Wikipedia[4]. The Wikipedia style guide discourages contraction usage and provides a dictionary with contractions that should be avoided.[5] We use an adapted version[6] to select 100 sentences where an apostrophe is present and a contraction is possible. Such a sentence could be “It is near Thomas’s car”. For each sentence, we create a parallel sentence with a contraction, e.g., “It’s near Thomas’s car”, cf. Table 4.1.

Number substitution feature The character by number substitution task instances are semi-automatically created from the Reddit comment corpus of the months May 2007-September 2007, June 2012, June 2016 and June 2017 taken from the Pushshift dataset (Baumgartner et al., 2020). We select a pool of potential sentences where words contained common character substitution symbols (4,3,1,!,0,7,5) or are part of a manually selected list of number substitution words (see Appendix B.1.2). Then, we manually filtered out sentences without number substitutions (e.g., common measuring units or product numbers). We selected 100 sentences, 50 of which were selected to contain at least one additional number that is not part of a number substitution word (e.g., Anchor 1 in Table 4.1). This setup ensures that the task is not as simple as checking whether there are numbers present in the sentence. To create the parallel phrases, we manually translated the sentences to contain no number substitutions. As we looked for naturally occurring number substitution words, we decided to keep words that contain additional changes besides number substitution. For example, the swapping of characters (t3h, the), generally different spelling (e.g., ‘d00d’, ‘dude’) or phonetic spelling (e.g., ‘str8’, ‘straight’). We decided to replace the number substitution symbols with alphabetic characters only and not consider other character types

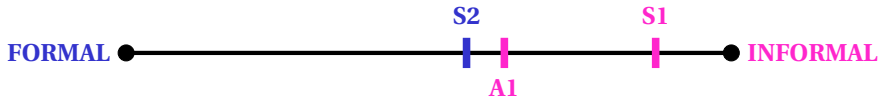
[4] https://archive.org/details/enwiki-20181220

[5] https://en.wikipedia.org/wiki/Wikipedia:List_of_English_contractions

[6] We removed less common contractions like ain’t or ’twas. See Appendix B.1.1 for the used contraction list.



(a) Quadruple Setup: All four sentences are available.



(b) Triple Setup: A2 is removed, the gold label is wrong.

Figure 4.2: Triple Problem. Quadruple tasks are created from sentence pairs (A1, A2) and (S1, S2) that are split along the same style dimension (e.g., formal/informal). See an example in (a). For each pair, only the relative order on the axis (here: S1 is more informal than S2 and A1 is more informal than A2) but not the absolute localization is known. This might lead to a wrong generated label for the triple setup: For the triple setup in (b), A2 is removed. Here, S2 is stylistically closer to A1, whereas the generated "ground truth" label would be S1 is closer to A1, because A1 and S1 were both more informal than their counterpart.

III

like punctuation marks. For example, we convert a string like 's1de!!!!!1!' to 'side!!!!!1!', replacing 1 with i but not replacing 1 with !.[7]

4.4.2. Ambiguity

Manual inspection shows that the created potential task instances of the formal/informal and simple/complex dimension contain ambiguities: (i) Some are a result of unclear or very fine distinctions between the two parallel styles in the original data. For example, consider *"There he died six weeks later, on 13 January 888."* and *"here he died six weeks later, on 13 January 888."*. The second is labeled as written in a simpler style in Xu et al. (2016). However, replacing "there" with "here" arguably does not simplify the sentence. After manual inspection, ambiguities like this, where more than one label could be justified for a given sentence, seem to be more prevalent for the simple/-complex than the formal/informal dimension. (ii) Other ambiguities are the result of entangled additional linguistic components. For example, consider the potential task instance (A1) *"He's supposed to be in jail!"*, (A2) *"I understood he was still supposed to be incarcerated."* and (S1) *"green day is the best i think"*, (S2) *"I think Green Day is the best."*. The sentences are clearly split along the formal/informal dimension leading to the label S1-S2. Still, A1 and S2 could also be understood as being written in a more decisive tone than A2 and S1 leading to the order S2-S1.

We find that using parallel corpora to create the triple instances has additional theoretical limitations that can lead to ambiguity: Consider the *Triple Problem* in Figure 4.2: Here, (A1, A2) and (S1, S2) are two paraphrase pairs taken from the parallel informal/-formal dataset, where both A1 and S1 are labelled as more informal than their counterpart. To get from the quadruple to a triple setup, we remove A2 and automatically create the "gold label" that says that A1 and S1 match in style, as they were both on the

[7] Note, substituting ! for 1 is a somewhat common practice online. See for example Androutsopoulos (2023).

	Sample				Total		
	n	Triple		Quadruple	n	Quadruple	
Dim		κ	acc.	κ	acc.		κ	acc.
all	602	0.29	0.62	0.35	0.78	2113	0.30	0.77
c	301	0.19	0.51	0.16	0.68	918	0.17	0.68
f	301	0.39	0.74	0.51	0.89	1195	0.48	0.90

(a) Results on the sample and total of task instances

Triple	Quad	Dim	Share
✗	✓	f	**0.196**
		c	**0.312**
✓	✗	f	0.047
		c	0.140
✗	✗	f	0.066
		c	0.179
✓	✓	f	**0.691**
		c	**0.369**

(b) Sample analysis

Table 4.2: Annotation Results. We filter out ambiguous task instances via annotations. In (a), we display inter-annotator agreement (Fleiss's κ) and annotation accuracy (acc.) for the sample and total of potential task instances on the quadruple and triple setup for the simple/complex (c) and the formal/informal (f) dimensions. We also display the number of task instances per dimension (n). In (b), we display the share of all combinations of "correct" (✓) and "wrong" (✗) annotations w.r.t. the generated ground truth per dimension and task setup. The union of ✓✓and ✗✓cases (bold) make up a majority.

III

more informal side. However, due to the placement of (A1, A2) on the formal/informal axis (see Figure 4.2), A1 is actually closer in style to S2 (compare also to the text example in the paragraph above). Our generated "gold label" would be wrong for this toy example. Additionally, having fewer sentences in the triple setup increases the chance of a random correlation with a different linguistic component (similar to the 'decisive tone' in the previous example).

4.4.3. Removing Ambiguity

Using crowd-sourced annotations, we filter the previously discussed ambiguity out of the potential formal/informal and simple/complex task instances. We do this by removing instances, where annotations do not align with our generated "ground truth". We do not filter the created simpler style features (contraction and number substitution) as those, by design, differ only in number substitution and contraction use (e.g., Table 4.1) and should contain few if any ambiguities.

Annotation tasks For both the triple and quadruple setup we collected annotations on a sample of all created task instances (301 simple/complex and formal/informal instances respectively). Then, motivated by the better agreement with the quadruple setup on the sample (see Table 4.2a), we annotated a larger total set of task instances on the quadruple setup alone. We had 617 and 894 more task instances annotated for the formal/informal and simple/complex dimension respectively.

Annotation setup We used annotations from 839 different `Prolific`[8] crowd workers with 5 distinct annotators per potential task instance. We paid participants 10.21£/h[9]

[8] `https://www.prolific.co/`

[9] above UK minimum wage of 8.91£/h at the time of the study (April 2021), see `https://www.gov.uk/national-minimum-wage-rates`

on average. All annotators were native English speakers as we assume them to have a better intuition about their language. See Appendix B.3 for further detail.

Annotator agreement In Table 4.2a, we report inter-annotator agreement with Fleiss's κ (Fleiss, 1971) as κ allows different items to be rated by different sets of raters. Inter-annotator agreement is only moderate. This does not mean that the annotations are of poor quality. As discussed in Section 4.4.2, our created data contains ambiguous, noisy or faulty task instances. Manual inspection confirms that low annotator agreement is a sign of ambiguity (see also 'Annotation Analysis' Table B.2 in Appendix B.3). This problem is more pronounced for the simple/complex than the formal/informal dimension. We ensured annotator quality with screening questions (Appendix Table B.1) and by selecting annotators with the highest platform-internal rating.

Annotation results Results are reported in Table 4.2a. Annotation accuracy is the share of "correctly" annotated task instances out of all potential task instances. We count an instance as having been annotated correctly if a majority of at least 3 annotators align with the generated ground truth.

The accuracy and the inter-annotator agreement are considerably higher for the formal/informal dimension (Table 4.2a) than for the simple/complex dimension. This aligns with our expectation of more ambiguity in the simple/complex task instances, cf. Section 4.4.2(i). Similarly, our expectations regarding theoretical problems with the triple setup (Section 4.4.2) are confirmed: Accuracy in the sample is generally higher for the quadruple than the triple setting. There are more examples where the quadruple setup was correctly annotated but the triple setup was not (✗✓in Table 4.2b), than there are for the opposite kind (✓✗).

As a consequence, the annotation of the bigger set of task instances was only done on the quadruple setup. On the total set of potential task instances (which includes the sample), we obtained similar accuracy and annotator agreement as on the sample (see Table 4.2a). We filter the potential task instances by only keeping those that were correctly annotated by a majority (i.e., at least 3/5). This leaves 822 task instances for the formal/informal and 815 for the simple/complex dimension. We randomly remove 7 task instances from the formal/informal dimension for equal representation of the two style dimensions. In the following, and under the name STEL, we will only consider the quadruple setup on the 1830 filtered task instances (i.e., 815, 815, 100 and 100 for simple/complex, formal/informal, number substitution and contraction respectively).

4.5. Evaluation

We use our STEL framework to test several models and methods that could be expected to capture style information (Section 4.5.1). We describe how the models decide the STEL tasks (Section 4.5.2) and discuss their performance on STEL (Section 4.5.3).

4.5.1. Style Measuring Methods

We describe methods and models that could be used to calculate a (style) similarity. Given two sentences, the methods return a similarity value between 0 and 1 or -1 and 1 (when using cosine similarity), where 1 represents the highest similarity.

Language models We use the base BERT UNCASED and base BERT *cased* model (Devlin et al., 2019). We calculate the mean over the subwords in the last hidden layer to generate two sentence embeddings. Then, we use cosine similarity to compare the sentences. We do the same with the cased ROBERTA base model (Liu et al., 2019). ROBERTA encompasses BERT's pretraining dataset and removes the next sentence prediction (NSP) task. However, closer sentences could generally be more similar in style than a different random sentence—possibly making the NSP a valuable learning objective for style similarity learning. To look at this, we experiment with the BERT NSP head on the cased and uncased base model. As the prediction head does not provide an obvious vector representation, we calculate "similarity" value sim(A1, S1) by using the predicted softmax probability that A1 is followed by S1. Additionally, we compare to the sentence BERT 'all-mpnet-base-v2' (SBERT MPNET)[10] and 'paraphrase-multilingual-mpnet-base-v2' (SBERT PARA-MPNET)[11] models (Reimers and Gurevych, 2019, 2020). Like BERT, MPNET uses a transformer architecture, but with a permuted instead of a masked language modeling pre-training task (Song et al., 2020). Furthermore, we experiment with the universal sentence encoder (USE) from Cer et al. (2018).

Authorship attribution methods The following methods are inspired by successful or commonly used approaches in authorship attribution (Neal et al., 2017; Sari et al., 2018). We use character 3-gram similarity by calculating the cosine similarity between the frequencies of all *character 3-gram*s. We calculate the *word length* similarity via the average word lengths a and b of two sentences: $1-|a-b|/\max(a,b)$. We calculate the *punctuation* similarity by using the cosine similarity between the frequencies of punctuation marks {',:,‘,’,_,!,?,;,.,",(,),-}, taken from Sari et al. (2018).

LIWC-based style measuring methods LIWC categories have previously been used as style features (Niederhoffer and Pennebaker, 2002). We use LIWC 2015 (Pennebaker et al., 2015) for (a) *LIWC* similarity by taking the cosine similarity between the complete LIWC frequency vectors, (b) *LIWC (style)* similarity by taking the cosine similarity between the 8 dimensional binary LIWC style vectors (1 if a word of the category is present in the sentence, 0 otherwise) proposed in Danescu-Niculescu-Mizil et al. (2012), (c) *LIWC (function)* similarity by taking 1− the difference between the relative frequencies of function words. Function words have previously been used as a proxy for style (Neal et al., 2017).

Other methods We also experiment with the "*deepstyle*" model (Hay et al., 2020) by taking the cosine similarity between the style vector representations. Additionally, we

[10] best performing sentence embedding in September 2021, according to `https://www.sbert.net`
[11] best performing embedding trained on paraphrase data

consider the following sentence features: NLTK *POS Tags* (Bird et al., 2019) and *share of cased* characters (e.g., Sari et al., 2018) via the cosine similarity between the frequency vectors and 1 - the difference between the proportion of cased characters respectively. We also include the *edit dist*ance as a simple baseline.

4.5.2. Similarity-based Decision

To determine an answer for a STEL task in the quadruple setup, the methods need to order two sentences (Figure 4.1). We do this by calculating the similarities (sim) between Anchor 1 (A1), Anchor 2 (A2), Sentence 1 (S1) and Sentence 2 (S2). We decide for the order S1-S2 if

$$\begin{aligned}&(1-\text{sim}(\text{A1},\text{S1}))^2+(1-\text{sim}(\text{A2},\text{S2}))^2 <\\&(1-\text{sim}(\text{A1},\text{S2}))^2+(1-\text{sim}(\text{A2},\text{S1}))^2\end{aligned} \tag{4.1}$$

For the '>' case we use the order S2-S1, for '=' ordering is settled randomly (cf., 'random' in Table 5.3). See the Appendix Figure B.1 for a proof sketch after transforming similarities to distances.

III

4.5.3. Results

Performance results are shown in Table 4.3. Random guessing would show an accuracy of 0.5 exactly. Stylistic differences can be subtle for the STEL dimensions and we expect this to be a hard task to solve. In contrast, the STEL features (i.e., contraction and number substitution) should be easier to solve (via detecting an additional apostrophe or number) and are especially interesting for model error analysis. Note: We do not make general quality judgements because methods were not trained on the components of STEL and were often not even meant to measure style directly.

RoBERTa and BERT encode style information The best performing models are the cased BERT base model (accuracy of 0.77) and RoBERTa (0.80) (Liu et al., 2019), a successor of BERT . The BERT NSP heads seem to retain some information related to style, however, the performance remains below the mean pooled BERT representations. The effect of training objectives on learning style information could be explored further in future work.

(Semantic) sentence embedding methods perform well SBERT PARA-MPNET (0.68) trained on the paraphrase data performs better than SBERT MPNET (0.61) and USE (0.59). Overall, SBERT PARA-MPNET is the fourth best performing model after the BERT/RoBERTa models and the best performing model in the nb3r dimension. In future work, it could be interesting to explore the effect of different training data on the performance of embedding models.

[12] RoBERTa results were updated compared to the publication. There was an issue with batch size and my original implementation. The updated results show a higher RoBERTa than BERT performance and led to small changes in the result discussion.

[13] NSP results were updated as well. They remain very similar to the original results.

	all	formal		complex		nb3r	c'tion	random	
		filter	full	filter	full			filter	full
BERT UNCASED	0.74	0.79	**0.77**	0.65	0.63	0.90	0.90	**0**	0
BERT CASED	**0.77**	**0.82**	**0.81**	**0.68**	0.64	0.92	**1.0**	**0**	0
ROBERTA[12]	**0.80**	**0.83**	**0.81**	**0.73**	0.67	**0.94**	**1.0**	**0**	0
SBERT MPNET	0.61	0.64	0.62	0.53	0.52	0.71	0.84	**0**	0
SBERT PARA-MPNET	0.68	0.73	0.72	0.55	0.54	**0.95**	**1.0**	**0**	0
USE	0.59	0.59	0.58	0.55	0.52	0.58	0.85	**0**	0
BERT UNCASED NSP[13]	0.67	0.72	0.71	0.60	0.57	0.67	0.76	0	0
BERT cased NSP	0.71	0.79	**0.77**	0.60	0.58	0.77	0.97	0	0
char 3-gram	0.55	0.58	0.57	0.52	0.50	0.50	0.64	0.05	0.05
word length	0.58	0.53	0.53	0.59	0.57	0.50	0.94	0.08	0.08
punctuation	0.56	0.58	0.58	0.50	0.49	0.50	0.92	0.38	0.39
LIWC	0.55	0.52	0.52	0.52	0.52	0.50	**0.99**	0.09	0.09
LIWC (style)	0.50	0.52	0.52	0.50	0.50	0.50	0.50	0.62	0.64
LIWC (function)	0.53	0.48	0.48	0.52	0.51	0.50	**1.0**	0.28	0.28
deepstyle	0.66	0.71	0.70	0.55	0.52	0.84	0.96	**0**	0
POS Tag	0.52	0.53	0.53	0.52	0.52	0.50	0.50	0.20	0.20
share cased	0.56	0.55	0.54	0.53	0.51	0.50	**1.0**	0.08	0.08
edit dist	0.54	0.56	0.56	0.52	0.51	0.50	0.39	0.08	0.07

Table 4.3: STEL Results. We display STEL accuracy for different language models and methods. Random performance is at 0.5. The share of task instances for which a method decides randomly as it can not decide between the two options ('=' in Equation 4.1) is given in the 'random' column. Both the performance on the set of task instances before (full) and after crowd-sourced filtering (filter) is displayed. The two best accuracies are boldfaced. NSP stands for the next sentence prediction head. The BERT and ROBERTA models perform the best. On average, methods perform best for the c'tion and worst for the simple/complex dimension.

Off-the-batch LIWC vectors do not perform well On the style dimensions LIWC performs similar to the random baseline, possibly because the LIWC methods often find no difference between the two possible orderings (9%, 62% and 28% of the tasks). Future work could explore models that weigh different LIWC categories against each other or consider more fine-grained differences between LIWC categories.

Authorship attribution methods perform better than random Character 3-grams and punctuation perform at 0.58 accuracy on the formal/informal dimension. Considering some of the informal examples, punctuation seems to be one of the most prominent visible changes from a formal to an informal style (see Appendix Table B.2). Interestingly, word length is the method that most clearly performs better on the simple/complex than the formal/informal dimension. This aligns with the intuition that shorter words are a sign of a simpler style as found in Paetzold and Specia (2016).

Casing encodes style information The uncased performs worse than the cased BERT model (0.74 vs. 0.77). Additionally, the cased letter ratio performs slightly better than random for the formal/informal dimension (0.55) and perfect for the contraction

feature (1.0): When the sentence consists of fewer lower cased characters (as a result of removing them when using contractions), the share of upper cased characters increases.

Style embedding yields promising results The method "deepstyle" (Hay et al., 2020) performs well across STEL components (0.66). It performs the worst on the simple/-complex dimension (0.55). The method embeds sentences in a vector space where texts by "similar" authors are similarly embedded. In the training data (blog and news articles), authors might not consistently use one style over the other. The difference between same author and same style could be explored in future work.

Less ambiguous task instances reach higher accuracy values Table 5.3 (cf. 'full') shows the accuracy of the style measuring methods for the complete set of potential task instances before filtering out ambiguity (Section 5.3.2). The accuracies are the same or lower than the crowd-validated task instances in STEL. The differences are more pronounced for the simple/complex than the formal/informal dimension. This aligns with the higher (expected) ambiguity in the simple/complex dimension (Section 4.4.2 and Section 4.4.3). In general, we recommend to use the filtered STEL task with less ambiguity for testing.

4.6. Limitations and Future Work

Our illustrative set of task instances does not cover all possibilities of style variation. Future work could extend STEL to cover additional style dimensions or more fine-grained task instances using several sources of data.

The STEL task instances for one style component can contain correlations with unconsidered (style) components. Consider the following task instance (shortened for readability): (A1) *"Forty-nine species of pipefish [...] have been recorded."*, (A2) *"Forty-nine type of pipefish [...] have been found"*, (S1) *"Patients [...] must have their liver checked for damage and other side effects."* and (S2) *"[...] patients [...] must be monitored for liver damage and other possible side effects."*. (A2) and (S1) are the simpler version of (A1) and (S2) (Xu et al., 2016). Additionally, the sentences vary along other aspects: (A2) is missing the punctuation mark and includes a misspelling. (S1) is different in content from (S2) as (S1) is only considering effects on the liver while (S2) also includes other side effects. However, those aspects did not change the label given by the annotators (S2-S1) and should mostly be secondary to the considered style dimension.

With STEL, language models and methods are tested only on whether they capture clear differences in style when content is approximately the same. When there are also content differences, such models might put more emphasis on content than stylistic aspects. Our framework could be extended to allow testing for whether a model prefers style over content (e.g., with a new task format where sentence 1 is closer in content to anchor 1 but closer in style to anchor 2, cf. Figure 4.1).

STEL could also be extended to test for individual author styles and style variation related to the social or regional background of authors (e.g., different age groups). For

example, by including sentence pairs with the same content but written by different authors. Current and future dimensions could also be extended by a train/dev/test split to enable training on the task directly. Further, STEL could be enriched by including longer texts (e.g., paragraphs or documents) as anchor and alternative sentences.

4.7. Conclusion

Style is an integral part of language. However, there are only few benchmarks for linguistic style. In this work, we introduce STEL, a modular, content controlled and fine-grained similarity-based style evaluation framework. We provide task instances on well-established dimensions of style (formal/informal and simple/complex) as well as more fine-grained style features (contraction usage and numb3r substitutions). We control for content with the help of paraphrases and a quadruple setup.

On STEL, we test several common approaches that have been used as a proxy for style in the past. Several results are expected: Punctuation and case-sensitive approaches help for the formal/informal dimension (Pavlick and Tetreault, 2016), word length helps to recognize simple/complex styles (Paetzold and Specia, 2016) and common authorship attribution methods like character n-grams help on STEL. Surprisingly, the common LIWC dictionary (Pennebaker et al., 2015) based approach does not work well on STEL off the batch. It might need more specific weighing of the different LIWC categories against each other. Newer neural and transformer-based approaches outperform feature-based approaches even when they were only trained on MLM or related semantic training tasks. Out of the evaluated language models and methods, the RoBERTa base model performs the best.

We hope that this framework will grow to include an even more exhaustive representation of linguistic style and will facilitate the development of improved style(-sensitive) measures.

Task Usage

When using this task, please also cite the original datasets from which the tasks were created: (1) Rao and Tetreault (2018) for the formal/informal component and (2) Xu et al. (2016) for the simple/complex component. (1) also needs the permission for usage of the "L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi part)"[14].

Ethical Considerations

The STEL tasks are based on datasets (Rao and Tetreault, 2018; Baumgartner et al., 2020; Xu et al., 2016) from popular online forums and web pages (Yahoo! Answers, Reddit, Wikipedia). However, the user demographics on these platforms are often skewed towards particular demographics. For example, Reddit users are more likely

[14] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l

to be young and male.[15] Thus, our dataset might not be representative of (English) language use across different social groups. Further, the usage of posts from online platforms without explicit consent from users might lead to (among others) privacy concerns. The Wikipedia simplifications and formal Yahoo! Answers paraphrases were created by consenting crowd workers (Xu et al., 2016; Rao and Tetreault, 2018). We expect the sentences that were extracted from Wikipedia for the contraction dimension and for the complex/simple dimension to lead to minimal privacy concerns as they were meant to be read and copied by a broader public.[16] Rao and Tetreault (2018) and the nb3r dimension do not include user names. However, we acknowledge that users might be identifiable from the exact wording of posts. We removed nb3r substitution instances that included Reddit user names. We hope the ethical impact of reusing the already published Rao and Tetreault (2018) dataset to be small.

Acknowledgements

We thank the anonymous EMNLP conference reviewers for their helpful feedback. We thank Yupei Du and Qixiang Fang for the productive discussions and their equally helpful feedback. We thank Kees van Deemter for his helpful feedback. This work was supported by the "Digital Society - The Informed Citizen" research program, which is (partly) financed by the Dutch Research Council (NWO), project 410.19.007. Dong Nguyen was supported by the research program Veni with project number VI.Veni.192.130, which is (partly) financed by the Dutch Research Council (NWO).

[15] https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

[16] `https://en.wikipedia.org/wiki/Wikipedia:Copyrights`

5

Same Author or Just Same Topic? Towards Content-Independent Style Representations

In this second chapter of Part III on Building Neural Style Representations, I discuss methods for training neural text representations of "linguistic style". We evaluate the newly trained text representations on STEL, which was introduced in Chapter 4. This chapter is based on Wegmann, A., Schraagen, M. & Nguyen, D. (2021). Same Author or Just Same Topic? Towards Content-Independent Style Representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP co-located with ACL 2022* (pp. 249–268). https://doi.org/10.18653/v1/2022.repl4nlp-1.26 See a video of the conference presentation here: https://youtu.be/QHW7pfwJ56E.

Abstract

Linguistic style is an integral component of language. Recent advances in the development of style representations have increasingly used training objectives from *authorship verification* (AV): Do two texts have the same author? The assumption underlying the authorship verification training task (same author approximates same writing style) enables self-supervised and, thus, extensive training. However, a good performance on the authorship verification task does not ensure good "general-purpose" style representations. For example, as the same author might typically write about certain topics, representations trained on authorship verification might also encode content information instead of style alone. We introduce a variation of the authorship verification training task that controls for content using conversation or domain labels. We evaluate

Author contributions: AW developed the idea, prepared the data, implemented the experiments, and wrote the manuscript. MS and DN provided supervision and feedback throughout the entire research process.

whether known style dimensions are represented and preferred over content information through an original variation to the recently proposed STEL framework. We find that representations trained by controlling for conversation are better than representations trained with domain or no content control at representing style independent from content.

5.1. Introduction

Linguistic style (i.e., how something is said) is an integral part of natural language. Style is relevant for natural language understanding and generation (Nguyen et al., 2021; Ficler and Goldberg, 2017) as well as the stylometric analysis of texts (El Bouanani and Kassou, 2014; Goswami et al., 2009). Applications include author profiling (Rao et al., 2010) and style preservation in machine translation systems (Niu et al., 2017; Rabinovich et al., 2017).

While authors are theoretically able to talk about any topic and (un-)consciously choose to use many styles (e.g., designed to fit an audience, see Bell, 1984), it is typically assumed that there are combinations of style features that are distinctive for an author (sometimes called an author's idiolect). Based on this assumption, the *authorship verification* task (AV) aims to predict whether two texts (A, T_1) have been written by the same author (Martindale and McKenzie, 1995; Coulthard, 2004; Neal et al., 2017). Recently, training objectives based on the authorship verification task have been used to train neural vector representations of text that are sensitive to style (Boenninghoff et al., 2019b; Hay et al., 2020; Zhu and Jurgens, 2021), we call them *style representations*. Training objectives based on authorship verification are especially promising because they do not require any additional labeling when author identifiers are available. Similar to the distributional hypothesis, the assumption underlying the authorship verification training task (same author approximates same writing style) enables extensive self-supervised learning.

Unfortunately, style representations trained on the authorship verification task might suffer from not just being sensitive to changes in style, but also to changes in content. This is because style and content are often correlated (Gero et al., 2019; Bischoff et al., 2020). One author in our dataset might always write about their professional career, while another might mostly write about a personal hobby. As a result, style representations might encode spurious content correlations (Poliak et al., 2018) because it helps to solve the authorship verification task. A solution could be to control the authorship verification task for content. For example, by having two texts have similar content also when they were written by different authors. Current style representation learning methods either use no (Halvani et al., 2019; Sundararajan and Woodard, 2018) or only limited control for content (Hay et al., 2020) or use domain labels to approximate topic (Boenninghoff et al., 2019a). Zhu and Jurgens (2021) work with 24 domain labels (here: product categories) for more than 100k Amazon reviews. However, using a small set of labels might be too coarse-grained to fully represent and thus control for content. We introduce another level of content control based on conversations and test whether

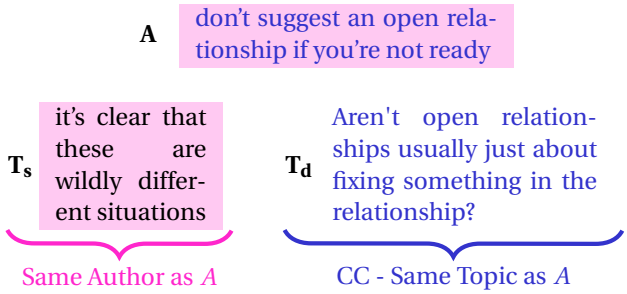


Figure 5.1: Content Control (CC) and Contrastive (CAV) Variant of the Authorship Verification Task. The CAV task is to match A with the utterance T_s that was written by the same author. Contrary to the traditional binary authorship verification task (A, T_1), this includes a third "constrastive" utterance T_2 that was written by a different author than T_1. In addition to the contrastive variation to authorship verification, we experiment with content control (CC) by selecting T_d and A to have the same approximate content with the help of a topic proxy. The topic proxy approximates the topic of an utterance more or less well. The topic proxies we experiment with are (1) random or no topic proxy, (2) the domain the utterance was written in and (3) the conversation the utterance was written in.

the resulting representations are more sensitive to changes in style than to changes in content.

Approach We introduce two independent variants to an authorship verification task (A, T_1): the contrastive authorship verification or CAV variant and the content control or CC variant. For CAV, we add a contrastive sentence T_2 such that there exists exactly one T_d with $d \in \{1,2\}$ that was written by a different author than A. For CC, we select all T_d to have the same approximate content. An example of both variants can be found in Figure 5.1. We fine-tune several siamese RoBERTa-based neural networks (Reimers and Gurevych, 2019) to evaluate style representations trained with the new variants for the authorship verification task. We train on utterances from the `Reddit` platform but our approach could be applied to any other conversation dataset. While previous work mainly aimed at learning representations that represent an author's individual style (Boenninghoff et al., 2019b; Hay et al., 2020; Zhu and Jurgens, 2021), we target general-purpose style representations. Therefore, we ask whether the trained representations (a) represent known style dimensions (e.g. formal vs. informal) in the embedding space (Section 5.4.2), (b) favor style information over content information (Section 5.4.3) and (c) distance utterances written by different authors further from each other, even if they have the same approximate content (Section 5.4.1).

Result When the content stays the same, we find that representations fine-tuned on authorship verification are less sensitive to known style dimensions than the representations extracted from the RoBERTa base model. However, when content varies, representations fine-tuned on authorship verification favor style information over content information more often than the RoBERTa base model. The representations trained with the combination of our CAV and CC variant are better than all other approaches

III

at consistently distancing utterances written by the different authors further from each other, even if they have the same approximate content. We show that our best representations are sensitive to stylistic features like punctuation and apostrophe types such as ’ vs. ' using agglomerative clustering.

Contribution With this paper, we (1) contribute an extension of the authorship verification task that aims to control for content (CC) with conversation labels and (2) introduce a novel variant of the authorship verification setup by adding a contrastive utterance (CAV setup). Further, we (3) introduce a variation of the STEL framework (Wegmann and Nguyen, 2021) to evaluate whether representations favor content over style information. Finally, we (4) release the first style embedding model on the Hugging Face hub[1] and (5) demonstrate found stylistic features via agglomerative clustering. We hope to further the development of content-controlled style representations. Our code and data are available on GitHub.[2]

5.2. Related Work

Semantic representation learning often uses the approach of self-supervised training with *contrastive learning objectives* (Reimers and Gurevych, 2019; Gao et al., 2021). Contrastive learning objectives (Hadsell et al., 2006) for semantic representations push semantically distant sentence pairs apart and pull semantically close sentence pairs together. Different strategies for selecting sentence pairs have been used. Recently, SimCSE used same sentences with dropout as semantically close and randomly sampled sentences as semantically distant sentences (Gao et al., 2021). Loss functions that are known from semantic embedding learning have increasingly been used to learn style representations as well (Boenninghoff et al., 2019a; Hay et al., 2020; Zhu and Jurgens, 2021). “Semantically close” sentences are then replaced by sentences written by the same author, i.e., style representations are often trained and evaluated on the authorship verification task. As a result, deep learning approaches have been used to learn text representations that solve the authorship verification task have been successful (Shrestha et al., 2017; Litvak, 2019; Boenninghoff et al., 2019a; Saedi and Dras, 2021; Hay et al., 2020; Hu et al., 2020; Rivera-Soto et al., 2021; Zhu and Jurgens, 2021). Fine-tuning RoBERTa has been shown to be competitive with other neural as well as non-neural approaches (Zhu and Jurgens, 2021). Probably because the RoBERTa base model already represents style information better than other feature based approaches (Wegmann and Nguyen, 2021). As a result, we also chose the current SOTA approach and fine-tune RoBERTa models to solve the authorship verification task. Instead of sentence pair based loss functions, Reimers and Gurevych (2019) also experiment with a *triplet loss*, which pushes an anchor closer to a semantically close sentence and pulls the same anchor apart from a semantically distant sentence. We use the same approach for our contrastive authorship verification (CAV) setup.

[1] https://huggingface.co/AnnaWegmann/Style-Embedding
[2] https://github.com/nlpsoc/Style-Embeddings

It is a well known problem that style and content are often correlated (Gero et al., 2019; Bischoff et al., 2020). Feature-based authorship verification methods have controlled for content by restricting the feature space to contain "content-independent" features like function words or character n-grams (Grieve, 2007; Neal et al., 2017; Stamatatos, 2017; Sundararajan and Woodard, 2018). However, even these features have been shown to not necessarily be content-independent (Litvinova, 2020). Deep learning approaches have used domain labels to approximate content and control for it (Boenninghoff et al., 2019a; Zhu and Jurgens, 2021). Zhu and Jurgens (2021) address possible spurious correlations by sampling half of the different and same author utterances from the same and the other half from different domains (e.g., subreddits for `Reddit`). We introduce a new method to control for content: using conversation labels.

5.3. Style Representation Learning

We aim to learn neural vector representations of style. Our approach is to fine-tune BERT-based encoder models and use the mean-pooled representations in the last layer as representations. We describe the fine-tuning task in Section 5.3.1, the dataset in Section 5.3.2 and the model training in Section 5.3.3.

5.3.1. Training Task

The authorship verification (AV) task is the task of predicting whether two texts are written by the same or different authors. It has recently been used to train neural vector representations that are sensitive to style Boenninghoff et al. (2019b); Hay et al. (2020); Zhu and Jurgens (2021). In the following, we introduce two independent variations to the authorship verification task: Adding (1) content control via topic proxies and (2) contrastive information with the CAV setup.

Content control (CC) Models optimized for authorship verification have been known to make use of semantic information (Sari et al., 2018; Sundararajan and Woodard, 2018; Potha and Stamatatos, 2018) and to perform badly in cross-topic settings (Halvani et al., 2019; Bischoff et al., 2020). Recent studies use authorship verification tasks to train style representations and address possible correlations by controlling for domain (Zhu and Jurgens, 2021; Boenninghoff et al., 2019b). The idea is that texts from the same domain are more likely to be about a similar topic than ones from different domains. However, using a (usually small set of) domains might be too coarse-grained to fully control for content. We compare three different levels of content control by approximating content with the help of a topic proxy. We sample the utterance pairs written by different authors (T_d and A, cf. Figure 5.1) (i) from the same *conversation*, (ii) from the same *domain* (e.g., subreddit for `Reddit` as in Zhu and Jurgens, 2021) or (iii) *randomly* (as a baseline, similar to Hay et al., 2020). We hypothesize that two utterances from the same conversation are more likely to be about the same topic than two utterances from different conversations. Our newly proposed use of the same conversation "topic proxy" is inspired by semantic sentence representation learning, where conversations have previously been used as a proxy for semantic information encoded in utterances (Yang et al., 2018; Liu et al., 2021). We test to what extent the three differ-

ent topic proxies are contributing to content-independent style representations during evaluation (Section 5.4.3).

CAV setup We introduce an adaption of the Authorship Verification task—the Contrastive Authorship Verification setup (CAV, Figure 5.1): Given an anchor utterance A and two other utterances $\{T_1, T_2\} = \{T_s, T_d\}$, the task is to identify which of the two texts T_1, T_2 equals T_s and was thus written by the same author as A. We experiment with both CAV and binary AV setups for style representation learning. One task with the CAV setup, which consists of three utterances (A, T_1, T_2), can be split up into two AV tasks: (A, T_1) and (A, T_2). Using a contrastive authorship verification setup adds learnable information to the task. Namely, the contrast between the commonalities between (A, T_s) and (A, T_d). It might be easier for the model to learn the distinctive style of A's author by seeing both at the same time—a text written by the same author and a text written by a different author from A. Note that the CAV setup by itself is independent from the content control variation (explained in the previous paragraph), so we do not make any restrictions about the content similarity between A, T_1 and T_2. However, we suspect that CAV is especially useful when T_d has a similar content as A, forcing the model to look at style. Such a different author utterance can be called a "hard negative". In the future, it is also possible to adapt this setup to include several instead of just one contrastive "hard negative" different author utterance. Such contrastive approaches have been successful in semantic embedding learning (Gao et al., 2021; Reimers and Gurevych, 2019).

5.3.2. Dataset

We use a 2018 `Reddit` sample with utterances from 100 active subreddits[3] extracted via `ConvoKit`[4] (Chang et al., 2020). Per subreddit, we sample 600 conversations with at least 10 posts (which we call utterances). All subreddits are directed at an English audience, which we infer from the subreddit descriptions. We removed all invalid utterances.[5] Then, we split the set of authors into a non-overlapping 70% train, 15% development and 15% test author split. For each CC level (conversation, domain, no) and each author split, we created a set of triples (A, T_s, T_d), i.e., nine sets in total (see Table 5.1).

First, we created the triples for the train split of the dataset with conversation content control. We sampled 210k distinct utterances A from the train author split. We use a weighted sampling process to not overrepresent authors that wrote more utterances than others. As a result, an author wrote text A at most 9 times (cf. "ma" in Table 5.1). Then, for each utterance A, we randomly sampled an utterance T_d that was part of the same conversation as A but written by a different author. Then, for all 210k (A, T_d)-pairs, an utterance T_s was sampled randomly from all utterances written by the same

[3] `https://zissou.infosci.cornell.edu/convokit/datasets/subreddit-corpus/subreddits_small_sample.txt`

[4] MIT license

[5] Utterance of only spaces, tabs, line breaks or of the form: `""`, `" [removed] "`, `"[ removed ]"`, `"[removed]"`, `"[ deleted ]"`, `"[deleted]"`, `" [deleted] "`

CC level	Data Split	Setup # AV	Setup # CAV	Utt. #	Author #	Author max	(A, T_s) co	(A, T_s) do	(A, T_d) co	(A, T_d) do
Conversation	train set	420,000	210,000	546,757	194,836	9	0.27	0.56	1.00	1.00
	dev set	90,000	45,000	116,451	41,848	8	0.26	0.55	1.00	1.00
	test set	90,000	45,000	116,621	41,902	8	0.27	0.55	1.00	1.00
Domain	train set	420,000	210,000	544,587	240,065	9	same		0.01	1.00
	dev set	90,000	45,000	116,490	50,939	8	as		0.02	1.00
	test set	90,000	45,000	116,586	51,182	8	conversation		0.02	1.00
No	train set	420,000	210,000	548,082	270,079	9	same		0.00	0.01
	dev set	90,000	45,000	117,149	57,352	8	as		0.00	0.01
	test set	90,000	45,000	117,434	57,726	8	conversation		0.00	0.02

Table 5.1: Data Split Statistics. Per content control (CC) level, we display the number of tasks per setup (# CAV, # AV), unique utterances (Utt.) and authors for each split. We also show the maximum number of times an author occurs as A's author (max) and the fraction of same author utterances (A, T_s) and different authors utterances (A, T_d) that occur in the same conversation (co) and domain (do). The choice of (A, T_s) is the same for all CC levels for comparability.

author as A and for which $A \neq T_s$ holds. We equivalently sampled 45k tasks for the dev and test.

For the domain and no CC level, we reuse A and T_s, to keep as many correlating variables constant as possible. Thus, we only resampled 210k utterances T_d written by a different author from A by sampling from the same domain or randomly.

We make sure that each combination of (A, T_s, T_d) occurs only once. Thus there are no repeating CAV tasks.[6] However, it is possible that some utterances occur more than once across tasks. We randomly order T_s and T_d to form triples (A, T_1, T_2) where $\{T_1, T_2\} = \{T_s, T_d\}$. In total, we generate 210k train, 45k dev and 45k test tasks for each CC level (see Table 5.1), corresponding to a total of 420k, 90k and 90k AV-pairs when splitting the CAV task into (A, T_1) and (A, T_2) pairs (cf. Section 5.3.1).

5.3.3. Training

We use the `Sentence-Transformers`[7] Python library (Reimers and Gurevych, 2019)[8] to fine-tune several siamese networks based on (1) 'bert-base-uncased', (2) 'bert-base-cased' (Devlin et al., 2019) and (3) 'roberta-base' (Liu et al., 2019). We expect those to perform well based on previous work (Rivera-Soto et al., 2021; Zhu and Jurgens, 2021; Wegmann and Nguyen, 2021). We compare using (a) contrastive loss (Hadsell et al., 2006) with the AV setup (Section 5.3.1) tasks and (b) triplet loss (Reimers and Gurevych, 2019) with the CAV setup (Figure 5.1). The binary contrastive loss function uses a pair of sentences as input while the triplet loss expects three input sentences. For the loss functions, we experiment with three different values for the margin hyperparameter (i) 0.4, (ii) 0.5, (iii) 0.6. We train with a batch size of 8 over 4 epochs using 10% of the

[6] There might be same author (A, T_s) pairs that occur twice due to our specific sampling process. However, this remains unlikely due to the high number of authors and utterances. Overall, the share of repeating pairs remains lower than 1%.

[7] `https://sbert.net/`

[8] with Apache License 2.0

III

Training Task		Testing Task					
			AV			CAV	
		Conversation	Domain	No	Conversation	Domain	No
Setup	CC level	AUC $\pm\sigma$	AUC $\pm\sigma$	AUC $\pm\sigma$	acc $\pm\sigma$	acc $\pm\sigma$	acc $\pm\sigma$
RoBERTa base		.53	.57	.61	.53	.58	.63
AV	Conversation	**.69** $\pm$ .02	.70 $\pm$ .02	.71 $\pm$ .02	**.68** $\pm$ .02	.69 $\pm$ .02	.70 $\pm$ .02
	Domain	.68 $\pm$ .01	**.71** $\pm$ .01	.73 $\pm$ .02	.67 $\pm$ .01	**.70** $\pm$ .01	.73 $\pm$ .00
	No	.58 $\pm$ .01	.63 $\pm$ .02	**.79** $\pm$ .00	.59 $\pm$ .01	.66 $\pm$ .01	**.78** $\pm$ .00
CAV	Conversation	**.69** $\pm$ .00	.70 $\pm$ .00	.71 $\pm$ .00	**.68** $\pm$ .00	.69 $\pm$ .00	.70 $\pm$ .00
	Domain	.68 $\pm$ .00	.70 $\pm$ .00	.72 $\pm$ .00	**.68** $\pm$ .00	**.70** $\pm$ .00	.72 $\pm$ .01
	No	.58 $\pm$ .00	.63 $\pm$ .03	.77 $\pm$ .00	.59 $\pm$ .00	.65 $\pm$ .00	.77 $\pm$ .00

Table 5.2: Test Results. Results for 6 different fine-tuned RoBERTa models on the test sets. We display the AUC for the authorship verification task (AV) and the accuracy of the models for the contrastive authorship verification setup (CAV) with different content control approaches (CC). We display the standard deviation (σ) over three different seeds. Best performance per column is boldfaced. Models generally outperform others on the CC level they have been trained on.

training data as warm-up steps. We use the Adam optimizer with the default learning rate (0.00002). We leave all other parameters as default. We use the BinaryClassificationEvaluator on the AV setup with contrastive loss and the TripletEvaluator on the CAV setup with triplet loss from `Sentence-Transformers` to select the best model out of the 4 epochs. The BinaryClassificationEvaluator calculates the accuracy of identifying similar and dissimilar sentences, while the TripletEvaluator checks if the distance between A and T_s is smaller than the distance between A and T_d. We use cosine distance as the distance function.

5.4. Evaluation of Style Representations

We evaluate the learned style representations on the Authorship Verification task (i.e., the training task) in Section 5.4.1. Then, we evaluate whether models learn to represent known style dimensions via the performance on the `STEL` framework (Wegmann and Nguyen, 2021) in Section 5.4.2. Last, we evaluate representations on their content-independence with an original manipulation of `STEL` (Section 5.4.3). We find that all investigated approaches perform similarly in representing style when content stays the same, but using our proposed approach for style representation learning (CAV with conversation level content control) makes style representations more independent from content than other training approaches.

5.4.1. Authorship Verification

We display the AV and CAV performance of trained models in Table 5.2. On the development sets, RoBERTa models consistently outperformed the cased and uncased BERT models. Also, different margin values only led to small performance differences (Appendix C.1). Consequently, in Table 5.2, we only display the performance of the six fine-tuned RoBERTa models on the test sets using the three different content controls

(CC) and two different task setups (AV and CAV setups) with constant margin values of 0.5.

AV performance is usually calculated with either (i) AUC or (ii) accuracy using a predetermined threshold (Zhu and Jurgens, 2021; Kestemont et al., 2021). We use cosine similarity to calculate the similarity between sentence representations. Thus, there is no clear constant default threshold to decide between same and different author utterances. A threshold could be fine-tuned on the development set, however, for simplicity we use AUC to calculate AV performance instead. We use accuracy for the CAV task—here no threshold is necessary (cosine similarity is calculated between A, T_1 and A, T_2 and the highest similarity utterance is chosen). This makes the performance scores on the test sets less comparable across setups—however, comparability of the CAV and AV performance scores are limited in any case as the AV vs. CAV setups are fundamentally different. Performance scores can be compared across the same column, i.e., within the same AV and CAV setup. We aggregate performance with mean and standard deviation for three different random seeds per model parameter combination.[9]

Overall, the AV & CAV training task setup (rows in Table 5.2) lead to similar performance on the test sets. As a result, we do not distinguish between them in this section's discussion. Generally, the representations tested on the CC level they were trained on (diagonal) outperform other models that were not trained with the same CC level. For example, representations trained with the conversation CC level, perform better on the test set with the conversation CC than representations trained with the domain or no CC.

Tasks with the conversation label are hardest to solve For all models, the performance is lowest on the conversation test set and increases on the domain and further on the random test set. This is in line with our assumption that the conversation test set has semantically closer different author utterance (A, T_d)-pairs that make the authorship verification task harder due to reduced spurious content cues (Section 5.3.1).

Representations trained with the conversation CC might encode less content information For the three test sets with the different CC levels, the standard deviation of performance is largest for models trained without CC and smallest for models trained with the conversation CC. This aligns with our expectation that texts in the same conversation better approximate same content than two random texts or texts taken from the domain domain (Section 5.3.1). Representations trained with domain or no CC can rely more heavily on semantic features, which is advantageous for the no and domain CC test sets. In contrast, models trained with conversation CC may develop more content-agnostic representations that are similarly helpful in the no and domain CC test sets. Good CAV & AV performance alone is not necessarily indicative of a good representation of style. In Sections 5.4.2 and 5.4.3, we will further investigate the quality of style representations and their content independence.

[9] We used seeds 103-105. A total of 5 out of 18 models did not learn. We re-trained those with different seeds.

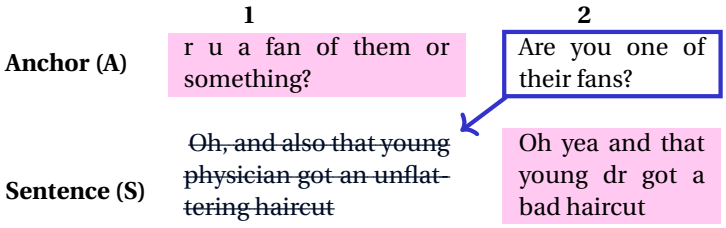


Figure 5.2: STEL-Or-Content Task. The task is to match the anchor sentence (A1) to the sentence that is written in the same style (S2), but is complicated by having to decide against the sentence that has the same content (the new S1, i.e., "Are you one of their fans?"). To create STEL-Or-Content instances, we take the original STEL instances (original A1-2, S1-2, i.e., the figure without manipulations) and move A2 to the sentence position with the different style (here: the more formal A2 replaces the more formal S1).

5.4.2. STEL Task

We calculate the performance of the representations on the STEL framework (Wegmann and Nguyen, 2021).[10] Here, models are evaluated on whether they are able to measure differences in style across four known dimensions of style (formal vs. informal style, complex vs. simple style, contraction usage and number substitution usage). Models are tested on 1830 tasks of the same setup: Two "sentences" S1 and S2 have to be matched to the style of two given "anchor" sentences A1 and A2. The task is binary. Sentences can either be matched without reordering (A1-S1 & A2-S2) or with reordering (A1-S2 & A2-S1). For example, consider the sentences in Figure 5.2 before alterations. The correct solution to the task is to reorder the sentences, i.e., to match A1 with S2 because they both exhibit a more informal style and A2 with S1 because they both exhibit a more formal style. The STEL sentence pairs (S1, S2) and (A1, A2) are always paraphrases of each other (in contrast to A and T_d for the authorship verification task which are only chosen to be about the same approximate topic, cf. 5.3.1). The anchor pairs and sentence pairs are randomly matched and are thus otherwise expected to have no connection in content or topic. Representations can thus not make use of learned content features to solve the task.

We display the STEL results for the RoBERTa models in Table 5.3. **STEL performance is comparable across all fine-tuned models—for all different CC levels and AV & CAV setups.** Compared to common non-neural representations (e.g., character 3-grams) our fine-tuned RoBERTa model performs better (Wegmann and Nguyen, 2021). Surprisingly, the overall STEL performance for the fine-tuned models is lower than that of the original RoBERTa base model (Liu et al., 2019) and also than the cased BERT base model (Wegmann and Nguyen, 2021). We would have expected the fine-tuned models to improve not worsen their performance in representing common style dimensions. Thus, our fine-tuned models may have 'unlearned' some style information.

[10] `https://github.com/nlpsoc/STEL`, with data from Rao and Tetreault (2018) and Xu et al. (2016) and with permission from Yahoo for the "L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi part)": `https://webscope.sandbox.yahoo.com/catalog.php?datatype=l`. Data and code available with MIT License with exceptions for proprietary Yahoo data.

		all		formal, n = 815		complex, n = 815		nb3r, n = 100		c'tion, n = 100	
		o	o-c	o	o-c	o	o-c	o	o-c	o	o-c
				acc±σ	acc±σ	acc±σ	acc±σ	acc±σ	acc±σ	acc±σ	acc±σ
org		**.80**	.05	.83	.09	**.73**	.01	**.94**	**.13**	**1.0**	.00
A	c	.71	.35	.83 ± .02	.64 ± .00	.57 ± .02	.13 ± .04	.61 ± .02	.04 ± .01	.91 ± .10	.00 ± .01
	d	.73	.28	.84 ± .01	.56 ± .04	.69 ± .05	.05 ± .02	.61 ± .02	.03 ± .02	.98 ± .03	.00 ± .00
	n	.72	.22	**.85** ± .01	.46 ± .04	.57 ± .01	.03 ± .01	.62 ± .04	.05 ± .02	.98 ± .01	.00 ± .00
C	c	.71	**.42**	.81 ± .02	**.69** ± .02	.59 ± .01	**.24** ± .02	.65 ± .09	.03 ± .01	.99 ± .02	**.04** ± .02
	d	.71	.32	.82 ± .01	.61 ± .02	.57 ± .01	.12 ± .01	.64 ± .05	.03 ± .01	.99 ± .01	.01 ± .01
	n	.71	.24	**.85** ± .00	.50 ± .02	.56 ± .01	.04 ± .01	.59 ± .03	.06 ± .01	.98 ± .04	.00 ± .00

Table 5.3: STEL and STEL-Or-Content Results. We display STEL accuracy across 4 style dimensions (n = number of instances) for the same ROBERTA models as in Table 5.2: Per task setup (AV - A, CAV - C) and content control level (conversation - c, domain - d, none - n), the performance on the original (o) and the STEL-Or-Content task instances (o-c) are displayed. Per column, the best performance is boldfaced. For the fine-tuned ROBERTA models, performance generally increases on the STEL-Or-Content task compared to the original ROBERTA model (org).

In the remainder of this subsection, we analyze possible reasons for this STEL performance drop.

Performance stays approximately the same or improves for the formal/informal and the contraction dimensions but drops for the complex/simple and the nb3r substitution dimensions. Based on manual inspection, we notice nb3r substitution to regularly appear in specific conversations and for specific topics. Future work could investigate whether the use of nb3r substitution is less consistent for one author than other stylistic dimensions. As the nb3r dimension of STEL only consists of 100 instances, future work could increase the number of instances. Further, we perform an error analysis to investigate the STEL performance drop in the complex/simple dimension. We manually look at consistently unlearned (i.e., wrongly predicted by the fine-tuned but correctly predicted by the original ROBERTA model) or learned (i.e., wrongly predicted by the ROBERTA model and correctly predicted by the fine-tuned model) STEL instances (see details in Appendix C.2.1). We find several problematic examples where the correct solution to the task is at least ambiguous. We display two such examples in Table 5.4. The share of examples with problematic ambiguities is higher for the unlearned (50/55) than for the newly learned STEL instances (29/41). Generally, the number of complex/simple STEL instances with ambiguities is surprisingly high for both the learned as well as the unlearned instances, consistent with the lower performance of the models in this category. Several of the ambiguities found should be relatively easy to correct in the future (e.g., spelling mistakes or punctuation differences).

5.4.3. Content-Independence of Style Representations

We are interested in representations that represent the style of a text. Since style and content are often correlated (Gero et al., 2019; Bischoff et al., 2020), we not only tested whether models are able to distinguish between texts written by different authors (in Section 5.4.1) but also whether models are able to represent styles when the content remains the same (Section 5.4.2). However, we have not tested whether models are

Agg.	GT	Anchor 1 (A1)	Anchor 2 (A2)	Sentence 1 (S1)	Sentence 2 (S2)	Ambiguity
un	✓	TDL Group announced in March 2006, in response to a request [...]	[...] storm names Alberto Helene Beryl Isaac Chris [...]	Palestinian voters in the Gaza Strip [...] were eligible to participate in the election.	1. Palestinian voters in the Gaza Strip [...] were eligible to participate in the election.	A1/A2 have different content
l	✗	[...] 51 Phantom [...] received nominations in that same category.	[...] 1 phantom [...] received nominations in the same category.	[...] the Port Jackson District Commandant could exchange with all military land with buildings on the harbor.	[...] the Port Jackson District Commandant could communicate with all military installations on the harbour.	A2 spelling mistake: 1 instead of 51, S1 sounds unnatural

Table 5.4: STEL Error Analysis. For the complex/simple STEL dimension, we display examples of instances that were learned (l) or unlearned (un) by the fine-tuned RoBERTa models, but upon manual inspection seemed potentially ambiguous. Under the column "Ambiguity", we describe why we think that the ground truth (GT) label might be incorrect or at least subject to discussion. A ground truth (GT) of ✓ means that S1 matches with A1 and S2 with A2 in style, while ✗ means S1 matches with A2 and S2 with A1.

more sensitive to style than content information. A style representation that is content-independent would always be more sensitive to style than content information.

III

Different approaches have been used to test whether style representations encode unwanted content information, including (a) comparing performance on the authorship verification task across domain (Boenninghoff et al., 2019b; Zhu and Jurgens, 2021), (b) assessing performance on function vs. content words (Hay et al., 2020; Zhu and Jurgens, 2021) and (c) assessing performance on predicting domain labels using style representations (Zhu and Jurgens, 2021). However, these evaluation methods have limitations: Domain labels usually come from a small set of coarse-grained labels and function words have been shown to not necessarily be content-independent (Litvinova, 2020). Additionally, next to content, authorship verification might include other spurious features that help increase performance without representing style.

To test if models learn to prefer style over content, we introduce a variation to the STEL framework—the *STEL-Or-Content* task. In the STEL-Or-Content task, we assess whether models are more sensitive to style or content by presenting two text options—one that closely matches the content and another that closely matches the style—and evaluating which one the model selects. See an example in Figure 5.2. We create the STEL-Or-Content instances from the original STEL instances. From one original STEL instance (Section 5.4.2), we take the sentence that has the same style as A2 (here: S1, also in more formal style) and replace it with A2. In Figure 5.2. The new task is to decide whether A1 matches in style with the new S1 (i.e., "Are you one of their fans?") or with S2. The task is more difficult than the original STEL task as the new S1 is written in a different style but has the same content and potentially high lexical overlap. The representations will have to decide between giving 'style or content' more weight. This setup is similar to the CAV task (Figure 5.1). The main differences to the CAV task are (i) that we do not use same author as a proxy for same style but instead use the theory-derived style dimensions from the STEL framework and (ii) that we control for content with the help of paraphrases (instead of using only a topic proxy).

We display the STEL-Or-Content results in Table 5.3. The performance for the new task is low (< 0.5 which corresponds to a random baseline). However, the task is also very difficult as lexical overlap is usually high between the anchor and the false choice (i.e., the sentence that was written in a different style but has the same content). Nevertheless, performance should only be considered in combination with other evaluation approaches (Sections 5.4.1 and 5.4.2) as on this task alone models might perform well because they punish same content information.

Models trained on the CAV task with the conversation CC level are the best at representing style independent from content The performance increases from an accuracy of 0.05 for the original RoBERTa model to up to $0.42 \pm .01$ for the representation trained with the CAV task and the conversation CC. This 'CAV conversation representation' did not just learn to punish same content cues as demonstrated by its performance on the AV task and the STEL framework: (1) On the AV task, the representation performed similarly on all three test sets. If the model had merely learned punish same content cues, we would expect a clearer difference in performance, particularly because confounding content information should be more prevalent in the random test set compared to the conversation test set. (2) The representation performed comparably to the other representations on the STEL framework, where style information is needed to solve the task but content information cannot be used.

III

5.5. Style Representation Analysis

To better understand what the learned style representations consider to be similar styles, we analyze the best-performing style representation (RoBERTa trained on the CAV task with the conversation CC and seed 106). We apply agglomerative clustering to a sample of 5,000 CAV tasks of the conversation test set resulting in 14,756 unique utterances. Based on an analysis of Silhouette scores (Appendix C.3), we group the utterances into 7 clusters.

We find that 46.2% of utterance pairs written by the same author fall into the same cluster, compared to $20.1\% \pm .00$[11] expected from random assignments among 7 clusters. As authors will have a certain variability to their style, a perfect clustering according to general linguistic style would not assign all same author pairs to the same cluster. In Table 5.5, we display examples for 4 out of 7 clusters. We manually looked at a few hundred examples per cluster to find consistencies. We found clear consistencies within clusters in the punctuation (e.g., 97% of utterances have no last punctuation mark in Cluster 3 vs. an average of 37% in the other clusters), casing (e.g., 67% of utterances that use *i* instead of *I* appear in Cluster 4), contraction spelling (e.g., 22 out of 27 utterances that use *didnt* instead of *didn't* appear in Cluster 4), the type of apostrophe used (e.g., 90% of utterances use ‘ vs ' in Cluster 5 vs. an average of 0% in the other clusters) and line breaks within an utterance (e.g., 72% of utterances in Cluster 7 include line breaks vs. an average of 22% in the other clusters). We mostly found such

[11] Calculated mean and standard deviation over 100 runs when randomly assigning the 14,756 utterances to 7 clusters of the same size.

C #	Consistencies	Example
3	no last punctuation mark	I am living in china, they are experiencing an enormous baby boom
4	punctuation / casing	huh thats odd i'm in the 97% percentile on iq tests, the sat, and the act
5	’ vs '	I assume it’s the blind lady?
7	linebreaks	I admire what you're doing but [...]
		I know I'm [...]

Table 5.5: Clusters for RoBERTa Trained on CAV with Conversation Content Control. We display one example for 4 out of 7 clusters. We mention noticeable consistencies within the cluster (“Consistencies” column).

character-level consistencies—likely because they are easiest to spot manually. We expect representations to also capture more complex stylometric information because of their performance on the authorship verification and STEL tasks (Section 5.4). Future work could analyze whether and what other stylistic consistencies are represented by the models.

For comparison, we also cluster with the base RoBERTa model (see Appendix C.4). The only three interesting RoBERTa clusters (i.e., clusters 2,3,4 that contain more than three elements and not as many as 86.7% of all utterances), seem to mostly differ in utterance length (average number of characters are 15 in Cluster 2 vs. in 1278 in Cluster 3) and in the presence of hyperlinks (84% of utterances contain ‘https://’ in Cluster 4 vs. an overall average of 2%). Average utterance lengths are not as clearly separated by the clusters of the trained style representations.

III

5.6. Limitations and Future Work

We propose several directions for future research:

First, conversation labels are already inherently available in conversation corpora like `Reddit`. However, it remains a difficulty to transfer the conversation CC to non-conversation datasets. Moreover, even when using the conversation CC, content information might still be useful for authorship verification: If one person writes “my husband” and another writes “my wife” within the same conversation, it is highly unlikely that those utterances have been generated by the same person. With the recent advances in semantic sentence embeddings, it might be interesting to train style representations on CAV tasks with a new content control level: Two utterances could be labelled as having the same content if their semantic embeddings are close to each other (e.g., when cosine similarity is above a certain threshold).

Second, for the STEL-Or-Content task, the so-called “triplet problem” (Wegmann and Nguyen, 2021) remains a potential problem. Consider the example in Figure 5.2. Here, the STEL framework only guarantees that A1 is more informal than A2 and S2 is more informal than S1. Thus, in some cases A2 can be stylistically closer to A1 than S2. However, we expect this case to be less prevalent: A2 would need to be already pretty close in style to A1, or both S2 and S1 would need to be substantially more informal or formal

than A1. In the future, removing problematic instances could alleviate a possible maximum performance cap.

Third, the representation models may learn to represent individual stylistic variation as we use utterances from the same individual author as positive signals (cf. Zhu and Jurgens, 2021). However, because the representation models learn with same author pairs that are generated from thousands of authors, it is likely that they also learn consistencies along groups of authors that use similar style features (e.g., demographic groups based on age or education level, or subreddit communities). Future work could explore how different CC levels and training tasks influence the type of styles that are learned.

5.7. Conclusion

Recent advances in the development of style representations have increasingly used training objectives from authorship verification (Hay et al., 2020; Zhu and Jurgens, 2021). However, representations that perform well on the Authorship Verification (AV) task might do so not because they represent style well but because they latch on to spurious content correlations. We train different style representations by controlling for content (CC) using conversation or domain membership as a proxy for topic. We also introduce the new Contrastive Authorship Verification setup (CAV) and compare it to the usual AV setup. We evaluate our fine-tuned RoBERTa models on the recent STEL framework (Wegmann and Nguyen, 2021) and surprisingly find that the fine-tuned models might have unlearned some style information compared to the RoBERTa base model. We propose an original adaptation of STEL to test whether learned representations favor style over content information. We find that representations that were trained on the CAV setup with conversation CC represent style in a way that is more independent from content than other fine-tuned models or the RoBERTa base model. We demonstrate some of the learned stylistic differences via agglomerative clustering—e.g., the use of a right single quotation mark vs. an apostrophe in contractions.

In this work, we have introduced a successful approach for training neural text representations that are independent from content. Our findings demonstrate that style representations trained with the CAV setup and conversation CC capture stylistic variation more independently from content than previous approaches. To the best of our knowledge, this is the first work to evaluate neural text representations on their ability to encode linguistic style while disentangling it from content. Additionally, are are the first to release stylistic text representations for broad use on the Hugging Face Hub.

Beyond the technical contributions, this work underscores the broader need for better style representations in NLP. Future research should continue refining methods for isolating style from content and improving evaluation methods like STEL to better represent stylistic nuances.

Ethical Considerations

We use utterances taken from 100 subcommunities (i.e., subreddits) of the popular online platform Reddit to train style representations with different training tasks and compare their performance. With our work, we aim to contribute to the development of general style representations that are disentangled from content. Style representations have the potential to increase classification performance for diverse demographics and social groups (Hovy, 2015).

The user demographics on the selected 100 subreddits are likely skewed towards particular demographics. For example, locally based subreddits (e.g., Canada, Singapore) might be over-represented. Generally, the average Reddit user is typically more likely to be young and male.[12] Thus, our representations might not be representative of (English) language use across different social groups. However, experiments on the set of 100 distinct subreddits should still demonstrate the possibilities of the used approaches and methods. We hope the ethical impact of reusing the already published Reddit dataset (Baumgartner et al., 2020; Chang et al., 2020) to be small but acknowledge that reusing it will lead to increased visibility of data that is potentially privacy infringing. As we aggregate the styles of thousands of users to calculate style representations, we expect it to not be indicative of individual users.

We confirm to have read and that we abide by the ACL Code of Ethics.

Acknowledgements

We thank the anonymous ARR reviewers for their helpful comments. This research was supported by the "Digital Society - The Informed Citizen" research programme, which is (partly) financed by the Dutch Research Council (NWO), project 410.19.007. We thank the Utrecht NLP Group for discussions and feedback on writing and presentation. Dong Nguyen was supported by the research programme Veni with project number VI.Veni.192.130, which is (partly) financed by the Dutch Research Council (NWO).

[12] https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Impact of STEL and Style Embedding Work

Since the STEL framework (Wegmann and Nguyen, 2021) and our Style Embedding work (Wegmann et al., 2022) were published, our trained model has been actively used, with an average of 1-4k model downloads per month, reaching up to 400k before the ACL 2025 submission deadline. Our work attracted some attention from NLP researchers: Some work use STEL for evaluation (Chim et al., 2025; Patel et al., 2023), e.g., of their own style embeddings. Others use our model to automatically evaluate if a generated text is close to a target style (Liu et al., 2023; Khan et al., 2023; Horvitz et al., 2024a). Several works use our model as a baseline on different tasks relating to authorship attribution (Saxena et al., 2023; Huertas-Tato et al., 2024; Soto et al., 2024; Aggazzotti et al., 2024; Michel et al., 2024), e.g., detection of machine-generated texts. Some works incorporate our model in text generation pipelines to manipulate the style of the generated text (Horvitz et al., 2024a,b; Liu et al., 2023), e.g., for style transfer.

Part IV

Paraphrasing Across Speakers

So far, I have defined NLP tasks requiring robustness (e.g., paraphrase classification) and NLP tasks requiring sensitivity to language variation (e.g., authorship verification). I showed that language variation might be important to consider at all stages of building LLMs—including tokenizers—and that neural text representations have the potential to better represent at least one aspect of language variation: linguistic style. In Part IV of this dissertation, I focus on a task requiring robustness to language variation: detecting paraphrases across speaker turns in dialog. I provide annotation procedures that deal with disagreements among annotators, provide a dataset and experiment with decoder and encoder models to detect paraphrases computationally. For paraphrase span identification, encoder models profit from not being able to hallucinate quotes. This part demonstrates that both humans and NLP models face significant challenges when finding similarities in the referential meaning of utterances that vary in language across speakers.

Contents

IV

6

What's Mine becomes Yours: Defining, Annotating and Detecting Context-Dependent Paraphrases in News Interview Dialogs

This chapter is based on Wegmann, A., van den Broek, T. & Nguyen, D. (2024). What's Mine becomes Yours: Defining, Annotating and Detecting Context-Dependent Paraphrases in News Interview Dialogs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 882-–912). https://doi.org/10.18653/v1/2024.emnlp-main.52. See a video of the conference presentation here: https://youtu.be/9H-YD7J0kvM.

Abstract

Best practices for high conflict conversations like counseling or customer support almost always include recommendations to paraphrase the previous speaker. Although paraphrase classification has received widespread attention in NLP, paraphrases are usually considered independent from context, and common models and datasets are not applicable to dialog settings. In this work, we investigate paraphrases across turns in dialog (e.g., Speaker 1: "That book is mine." becomes Speaker 2: "That book is yours."). We provide an operationalization of context-dependent paraphrases, and develop a training for crowd workers to classify paraphrases

Author contributions: AW developed the idea, prepared the data, implemented the experiments, and wrote the manuscript. TB and DN provided supervision and feedback throughout the entire research process.

Guest: And people always prefer, of course, to see the pope as the principal celebrant of the mass. So that's good. That'll be tonight. And it will be his 26th mass and it will be the 40th or, rather, the 30th time that this is offered in round the world transmission. And it will be my 20th time in doing it as a television commentator from Rome so.
Host: Yes, you've been doing this for a while now.

Figure 6.1: Context-Dependent Paraphrase in a News Interview. The interview host paraphrases part of the guest's utterance. It is only a paraphrase in the current context (e.g., *doing something 20 times* and *doing something for a while* are not generally synonymous). Our annotators provide word-level highlighting. The color's intensity shows the share of annotators that selected the word. Here, most annotators selected the same text spans; some included "from Rome" as part of what is paraphrased by the host. We underline the paraphrase identified by our fine-tuned DeBERTa token classifier.

in dialog. We introduce `ContextDeP`, a dataset with utterance pairs from NPR and CNN news interviews annotated for context-dependent paraphrases. To enable analysis on label variation, the dataset contains 5,581 annotations on 600 utterance pairs. We present promising results with in-context learning and with token classification models for automatic paraphrase detection in dialog.

6.1. Introduction

IV

Repeating or paraphrasing what the previous speaker said has time and time again been found to be important in human-to-human or human-to-computer dialogs: It encourages elaboration and introspection in counseling (Rogers, 1951; Miller and Rollnick, 2013; Hill, 1992; Shah et al., 2022), can help deescalate conflicts in crisis negotiations (Vecchi et al., 2005; Voss and Raz, 2016; Vecchi et al., 2019), can have a positive impact on relationships (Weger Jr. et al., 2010; Roos, 2022), can increase the perceived response quality of dialog systems (Weizenbaum, 1966; Dieter et al., 2019) and generally provides tangible understanding-checks to ground what both speakers agree on (Clark, 1996; Jurafsky and Martin, 2025).

Fortunately, in NLP, paraphrases have received wide-spread attention: Researchers have created numerous paraphrase datasets (Dolan and Brockett, 2005; Zhang et al., 2019; Dong et al., 2021; Kanerva et al., 2023), developed methods to automatically identify (Zhang et al., 2019; Wei et al., 2022a; Zhou et al., 2025), and generate paraphrases (Wubben et al., 2010; Zhou and Bhat, 2021), and used paraphrase datasets to train semantic sentence representations (Reimers and Gurevych, 2019; Gao et al., 2021) and benchmark LLMs (Wang et al., 2018; Srivastava et al., 2023). However, most previous work (1) has focused on context-independent paraphrases, i.e., texts that are semantically equivalent independent from the given context, and has not investigated the automatic detection of paraphrases across turns in dialog, (2) has classified paraphrases at the level of full texts even though paraphrases often only occur in portions of larger texts (see also Figure 6.1), (3) uses a small number of 1–3 annotations per paraphrase pair (Dolan and Brockett, 2005; Kanerva et al., 2023), (4) only annotate text

Dataset	Agreement Acc.	Agreement α	Single Example with High Variation Shortened Example	Vote
BALANCED	0.71	0.32	**Guest:** [...] Maybe the money will help. **Host:** It can't hurt, let's put it that way.	9/20
RANDOM	0.72	0.23	**G:** So both parties agree that we need to stop horrific acts of violence against animals. But everyone is standing behind this. It is time to stop horrific acts of brutality on animals. **H:** Britain's Queen Elizabeth's senior dresser writes "If her majesty is due to attend an engagement in particularly cold weather from 2019 onwards fake fur will be used to make sure she stays warm." it's a very stark example of a monarch following public opinion in the U.K. which is moving away from fur and it very much embraces orange85basehl!85.714285714285714prevention of cruelty to the animals.	7/15
PARA	0.65	0.19	**G:** [...] it could be programmed in. But again, you'd have to set that up as part of your flight plan. **H:** So you'd have to say I'm going to drop to 5,000 feet, then go back up to 35,000 feet, and you would have had to have done that at the beginning.	8/15

Table 6.1: Agreement Scores as an Indicator of Plausible Variation. For each dataset, we display the "accuracy" with the majority vote (Acc.) which is the mean overlap of a rater's classification with the majority vote classification excluding the current rater and Krippendorff (2019)'s alpha (α) for the binary classifications by all raters over all pairs. The relatively low K's α scores can be explained by pairs where either label is plausible. We display such an example for each dataset with the share of annotators classifying it it as a paraphrase (Vote).

IV

pairs that are "likely" to include paraphrases using heuristics such as lexical similarity (Dolan and Brockett, 2005), although, especially for the dialog setting, we can not expect lexical similarity to be high for all or even most paraphrase pairs (e.g., the pair in Figure 6.1 only overlaps in two words) and (5) either use short annotation instructions (Dolan and Brockett, 2005) that rely on annotator intuitions or long and complex instructions (Kanerva et al., 2023) that limit the total number of annotators.

We address all five limitations with this work. First, we are, to the best of our knowledge, the first to focus on operationalizing, annotating and automatically detecting **context-dependent paraphrases across turns in dialog**. Dialog is a setting that is uniquely sensitive to context (Grice, 1957, 1975; Davis, 2003), e.g., "doing this for a while now" and "20th time [...] as a television commentator" in Figure 6.1 are not generally semantically equivalent. Second, instead of classifying whether two complete texts A and B are paraphrases of each other, we focus on classifying whether there exists a selection of a text B that paraphrases a selection of a text A, and **identifying the text spans that constitute the paraphrase pair** (e.g., Figure 6.1). Third, we collect a **larger number of annotations** of up to 21 per item in line with typical efforts to address plausible human label variation (Nie et al., 2020b; Sap et al., 2022). Even though context-dependent paraphrase identification in dialog might at first seem straight forward with a clear ground truth, similar to other "objective" tasks in NLP (Uma et al., 2021), human annotators (plausibly) disagree on labels (Dolan and Brockett, 2005; Kanerva et al., 2023). For example, consider the first text pair in Table 6.1. "[The money] can't hurt" can be

interpreted in at least two different ways: as a statement with approximately the same meaning as "the money will help" or as an opposing statement meaning the money actually won't help but at least "It can't hurt" either. Fourth, instead of using heuristics to select text pairs for annotations, we choose a dialog setting where paraphrases are relatively likely to occur: transcripts of **NPR and CNN news interviews** (Zhu et al., 2021) since in (news) interviews paraphrasing or more generally active listening is encouraged (Clayman and Heritage, 2002; Hight and Smyth, 2002; Sedorkin, 2020). While the interview domain shows some unique characteristics limiting generalizability (e.g., hosts using paraphrases to simplify the guest's statements for the audience), the interview domain is is suitable to demonstrate our new task and includes a diverse set of topics and guests. Fifth, we develop an annotation procedure that goes beyond relying on intuitions and is scalable to a large number of annotators: an accessible **example-centric, hands-on, 15-minute training** before annotation.

In short, we operationalize context-dependent paraphrases in dialog with a definition and an iteratively developed hands-on training for annotators. Then, annotators classify paraphrases and identify the spans of text that constitute the paraphrase. We release `ContextDeP` (Context-Dependent Paraphrases in news interviews), a dataset with 5,581 annotations on 600 utterance pairs from NPR and CNN news interviews. We use in-context learning (ICL) with generative models like LLAMA 2 or GPT-4 and fine-tune a DEBERTA token classifier to detect paraphrases in dialog. We reach promising results of F1 scores from 0.73 to 0.81. Generative models perform better at classification, while the token classifier provides text spans without parsing errors. We hope to advance dialog based evaluations of LLMs and the detection of paraphrases in dialog. Code[1], annotated data[2,3] and the trained model[4] are publicly available for research purposes.

IV

6.2. Related Work

Paraphrases have most successfully been classified by encoder architectures with fine-tuned classification heads (Zhang et al., 2019; Wahle et al., 2023) and more recently using in-context learning with generative models like GPT-3.5 and LLAMA 2 (Wei et al., 2022a; Wang et al., 2022b; Wahle et al., 2023). To the best of our knowledge, only Wang et al. (2022a) go beyond classifying paraphrases at the complete sentence level. They use a DEBERTA token classifier to highlight text spans that are not part of a paraphrase, i.e., the reverse of our task. However, they only consider context-independent paraphrases taken and generated based on the MRPC dataset (Dolan and Brockett, 2005).

Paraphrase taxonomies commonly go beyond binary classifications to make finer distinctions between different types of paraphrases. These distinctions sometimes depend on the context in which the text pair appears. Context can refer to different

[1] `https://github.com/nlpsoc/Paraphrases-in-News-Interviews`
[2] `https://huggingface.co/datasets/AnnaWegmann/Paraphrases-in-Interviews`
[3] This is in line with the license from the original data publication Zhu et al. (2021).
[4] `https://huggingface.co/AnnaWegmann/Highlight-Paraphrases-in-Dialog`

What?	Shortened Examples
Clear Contextual Equivalence ⊆ CP	Guest: I know they are cruel. Host: You know they are cruel.
	Guest: We have been the punching bag of the president. Host: The president has been using Chicago as a punching bag.
Approximate Contextual Equivalence ⊆ CP	Guest: I'm like, "Fortnite", what is that? I don't even know what it is – Host: So, you weren't even familiar?
	Guest: My wife is going through the same thing herself. Host: She's also looking for work.

Table 6.2: Contextual Paraphrases (CP). We include text spans (⊆ CP) that range from clear to approximate equivalence for the given context. Few examples are very clear. Deciding between approximate equivalence and non-equivalence turns out to be a difficult task. In our dataset, annotator agreement scores can be used as a proxy for the ambiguity of an item.

things, for example, the document a sentence appears in (Kanerva et al., 2023), general encyclopedic knowledge (e.g., "Penelope" was the queen of Ithaca and wife of Odysseus in Greek mythology in Vila et al., 2014) and situational knowledge (e.g., "here" refers to "Paris" in Vila et al., 2014). Bhagat and Hovy (2013) and Kovatchev et al. (2018) emphasize that the context matters in determining what lexical operations, such as word substitutions, create paraphrases. Shwartz and Dagan (2016) show that context information can even reverse semantic relationships between phrases. Vila et al. (2014) discuss text pairs that are equivalent only when one presupposes encyclopedic or situational knowledge, but exclude them as non-paraphrases. Further, to the best of our knowledge, most previous work annotate sentence pairs without considering the document context, with Kanerva et al. (2023) being the only exception, and no previous work looking at detecting paraphrases in dialog.

Dialog act taxonomies aim to classify the communicative function of an utterance in dialog and commonly include acts such as `Summarize/Reformulate` (Stolcke et al., 2000; Core and Allen, 1997). However, generally, communicative function can be orthogonal to meaning equivalence. For example, the paraphrase from Table 6.2 "So you weren't even familiar?" would probably be a `Declarative Yes-No-Question` dialog act (Stolcke et al., 2000), while the non-paraphrase "So you don't have a problem with ... ?" in Table 6.3 would also be a `Declarative Yes-No-Question`. We see paraphrase detection in dialog as complementary to investigating communicative function of utterances.

6.3. Context-Dependent Paraphrases in Dialog

In NLP, paraphrases typically are pairs of text that are approximately equivalent in meaning (Bhagat and Hovy, 2013), since full equivalence usually only applies for practically identical strings (Bhagat and Hovy, 2013; Dolan and Brockett, 2005) —with some scholars even claiming that different sentences can never be fully equivalent in meaning (Bhagat and Hovy, 2013; Clark, 1992; Bolinger, 1974). The field of NLP has mostly focused on paraphrases that are *context-independent*, i.e., approximately equivalent

What?	Shortened Example
Additional Conclusions or Facts $\not\subseteq$ CP	Guest: If you're not in our country, there are no constitutional protections for you. Host: So, you don't have a problem with Facebook giving the government access to the private accounts of people applying to enter the U.S.?
Isolated Equivalence $\not\subseteq$ CP	Guest: There are militant groups out there firing against the military. Host: Why did the army decide today to move in and clear out the camp?

Table 6.3: Non-Paraphrases in Dialog. We do not include text pairs ($\not\subseteq$ CP) that are semantically related but where the second speaker does not actually rephrase a point the first speaker makes. Frequent cases are text spans that might only be considered approximately equivalent when taken out of context (underlined) and pairs that have too distant meanings, for example, when the interviewer continues with the same or a related topic but adds further-reaching conclusions or new facts.

without considering a given context (Dolan and Brockett, 2005; Wang et al., 2018; Zhang et al., 2019). Some studies have operationalized paraphrases using more fine-grained taxonomies, where context is sometimes considered (Bhagat and Hovy, 2013; Vila et al., 2014; Kovatchev et al., 2018). However, only a few datasets include such paraphrases (Kovatchev et al., 2018; Kanerva et al., 2023) and to the best of our knowledge none that focus on context-dependent paraphrases or dialog data.

We define a **context-dependent paraphrase** as two text excerpts that are at least approximately equivalent in meaning in a given situation but not necessarily in all non-absurd situations.[5] For example, consider the first exchange in Table 6.2. In this situation, "I" uttered by the first speaker and "You" uttered by the second speaker are clearly signifying the same person. However, if uttered by the same speaker "I" and "you" probably do not signify the same person. The text pair in Table 6.2 is thus equivalent in at least one but not in all non-absurd situations. The text excerpts forming context-dependent paraphrases do not have to be complete utterances. In many cases they are portions of utterances, see the highlights in Figure 6.1. Note that in dialog settings, the second speaker should rephrase part of the first speaker's point in the given situation (*context* condition) and not just talk about something semantically related (*equivalence* condition).

Context-dependent paraphrases range from clear (first example in Table 6.2) to approximate contextual equivalence (last example in Table 6.2). When the guest says "My wife is going through the same thing", it seems reasonable to assume that the host is using contextual knowledge to infer that "the same thing" and "looking for a job" are equivalent for the given exchange. Even though in this last example the meaning of the two utterances could also be subject to different interpretations, we still consider such cases to be context-dependent paraphrases for two reasons: (1) similar to findings in context-independent paraphrase detection, limiting ourselves to very clear cases would mostly result in uninteresting, practically identical strings and (2) we ultimately want to identify paraphrases in human dialog, which is full of implicit contextual meaning (Grice, 1957, 1975; Davis, 2003).

[5] This definition combines elements from Kanerva et al. (2021) and Bhagat and Hovy (2013)

	Preprocessed		Sampled		Released	
	# i	# gh	# i	# gh	# i	# gh
all	34419	148522	1304	4450	480	600
NPR	11506	49065	423	1550	167	218
CNN	22913	99457	881	2900	313	382

Table 6.4: Dataset Statistics. Number of interviews (#i) and (guest, host)-pairs (# gh) respectively after preprocessing (Section 6.4.1), random sampling (Section 6.4.2) and the selection of paraphrase candidates for annotation (Section 6.4.2).

We specifically exclude common cases of disagreements between annotators[6] that we consider not to be context-dependent paraphrases in dialog, see Table 6.3. First, we exclude text spans that might be considered approximately equivalent when they are looked at in isolation but do not represent a paraphrase of the guest's point in the given situation (e.g., "the military" and "the army" in Table 6.3). Second, we exclude text pairs that diverge too much from the original meaning when the second speaker adds conclusions, inferences or new facts. In an interview setting, journalists make use of different question types and communication strategies relating to their agenda (Clayman and Heritage, 2002) that can sometimes seem like paraphrases. For example in Table 6.3, the host's question "So, you ...?" could be read as a paraphrase with the goal of checking understanding with the guest. However, it is more likely to be a declarative conclusion that goes beyond what the guest said.

IV

6.4. Dataset

Generally, people do not paraphrase each other in every conversation. We focus on the news interview setting, because paraphrasing, or more generally active listening, is a common practice for journalists (Clayman and Heritage, 2002; Hight and Smyth, 2002; Sedorkin, 2020). We therefore also only consider whether the journalist (the interview host) paraphrases the interview guest and not the other way around. We use Zhu et al. (2021)'s *MediaSum* corpus which consists of over 450K news interview transcripts and their summaries from 1999–2019 NPR and 2000–2020 CNN interviews.[7] The topics of the CNN and NPR news interviews (Zhu et al., 2021) are broadly centered around U.S. politics (e.g., presidential or local elections, 9/11, foreign policy in the middle east), sports (e.g., baseball, football), domestic natural disasters or crimes and popular culture (e.g., interviews with book authors).

We perform several preprocessing and sampling steps on the NPR and CNN news interview dataset. In Table 6.4, we display the number of interviews and speaker turns remaining after preprocessing (Section 6.4.1), after random sampling for lead author annotations to find likely paraphrases ("Lead Author Annotation" in Section 6.4.2), and

[6] derived from pilot studies, see also Appendix D.3.1 and specifically Appendix Table D.3

[7] Released for research purpose, see `https://github.com/zcgzcgzcg1/MediaSum?tab=readme-ov-file`.

after random sampling for final annotations by crowd workers ("Paraphrase Candidate Selection" in Section 6.4.2).

6.4.1. Preprocessing

We only include two-person interviews, i.e., a conversation between an interview host and a guest. We remove interviews with fewer than four turns, utterances that only consist of two words or of more than 200 words, and the first and last turns of interviews (often welcoming addresses and goodbyes). Overall, this leaves 34,419 interviews with 148,522 (guest, host)-pairs. See Appendix D.2.1 for details on how we execute these preprocessing steps.

6.4.2. Data Samples for Annotation

Even though paraphrases are relatively likely in the news interview setting, most randomly sampled text pairs still do not include paraphrases. To distribute annotation resources to text pairs that are likely to be paraphrases, previous work usually selects pairs based on heuristics like textual similarity features, e.g., word overlap, edit distance, or semantic similarity (Dolan and Brockett, 2005; Su and Yan, 2017; Dong et al., 2021). These approaches are systematically biased towards selecting more obvious cases, e.g., text pairs that are lexically similar. However, this might exclude many context-dependent paraphrases. For example, the guest and host utterance in Figure 6.1 have varying lengths, only overlap in three words and have a semantic similarity score of only 0.13[8]. Similar to Kanerva et al. (2023), we instead use a manual selection of promising text pairs for annotation: We (1) randomly sample a set of text pairs and (2) manually classify each of them to (3) select three sets of text pairs that vary in their paraphrase distribution for the more resource-intensive crowd-sourced annotations: the RANDOM, BALANCED and PARA set.

Lead author annotation We shuffle and uniformly sample 1,304 interviews. For each interview, we sample a maximum of 5 consecutive (guest, host)-pairs. To select promising paraphrase candidates, the lead author then manually classifies all 4,450 text pairs as paraphrases vs. non-paraphrases (see Appendix D.2.2 for details).[9] In total, about 14.9% of the sampled text pairs are classified as paraphrases by the lead author. On a random set of 100 (guest, host)-pairs (RANDOM), we later compare the lead author's classifications with the crowd-sourced paraphrase classifications (see Appendix D.2.2). 89% of the lead author's classifications are the same as the crowd majority. Note that the lead author's classifications do not affect the quality of the annotations released with the dataset but only the text pairs that are selected for annotation. However, using lead author annotations instead of lexical level heuristics should increase paraphrase diversity in the released dataset beyond high lexical similarity pairs.

[8] using cosine-similarity and encodings from `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

[9] After experimenting with crowd workers, having a first pass for selection done by one of our team seemed the best considering cost-performance trade-offs.

Dataset	size	# paraphrases	# anns/item
BALANCED	100	54	20.1
RANDOM	100	13	5.7
PARA	400	254	7.5
Total	600	321	9.3

Table 6.5: Dataset Statistics. For each dataset, we display the size, the number of paraphrases according to the majority vote and the average annotations per text pair.

Paraphrase candidate selection We sample three datasets for annotation that differ in their estimated paraphrase distributions (based on the lead author annotations): **BALANCED** is a set 100 text pairs sampled for equal representation of paraphrases and non-paraphrases. We annotate this dataset first with a high number of annotators per (guest, host)-pair, to decide on a crowd worker allocation strategy that performs well for paraphrases as well as non-paraphrases. **RANDOM** is a uniform random sample of 100 text pairs. One main use of the dataset is to evaluate the quality of crowd worker annotations on a random sample. **PARA** is a set of 400 text pairs with an estimated 84% of paraphrases according to lead author annotations designed to increase the variety of paraphrases in our dataset. Details on the sampling of the three datasets can be found in Appendix D.2.3.

6.5. Annotation

We first describe the annotation task (Section 6.5.1). Then, we discuss why the annotation task is difficult and a clear ground truth classification might not exist in many cases (Section 6.5.2). Therefore, we dynamically collect many judgments for text pairs that show initial high disagreements (Section 6.5.4). The annotation of utterance pairs takes place in two rounds with `Prolific` crowd workers: (1) training crowd workers (Section 6.5.3) and (2) annotating paraphrases with trained crowd workers (Section 6.5.4 and Section 6.5.5).

6.5.1. Annotation Task

Given a (guest, host) utterance pair, annotators (1) classify whether the host is paraphrasing any part of the guest's utterance and, if so, (2) highlight the paraphrase in the guest and host utterance. This results in data points like the one in Figure 6.1. Note that our setup differs from prior work, which usually involves classifying whether an entire text B is a paraphrase of an entire text A (e.g., Dolan and Brockett, 2005). Instead, given texts A and B, our task is to determine whether there exists a selection of words from text B that is a paraphrase of a selection of text A. Our annotators are not only performing binary classification, but they also highlight the position of the paraphrase. To the best of our knowledge, we are the first to approach paraphrase detection in this way. Moreover, in contrast to previous work, the considered text pairs are usu-

ally longer than just one sentence and are dialog turns. We also provide context in the form of a date, names and an interview summary.

6.5.2. Plausible Label Variation

There can already be disagreements for relatively "easy" semantic tasks like annotating whether a sentence is about a certain topic (Andresen et al., 2020). An even more difficult task is annotating context-independent paraphrases. Disagreements between human annotators are common (Dolan and Brockett, 2005; Krishna et al., 2020; Kanerva et al., 2023)—even with extensive manuals for annotators (Kanerva et al., 2023). In related semantic tasks like textual entailment,[10] disagreements have been linked to plausible label variations inherent to the task (Pavlick and Kwiatkowski, 2019; Nie et al., 2020b; Jiang and de Marneffe, 2022).

Our task setup adds further challenges: First, instead of classifying full sentence pairs, annotators have to read relatively long texts and decide whether any portion of the text pair is a paraphrase. Second, while in previous work annotators usually had to decide if two texts are generally approximately equivalent, they now need to identify paraphrases in a highly contextual setting with often incomplete information.

As a result, similar to the task of textual entailment, we expect classifying context-dependent paraphrases in dialog to not always have a clear ground truth. We display examples of plausible label variation in Table 6.1. To handle label variation, common strategies are performing quality checks with annotators (Jiang and de Marneffe, 2022) and recruiting a larger number of annotators for a single item (Nie et al., 2020b; Andresen et al., 2020; Sap et al., 2022). We do both, see our approach in Section 6.5.3 and Section 6.5.4.

6.5.3. Annotator Training

When annotating paraphrases, the instructions for annotators are often short, do not explain challenges and rely on annotator intuitions (Dolan and Brockett, 2005; Lan et al., 2017).[11] In contrast, Kanerva et al. (2023) recently used an elaborate 17-page manual. However, they relied on a total of only 6 expert annotators across all tasks with up to three annotations per task (Section 6.5.2). We aim for a trade-off between short intuition-based and long complex instructions that facilitates recruitment of a larger number of annotators: an accessible example-centric, hands-on 15-minute training of annotators that teaches our operationalization of context-dependent paraphrases (Section 6.3). See Appendix Section D.3.1 for the iterative development process of the annotator training, among others based on disagreements between lab members and crowd workers in several pilot rounds.

See Appendix Section D.3.2 for the exact instructions we settled on for annotator training. We provide (1) a short paraphrase definition, (2) examples of context-dependent

[10] Paraphrase classification has been frequently equated to (bi-)directional entailment classification (Dolan and Brockett, 2005; Androutsopoulos and Malakasiotis, 2010)

[11] For example, instructions are to rate if two sentences "mean the same thing" (Dolan and Brockett, 2005) or are "semantically equivalent" (Lan et al., 2017).

Shortened Examples
Guest: we don't really know what went into their algorithm to make it turn out that way. **Host:** We're talking about algorithms, but should we be talking about the humans who design the algorithms?
Guest: In Harrison County. **Host:** In Harrison County. Are you [...]

Table 6.6: Low Quality Annotations. We show human highlights that can be considered wrong or noisy. When absent, we underline the correct highlights.

paraphrases showing clear and approximate equivalence (cf. Table 6.2), (3) examples of common difficulties with paraphrase classification in dialog (cf. Table 6.3 and Section 6.3), and use (4) a hands-on approach where annotators have to show understanding by immediately classifying and highlighting paraphrases after receiving a new set of instructions. Only once they make the right choice on what is (Table 6.2) and is not a paraphrase (Table 6.3) and highlight the correct spans they are shown the next set of instructions. Only annotators that undergo the full training and pass two comprehension and two attention checks are part of our released dataset. Overall, 49% of the annotators who finished the training passed it.

6.5.4. Annotator Allocation

To the best of our knowledge, text pairs in paraphrase datasets receive a fixed number of 1, up to a maximum of 5 annotations (Kanerva et al., 2023; Zhang et al., 2019; Lan et al., 2017; Dolan and Brockett, 2005). However, this might not be enough to represent the inherent plausible variation to the task (Section 6.5.2). We have each pair in BALANCED annotated by 20–21 trained annotators to simulate different annotator allocation strategies (Appendix D.3.5). Then, for RANDOM and PARA, we use a dynamic allocation strategy: Each pair receives at least 3 annotations. We dynamically collect more annotations, up to 15, on pairs with high disagreement (i.e., entropy > 0.8). Overall, this results in an average of 9 annotations per text pair across our released dataset.

6.5.5. Results

We discuss annotations results (Tables 6.1, 6.5, 6.7) on our datasets BALANCED, RANDOM and PARA.

Classification agreement as an indicator of variation The agreement between annotators on whether a paraphrase is present is relatively low (Table 6.1). We inspect a sample of 100 annotations on the RANDOM set and manually assess annotation quality. 90% of the annotations can be said to be at least plausible (see Table 6.6 for low quality and Table 6.1 for plausible variation examples), which is in line with the fact that we only use high quality annotators (Section 6.5.3). Further, we manually analyze the 42 annotations of ten randomly sampled annotators: Nine annotators consistently provide high quality annotations, while the other annotator chooses "not a

Dataset	Guest		Host	
	α	$\frac{A \cap B}{A \cup B}$	α	$\frac{A \cap B}{A \cup B}$
BALANCED	0.42	0.51	0.48	0.63
RANDOM	0.53	0.63	0.53	0.64
PARA	0.43	0.50	0.50	0.64

Table 6.7: Agreement on highlights. For pairs that at least two annotators classified a paraphrase, we display the average lexical overlap between the highlights (Jaccard Index displayed as $\frac{A \cap B}{A \cup B}$) and Krippendorff's unitizing α over all words for guest and host highlights, see Krippendorff (1995).

paraphrase" a few times too often (see Appendix D.3.8 for details). As a result, we assume that most disagreements are due to the inherent plausible label variation of the task (Section 6.5.2).

Higher agreement on paraphrase position Krippendorff's unitizing α on the highlights is higher than in other tasks[12] (see Table 6.7). We also calculate the "Intersection-over-union" between the highlighted words (i.e., Jaccard Index), a common and interpretable evaluation measure for annotator highlights (Herrewijnen et al., 2024; Mendez Guzman et al., 2022; Mathew et al., 2021; Malik et al., 2021). It seems that while annotations vary on whether there is a paraphrase or not, they agree frequently on the position of the possible paraphrase. On average, at least 50% of the highlighted words are the same between annotations.[13] Agreement is higher on the host utterance, because on average the host utterance is shorter than the guest utterance (33 < 85 words).

Label variation is highest for paraphrases Between the datasets, classification agreement is lowest for PARA. This is what we expected since it has the largest portion of "hard" non-repetition paraphrases (see Appendix D.2.3). Krippendorff's α is lower for the RANDOM than the BALANCED set, even though we expected the RANDOM set to include easier decisions for annotators (RANDOM includes more unrelated non-paraphrases, see Appendix D.2.3). As the other agreement heuristic is relatively high on RANDOM, the lower α values could be a result of Krippendorff's measure being sensitive to imbalanced label distributions (Riezler and Hagmann, 2022), see also Table 6.5 displaying the imbalanced distribution for RANDOM.

[12] E.g., 0.41 for hate speech (Carton et al., 2018) or 0.35 for sentiment analysis (Sullivan Jr. et al., 2022). Because of the different tasks these values are not exactly comparable.

[13] 100% overlap in highlighting is uncommon. DeYoung et al. (2020) consider two highlights a match if Jaccard is greater than 50%.

Split	# (guest, host)-pairs	# annotations
Train	420	3896
Dev	88	842
Test	92	843
Total	600	5,581

Table 6.8: Split of Dataset. For each set, we show the number of text pairs and the total number of annotations.

6.6. Modeling

In Table 6.8, we do a random 70, 15, 15 split of our 5,581 annotations, along the 600 unique pairs.

Token classifier Similar to Wang et al. (2022a), we fine-tune a large DeBERTa model[14] (He et al., 2021) on token classification to highlight the paraphrase positions (for hyperparameters, see Appendix D.4.2). We train two models: using all 3,896 training annotations ("ALL" in Table 6.9) and using deduplicated training annotations over the 420 unique (guest, host) training pairs ("AGGREGATED" in Table 6.9). We consider a model to have predicted a paraphrase for a pair if at least one token is highlighted with softmax probability ≥ 0.5 in both texts. For each model, we average performances over three seeds.

In-context learning We further prompt the following generative models (see URLs in Appendix D.4.1) to both classify and highlight the position of paraphrases: Llama 2 7B and 70B (Touvron et al., 2023), Vicuna 7B (Zheng et al., 2023), Mistral 7B Instruct v0.2 (Jiang et al., 2023), OpenChat 3.5 (Wang et al., 2024), Gemma 7B (Mesnard et al., 2024), Mixtral 8x7B Instruct v0.1 (Jiang et al., 2024) and GPT-4[15] (Achiam et al., 2023). We design the prompt to be as close as possible to the annotator training using a few-shot setup (Brown et al., 2020; Zhao et al., 2021) with all 8 examples shown during annotator training. We also provide explanations in the prompt (Wei et al., 2022b; Ye and Durrett, 2022) and use self-consistency by prompting the models 10 (GPT-4 and Llama 70B: 3) times (Wang et al., 2023b). For the prompt and further hyperparameter settings see Appendix D.4.1.

IV

[14] `microsoft/deberta-v3-large`
[15] API calls where performed using the "gpt-4" model id in March 2024.

	Classification				Highlighting		
Model	Extract ↓	F1 ↑	Prec ↑	Rec ↑	Extract ↓	Jacc Guest ↑	Jacc Host ↑
LLAMA 2 7B	1%	0.66	0.49	**0.98**	59%	0.34	0.44
VICUNA 7B	**1%**	0.29	0.67	0.19	32%	0.30	0.46
MISTRAL 7B INST.	3%	0.62	0.66	0.58	66%	0.40	0.51
OPENCHAT 3.5	**0%**	0.66	0.76	0.58	64%	0.46	0.50
GEMMA 7B	1%	0.64	0.66	0.63	48%	0.24	0.51
MIXTRAL 8X7B INST.	**0%**	0.74	0.73	0.74	65%	0.35	0.52
LLAMA 2 70B	**0%**	0.66	0.72	0.61	71%	0.29	0.56
GPT-4	**0%**	**0.81**	0.78	0.84	17%	**0.67**	**0.71**
DEBERTA AGG.	-	0.73	0.67	0.81	-	0.52	0.66
DEBERTA ALL	-	0.66	**0.82**	0.56	-	0.45	0.64

Table 6.9: Modeling Results. We **boldface** the best performance and underline the second best. We report the extraction error of predictions from generative models. For classification, we provide the F1, precision and recall score. For highlights, we include the Jaccard Index for both guest and host utterances. Higher values are better (↑) except for extraction errors (↓). GPT-4 is the best classification model. However, DEBERTA is the best highlight model as it does not lead to any extraction errors.

Results For evaluation, we consider a pair to contain a paraphrase if it has been classified by a majority of crowd workers and a word to be part of the paraphrase if it has been highlighted by a majority of crowd workers. We leave soft-evaluation approaches to future work (Uma et al., 2021), among others because of challenges in extracting label distributions for in-context learning in a straight-forward way (Hu and Levy, 2023; Lee et al., 2023). See Table 6.9 for test set performances. Performances for the token classifier are the mean over three seeds. Performances for the generative models are the majority vote for the 3–10 self-consistency calls. We display the F1 score for classification and, as before (Section 6.5.5), Intersection-Over-Union of the highlighted words for guest and host utterance highlights (Jaccard Indices), see, for example, DeYoung et al. (2020). For in-context learning, we also report how often we could not extract the highlights or classifications from model responses. Note that the test set contains 93 elements, so differences between models might appear bigger than they are.

Overall, GPT-4 and MIXTRAL 8X7B achieve the best results in paraphrase classification. For highlighting, our DEBERTA token classifiers and GPT-4 achieve the best overlap with human annotations. However, due to problems with extracting highlights from model responses (e.g., hallucinations, see Appendix D.4.3), our fine-tuned DEBERTA token classifiers are probably the best choice to extract the position of paraphrases. While the DEBERTA AGGREGATED model achieves higher F1 scores, the DEBERTA ALL model has the highest precision out of all models. We provide our best-performing DEBERTA AGGREGATED model (model with seed 202 and F1 score of 0.76) on the Hugging Face Hub[16] and use it in the following error analysis.

[16] https://huggingface.co/AnnaWegmann/Highlight-Paraphrases-in-Dialog

Predictions			Shortened Examples
crowd majority	GPT-4	DeBERTa	
✗	✗	✓	**Guest:** He was the most famous guy in the world of sports... **Host:** The most famous Italian...
✓	✗	✓	**Guest:** A lot of them were the Bay Area influx that came up and bought homes to flip. You know what flipping is, right? **Host:** Mm-hmm. Buying a house, improving it, selling it out of profit.

Table 6.10: Model Errors. We show examples of prediction errors made by DeBERTa and GPT-4. We display model predictions for paraphrases (✓) and non-paraphrases (✗) and compare it to the crowd majority. If one model predicted a paraphrase the corresponding text spans are underlined. For comparison, we also display the crowd majority highlights.

Error analysis We consider the best-performing classification and highlighting models for error analysis, i.e., GPT-4 and DeBERTa AGGREGATED. We manually analyze a sample of misclassifications, for examples see Table 6.10. Overall, the classification quality is better for GPT-4. The DeBERTa classifier finds more paraphrases (note that DeBERTa AGGREGATED for seed 202 has a recall of 0.86) but also predicts more false positives than GPT-4. For both models, the items with incorrect predictions also show higher human disagreement. The average entropy for human classifications is lower for the correct (0.45 for DeBERTa, 0.45 for GPT-4) than for the incorrect model predictions (0.59 for DeBERTa, 0.67 for GPT-4). DeBERTa highlights shorter spans of text (on average 6.6/6.2 words, compared to 16.7/10.9 for GPT-4 for guest/host respectively), while GPT-4 usually highlights complete (sub-)sentences. GPT-4 highlights are largely of good quality, however they often can not be extracted (see Appendix D.4.3). The DeBERTa highlights can seem "chopped up" and missing key information (e.g., the original host highlights in Table 6.11 are just "Rudy Giuliani", "coming" and "conversation"). We recommend classifying utterance pairs as a paraphrase when there exist softmax probabilities ≥ 0.5 for both guest and host utterance, but then selecting the highlights based on softmax probabilities lower than 0.5. Alternatively, the best DeBERTa ALL model[17] provides fewer but seemingly more consistent highlights (see Appendix D.4.3).

IV

[17] https://huggingface.co/AnnaWegmann/Highlight-Paraphrases-in-Dialog-ALL

Shortened Example
Guest: ... then he goes on and references and **makes mention of Rudy Giuliani three times** in this conversation **Host:** And **Rudy Giuliani** was a private lawyer not a government official, so why is **he coming up so much in this conversation** between two world leaders?

Table 6.11: Highlighting Differences. We show examples of highlights made by DeBERTa, **GPT-4** and human highlights. Lower intensity means fewer human annotators selected the word. While GPT-4 struggles with providing highlights at all (cf. extraction error in Table 6.9), DeBERTa highlights tend to be too sparse (just "Rudy Giuliani", "coming" and "conversation" in the host utterance). Here, we also highlight words, when the softmax probability is > 0.44[18] instead of ≥ 0.5. On the complete test set, this also increases the mean Jaccard Index (by 0.06/0.01 for guest/host compared to Table 6.9).

6.7. Conclusion

A majority of work on paraphrases in NLP has looked at the semantic equivalence of sentence pairs in context-independent settings. However, the human dialog setting is highly contextual and typical methods fall short. We provide an operationalization of context-dependent paraphrases and an up-scalable hands-on training for annotators. We demonstrate the annotation approach by providing 5,581 annotations on a set of 600 turn pairs from news interviews. Next to paraphrase classifications, we also provide annotations for paraphrase positions in utterances. In-context learning and token classification both show promising results on our dataset. With this work, we contribute to the automatic detection of paraphrases in dialog. We hope that this will benefit both NLP researchers in the creation of LLMs and social science researchers in analyzing paraphrasing in human-to-human or human-to-computer dialogues on a larger scale.

IV

Limitations

Even though the number of our unique text pairs is relatively small, we release a high number of high quality annotations per text pair (5,581 annotations on 600 text pairs). Releasing more annotations on fewer "items" (here: text pairs), has increasingly been more common in NLP (Nie et al., 2020b; Sap et al., 2022). Further, big datasets become less necessary with better generative models: Using only eight paraphrases pairs in our prompt already led to promising results. We further use the full 3,896 annotations from the training set to train a token classifier showing competitive results with the open generative models. However, the token classifier and other potential fine-tuning approaches would probably profit from a bigger dataset.

Even though our dataset of news interviews showed frequent, different and diverse occurrences of paraphrasing, it is likely not representative of paraphrasing behavior

[18] This is only for illustrative purposes. For the main results we used ≥ 0.5. For this figure, we selected 0.44 as it led to the biggest gain in the Jaccard Index on the test set.

in conversations across different contexts and social groups. In the future, we aim to expand our dataset with further out-of-domain items.

Our data creation process was not aimed at scalability. While our developed annotator training procedure can easily be scaled to a larger group of crowd workers, we manually selected text pairs for annotation. Future work could scale this by skipping manual selection and accepting a more imbalanced dataset or using our trained classifiers as a heuristic to identify likely paraphrases.

Even though we carefully prepared the annotator training and took several steps to ensure high-quality annotations, there remain several choices that were out of our scope to experiment with, but might have improved quality even more. For example, experimenting with different visualizations of paraphrase highlighting, text fonts, giving annotators an option to add confidence scores for classifications (e.g., as done in Andresen et al., 2020) and so on.

We only use one prompt that is as close as possible to the instructions the human annotators receive. We use the same prompt with the exact same formatting for all different generative LLMs. However, experimenting with different prompts might improve performance (Weng, 2023) and some models might benefit from certain formatting or phrasing. We leave in-depth testing of prompts to future work. Further, it might be possible to improve the performance of our DeBERTa model, through providing contextual information (like speaker names and interview summary). Currently, these are only provided to the generative models.

In this work we collect a high number of human annotations per item and highlight the plausible label variation in our dataset. However, we use hard instead of soft-evaluation approaches (Uma et al., 2021) for the computational models. We do this because, among others, extracting label distributions for in-context learning is challenging (Hu and Levy, 2023; Lee et al., 2023). We leave the development of a soft evaluation approach to future work but want to highlight the potential of our dataset here: The high number of annotations per item enables the modeling of classifications and text highlights as distributions, similar to Zhang and de Marneffe (2021). Further, our dataset provides anonymized unique ids for all annotators and enables modeling of different perspectives, e.g., with similar methods to Sachdeva et al. (2022) and Deng et al. (2023).

We do not differentiate between different communicative functions, intentions or strategies that affect the presence of paraphrases in a dialog. This is relevant as paraphrases might, for example, be a conscious choice by interviewers (Clayman and Heritage, 2002) or an unconscious occurrence similar to how speakers align their linguistic choices when referring to objects discussed (Xu and Reitter, 2015; Garrod and Anderson, 1987). We do not differentiate between intentional and unintentional[19] paraphrases and do not ask why a speaker utters a paraphrase in a given situation. Instead,

[19] We address an aspect of intentionality in our annotation process. In pilot studies, annotators occasionally labeled text spans as paraphrases merely because they referred to the same object with the same or a related term (cf. Appendix Table D.3). While these references might appear to be paraphrases when studied in isolation, they more likely present a practical conversation strategy to discuss similar topics rather than a deliberate attempt to rephrase what the guest said. To prevent such overidentification, our annotator

we provide an outline of the general class of context-dependent paraphrases in dialog that lays the groundwork for further, fine-grained distinctions.

Ethical Considerations

We hope that the ethical concerns of reusing a public dataset (Zhu et al., 2021) are minimal, especially since the CNN and NPR interviews are between public figures and were broadcast publicly, with consent, on national radio and TV.

Our dataset might not be representative of English paraphrasing behavior in dialogs across different social groups and contexts as it is taken from U.S. news interviews with public figures from two broadcasters. We caution against using our models without validation on out-of-domain data.

We performed several studies with U.S.-based crowd workers. We paid participants a median of ≈ 11.41$/h which is above federal minimum wage. crowd workers consented to the release of their annotations. We do not release identifying ids of crowd workers.

We confirm to have read and that we abide by the ACL Code of Ethics. Beside the mentioned ethical considerations, we do not foresee immediate risks of our work.

Acknowledgements

We thank the anonymous ARR reviewers for their constructive comments. Further, we thank the NLP Group at Utrecht University and, specifically, Elize Herrewijnen, Massimo Poesio, Kees van Deemter, Yupei Du, Qixiang Fang, Melody Sepahpour-Fard, Shane Kaszefski Yaschuk, Pablo Mosteiro, and Albert Gatt, for, among others, feedback on writing and presentation, discussions on annotator disagreement and testing multiple iterations of our annotation scheme. We thank Charlotte Vaaßen, Martin Wegmann and Hella Winkler for feedback on our annotation scheme. We thank Barbara Bziuk for feedback on presentation. This research was supported by the "Digital Society - The Informed Citizen" research programme, which is (partly) financed by the Dutch Research Council (NWO), project 410.19.007.

training instructs annotators to not highlight what we call "isolated" or "partial" equivalence (cf. annotator instruction in Appendix Figure D.7).

Part V

Closing Remarks

Contents

7

Conclusion

In this section, I summarize the main findings and implications of my dissertation (Section 7.1 & Section 7.2) and discuss future research directions inspired by these findings—including those that I plan to pursue in my postdoc (Section 7.3).

7.1. Main Findings

In this section, I summarize the main findings of this dissertation, following the order of the research questions introduced in Section 1.1 of Chapter 1.

> **RQ1**: How do different key algorithmic decisions for tokenizers influence the performance on downstream tasks: Tasks requiring robustness to language variation and tasks requiring sensitivity to language variation?

In Chapter 3, I introduce the notion of tasks that require robustness to language variation (e.g., for semantic tasks like natural language inference, labels do not depend on whether a text uses British or American spelling) and tasks that require sensitivity to language variation (e.g., for form-based tasks like authorship verification, labels depend on whether a text uses British or American spelling). Then, I investigate how key algorithmic decisions for tokenizers (i.e., fitting corpus, pre-tokenizer and vocabulary size) impact downstream model performance. I make three practical suggestions for selecting tokenizer settings: (1) Pay the most attention to the pre-tokenizer. It influences how Unicode Character Categories can be combined, and how many different words are ultimately part of the vocabulary. (2) Choose a bigger vocabulary size for settings that require sensitivity to language variation. (3) Use a small machine learning classifier to test the effect of different tokenizers on tasks robust and sensitive to language variation.

My experiments show that language variation should be considered at all stages of building LLMs, down to the very basic building blocks. I motivate why tokenizers specifically are likely to be sensitive to language variation. The distinction between

tasks requiring robustness to language variation and sensitivity to language variation demonstrate the complexity of juggling different requirements for varying task types when considering language variation.

> **RQ2a**: How can we evaluate whether text representations are sensitive to changes in linguistic style?

In Chapter 4, I investigate how to evaluate whether linguistic style is encoded in text representations. I propose STEL, a task framework to evaluate NLP methods on their sensitivity to style shifting, when referential meaning is constant. Drawing from style literature, I create four set of tasks relating to four dimensions of style. I find that neural representations outperform vanilla feature-based representations like character 3-grams and function word frequencies in their sensitivity to well-established dimensions of style. Out of the investigated text representation methods, RoBERTa (Liu et al., 2019) is the most sensitive to style when referential meaning stays the same.

In this work, I identify a lack of evaluation approaches that measure the sensitivity of NLP methods to changes in linguistic style. The few existing evaluation approaches are typically application-specific, often not based in style literature and tend to be correlated with referential meaning. With STEL, I present the first evaluation approach that assesses NLP methods on well-established dimensions of style while controlling for referential meaning correlations. To the best of my knowledge, I am the first to systematically compare a variety of different style measuring methods on a linguistically informed benchmark on style. Any NLP method that can compare texts can be evaluated on STEL.

> **RQ2b**: How can we build neural representations of linguistic style that are disentangled from referential meaning?

In Chapter 5, I investigate how to build neural text representations that are sensitive to linguistic style but not sensitive to referential meaning. I use a contrastive fine-tuning task on RoBERTa (Liu et al., 2019) that learns to place texts written by the same author closer together and texts written by different authors further apart. The assumption is that two texts written by the same author are more likely to be written in the same style than two texts written by different authors. In contrast to previous approaches, I aim to train text representations that encode well-established dimensions of style using texts written by different authors extracted from the same conversation as "hard negatives". The underlying assumption is that texts written in the same conversation are more likely to be about the same topic. I evaluate the representations on their sensitivity to linguistic style with the STEL framework. I further introduce a variation to STEL that tests whether representations are more sensitive to shifts in linguistic style or to shifts in referential meaning. Compared to other approaches, I reach a higher independence from referential meaning with my contrastive approach using hard negatives.

With this chapter, I provide a promising approach to train neural text representations that are independent from referential meaning. To the best of my knowledge, I am the first to release stylistic text representations for broad use on the Hugging Face hub. My style embeddings have been appreciated by the community[1]—

[1] Reaching 400k downloads on Hugging Face before the ACL 2025 submission deadline.

in part due to their improved independence from referential meaning compared to previous methods (Patel et al., 2023, 2024; Horvitz et al., 2024b).

RQ3: How can we detect paraphrases across speakers in dialog?

In Chapter 6, I introduce the task of context-dependent paraphrase detection across speaker turns in dialog. I motivate the task in relation to active listening, give a definition and explain ambiguities in annotation. I iteratively develop and provide a training for crowd workers to classify paraphrases in dialog and introduce a dataset with utterance pairs from NPR and CNN news interviews annotated by up to 21 annotators for context-dependent paraphrases. I reach promising results with fine-tuned encoder models as well as with in-context learning with decoder models. When identifying text spans that constitute paraphrase pairs, encoder models profit from not being able to hallucinate quotes.

Dialog is a setting that is uniquely sensitive to context (Grice, 1957, 1975; Davis, 2003) and makes matching the same referential meaning across speaker turns especially difficult. To the best of my knowledge, I am the first to operationalize, annotate and automatically detect context-dependent paraphrases across turns in dialog. I provide the annotator instructions, the annotated dataset and the best performing encoder models to the NLP community. My work demonstrates that both humans and NLP models face significant challenges when finding similarities in the referential meaning of utterances that vary across speakers.

7.2. Summing Up

In my dissertation, I have worked with the following overarching motivation

Motivation: Develop NLP methods that account for language variation.

I address this motivation by (Part II) evaluating key algorithmic choices of a basic LLM building block—tokenizers—on their sensitivity and robustness to language variation; (Part III) developing encoder models that are sensitive to one aspect of language variation: linguistic style; and (Part IV) detecting paraphrases in dialog when language varies across speakers. Thus, with this dissertation, I contribute to efforts that make NLP models more sensitive to language variation (in Parts II and III) and more robust to language variation (in Part II and IV). However, this dissertation is just one drop in the proverbial sea of open research questions when it comes to accounting for language variation in NLP—that is, (1) making models sensitive to form- and style-based differences and (2) making models robust to form and style-based differences. I hope that this work encourages the NLP community to dedicate more attention and effort to accounting for language variation in NLP methods.

7.3. Future Research

I provide some ideas and plans for future research based on the work in and insights of my dissertation.

Fostering constructive online conversations This dissertation has been funded by the NWO Digital Society research program under the title "The Power of Words: The Role of Mediators' Language in Increasing Intergenerational Empathy in Online Discussions".[2] The overarching aim of the project is to develop computational interventions that can bring different social groups together online. In line with this goal, Part III of this dissertation focuses on computationally modeling the linguistic styles of texts, motivated by the idea that helping social groups adapt to each other's writing styles in online conversations might help reduce polarization. More broadly, modeling linguistic style can offer insights into the community dynamics in which online conversations are embedded. For instance, it can help track the emergence of linguistic norms in a community and measure how closely a user aligns with them (Danescu-Niculescu-Mizil et al., 2013b; Gelfand et al., 2024). Further, I study paraphrases in dialog with the motivation that paraphrasing in online conversations might help people to listen to each other more actively. I develop novel methods to detect when people are saying the same thing in different ways (cf. Paraphrase detection across speaker turns in Chapter 6). In the future, these methods could enable computational interventions that might help foster more constructive online conversations: Adapting one's speech to align with that of a conversation partner has been shown to promote more positive evaluations such as perceived cooperativeness in verbal interactions (Giles et al., 1991). Further, I am particularly optimistic about paraphrasing-based interventions, as paraphrasing is a proven practice in verbal conflict resolution (cf. Chapter 6).

V

Applying developed methods to research questions in sociolinguistics In this dissertation, I discuss linguistic style and language variation in ways common to sociolinguistics (cf. Introduction Chapter 1 and Background Chapter 2.2). While my research is also motivated by typical sociolinguistic goals (e.g., understanding conversational interactions and identity construction, cf. Chapter 4), I focus on developing NLP methods to measure when language varies and referential meaning stays the same. However, I do not yet explicitly connect it to the social aspects that sociolinguistics is centrally concerned with. Future work could further develop this direction as a way that NLP methods can contribute to sociolinguistic research (Nguyen, 2025). For example, my style embedding model could be used to measure and study language accommodation[3]—that is, how speakers adapt their speech to their conversation partners (Giles et al., 1991). Much of the existing work on accommodation focuses on a narrow set of linguistic features (e.g., pronoun usage, part-of-speech categories or utterance length) (Danescu-Niculescu-Mizil et al., 2011; Giles et al., 1991) which might overlook various types of shift in linguistic style. In contrast, a style embedding model

[2] with grant number 410.19.007, see also `https://www.nwo.nl/en/projects/41019007`

[3] A related term is "alignment". For example, Garrod and Anderson (1987) study the alignment between interlocutors in referring expressions when referring to moving pieces in a maze.

could capture a wider range of stylistic variation. Similar to theories and findings in verbal settings—or studies limited to narrower linguistic indicators—one could investigate whether adapting to another's linguistic style increases social approval online. Apart from linguistic style, accommodation can also occur at the level of content (Ferrara, 1991). My paraphrase classification model (cf. Chapter 6) could support research into whether rephrasing a speaker's statement lead to elaboration by discussion partners in online contexts. Finally, my work on tokenization (cf. Chapter 3) can inform the design of variation-sensitive tokenizers that might improve a models' ability to automatically annotate texts (Nguyen, 2025; Pavlovic and Poesio, 2024), particularly those containing certain types of language variation, such as dialectal variation.

Evaluating how NLP methods account for language variation Over the course of my dissertation, I identified a significant gap in NLP literature: the lack of evaluation methods that assess how well NLP models account for language variation (cf. Chapter 4). I address a part of this problem in Chapter 4 and Chapter 5 when developing STEL to evaluate NLP models in their sensitivity to shifts in linguistic style. I further introduce a task suite that allows to test NLP methods in their sensitivity and robustness to language variation in in Chapter 3. In the months after finishing my PhD, I aim to valorize these efforts by releasing a Python package that make my evaluation methods easily accessible to the NLP community. Overall, evaluation methods are crucial to develop NLP methods that are more sensitive and robust to language variation in the future. Unfortunately, my developed evaluation methods do not cover the wide range of language variation in English: For example, one could consider a wider range of narrow style features (e.g., inspired by features in Table 2.1 in Section 2.2) and systematically evaluate how they are considered in NLP models. Additionally, it would be interesting to investigate how models handle language varieties associated with different geographical regions, ethnicities, age groups, genders and social classes. Further, one could also consider language variation for more languages than English.

Going beyond language variation in English In this dissertation, I focus on models primarily trained on English data and consider variation only within the English language. This aligns with a broader pattern in NLP that makes English the most studied and represented language by far—however, other languages have recently received more attention (Joshi et al., 2020; Ranathunga and de Silva, 2022; Ranathunga et al., 2023). Of course language variation is pervasive in all natural languages (Ball et al., 2023). Future work could extend evaluation methods to other languages—including STEL (Chapter 4) and my task-based evaluation suite (Chapter 3)—as well as develop style embeddings (Chapter 5) and paraphrase detection methods (Chapter 6) for more languages. Further, it could be interesting to evaluate the capabilities of other types of NLP methods and models—such as multilingual or multimodal language models—on tasks requiring sensitivity and robustness to language variation—within English and across languages and modalities. Additionally, code-switching (Doğruöz et al., 2021)—where speakers alternate between two or more languages within a single utterance (e.g., "Thanks voor de steun tijdens mijn PhD.")—brings unique challenges for language models and evaluation approaches. I encourage future work in NLP to account for language variation for more languages and modalities.

Privacy and integrating social factors into NLP Integrating more language diversity and with it social factors into NLP is a double-edged sword: There are clear advantages of integrating more diversity in NLP models and specifically representing the varieties of minorities to increase fairness and representativeness of NLP models (Hovy and Yang, 2021; Bird and Yibarbuk, 2024; Markl et al., 2024; Grieve et al., 2025). However, making NLP models more sensitive to social factors could also make them a threat to privacy across social groups. The performance of machine learning approaches for tasks like author profiling could increase. This results in a large potential for misuse, for example: (1) Author profiles could be used to identify and profile individuals or political dissenters (Hovy and Spruit, 2016), (2) Author profiling could be used for predatory ad targeting, which might show gambling ads to vulnerable groups or not show job postings to certain social groups (Dudy et al., 2021). (3) Author profiles could lead to data leakage, for example making health conditions recoverable for insurance companies that might increase their rates for certain individuals (Dudy et al., 2021). This conflict between privacy and fairness has been described as one of the "dual-use problems" in NLP by Hovy and Spruit (2016). In my work, I aim to improve fairness without compromising individual privacy and safety, but I acknowledge that progress in one might sometimes come at the expense of the other. I want to encourage researchers in the NLP community to engage with the dual-use problem more actively and work on techniques to make the design of language models more sensitive to human values as suggested in Dudy et al. (2021).

Language variation in datasets There are several aspects of NLP model construction that I have not investigated, but that are crucial when accounting for language variation in NLP. One key aspect is the datasets used at multiple stages for NLP models, including pre-training, fine-tuning, post-training and evaluation. Many current dataset construction strategies rely on dataset size rather than quality or diversity. However, dataset composition for (post-)training might lead to biases (Bender and Friedman, 2018; Bukharin et al., 2024) and a lack of robustness in how different social groups are treated. This is especially relevant in the age of CHATGPT, where increasing amounts of text data are now generated by language models and circulated online. As generated data becomes increasingly prevalent, it risks being re-ingested—both intentionally (Wang et al., 2023c) and unintentionally—into future training datasets. If linguistic diversity is not explicitly considered, this could lead to a narrowing of the linguistic range represented in models: models trained on increasingly homogeneous, generated data might reproduce and reinforce dominant varieties, ultimately crowding out less common linguistic forms (Guo et al., 2024). To address this, I will join Dong Nguyen's DataDivers ERC project[4] for a two-year postdoc where we will investigate how dataset diversity influences model representativeness, fairness and robustness. During the postdoc, I am aiming to specifically examine the role of linguistic and stylistic variation in datasets.

[4] https://datadivers-erc.github.io/

Part VI

Appendices

Contents

Bibliography

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *Computing Research Repository*, arXiv:2303.08774.

Cristina Aggazzotti, Nicholas Andrews, and Elizabeth Allyn Smith. 2024. Can authorship attribution models distinguish speakers in speech transcripts? *Transactions of the Association for Computational Linguistics*, 12:875–891.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? Tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore, Singapore. Association for Computational Linguistics.

Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. 2018. Unsupervised learning of style-sensitive word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 572–578, Melbourne, Australia. Association for Computational Linguistics.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys*, 54(2):1–36.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924, Mexico City, Mexico. Association for Computational Linguistics.

Melanie Andresen, Michael Vauth, and Heike Zinsmeister. 2020. Modeling ambiguity with many annotators and self-assessments of annotator certainty. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 48–59, Barcelona, Spain. Association for Computational Linguistics.

Melanie Andresen and Heike Zinsmeister. 2017. Approximating style by n-gram-based annotation. In *Proceedings of the Workshop on Stylistic Variation*, pages 105–115, Copenhagen, Denmark. Association for Computational Linguistics.

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

VI

Jannis Androutsopoulos. 2023. Punctuating the other: Graphic cues, voice, and positioning in digital discourse. *Language & Communication*, 88:141–152.

Ehsan Arabnezhad, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, and Julinda Stefa. 2020. A light in the dark web: Linking dark web aliases to real internet identities. In *Proceedings of the 40th International Conference on Distributed Computing Systems (ICDCS)*, pages 311–321, Singapore, Singapore. Institute of Electrical and Electronics Engineers.

Inbal Arnon and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

Martin J. Ball, Rajend Mesthrie, and Chiara Meluzzi. 2023. *The Routledge Handbook of Sociolinguistics Around the World*, 2nd edition. Routledge, London, UK.

John Bandy and Nicholas Vincent. 2021. Addressing "documentation debt" in machine learning: A retrospective datasheet for BookCorpus. In *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*, Online. Curran Associates, Inc.

Rusty Barrett. 2006. Language ideology and racial inequality: Competing functions of Spanish in an Anglo-owned Mexican restaurant. *Language in Society*, 35(2):163–204.

Fabian Barteld, Chris Biemann, and Heike Zinsmeister. 2018. Variations on the theme of variation: Dealing with spelling variation for finegrained POS tagging of historical texts. In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS)*, Vienna, Austria. Austrian Academy of Sciences.

Fabian Barteld, Ingrid Schröder, and Heike Zinsmeister. 2016. Dealing with word-internal modification and spelling variation in data-driven lemmatization. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 52–62, Berlin, Germany. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 830–839, Atlanta, USA. Association for the Advancement of Artificial Intelligence.

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

Allan Bell. 2014. *The Guidebook to Sociolinguistics*. John Wiley & Sons, Chichester, UK.

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

Basil Bernstein. 2003. *Class, Codes and Control: Applied Studies Towards a Sociology of Language*, volume 2. Routledge, London, UK.

Janek Bevendorff, Khalid Al Khatib, Martin Potthast, and Benno Stein. 2020a. Crawling and preprocessing mailing lists at scale for dialog analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1151–1158, Online. Association for Computational Linguistics.

Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Paolo Rosso, Alisa Smirnova, Efstathios Stamatatos, Benno Stein, Mariona Taulé, Dmitry Ustalov, Matti Wiegmann, and Eva Zangerle. 2024. Overview of PAN 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative AI authorship verification. In *Advances in Information Retrieval at ECIR*, pages 3–10, Glasgow, UK. Springer.

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020b. Overview of PAN 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on Twitter, and style change detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction at CLEF*, pages 372–383, Thessaloniki, Greece. Springer.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge, UK.

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

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2nd edition. Cambridge University Press, Cambridge, UK.

Steven Bird, Ewan Klein, and Edward Loper. 2019. *Natural Language Processing with Python*.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2020. The importance of suppressing domain style in authorship analysis. *Computing Research Repository*, arXiv:2005.14714.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

VI

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 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*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M. Nickel. 2019a. Explainable authorship verification in social media via attention-based similarity learning. In *Proceedings of the International Conference on Big Data (Big Data)*, pages 36–45, Los Angeles, USA. Institute of Electrical and Electronics Engineers.

Benedikt Boenninghoff, Robert M. Nickel, Steffen Zeiler, and Dorothea Kolossa. 2019b. Similarity learning for authorship verification in social media. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2457–2461, Brighton, UK. Institute of Electrical and Electronics Engineers.

Benedikt Boenninghoff, Julian Rupp, Robert M. Nickel, and Dorothea Kolossa. 2020. Deep Bayes factor scoring for authorship verification. In *Working Notes of Conference and Labs of the Evaluation Forum (CLEF)*, Thessaloniki, Greece. CEUR Workshop Proceedings.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Dwight Bolinger. 1974. Meaning and form. *Transactions of the New York Academy of Sciences*, 36(2 Series II):218–233.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Richard Y. Bourhis and Howard Giles. 1977. The language of intergroup distinctiveness. In Howard Giles, editor, *Language, Ethnicity and Intergroup Relations*, pages 119–135. Academic Press, London, UK.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. In *ACM Transactions on Information and System Security (TISSEC)*, volume 15, pages 1–22, New York, USA. Association for Computing Machinery.

Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang. 2013. Authorship verification for short messages using stylometry. In *International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6, Athens, Greece. Institute of Electrical and Electronics Engineers.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, UK. Association for Computational Linguistics.

Julian Brooke and Graeme Hirst. 2013. Hybrid models for lexical acquisition of correlated styles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 82–90, Nagoya, Japan. Asian Federation of Natural Language Processing.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, Online. Curran Associates, Inc.

Mikael Brunila and Jack LaViolette. 2022. What company do words keep? Revisiting the distributional semantics of J.R. Firth & Zellig Harris. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4403–4417, Seattle, USA. Association for Computational Linguistics.

Mary Bucholtz. 1999. You da man: Narrating the racial other in the production of white masculinity. *Journal of Sociolinguistics*, 3(4):443–460.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. 2024. Data diversity matters for robust instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3411–3425, Miami, USA. Association for Computational Linguistics.

John Burrows. 2002. 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Kathryn Campbell-Kibler. 2007. Accent,(ING), and the social logic of listener perceptions. *American Speech*, 82(1):32–64.

Kathryn Campbell-Kibler. 2009. The nature of sociolinguistic perception. *Language Variation and Change*, 21(1):135–156.

Kathryn Campbell-Kibler. 2011. The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*, 22(3):423–441.

Kathryn Campbell-Kibler, Penelope Eckert, Norma Mendoza-Denton, and Emma Moore. 2006. The elements of style. In *Poster Session at New Ways of Analyzing Variation (NWAV)*, Columbus, USA.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3497–3507, Brussels, Belgium. Association for Computational Linguistics.

VI

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, Online. Association for Computational Linguistics.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. 2024. On the diversity of synthetic data and its impact on training large language models. *Computing Research Repository*, arXiv:2410.15226.

Jenny Chim, Julia Ive, and Maria Liakata. 2025. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, 51(1):191–233.

Tanya Karoli Christensen and Torben Juel Jensen. 2022. *When Variants Lack Semantic Equivalence: Adverbial Subclause Word Order*, pages 171–206. Cambridge University Press, Cambridge, UK.

Eve V. Clark. 1992. Conventionality and contrast: Pragmatic principles with lexical consequences. In Adrienne Lehrer and Eva Feder Kittay, editors, *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, pages 171–188. Routledge, New York, USA.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge, UK.

Isobelle Clarke and Jack Grieve. 2017. Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Isobelle Clarke and Jack Grieve. 2019. Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PloS ONE*, 14(9):e0222062.

Steven Clayman and John Heritage. 2002. *The News Interview: Journalists and Public Figures on the Air*. Cambridge University Press, Cambridge, UK.

Marco Cognetta, Vilém Zouhar, Sangwhan Moon, and Naoaki Okazaki. 2024. Two counterexamples to tokenization and the noiseless channel. In *Proceedings of the*

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 16897–16906, Torino, Italia. ELRA and ICCL.

Jeff Collins, David Kaufer, Pantelis Vlachos, Brian Butler, and Suguru Ishizaki. 2004. Detecting collaborations in text comparing the authors' rhetorical language choices in the federalist papers. *Computers and the Humanities*, 38:15–36.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *Proceedings of the AAAI Fall Symposium on Communicative Aaction in Humans and Machines*, pages 28–35, Cambridge, USA. Association for the Advancement of Artificial Intelligence.

Malcolm Coulthard. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4):431–447.

Nikolas Coupland. 2007. *Style: Language Variation and Identity*. Cambridge University Press, Cambridge, UK.

Justin T. Craft, Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. 2020. Language and discrimination: Generating meaning, perceiving identities, and discriminating outcomes. *Annual Review of Linguistics*, 6(2020):389–407.

Evan N. Crothers, Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

David Crystal. 2008. *Txtng: The gr8 db8.* Oxford University Press, Oxford, UK.

David Crystal. 2011. *A Dictionary of Linguistics and Phonetics*, 6th edition. Blackwell Publishing, Malden, USA.

David Crystal and Derek Davy. 1969. *Investigating English Style*. Routledge, London, UK.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 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*, pages 22–31, Atlanta, USA. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*, page 745–754, Hyderabad, India. Association for Computing Machinery.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, USA. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In

VI

Proceedings of the 21st International Conference on World Wide Web (WWW), page 699–708, Lyon, France. Association for Computing Machinery.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013a. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013b. No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 307–318, Rio de Janeiro, Brazil. Association for Computing Machinery.

Wayne A. Davis. 2003. *Meaning, Expression, and Thought*. Cambridge University Press, New York, USA.

Ferdinand de Saussure. 1916. *Cours de Linguistique Générale*. Bayot, Paris, France.

William de Vazelhes, CJ Carey, Yuan Tang, Nathalie Vauquier, and Aurélien Bellet. 2020. metric-learn: Metric learning algorithms in Python. *Journal of Machine Learning Research*, 21(138):1–6.

Stefania Degaetano-Ortlieb. 2018. Stylistic variation over 200 years of court proceedings according to gender and social class. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 1–10, New Orleans, USA. Association for Computational Linguistics.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2018. *An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English*, pages 258–281. Brill, Leiden, The Netherlands.

Stefania Degaetano-Ortlieb, Tanja Säily, and Yuri Bizzoni. 2021. Registerial adaptation vs. innovation across situational contexts: 18th century women in transition. *Frontiers in Artificial Intelligence*, 4:609970.

Stefania Degaetano-Ortlieb and Elke Teich. 2017. Modeling intra-textual variation with entropy and surprisal: topical vs. stylistic patterns. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 68–77, Vancouver, Canada. Association for Computational Linguistics.

Enrique Dehaerne, Bappaditya Dey, Sandip Halder, Stefan De Gendt, and Wannes Meert. 2022. Code generation using machine learning: A systematic review. *IEEE Access*, 10:82434–82455.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 12475–12498, Singapore, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, USA. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Justin Dieter, Tian Wang, Arun Tejasvi Chaganty, Gabor Angeli, and Angel X. Chang. 2019. Mimic and rephrase: Reflective listening in open-ended dialogue. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 393–403, Hong Kong, China. Association for Computational Linguistics.

Steven H. H. Ding, Benjamin C. M. Fung, Farkhund Iqbal, and William K. Cheung. 2019. Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*, 49(1):107–121.

Robin Dodsworth and Richard A. Benton. 2020. *Language Variation and Change in Social Networks: A Bipartite Approach*. Routledge, New York, USA.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pages 9–16, Jeju Island, Korea. Asian Federation of Natural Language Processing.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. ParaSCI: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 424–434, Online. Association for Computational Linguistics.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2023. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.

VI

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Penelope Eckert. 1989. *Jocks and Burnouts: Social Categories and Identity in the High School*. Teachers College Press, New York, USA.

Penelope Eckert. 2003. Elephants in the room. *Journal of Sociolinguistics*, 7(3):392–397.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41(1):87–100.

Penelope Eckert and John R. Rickford. 2001. *Style and Sciolinguistic Variation*. Cambridge University Press, Cambridge, UK.

Maciej Eder. 2013. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2):167–182.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. 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 1634–1647, Minneapolis, USA. Association for Computational Linguistics.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, USA. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Sara El Manar El Bouanani and Ismail Kassou. 2014. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12):22–29.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Sean P. Engelson and Ido Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Santa Cruz, USA. Association for Computational Linguistics.

Susan M. Ervin-Tripp. 2001. Variety, style-shifting, and ideology. In Penelope Eckert and John R. Rickford, editors, *Style and Sociolinguistic Variation*, pages 44–56. Cambridge University Press, Cambridge, UK.

Alex Chengyu Fang and Jing Cao. 2009. Adjective density as a text formality characteristic for automatic text classification: A study based on the British National Corpus. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 130–139, Hong Kong, China. City University of Hong Kong.

Kathleen Ferrara. 1991. Accommodation in therapy. In Howard Giles, Justine Coupland, and Nikolas Coupland, editors, *Contexts of Accommodation: Developments in Applied Sociolinguistics*, pages 187–222. Cambridge University Press, Cambridge, UK.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

John R. Firth. 1957. *Studies in Linguistic Analysis*. Basil Blackwell, Oxford, UK.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, USA. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, Berlin, Germany. Association for Computational Linguistics.

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

Matthias Gallé. 2019. Investigating the effectiveness of BPE: The power of shorter sequences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

VI

Processing (EMNLP-IJCNLP), pages 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Lesya Ganushchak, Andrea Krott, and Antje Meyer. 2012. From gr8 to great: Lexical access to sms shortcuts. *Frontiers in Psychology*, 3(150).

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *Computing Research Repository*, arXiv:2101.00027.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Computing Research Repository*, arXiv:2312.10997.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Albert Gatt and Emiel Krahmer. 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.

Michele J. Gelfand, Sergey Gavrilets, and Nathan Nunn. 2024. Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change. *Annual Review of Psychology*, 75:341–378.

Katy Gero, Chris Kedzie, Jonathan Reeve, and Lydia Chilton. 2019. Low level linguistic controls for style transfer and content preservation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 208–218, Tokyo, Japan. Association for Computational Linguistics.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In *Contexts of Accommodation: Developments in Applied Sociolinguistics*, pages 1–68. Cambridge University Press, Cambridge, UK.

Howard Giles and Peter F. Powesland. 1975. *Speech Style and Social Evaluation*. Academic Press, London, UK.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 214–217, San Jose, USA. Association for the Advancement of Artificial Intelligence.

Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. 2022. Improving tokenisation by alternative treatment of spaces. In *Proceedings*

of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11430–11443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Tanya Goyal, Nazneen Rajani, Wenhao Liu, and Wojciech Kryscinski. 2022. HydraSum: Disentangling style features in text summarization with multi-decoder models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 464–479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *Computing Research Repository*, arXiv:2407.21783.

H. Paul Grice. 1957. Meaning. *The Philosophical Review*, 66(3):377–388.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Volume 3: Speech Acts*, pages 41–58. Academic Press, New York, USA.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Jack Grieve. 2011. A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics*, 16(4):514–546.

Jack Grieve. 2023. Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, 19(1):47–77.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7:1472411.

Jack Grieve, Douglas Biber, Eric Friginal, and Tatiana Nekrasova. 2011. Variation among blogs: A multi-dimensional analysis. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web*, pages 303–322. Springer, Dordrecht, the Netherlands.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *Computing Research Repository*, arXiv:2203.05794.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):2.

VI

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Computing Research Repository*, arXiv:2501.12948.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? Measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ton van Haaften and Maarten van Leeuwen. 2021. On the relation between argumentative style and linguistic style: Integrating linguistic-stylistic analysis systematically into the analysis of argumentative style. *Journal of Argumentation in Context*, 10(1):97–120.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, pages 1735–1742, New York, USA. Institute of Electrical and Electronics Engineers.

Oren Halvani, Christian Winter, and Lukas Graner. 2019. Assessing the applicability of authorship verification methods. In *Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES '19)*, Canterbury, UK. Association for Computing Machinery.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, USA. Association for Computational Linguistics.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585:357–362.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2–3):146–162.

Julien Hay, Bich-Lien Doan, Fabrice Popineau, and Ouassim Ait Elhara. 2020. Representation learning of writing style. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 232–243, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*, Online. Curran Associates, Inc.

Juan Manuel Hernández-Campoy. 2016. *Sociolinguistic styles*. John Wiley & Sons, Chichster, UK.

Elize Herrewijnen, Dong Nguyen, Floris Bex, and Kees van Deemter. 2024. Human-annotated rationales and explainable text classification: a survey. *Frontiers in Artificial Intelligence*, 7:1260952.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Internal Report, Center "Leo Apostel", Vrije Universiteit Brussels*.

Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7:293–340.

Joe Hight and Frank Smyth. 2002. *Tragedies & Journalists: A guide for more effective coverage*. Dart Center for Journalism & Trauma, New York, USA.

Clara E. Hill. 1992. An overview of four measures developed to test the Hill process model: Therapist intentions, therapist response modes, client reactions, and client behaviors. *Journal of Counseling & Development*, 70(6):728–739.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2024. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, USA. Curran Associates Inc.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154.

Nicole Holliday. 2021. Intonation and referee design phenomena in the narrative speech of Black/biracial men. *Journal of English Linguistics*, 49(3):283–304.

David I. Holmes. 1985. The analysis of literary style – a review. *Journal of the Royal Statistical Society: Series A (General)*, 148(4):328–341.

Janet Holmes and Nick Wilson. 2017. *An Introduction to Sociolinguistics*, 5th edition. Routledge, London, UK.

Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen McKeown. 2024a. ParaGuide: Guided diffusion paraphrasers for plug-and-play textual style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18216–18224, Vancouver, Canada. Association for the Advancement of Artificial Intelligence.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. 2024b. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13376–13390, Miami, USA. Association for Computational Linguistics.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *Computing Research Repository*, arXiv:2403.03952.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

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

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

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

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore, Singapore. Association for Computational Linguistics.

Zhiqiang Hu, Roy Ka-Wei Lee, Lei Wang, Ee-peng Lim, and Bo Dai. 2020. DeepStyle: User style embedding for authorship attribution of short texts. In *Web and Big Data*, pages 221–229, Tianjin, China. Springer International Publishing.

Weihang Huang and Jack Grieve. 2024. Authorial language models for AI authorship verification. In *Working Notes of Conference and Labs of the Evaluation Forum (CLEF)*, Grenoble, France. CEUR Workshop Proceedings.

Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2024. Understanding writing style in social media with a supervised contrastively pre-trained transformer. *Knowledge-Based Systems*, 296:111867.

Judith T. Irvine. 2001. "Style" as distinctiveness: the culture and ideology of linguistic differentiation. In Penelope Eckert and John R. Rickford, editors, *Style and Sociolinguistic Variation*, pages 21–43. Cambridge University Press, Cambridge, UK.

Sarah Jablotschkin, Elke Teich, and Heike Zinsmeister. 2024. DE-Lite - a new corpus of easy German: Compilation, exploration, analysis. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 106–117, St. Julian's, Malta. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Computing Research Repository*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Computing Research Repository*, arXiv:2401.04088.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

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

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Dan Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Jean Kaddour. 2023. The MiniPile challenge for data-efficient language models. *Computing Research Repository*, arXiv:2304.08442.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, and Otto Tarkka. 2021. Annotation guidelines for the Turku paraphrase corpus. *Computing Research Repository*, arXiv:2108.07499.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, and Otto

VI

Tarkka. 2023. Towards diverse and contextually anchored paraphrase modeling: A dataset and baselines for Finnish. *Natural Language Engineering*, 30(2):319–353.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (Male, Bachelor) and (Female, Ph.D) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1696–1706, Hong Kong, China. Association for Computational Linguistics.

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

Mahmut Kaya and Hasan Şakir Bilge. 2019. Deep metric learning: A survey. *Symmetry*, 11(9):1066.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at PAN 2020. In *Working Notes of Conference and Labs of the Evaluation Forum (CLEF)*, Thessaloniki, Greece. CEUR Workshop Proceedings.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2021. Overview of the cross-domain authorship verification task at PAN 2021. In *Working Notes of Conference and Labs of the Evaluation Forum (CLEF)*, pages 1743–1759, Bucharest, Romania. CEUR Workshop Proceedings.

Aleem Khan, Andrew Wang, Sophia Hager, and Nicholas Andrews. 2023. Learning to generate text in arbitrary writing styles. *Computing Research Repository*, arXiv:2312.17242.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Online. Curran Associates, Inc.

Hazen Kirk. 2023. Sociolinguistics in the USA. In Martin J. Ball, Rajend Mesthrie, and Chiara Meluzzi, editors, *The Routledge Handbook of Sociolinguistics Around the World*, pages 13–27. Routledge, London, UK.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, USA. Curran Associates, Inc.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4:155–190.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Bernd Kortmann, Kerstin Lunkenheimer, and Katharina Ehret, editors. 2020. *The Electronic World Atlas of Varieties of English*.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2018. ETPC - a paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Klaus Krippendorff. 1995. On the reliability of unitizing continuous data. *Sociological Methodology*, pages 47–76.

Klaus Krippendorff. 2019. Reliability. In *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications, Thousand Oaks, USA.

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.

Annelies Kusters and Ceil Lucas. 2022. Emergence and evolutions: Introducing sign language sociolinguistics. *Journal of Sociolinguistics*, 26(1):84–98.

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

William Labov. 2006. *The Social Stratification of English in New York City*, 2nd edition. Cambridge University Press, Cambridge, UK.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of English. *Language Resources and Evaluation*, 57:1045–1079.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Sander Land and Max Bartolo. 2024. Fishing for magikarp: Automatically detecting under-trained tokens in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11631–11646, Miami, USA. Association for Computational Linguistics.

Beatriz R. Lavandera. 1978. Where does the sociolinguistic variable stop? *Language in Society*, 7(2):171–182.

VI

Marc T. Law, Nicolas Thome, and Matthieu Cord. 2016. Learning a distance metric from relative comparisons between quadruplets of images. *International Journal of Computer Vision*, 121:65–94.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore, Singapore. Association for Computational Linguistics.

Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, USA.

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.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Jindřich Libovický, Helmut Schmid, and Alexander Fraser. 2022. Why don't people use character-level machine translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2470–2485, Dublin, Ireland. Association for Computational Linguistics.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Marina Litvak. 2019. Deep dive into authorship verification of email messages with convolutional neural network. In *Proceedings of the International Conference on Information Management and Big Data (SIMBig)*, pages 129–136, Lima, Peru. Springer.

Tatiana Litvinova. 2020. Stylometrics features under domain shift: Do they really "context-independent"? In *Proceedings of the International Conference on Speech and Computer (SPECOM)*, pages 279–290, St. Petersburg, Russia. Springer.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. DialogueCSE: Dialogue-based contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2396–

2406, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76, Portland, USA. Association for Computational Linguistics.

Hui Liu, Yongzheng Zhang, Yipeng Wang, Zheng Lin, and Yige Chen. 2020. Joint character-level word embedding and adversarial stability training to defend adversarial text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8384–8391, New York, USA. Association for the Advancement of Artificial Intelligence.

Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023. RECAP: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8404–8419, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Computing Research Repository*, arXiv:1907.11692.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Li Lucy, Su Lin Blodgett, Milad Shokouhi, Hanna Wallach, and Alexandra Olteanu. 2024. "One-size-fits-all"? Examining expectations around what constitute "fair" or "good" NLG system behaviors. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1054–1089, Mexico City, Mexico. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503, Prague, Czech Republic. Association for Computational Linguistics.

VI

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Pranav Maneriker, Yuntian He, and Srinivasan Parthasarathy. 2021. SYSML: StYlometry with Structure and Multitask Learning: Implications for Darknet forum migrant analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6844–6857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrei Manolache, Florin Brad, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. 2022. VeriDark: A large-scale benchmark for authorship verification on the dark web. In *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, USA. Curran Associates, Inc.

Nina Markl, Lauren Hall-Lew, and Catherine Lai. 2024. Language technologies as if people mattered: Centering communities in language technology development. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10085–10099, Torino, Italia. ELRA and ICCL.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.

Colin Martindale and Dean McKenzie. 1995. On the utility of content analysis in author attribution: The Federalist. *Computers and the Humanities*, 29:259–270.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14867–14875, Online. Association for the Advancement of Artificial Intelligence.

Jacob Matthews, John Starr, and Marten Schijndel. 2024. Semantics or spelling? Probing contextual word embeddings with orthographic noise. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4495–4504, Bangkok, Thailand. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Erick Mendez Guzman, Viktor Schlegel, and Riza Batista-Navarro. 2022. RaFoLa: A rationale-annotated corpus for detecting indicators of forced labour. In *Proceedings*

of the Thirteenth Language Resources and Evaluation Conference, pages 3610–3625, Marseille, France. European Language Resources Association.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, et al. 2024. Gemma: Open models based on gemini research and technology. *Computing Research Repository*, arXiv:2403.08295.

Miriam Meyerhoff. 2006. *Introducing Sociolinguistics*. Routledge, London, UK.

Gaspard Michel, Elena Epure, Romain Hennequin, and Christophe Cerisara. 2024. Distinguishing fictional voices: a study of authorship verification models for quotation attribution. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 160–171, St. Julians, Malta. Association for Computational Linguistics.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *Computing Research Repository*, arXiv:2112.10508.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, Stateline, USA. Curran Associates, Inc.

William R. Miller and Stephen Rollnick. 2013. *Motivational Interviewing: Helping People Change*, 3rd edition. Guilford Press, New York, USA.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Frederick Mosteller and David L. Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50(6):86.

Dong Nguyen. 2025. Collaborative growth: When large language models meet sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How old do you think I am?" A study of language and age in Twitter. In *Proceedings of the In-*

VI

ternational AAAI Conference on Web and Social Media, pages 439–448, Cambridge, USA. Association for the Advancement of Artificial Intelligence.

Dong Nguyen and Jack Grieve. 2020. Do word embeddings capture spelling variation? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 870–881, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Xing Niu and Marine Carpuat. 2017. Discovering stylistic variations in distributional vector space models via lexical paraphrases. In *Proceedings of the Workshop on Stylistic Variation*, pages 20–27, Copenhagen, Denmark. Association for Computational Linguistics.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, USA. Association for Computational Linguistics.

OpenAI. 2022. Introducing ChatGPT. *OpenAI blog*.

Shereen Oraby, Lena Reed, Shubhangi Tandon, Sharath T.S., Stephanie Lukin, and Marilyn Walker. 2018. Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 180–190, Melbourne, Australia. Association for Computational Linguistics.

Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. 2024. Tokenization matters: Navigating data-scarce tokenization for gender inclusive language technologies. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1739–1756, Mexico City, Mexico. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, USA. Association for Computational Linguistics.

Xudong Pan, Mi Zhang, Beina Sheng, Jiaming Zhu, and Min Yang. 2022. Hidden trigger backdoor attack on NLP models via linguistic style manipulation. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3611–3628, Boston, USA. USENIX Association.

Annaleena Parhankangas and Maija Renko. 2017. Linguistic style and crowdfunding success among social and commercial entrepreneurs. *Journal of Business Venturing*, 32(2):215–236.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore, Singapore. Association for Computational Linguistics.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2024. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. *Computing Research Repository*, arXiv:2410.12757.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

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

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

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher,

VI

Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin, Austin, USA.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, USA. Curran Associates, Inc.

Aleksandra Piktus, Necati Bora Edizel, Piotr Bojanowski, Edouard Grave, Rui Ferreira, and Fabrizio Silvestri. 2019. Misspelling oblivious word embeddings. 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 3226–3234, Minneapolis, USA. Association for Computational Linguistics.

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

Nektaria Potha and Efstathios Stamatatos. 2018. Intrinsic author verification using topic modeling. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence (SETN '18)*, Patras, Greece. Association for Computing Machinery.

Daniel Preoţiuc-Pietro, Wei Xu, and Lyle Ungar. 2016. Discovering user attribute stylistic differences via paraphrasing. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 3030–3037, Phoenix, USA. Association for the Advancement of Artificial Intelligence.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 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*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

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

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Online. Curran Associates, Inc.

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

Nived Rajaraman, Jiantao Jiao, and Kannan Ramchandran. 2024. Toward a theory of tokenization in LLMs. *Computing Research Repository*, arXiv:2404.08335.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):229.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at PAN 2013. In *Working Notes of Conference and Labs of the Evaluation Forum (CLEF)*, Valencia, Spain. CEUR Workshop Proceedings.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC)*, page 37–44, Toronto, Canada. Association for Computing Machinery.

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, USA. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 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)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

VI

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Ehud Reiter. 2025. *Natural Language Generation*. Springer, Cham, Switzerland.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore, Singapore. Association for Computational Linguistics.

John R. Rickford and McNair-Knox. 1994. Addressee-and topic-influenced style shift: A quantitative sociolinguistic study. In Douglas Biber and Edward Finegan, editors, *Sociolinguistic Perspectives on Register*, pages 235–276. Oxford University Press, New York, USA.

Stefan Riezler and Michael Hagmann. 2022. *Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science*. Springer, Cham, Switzerland.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carl R. Rogers. 1951. *Client-Centered Therapy: Its current practice, implications, and theory*. Houghton Mifflin Company, Boston, USA.

Carla Roos. 2022. *Everyday Diplomacy: dealing with controversy online and face-to-face*. Ph.D. thesis, University of Groningen, Groningen, the Netherlands.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Chakaveh Saedi and Mark Dras. 2021. Siamese networks for large-scale author identification. *Computer Speech & Language*, 70:101241.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5884–5906, Seattle, USA. Association for Computational Linguistics.

Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? Exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, USA. Association for Computational Linguistics.

Vageesh Saxena, Benjamin Ashpole, Gijs van Dijck, and Gerasimos Spanakis. 2023. IDTraffickers: An authorship attribution dataset to link and connect potential human-trafficking operations on text escort advertisements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8444–8464, Singapore, Singapore. Association for Computational Linguistics.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, USA. Curran Associates, Inc.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176B-parameter open-access multilingual language model. *Computing Research Repository*, arXiv:2211.05100.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Online. International Committee for Computational Linguistics.

Jonathan Schler, Moshe Koppel, and Shlomo Argamon. 2006. Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, Palo Alto, USA. Association for the Advancement of Artificial Intelligence.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, USA. Association for Computational Linguistics.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, pages 92–96, Austin, USA. SciPy.

Gail Sedorkin. 2020. *Interviewing: A guide for journalists and writers*, 2nd edition. Routledge, London, UK.

Gün R. Semin and Klaus Fiedler. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*, 54(4):558–568.

VI

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E. Kraut, and Diyi Yang. 2022. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):527.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott A. Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. 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 66–76, Hong Kong, China. Association for Computational Linguistics.

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2016. Adding context to semantic data-driven paraphrasing. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 108–113, Berlin, Germany. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, USA. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 16857–16867, Online. Curran Associates, Inc.

Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-shot detection of machine-generated text using style representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vienna, Austria. Curran Associates, Inc.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 5.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2017. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Wessel Stoop and Antal van den Bosch. 2014. Using idiolects and sociolects to improve word prediction. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 318–327, Gothenburg, Sweden. Association for Computational Linguistics.

Yu Su and Xifeng Yan. 2017. Cross-domain semantic parsing via paraphrasing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1235–1246, Copenhagen, Denmark. Association for Computational Linguistics.

Jamar Sullivan Jr., Will Brackenbury, Andrew McNutt, Kevin Bryson, Kwam Byll, Yuxin Chen, Michael Littman, Chenhao Tan, and Blase Ur. 2022. Explaining why: How instructions and user interfaces impact annotator rationales when labeling text data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–531, Seattle, USA. Association for Computational Linguistics.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.

Kalaivani Sundararajan and Damon Woodard. 2018. What represents "style" in authorship attribution? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2814–2822, Santa Fe, USA. Association for Computational Linguistics.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. 2024. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 114147–114179, Vancouver, Canada. Curran Associates, Inc.

VI

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Online. Curran Associates, Inc.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.

Henry Thompson. 1977. Strategy and tactics: A model for language production. In *Proceedings of the Regional Meeting of the Chicago Linguistics Society*, pages 651–668, Chicago, USA.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, Edmonton, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Computing Research Repository*, arXiv:2307.09288.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *Computing Research Repository*, arXiv:1908.08962.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Omri Uzan, Craig W. Schmidt, Chris Tanner, and Yuval Pinter. 2024. Greed is all you need: An evaluation of tokenizer inference methods. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 813–822, Bangkok, Thailand. Association for Computational Linguistics.

Johan van Benthem and Alice ter Meulen. 2011. *Handbook of Logic and Language*, 2nd edition. Elsevier, London, UK.

Kees van Deemter, Albert Gatt, Roger P.G. van Gompel, and Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4(2):166–183.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Ielka van der Sluis and Chris Mellish. 2009. Towards empirical evaluation of affective tactical NLG. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 146–153, Athens, Greece. Association for Computational Linguistics.

Maurice van Lieshout, Marijke Huisman, and Evert van der Veen. 2024. *Utrechtenaren een queer geschiedenis*. WBooks, Zwolle, the Netherlands.

Lindsey Vanderlyn and Ngoc Thang Vu. 2025. It's what you say and how you say it: Investigating the effect of linguistic vs. behavioral adaptation in task-oriented chatbots. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6120–6149, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, USA. Curran Associates, Inc.

Gregory M. Vecchi, Vincent B. Van Hasselt, and Stephen J. Romano. 2005. Crisis (hostage) negotiation: current strategies and issues in high-risk conflict resolution. *Aggression and Violent Behavior*, 10(5):533–551.

Gregory M. Vecchi, Gilbert K.H. Wong, Paul W.C. Wong, and Mary Ann Markey. 2019. Negotiating in the skies of Hong Kong: The efficacy of the behavioral influence stairway model (BISM) in suicidal crisis situations. *Aggression and Violent Behavior*, 48:230–239.

Menan Velayuthan and Kengatharaiyer Sarveswaran. 2025. Egalitarian language representation in language models: It all begins with tokenizers. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5987–5996, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marta Vila, M Antònia Martí, and Horacio Rodríguez. 2014. Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(1):205–218.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, An-

tônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, USA. Association for Computational Linguistics.

Chris Voss and Tahl Raz. 2016. *Never split the difference: Negotiating as if your life depended on it*. Random House, London, UK.

Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2023. Paraphrase types for generation and detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12148–12164, Singapore, Singapore. Association for Computational Linguistics.

Jan Philip Wahle, Terry Ruas, Yang Xu, and Bela Gipp. 2024. Paraphrase types elicit prompt engineering capabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11004–11033, Miami, USA. Association for Computational Linguistics.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? Probing numeracy in embeddings. 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 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys)*, pages 86–94, Vancouver, Canada. Association for Computing Machinery.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2605–2610, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023a. Can authorship representation learning capture

stylistic features? *Transactions of the Association for Computational Linguistics*, 11:1416–1431.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024. OpenChat: Advancing open-source language models with mixed-quality data. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vienna, Austria. Curran Associates, Inc.

Shuohang Wang, Ruochen Xu, Yang Liu, Chenguang Zhu, and Michael Zeng. 2022a. ParaTag: A dataset of paraphrase tagging for fine-grained labels, NLG evaluation, and data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7111–7122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Kigali, Rwanda. Curran Associates, Inc.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

VI

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Computing Research Repository*, arXiv:2412.13663.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Harry Weger Jr., Gina R. Castle, and Melissa C. Emmett. 2010. Active listening in peer interviews: The influence of message paraphrasing on perceptions of listening skill. *International Journal of Listening*, 24(1):34–49.

Anna Wegmann, Florian Lemmerich, and Markus Strohmaier. 2020. Detecting different forms of semantic shift in word embeddings via paradigmatic and syntagmatic association changes. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 619–635, Online. Springer International Publishing.

Anna Wegmann and Dong Nguyen. 2021. Does it capture STEL? A modular, similarity-based linguistic style evaluation framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7109–7130, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? Towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Online. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, USA. Curran Associates, Inc.

E. Judith Weiner and William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics*, 19(1):29–58.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Lilian Weng. 2023. Prompt engineering. *lilianweng.github.io*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, USA. Association for Computational Linguistics.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Ludwig Wittgenstein. 1953. *Philosophische Untersuchungen / Philosophical Investigations*. Blackwell, Oxford, UK.

Ka Wong and Praveen Paritosh. 2022. k-Rater Reliability: The correct unit of reliability for aggregated human annotations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–384, Dublin, Ireland. Association for Computational Linguistics.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, Online. Association for Computational Linguistics.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2010. Paraphrase generation as monolingual translation: Data and evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*, Trim, Ireland. Association for Computational Linguistics.

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

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

Wei Xu, 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.

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

Yang Xu and David Reitter. 2015. An evaluation and comparison of linguistic alignment measures. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–67, Denver, USA. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pretrained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

VI

Jinbiao Yang, Stefan L. Frank, and Antal van den Bosch. 2020. Less is better: A cognitively inspired unsupervised model for language segmentation. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 33–45, Online. Association for Computational Linguistics.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 164–174, Melbourne, Australia. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, USA. Curran Associates, Inc.

Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Wang Yongji, and Jian-Guang Lou. 2023. Large language models meet NL2Code: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7443–7464, Toronto, Canada. Association for Computational Linguistics.

Eva Zangerle, Maximilian Mayerl, Günther Specht, Martin Potthast, and Benno Stein. 2020. Overview of the style change detection task at PAN 2020. In *Working Notes of Conference and Labs of the Evaluation Forum (CLEF)*, Thessaloniki, Greece. CEUR Workshop Proceedings.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada. Curran Associates, Inc.

Rodolfo Zevallos and Nuria Bel. 2023. Hints on the data for language modeling of synthetic languages with transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12508–12522, Toronto, Canada. Association for Computational Linguistics.

Shiyue Zhang, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. 2022. How robust is neural machine translation to language imbalance in multilingual tokenizer training? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–116, Orlando, USA. Association for Machine Translation in the Americas.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, USA. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 12697–12706, Online. Proceedings of Machine Learning Research.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 46595–46623, New Orleans, USA. Curran Associates, Inc.

Chao Zhou, Cheng Qiu, and Daniel E. Acuna. 2025. Paraphrase identification with deep learning: A review of datasets and methods. *IEEE Access*, Early Access.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

Jian Zhu and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 19–27, Santiago, Chile. Institute of Electrical and Electronics Engineers.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

VI

Lal Zimman. 2019. Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language*, 2019(256):147–175.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

A

Additions to Chapter 3

Group	Choice	Examples of tokens
Corpus	Wikipedia	ebruary _Retrieved arliament
	Twitter	_[joy] [loudly-crying-face][loudly-crying-face] [heart] BTS
	PubMed	_effec _lymphadenopathy
	Misc.	\t\t \r \n _differe
Pre-Tok	NO	_that_ at_ in_this_case_
	WS	you're that's took develo
	_WS	\nI _I'm \nWhat _devices.
	GPT2	_..., ensional >"; _127
	LLAMA3	433 _{\n _*\n .apache
Size	500	! $ 0 1 2 A B C is he in
	4k	age very _will _would
	32k	_surveillance _Vietnam
	64k	CAN 322 _infuri
	128k	motherboard _narcotics

Table A.1: Examples of unique tokens for each tokenizer choice. The displayed tokens are unique to the given tokenizer in their respective group, except for the different vocabulary sizes. The tokens of the smaller vocabulary size is always included in the bigger vocabulary size. We represent whitespaces within tokens as _. We represent emojis within [] with their textual descriptions.

A.1. Tokenizer

In this section, we provide additional information on the tokenizer settings we investigate. See Table A.1 for examples of unique tokens for each setting.

A.1.1. Fitting Corpora

See an overview of the fitting dataset sizes in Table A.2. Usually, the fitting corpus for the tokenizer and the training corpus for the language model with that tokenizer are the same. As a result the size of the fitting corpus often varies as widely as the size

Source	Train	Dev	Test
Wikipedia	1,469,999,792	15,000,029	15,000,087
Twitter	1,470,004,662	15,000,048	15,000,057
PubMed	1,469,999,499	15,000,106	15,000,501
Miscellaneous	1,477,872,323	15,080,505	15,080,919

Table A.2: Fitting corpora with similar word counts. We compare three fitting corpora for tokenizers. Word count is calculated using white-space splitting. The size of the fitting corpora are not exactly the same when it comes to word count. But variations in word count are below 1% and should not affect the vocabulary of the tokenizer fitted on them.

of training datasets. We aim for 1.5 billion tokens for all fitting corpora. The MiniPile Kaddour (2023) dataset used for fitting in Schmidt et al. (2024) is of similar magnitude. We further display text examples for each dataset in Table A.3. Dataset sizes vary in less than 1% of word count. The variance in word count is an artifact of dataset creation from several documents with lenient word count limits.

Miscellaneous Miscellaneous consists of Reddit Baumgartner et al. (2020), literature sources (fanfictions from ao3[1], Gao et al., 2020's books before 1919 from the Gutenberg project), news articles and comments (Zellers et al., 2019's realnews, 2020 NYTimes articles and comments[2], Kolhatkar et al., 2020's sfu-socc), question answering (Gao et al., 2020's StackExchange), reviews (Hou et al., 2024's Amazon and Wan and McAuley, 2018's GoodReadsreviews), mails (Bevendorff et al., 2020a's Gmane), transcripts (YouTubeCommons[3] and Gao et al., 2020's OpenSubtitles), blogs (Schler et al., 2006's blogcorpus), raw text from webpages (Gao et al., 2020's Common Crawl), science articles (Lo et al., 2020's s2orc), code and mathematics (Gao et al., 2020's GitHub and DeepMind Mathematics). See the share of different domains in Table A.4.

A.1.2. Pre-Tokenizer

See the regular expressions defining the different considered pre-tokenizers in Table A.5. Differences affect mostly whitespace, contraction, punctuation and number handling.

[1] fanfictions until 2019 from Archive of Our Own `https://archiveofourown.org/`, downloaded from `https://archive.org/download/AO3_story_dump_continuing` in 2023, filtered for English language using AO3 tags. Dataset was removed but should be re-creatble using tools like `https://github.com/nianeyna/ao3downloader`.

[2] `https://www.kaggle.com/datasets/benjaminawd/new-york-times-articles-comments-2020` , minimum length filter of 250 was applied

[3] `https://huggingface.co/datasets/PleIAs/YouTube-Commons`

Source	Text	word count	domain
Wiki	Mary Jane Christie Serrano (c. 1840 – 1923) was a writer, poet and considered ...	24	-
Twitter	Where are the top places in Broward or Palm Beach? [thinking-face][eyes]	10	-
PMed	... Myoelectrical activity of the gut has been studied in the postoperative period ...	132	-
Misc.	Israel, as usual, wants American forces to fight a bloody war against Iran. ...	66	nytimes
	>In Israel, my grandfather fought for its life. The people down the street fought ...	581	reddit
	Q:\n\nHow can I determine the current focused process name and version in C#...	122	StackEx.
	I read the audio version of this story and loved it. ...	53	goodreads
	... always get 100 test cases (or whatever the default number of test is)?\n\nJanek ...	134	gmane
	there is Sydney waiting to enter. Their eyes meet. “Maggie! I was just ...	998	ao3
	... Abstract. How many n-orthants can be intersected in the n-dimensional ...	2048	s2orc
	from torch import optim as optim\n \n from geoopt.optim.mixin import ...	115	GitHub
	if you’re looking for enhancement cores in the game one really useful way ...	102	YouTube
	I have a lot of ties. But my favorite one. My favorite tie is owned by ...	183	blogcorpus
	Witnesses interviewed: 3 (N°1141, 1142, 1143). Nikolay S.(N°1142): "After the ...	72	Pile-CC
	It came in perfect condition and it is very soft.	9	amazon
	I’m sure that the procurement people are doing the best job they can ...	94	sfu-socc
	... Sort -11, -1, 0, -3, 5 in decreasing order ...	1645	DM Maths
	... "But the killer isn’t the Russian army." "It’s the subzero temperatures." ...	1978	OpenSubtitles
	... THE\n\n LIFE\n\n OF\n\n GEORGE WASHINGTON,\n\n COMMANDER IN ...	2048	Gutenberg
	Rumson, NJ – December 2013 What started in 2003 as a group of mostly Christian ...	592	realnews

Table A.3: Dataset Examples. We show text examples for all used fitting corpora. We display emojis within [] with their textual descriptions.

Genre	Domain	Train	Dev	Test
Forum	Reddit	245M	2.5M	2.5M
Literature	AO3	147M	1.5M	1.5M
Literature	Gutenberg before 1919	49M	0.5M	0.5M
News	Realnews	147M	1.5M	1.5M
News/Comments	NYTimes & Comments	24M	0.3M	0.3M
News/Comments	SFU-SOCC	3M	0.03M	0.02M
Q&A	StackExchange	196M	2.0M	2.0M
Reviews	Goodreads	49M	0.5M	0.5M
Reviews	Amazon	49M	0.5M	0.5M
Mails	Gmane	147M	1.5M	1.5M
Transcripts	YouTubeCommons	98M	0.9M	1.0M
Transcripts	OpenSubtitles	49M	0.5M	0.5M
Code	GitHub	49M	0.5M	0.5M
Science	S2ORC	98M	1.0M	0.9M
Blogs	BlogCorpus	10M	0.1M	0.1M
Raw Text Webpages	CommonCrawl	98M	1.0M	1.0M
Mathematics	DM Mathematics	20M	0.2M	0.2M
	Total:	**1,478M**	**15.1M**	**15.1M**

Table A.4: Miscellaneous Dataset Statistics

VI

name	RegEx	Example Text
NO	-	well..._$3000_for_a_tokenizer_isn't_cheapz_#lol_:)\n\nhttps://en.wikipedia.org/wiki/Sarcasm
WS	\s+	well... _ $3000 _ for _ a _ tokenizer _ isn't _ cheapz _ #lol _ :)\n\nhttps://en.wikipedia.org/wiki/Sarcasm
_WS	\s+(?!\S)|\s+	well... _$3000 _for _a _tokenizer _isn't _cheapz _#lol _:) \n \nhttps://en.wikipedia.org/wiki/Sarcasm
GPT2	's|'t|'re|'ve|'m|'ll|'d | ?\p{L}+| ?\p{N}+ | ?[^\s\p{L}\p{N}]+ |\s+(?!\S)|\s+	well ... _$ 3000 _for _a _tokenizer _isn ' t _cheapz _# lol _:) \n \n https :// en . wikipedia . org / wiki / Sarcasm
LLAMA3	(?i:'s|'t|'re|'ve|'m|'ll|'d) |[^\r\n\p{L}\p{N}]?\p{L}+ |\p{N}{1,3} | ?[^\s\p{L}\p{N}]+[\r\n]* |\s*[\r\n]+|\s+(?!\S)|\s+	well ... _$ 300 0 _for _a _tokenizer _isn 't _cheapz _# lol _:)\n\n https :// en .wikipedia .org /wiki /Sarcasm

Table A.5: Investigated Pre-tokenizers. The pre-tokenizers we investigate can be described with regular expressions. We investigate using no pre-tokenizer (NO), isolating whitespaces (WS), split on whitespaces including single leading whitespaces in non-whitespace tokens (_WS), the pre-tokenizer used by GPT-2 (GPT2) and the pre-tokenizer used by LLAMA 3 (LLAMA3). GPT-2 and LLAMA 3 mainly differ in contraction, URL, whitespace and number handling. We display how the investigated pre-tokenizer split an example text. We replace whitespaces with _ to highlight pre-token boundaries with whitespace.

A.2. Evaluation Tasks

A.2.1. Tasks Robust To Language Variation

GLUE task selection We originally planned to pre-train BERT models and test them on the same GLUE tasks (Wang et al., 2018) as the ones used for the original BERT model (Devlin et al., 2019), i.e., CoLA, SST-2, MRPC, STS-B, QQP, MNLI, QNLI and RTE. We removed CoLA (Warstadt et al., 2019) as it is the task of classifying linguistic acceptability. Models have to classify morphological, syntactic and semantic "violations" and the task can thus be expected to be sensitive to language variation. Further, we removed STS-B (Cer et al., 2017). STS-B is a regression task for the semantic similarity between two sentences. However, we aimed to focus on classification tasks that would be approachable with logistic regression. After pre-training and fine-tuning on the remaining GLUE tasks, we also removed MRPC (Dolan and Brockett, 2005) and RTE from the evaluated tasks. We removed RTE because the standard deviation of .04 makes the differences in performance for models trained with only one seed unclear, the variation could be related to the small training dataset of only 2.5k instances. We further removed MRPC as it showed almost no variation between different tokenizers—not even for the 500 vocabulary size. Removing MRPC had the additional advantage of removing another set with a small training dataset size (< 3.7k training tasks). All other tasks had a train set size of at least 67k. Note that SST-2 was released in a parsed format resulting in a lowercased and pre-tokenized text which might affect results.

GLUE-dialect We transform GLUE using `Multi-VALUE` (Ziems et al., 2023). `Multi-VALUE` has strong requirements on text formatting. When `Multi-VALUE` perturbations fail we leave the GLUE text in that row as is. Depending on the task this concerns between 3% and 18% of instances.

VI

Name	Task	Source	Text 1	Text 2	label
CORE	Register Classification	Laippala et al. (2023)	[...] What do you do when you just cant seem to mix something? Hi, You have mixed it...you just didn't know when to stop and move on! [...]	-	Interactive Discussion
AV	Authorship Verification	our contribution	Hi,\n\nI am currently evaluating OTRS for use as a Helpdesk System.\n\nHowever I am a little confused about how best to set it up. [...]	Hi,\n\nI am setting up Mailscanner not to deliver any spam to my users, but to\nsend a spam report once a day [...]	Same Author
PAN	Author Change	Bevendorff et al. (2024)	I'm not gonna watch the video. I gotta keep my sanity. With that being said, what could we call for that they haven't done? The cops have been fired and charged with murder. [...]	Maybe police agencies should be not federalized but under one agency. No more sherriffs no more local police. Just state police. [...]	Author Change
NUCLE	Error Classifications	Dahlmeier et al. (2013)	Chernobyl accident, happened in 1986, was a nuclear reactor accident.	-	[ArtOrDet, Trans]
Dialect	Dialect Classification	transformations with Ziems et al. (2023)	What the best things to do in Hong Kong one?	-	CollSgE

Table A.6: Tasks Sensitive to Language Variation. For each task, we show an example.

A.2.2. Tasks Sensitive To Language Variation

See an example for each task sensitive to language variation in Table A.6.

Authorship verification Similar to the Miscellaneous corpus in Section A.1.1, the authorship verification corpus consists of datasets taken from Amazon (Hou et al., 2024), AO3, GMane (Bevendorff et al., 2020a), 2020 NYTimes articles and comments, realnews (Zellers et al., 2019), Reddit (Baumgartner et al., 2020), StackExchange (Gao et al., 2020) and Wikipedia articles. Additionally, the corpus includes texts from BookCorpus (Zhu et al., 2015) and PubMed (Gao et al., 2020). It totals about 40.8k train pairs, 2.5k dev pairs and 4.8 test pairs.

PAN The dataset was extracted from Reddit and preprocessed by removing citations, markdown, emojis, hyperlinks, multiple line breaks and extra whitespace Bevendorff et al. (2024). Compared to the Authorship Verification task (where the classifiers may learn to rely on content cues, apostrophe encodings, whitespace encoding, etc., cf. Wegmann et al., 2022), this may be more difficult. The dataset was downloaded from `https://pan.webis.de/clef24/pan24-web/style-change-detection.html`.

NUCLE The dataset was downloaded from `https://www.comp.nus.edu.sg/~nlp/corpora.html`.

Text	label
... Sometimes, people just don't feel well. But if you don't feel well more than sometimes, it may be helpful to talk to someone about it. ...	opinion
... A transportation advocacy group is circulating a list of 100 questions aimed at broadening the British Columbia government's consultation on coastal ferry services. ...	narrative
... I'm sure many people have hit this brick wall. What do you do when you just cant seem to mix something? Hi, You have mixed it...you just didn't know when to stop and move on! ...	interactive discussion
'Always think of home': an introduction to the Buenos Ayres Notebook The Buenos Ayres Notebook takes its name from the city Buenos Aires ('Good Air' or 'Fair Winds'), ...	informational description/-explanation
... It would be a pleasure just to know just a little bit moreoh oh I could grow quite fond of your acquaintance ...	lyrical
An Acadian-style cabin constructed completely of rough sawed Southern Yellow Pine, surrounded by split rail fence. ...	informational persuasion
... it can be easy or even enjoyable. Here is a guide on how to give an oral presentation in front of your class. Decide on a topic ...	how-to/instructional
... Kareem Ettouney, the art director at Media Molecule always said, Mash up, not mish mash! ——— ——— What kinds of challenges did you face with replicating LBP's iconic 2D puppet aesthetic into a 3D space? ...	spoken

Table A.7: Examples for Main CORE Labels. We focus on the 8 main CORE Laippala et al. (2023) labels.

CORE We use 8 main register labels for multi-class prediction. We display an example for all considered CORE (Laippala et al., 2023) labels in Table A.7. The original CORE consists of main as well as sub-labels that make up a total of 56 labels in Laippala et al. (2023). However using all 56 in a multi-label setup proved too difficult for our BERT as well logistic regression models without further hyperparameter tuning. We decided for a multi-classifcation setup, limiting ourselves to 8 out of 9 main register labels, specifically we excluded the "OTHER" catgory. This reduced the train dataset from about 34k instances to 30k. We split up texts of length > 250 to chunks of a maximum of 250 using whitespace splitting. Then, we perform stratified sampling with replacement for each class to additionally upweigh small classes. Note that this results in duplicates for the small classes up to a maximum of 10 occurrences of the same text.

GPU h	# Params	# Tokens	loss	steps	batch size
0h	4.6M	10M	7.7	600	32
1h	42M	100M	6.5	6k	32
4h	110M	250M	6.2	15k	32
9.0	110M	750M	3.9	11k	128
11.0	110M	750M	**2.7**	45k	32
12.3	4.6M	330M	4.1	80k	256
13.2	110M	750M	3.2	14k	128
13.6	4.6M	3300M	4.1	75k	256
14.8	4.6M	3300M	4.1	80k	256
22.0	11.6M	3300M	3.2	75k	256

Table A.8: Hyperparameters for BERT Pre-Training. We compare the evaluation loss, and GPU hours while varying the number of parameters, tokens, steps and batch size. For similar GPU hours (between 11h-15h), using more pre-training tokens does not seem to improve performance as much as increasing model size. Balancing the number of model parameters and tokens, as well as the number of steps seems crucial.

A.3. Modeling

Compute optimal BERT We originally evaluated tokenizers on tiny BERT models, using 80k steps on a training dataset with 3.3B tokens during pre-training (second to last row in Table A.8). This corresponds to more than three epochs on the relatively large training dataset for a tiny BERT model with only 4.6M parameters. Using this setup we found that tokenizers with the largest vocabulary sizes repeatedly outperformed all other settings. For tiny BERT, models with larger vocabulary sizes also need to use orders of magnitude more parameters because of the larger embedding matrix. For tiny BERT, the model size rises to 17M for a vocabulary size of 128k. Were we using a non-optimal ratio between number of tokens and parameters for our invested 15 GPU hours? We experimented with different BERT model sizes P (tiny – 4.6M, small – 11.6M, base – 110M), number of tokens T, batch sizes and number of steps. See the results in Table A.8. We evaluate performance with the eval loss on the held out set of the pretraining corpus consisting equally of BookCorpus3 (Zhu et al., 2015) and OpenWebText2 (Gao et al., 2020). We use one tokenizer fit on the Miscellaneous corpus, using the GPT-2 pre-tokenizer and a vocabulary size of 32k tokens. For further hyperparameter choices, see the Hyperparameters paragraph. We make the following observations: For similar GPU hours (between 11h-15h), using exponentially more training corpus tokens T than parameters P for tiny BERT does not improve performance as much as increasing model size P. This and other model results might hint at Chinchilla's scaling law, that is, optimal token count scaling with the parameter size of a model for a fixed compute, specifically $T_{OPT} \approx P^{27/23}$ (Hoffmann et al., 2024), also holding for smaller encoder models. Further, for our low resource setting, using more steps, and thus weight updates, seems to be more important than using a large batch size. These conclusions seems promising but very tentative. Exhaustive pretraining experiments were out of scope for this work. Nevertheless, we think that finding compute-optimal

Source	Word Count	Excerpt
Book-Corpus	1,687,724,544	... visit and they all swore on a second blood oath that it wasn't them.\n\n"What about Phantom?" Hell Girl asked. ...
OpenWeb-Text2	1,650,000,384	... The future of SA-affiliated club sports, a cappella, and Greek groups is uncertain after the All-Campus Judicial Council ruled Friday ...
Total	3,337,724,928	

Table A.9: Pretraining Corpora. Word count is calculated using white-space splitting.

settings to train small transformer models is crucial to efficiently evaluate tokenizers going forward.

WebBook corpus Our training consists of equal parts BookCorpus3[4] (Zhu et al., 2015) and OpenWebText2 (Gao et al., 2020). See the statistics in Table A.9. We randomly sample sequences of 512 words, totaling 3.3 billion words as in the original BERT paper Devlin et al. (2019). We ensure English excerpts by removing books and web text that are not predicted as English language by using `langdetect`.[5] We split of 2% of the sampled data for held out dev and test sets to evaluate BERT pretraining.

Hyperparameters We keep the following pre-training settings the same as BERT (Devlin et al., 2019): Adam with learning rate of 1e-4, $\beta 1 = 0.9$, $\beta 2 = 0.999$, L2 weight decay of 0.01, learning rate warmup over the first 1% of steps, linear decay of the learning rate, dropout probability of 0.1 on all layers. However, we do not use the next sentence prediction (NSP) objective and only train on masked language modeling (MLM) with MLM probability at 15% as NSP proved to be inferior to MLM in later models (Liu et al., 2019). Originally, pre-training was performed on a set of 3.3 billion words over 40 epochs, we experiment with different number of tokens and steps in Section 3.5. We use the architecture of the tiny, small, medium and base BERT model (Turc et al., 2019) which consists of 4.6, 11.6, 42, and 110 million parameters respectively.[6]

A.3.1. Tasks Sensitive to Language Variation

Authorship verification For the contrastive training task, we use the Supervised Contrastive Loss[7] (Khosla et al., 2020) with a siamese setup, a batch size of 128, and a learning rate of 0.00001. We find the threshold best separating same author and distinct author pairs on the development set and report accuracy on the test set.

[4] `https://twitter.com/theshawwn/status/1320282149329784833`
[5] `https://github.com/Mimino666/langdetect`
[6] This might affect performances for models with especially big vocabulary sizes.
[7] SupConLoss as implemented in `pytorch_metric_learning`, see `https://kevinmusgrave.github.io/pytorch-metric-learning/losses/`

A.4. Intrinsic Evaluation

Renyi efficiency We record some obeservations: Using no pre-tokenizer consistently got the highest Rényi Efficency and the WS pre-tokenizer consistently got the lowest Rényi Efficency values, even though both were never among the best performing pre-tokenizers in the downstream tasks.

Varying vocabulary size Corpus Token Count is sensitive to the vocabulary size of the tokenizer. A tokenizer with a vocabulary size of 128k will almost always have a lower Corpus Token Count than a tokenizer with the vocabulary size of 32k. Independent of how many tokens might be a better fit for a given corpus or task. Similarly, Rényi efficiency (Zouhar et al., 2023) assumes the same vocabulary size to make the efficiency values comparable across tokenizers. Note that even when tokenizers have the same vocabulary size, the vocabulary coverage on the downstream corpus (i.e., the actual number of tokens that appear in the downstream corpus) might be smaller. For example, a tokenizer that was fitted on the Twitter corpus might include vocabulary that never appears in the original GLUE tasks. As a result Corpus Token Count and Rényi efficiency might be skewed for tokenizers that have a very low overlap in vocabulary with the downstream task corpus.

A.5. Modeling Results on Diverging Pre-training and Fitting Corpus

We experimented with diverging pre-training and fitting corpora. Specifically, we sampled 750 million words from WebBook (cf. Section A.3) to use as an alternative pre-training corpus. We expect WebBook to show less spelling and syntactic variation than Miscellaneous used in the main experiment. For the fitting corpus, we compared using PubMed, Wikipedia, Twitter and Miscellaneous. We further experimented with the same pre-tokenizer and vocabulary size settings as in the main experiment. See results in Table A.10, Table A.11 and Figure A.1.

For mismatched pre-training and fitting corpus, the vocabulary size needs to be higher for tasks requiring robustness to language variation. This is intuitive as more tokens increase the likelihood of more tokens being seen during pre-training.

Further, the pre-tokenizers NO, WS and _WS seem to have more trouble leveraging their tokens. Possibly because they have been fitted on a corpus different from the pre-training corpus. This is intuitive as their tokens can be expected to be especially dependent on the fitting corpus. This is probably also the explanation for LLAMA3 performing worse than GPT2 for tasks requiring sensitivity to language variation. GPT2 separates Unicode Character Categories the most and might thus have the largest overlap with the pre-training corous.

For mismatched pre-training and fitting corpus, the choice of fitting corpus seems to more influential for tasks requiring robustness to language variation. Potentially the lack of overlap between the fitting and the pre-training corpus change what tokens can be leveraged effectively. Wikipedia performs surprisingly well for tasks that require sensitivity to language variation. Wikipedia is a corpus using a very standardized version of English. We theorize that this might help with recognizing deviations from that norm.

VI

	Model	org	-typo	-dialect	AVG
Corpus	PMed	81.3	69.4	79.2	76.6
	Wiki	82.1	68.9	79.6	76.8
	Twitter	**82.4**	**70.4**	**80.5**	**77.8**
	Misc.	81.6	69.2	79.9	76.9
Pre-Tok	NO	72.3	61.6	70.9	68.3
	WS	79.5	66.9	78.1	74.8
	_WS	80.7	68.6	79.0	76.1
	LLAMA3	**82.2**	**69.4**	79.7	**77.1**
	GPT2	81.6	69.2	**79.9**	*76.9*
Size	500	79.6	**72.8**	78.1	*76.8*
	4k	80.9	70.8	78.9	*76.8*
	32k	81.6	69.2	79.9	76.9
	64k	**82.5**	69.3	**80.7**	**77.5**
	128k	81.9	68.5	79.9	76.7

Table A.10: Mismatched Pre-training and Fitting Corpus on Tasks Robust to Language Variation. We use the WebBook pre-training corpus and fit on the Miscellaneous, PubMed, Wikipedia and Twitter corpora.

	Model	AV	PAN	CORE	NUCLE	DIAL	Agg
Fitting Corpus	PubMed	80.6	65.5	55.6	23.8	88.3	62.8
	Wikipedia	80.9	67.2	56.8	**24.8**	88.7	63.7
	Twitter	80.8	64.7	57.2	23.5	**89.1**	63.0
	Misc.	**82.0**	**68.3**	**57.9**	24.1	89.0	**64.3**
Pre-Tokenizer	NO	79.8	52.4	51.5	16.9	76.6	55.4
	WS	74.9	65.2	55.7	16.6	85.3	59.5
	_WS	80.4	65.2	56.8	22.1	87.9	62.5
	LLAMA3	81.3	66.9	56.8	23.4	**89.0**	63.5
	GPT2	**82.0**	**68.3**	**57.9**	**24.1**	**89.0**	***64.3***
Vocab Size	500	77.6	62.7	53.2	15.0	86.6	59.0
	4k	80.3	65.1	56.0	21.3	88.1	62.2
	32k	82.0	**68.3**	57.9	24.1	89.0	**64.3**
	64k	82.1	67.2	**58.0**	**25.4**	88.9	**64.3**
	128k	**82.6**	66.6	54.3	24.1	**89.5**	63.4

Table A.11: Mismatched Pre-training and Fitting Corpus on Tasks Sensitive to Language Variation. We use the WebBook pre-training corpus and fit on the Miscellaneous, PubMed, Wikipedia and Twitter corpora. We display accuracy for AV, PAN and CORE and F1 for NUCLE and DIAL.

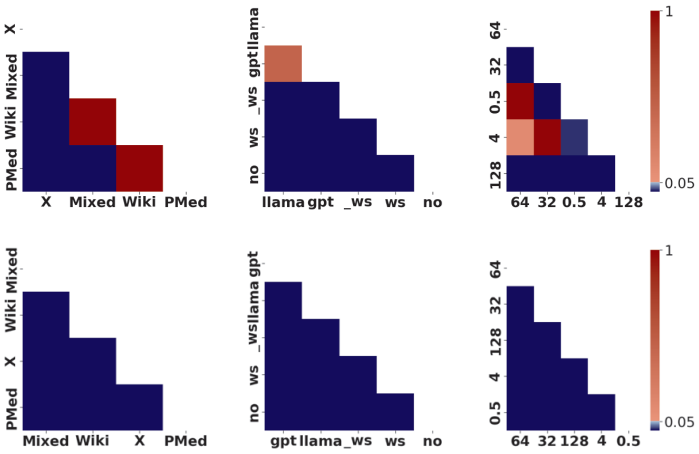


Figure A.1: Pairwise Significance Testing for Tasks Robust and Sensitive to Language Variation with Mismatched Pre-training and Fitting Corpus. We use the WebBook pre-training corpus and fit on the Miscellaneous, PubMed, Wikipedia and Twitter corpora. We use McNemar (1947)'s test to test how different the BERT models trained with different tokenizers classify the tasks robust to language variation (first row) and tasks sensitive to language variation (second row). Tokenizers are sorted by mean performance. Blue colors show statistical significance, while red colors are above the 0.05 threshold.

A.6. Intended Use and Licenses for used Datasets

VI

We discuss intended use and licenses for the datasets we re-used and created for this work.

A.6.1. Datasets curated from different sources

The Pile We include several datasets extracted from Gao et al. (2020)'s The Pile. The Pile consists of newly collected datasets as well as datastes from other sources. There is no license information included in the paper or original data release, but the Pile is described as "open source language modelling data"[8] and, even though not explicitly stated, the intended use should be for open source language modelling and research. We use the Pile to access datasets from originally other sources: books from the 1919 Gutenberg project (Rae et al., 2020)[9], StackExchange[10], OpenSubtitles (Tiedemann, 2016)[11], Common Crawl[12] and DeepMind[13] (Saxton et al., 2019). We further use the following datasets collected by the authors of the Pile: GitHub, OpenWebText2 and PubMed[14].

Reddit We use dataset originally downloaded from Pushshift (Baumgartner et al., 2020). While the original release was public, and used in many research publications, Reddit updated their terms and the original Pushshift releases are not publicly accessible anymore. Reddit mentions at least partial support of academic research after agreeing to their terms of service.[15]

Ao3 We downloaded a public release of Archive of Our Own from `https://archive.org/download/AO3_story_dump_continuing` in 2023. It did not include license nor intended use descriptions. The dataset was removed by now but should be re-creatable using tools like `https://github.com/nianeyna/ao3downloader`. Note that the license situation remains unclear. AO3 has a complex licenses where authors retain their rights and the website is granted a 'a world-wide, royalty-free, nonexclusive license to make your Content available'. AO3 terms of service forbids use of fanfictions for commercial generative AI.[16]

[8] `https://pile.eleuther.ai/`

[9] Project Gutenberg consists mostly of public US ebooks, see `https://www.gutenberg.org/policy/permission.html`

[10] anonymized data shared with a CC-BY-sa 4.0 license, see `https://archive.org/details/stackexchange`

[11] No licensing or intended use information included. Originally extracted from `https://www.opensubtitles.org/`.

[12] see Terms of Use: `https://commoncrawl.org/terms-of-use`

[13] No licensing information included in the paper.

[14] See terms of the original dataset here: `https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/README.txt`

[15] Developer Platform & Accessing Reddit Data

[16] `https://archiveofourown.org/tos_faq`

Amazon Reviews The Amazon Reviews dataset (Hou et al., 2024) was downloaded from https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/. While the dataset is publicly available, the license for the data remains with Amazon but the customers who wrote the reviews retain the copyright. There is no general site that provides guidance on the license and constraints for this data when used in the academic or research space. The guidelines for Amazon Services are noted here https://www.amazon.com/gp/help/customer/display.html?nodeId=508088

GoodReads GoodReads was publicly released with Wan and McAuley (2018); Wan et al. (2019) for academic use. We downloaded it through https://mengtingwan.github.io/data/goodreads. The GoodReads license is available at https://www.goodreads.com/about/terms. The license includes the text 'This license does not include any resale or commercial use of any part of the Service, or its contents; any collection and use of any book listings, descriptions, reviews or other material included in the Service; any derivative use of any part of the Service or its contents; any downloading, copying, or other use of account information for the benefit of any third party; or any use of data mining, robots, or similar data gathering and extraction tools.'

GMane Public mailing list emails collected from the gmane.io server, available at https://webis.de/data/webis-gmane-19.html. Released with Bevendorff et al. (2020a). Accessed through https://zenodo.org/records/3766985 after submitting a request. No license stated but not publicly available without request. Terms of use are documented on the dataset website https://zenodo.org/records/3766985.

Blogcorpus Released with Schler et al. (2006). The paper does not discuss license or intended use. Accessed through https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus. Schler et al. (2006) downloaded from https://www.blogger.com/.

BookCorpus3 Originally released with Zhu et al. (2015). Bandy and Vincent (2021) released a retrospective datasheet. Zhu et al. (2015) did not discuss intended use or licenses. However, the license for the data can be expected to remain with the original book copyright holders, except in cases where the copyright has expired.

NYTimes The dataset is publicly available at https://www.kaggle.com/datasets/benjaminawd/new-york-times-articles-comments-2020 shared with a CC BY-NC-SA 4.0 license. However, in all likelihood the license still belongs to the NYTimes while the copyright remains with the commenter. Some details are availble here https://help.nytimes.com/hc/en-us/articles/360039332111-The-New-York-Times-Content-Agreement.

Realnews Published with Zellers et al. (2019). Downloaded from https://github.com/rowanz/grover/tree/master/realnews. License can be found at this Google

Docs Form. It is intended only for research and education use and can not be distributed.

VI

SFU-Socc Released with Kolhatkar et al. (2020). Downloaded from https://github.com/sfu-discourse-lab/SOCC. Shared with Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

s2orc Released with Lo et al. (2020) mentioning research and development as intended use. Downloaded through https://github.com/allenai/s2orc/?tab=readme-ov-file. License is given as ODC-By 1.0.

YouTubeCommons Downloaded through https://huggingface.co/datasets/PleIAs/YouTube-Commons. Released with CC-BY license. All transcripts are part of a video shared under a CC-BY license.

GLUE Released a collection of tasks with Wang et al. (2018). Downloaded from https://huggingface.co/datasets/nyu-mll/glue. GLUE is a common public dataset used to evaluate language models. We use MNLI (Williams et al., 2018), SST-2 (Socher et al., 2013), QQP and QNLI.

PAN This PAN 2024 dataset was extracted from Reddit Bevendorff et al. (2024). We downloaded the dataset from https://zenodo.org/records/10677876. Reddit's terms of use might apply.

NUCLE We use the NUCLE 3.3 corpus Dahlmeier et al. (2013), downloaded from https://www.comp.nus.edu.sg/~nlp/corpora.html after submitting a request. It is available for for research purposes. License information can be found at https://sterling8.d2.comp.nus.edu.sg/nucle_download/nucle.php and does not allow for distribution of the corpus.

CORE Released with Laippala et al. (2023), downloaded from https://github.com/TurkuNLP/CORE-corpus. It is released with a CC BY-SA 4.0 license.

A.6.2. Datasets collected by us

Twitter Sampled in 2023 with Twitter research API access using the Decahose sampling stream. License to distribute tweet texts was not granted. ntended use was academic research.

A.7. Personally Identifying Information Or Offensive Content in Datasets

Some of the used datasets can be expected to include personally identifying information or offensive content. We did not take steps to remove identifiable cues or offensive

content. This was out of scope for the extensive amount of datasets used. We hope that the effect is negligible as for all datasets, except for Twitter, datasets were already publicly accessible. We acknowledge that re-distributing it might, however, make it more widely accessible. We do not release the Twitter dataset publicly.

A.8. Model Size and Budget

We used single A100s to run modeling. We pre-trained 24 distinct BERT models for our main experiments (taking less than 360 GPU hours), and fine-tuned each model for all evaluation tasks ($\approx 24 * (6h * 3$ [GLUE tasks] + $3h$ [tasks requiring sensitivity language variation]) $= 24 * 21h = 504h$).

A.9. Use of AI Assistants

We used ChatGPT and GitHub Copilot for coding, to look up commands and sporadically to generate individual functions. Generated functions were tested w.r.t. expected behavior. We used AI assistants for rephrasing and grammatical error correction.

B

Additions to Chapter 4

B.1. Task Creation

We provide details on the contraction and number substitution task creation.

B.1.1. Contraction dictionary

The Wikipedia style guide discourages contraction usage and provides a dictionary with contractions that should be avoided.[1] Some of those contractions are more colloquial (e.g., 'twas or ain't). We use an adapted version, removing colloquial and less common contractions: { "aren't": "are not", "can't": "cannot / can not", "could've": "could have", "couldn't": "could not", "didn't": "did not", "doesn't": "does not", "don't": "do not", "everybody's": "everybody is", "everyone's": "everyone is", "hadn't": "had not", "hasn't": "has not", "haven't": "have not", "he'd": "he had / he would", "he'll": "he will", "he's": "he has / he is", "here's": "here is", "how'd": "how did / how would", "how'll": "how will", "how's": "how has / how is", "I'd": "I had / I would / I should", "I'll": "I shall / I will", "I'm": "I am", "I've": "I have", "isn't": "is not", "it'd": "it would / it had", "it'll": "it shall / it will", "it's": "it has / it is", "mightn't": "might not", "mustn't": "must not", "must've": "must have", "needn't": "need not", "oughtn't": "ought not", "shan't": "shall not", "she'd": "she had / she would", "she'll": "she shall / she will", "she's": "she has / she is", "should've": "should have", "shouldn't": "should not", "somebody's": "somebody has / somebody is", "somebody'd": "somebody would / somebody had", "somebody'll": "somebody will", "someone's": "someone has / someone is", "someone'd": "someone would / someone had", "someone'll": "someone will", "something's": "something has / something is", "something'd": "something would / something had", "something'll": "something will", "that'll": "that will", "that's": "that has / that is", "that'd": "that would / that had", "there'd": "there had / there would", "there'll": "there shall / there will", "there's": "there has / there is", "there've": "there have", "these're": "these are", "they'd":

[1] https://en.wikipedia.org/wiki/Wikipedia:List_of_English_contractions

"they had / they would", "they'll": "they shall / they will", "they're": "they are", "they've": "they have", "wasn't": "was not", "we'd": "we had / we would / we should", "we'll": "we shall / we will", "we're": "we are", "we've": "we have", "weren't": "were not", "what's": "what has / what is / what does", "when's": "when has / when is", "who'd": "who would / who had", "who'll": "who will", "who's": "who has / who is", "won't": "will not", "would've": "would have", "wouldn't": "would not", "you'd": "you had / you would", "you'll": "you shall / you will", "you're": "you are", "you've": "you have"}

B.1.2. Number substitutions

We selected a pool of potential sentences where words contained character substitution symbols (4,3,1,!,0,7,5) or are part of a manually selected "seed list" of number substitution words[2]:
{ "2morrow": "tomorrow", "c00l": "cool", "n!ce":"nice", "l0ve":"love", "sw33t":"sweet", "l00k":"look", "4ever":"forever", "l33t":"leet", "1337":"leet", "sk8r":"skater" "n00b":"noob", "d00d":"dude", "ph34r":"fear", "w00t":"woot", "b4":"before", "gr8":"great", "2day":"today", "t3h":"teh", "m4d":"mad", "j00":"joo", "0wn":"own", "h8":"hate", "w8":"wait" }

Then, we manually removed sentences without number substitutions (e.g., common measuring units or product numbers). Our resulting list of 100 sentences pairs contains more substitution words than the above "seed list" (e.g., "d4rk", "appreci8", "h1m").

[2] Inspired by `https://www.gamehouse.com/blog/leet-speak-cheat-sheet/`, `https://simple.wikipedia.org/wiki/Leet`, Ganushchak et al. (2012), `https://h2g2.com/edited_entry/A787917` and manually looking at a few Reddit posts

B.2. Similarity-based Decision

In Figure B.1, we provide a proof sketch for the Formula (4.1) in Section 4.5.2, from

$$(1 - \text{sim}(A1, S1))^2 + (1 - \text{sim}(A2, S2))^2 <$$
$$(1 - \text{sim}(A1, S2))^2 + (1 - \text{sim}(A2, S1))^2$$

follows S1-S2.

VI

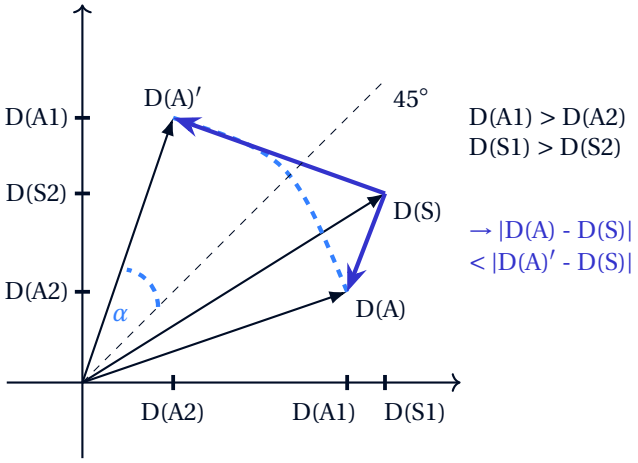


Figure B.1: Proof Sketch. Let D be the considered style component (e.g., formal/informal) and $D(\text{A1})$, $D(\text{A2})$, $D(\text{S1})$, $D(\text{S2})$ be the localization of A1, A2, S1, S2 along that component. We want to show that the inequality (4.1) holds when S1-S2 (i.e., $D(\text{A1}) > D(\text{A2})$ and $D(\text{S1}) > D(\text{S2})$ or $D(\text{A1}) < D(\text{A2})$ and $D(\text{S1}) < D(\text{S2})$, this was also denoted as ✓ before). With a similar approach one can show that inequality (4.1) with $>$ holds when S2-S1. As we are interested in the cases S1-S2 and S2-S1 exactly, we can use the formula to determine orderings. W.l.o.g., let $D(\text{S1}) > D(\text{S2})$. Let us assume that for all other style and content aspects $\widetilde{D}$ (e.g., simple/-complex), $\widetilde{D}(\text{A1}) = \widetilde{D}(\text{A2})$ and $\widetilde{D}(\text{S1}) = \widetilde{D}(\text{S2})$ hold. We define $D(\text{A}) := \begin{pmatrix} D(\text{A1}) & D(\text{A2}) \end{pmatrix}^\intercal$ and $D(\text{S}) := \begin{pmatrix} D(\text{S1}) & D(\text{S2}) \end{pmatrix}^\intercal$ as the style vectors of the combined anchor (A1 and A2) and alternative sentences (S1 and S2). Then, with the correct ordering being S1-S2, $D(\text{S1}) > D(\text{S2})$ holds iff $D(\text{A1}) > D(\text{A2})$. Thus, both $D(\text{A})$ and $D(\text{S})$ point to a coordinate below the 45°-axis when the first component of the respective vectors corresponds to the x-axis and the second to the y-axis (see sketch). Let $D(\text{A})'$ be the reflected vector of $D(\text{A})$ along the 45°-axis, i.e., $\begin{pmatrix} D(\text{A2}) & D(\text{A1}) \end{pmatrix}^\intercal$. Then, the angle between $D(\text{S})$ and $D(\text{A})$ will always be smaller than the one between $D(\text{S})$ and $D(\text{A})'$, because $D(\text{A})$ and $D(\text{S})$ are on the same side, and the reflection of $D(\text{A})$ has to be on the other side, with the same angle towards the 45°-axis. Then, the length of the vector $D(\text{A}) - D(\text{S})$ is smaller than $D(\text{A})' - D(\text{S})$, i.e., $(D(\text{A1}) - D(\text{S1}))^2 + (D(\text{A2}) - D(\text{S2}))^2 < (D(\text{A2}) - D(\text{S1}))^2 + (D(\text{A1}) - D(\text{S2}))^2$. This remaining step to inequality (4.1) is replacing the distance components (i.e., $(x-y)^2$) with the way we compare the location of x to y on the considered style axis D as we do not have concrete but only relative locations of x and y: cosine similarity-based distances (i.e., $(1-\text{sim}(x,y))^2$). Note: As only cosine 'angular distance' is a distance metric, we technically would need the angular cosine similarity to replace the squared euclidean distance in (4.1). However, angular cosine similarity can be replaced by cosine similarity in inequality (4.1) as relative ordering is the same for the two similarity metrics.

B.3. Removing Ambiguity

B.3.1. Annotation Setup

We display the task description (Figure B.2 and B.3) as well as the examples of the annotation task (Figure B.4 and B.5). `Prolific` crowd workers could participate up to five times in annotating different created tasks from the formal/informal and simple/-complex style dimensions. Each time a participant was asked to annotate 14 task instances.

Annotator screening Per study, the annotators saw two screening questions, randomly sampled from a list of 10 screening questions (see Table B.1). The screening questions were manually created and then unanimously and correctly answered by three lab-internal annotators in the triple setting. Crowd workers who answered any of the screening questions incorrectly were excluded.

Annotator payment Participants were paid 10,21£/hour on average (above 8.91£ UK minimum wage at the time of the study—April 2021—see `https://www.gov.uk/national-minimum-wage-rates`) and gave consent to the publication of their annotations.

Annotator requirements We required annotators to be native English speakers as we assume them to have a better intuition about their language than non-native speakers: During study design, we conducted a pilot study with 8 different non-native annotators. Several felt their English-speaking abilities were insufficient for the task. Further, the pilot study projected a higher perceived and measured difficulty of the simple/-complex dimension. As a result we required annotators to be native speakers and generated more potential simple/complex than formal/informal tasks.

VI

In this study, you will be expected to complete the following task:

Compare the linguistic style of text snippets.

Opposed to content, **style is not about "what" is said but about "how" it is said.**

The survey will take place in the form of multiple choice questions of the same setup. An example could be:

Question: Given the text snippet

It reminds me of an old song from the Beatles.

which of the following is more consistent in linguistic style?

Alternative A: KEVIN n nfnhfnigbubjbni.....I dunt really watch American Idol.....
Alternative B: Kevin, I am not exactly an 'American Idol' viewer.

Here, alternative B is more consistent in style as alternative A is noticeably more informal (e.g., 'nfnhf' or 'dunt really') than the other two text snippets.

Another example could be:

Question: Given the text snippet

This stamp became the standard for the remnant of Victoria's reign, and vast quantities were printed.

which of the following is more consistent in linguistic style?

Alternative A: Both names became defunct in 2007 when they were merged into The National Museum of Scotland.
Alternative B: Both names stopped being used in 2007 when they became a part of The National Museum of Scotland.

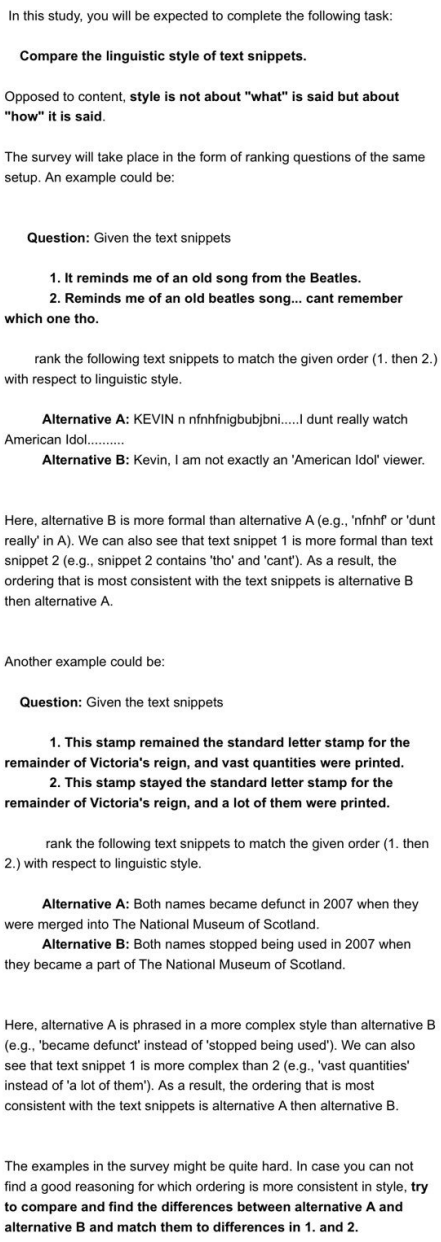
Here, alternative A is more consistent in style as alternative B is noticeably less complex (e.g., 'stopped being used' instead of 'became defunct') than the other two text snippets.

The examples in the survey might be quite hard. In case you can not find a good reasoning for which alternative is more consistent in style, **try to compare and find the differences between alternative A and alternative B**.

Figure B.2: Survey Task Description for the Triple Setup. This is a screenshot of what was shown to the crowd workers.

Comp	GT	Anchor 1 (A1)	Anchor 2 (A2)	Sentence 1 (S1)	Sentence 2 (S2)
formal/-informal	✓	They were engaging in intercourse.	They were having sex.	You do not have the perspective.	It's cause ya got no sense.
formal/-informal	✗	OH, REALLY?	Oh, is that so?	Girlfriends is one of my favorite shows on television.	GIRLFRIENDS IS ONE OF MY FAVORITE SHOWS.
simple/-complex	✓	Many species had vanished by the end of the nineteenth century.	Many animals had disappeared by the end of the 1800s.	They are culturally akin.	Their culture is like the other.
simple/-complex	✓	This stamp remained the standard letter stamp for the remainder of Victoria's reign, and vast quantities were printed.	This stamp stayed the standard letter stamp for the remainder of Victoria's reign, and a lot of them were printed.	Both names became defunct in 2007 when they were merged into The National Museum of Scotland.	Both names stopped being used in 2007 when they became a part of The National Museum of Scotland.
numb3r subs	✗	You are a n00b.	You are a noob.	This is cool.	This is c00l.
numb3r subs	✓	|-|0w n3rdy d0 y0u 1!k3 !7¿ ! d0 h4v3 4 107 0f !7 ;-)	How nerdy do you like it? I do have a lot of it ;-)	lol iM N0t CH3At1ng!	lol iM Not CHeating!
Shakespeare	✓	Why, uncle tis a shame.	It's a shame, uncle.	O, wilt thou leave me so unsatisfied?	Oh, you're gonna leave me unsatisfied, right?
formal/-informal	✗	i got limewire if i download songs on it will i get a ticket???	Will I get a ticket if I download songs?	The original song is very good.	The original song is like too good.....
formal/-informal	✗	I like the Click Five and enjoy their songs.	The Click Five...they totally rock!their songs are out of this world!!	i play guitar and some piano .....yet i cant read a note of music....lol	I can not read music, but I can play guitar and piano.
formal/-informal	✗	It reminds me of an old song from the Beatles.	Reminds me of an old beatles song... cant remember which one tho.	KEVIN n nfnhfnigbubjbni.....I dunt really watch American Idol.....	Kevin, I am not exactly an 'American Idol' viewer.

Table B.1: Screening Questions. List of manually created screening questions to ensure annotator quality. Anchor 2 is only used in the quadruple setup. The task is to match anchor 1 and anchor 2 with sentence 1 and sentence 2. The order is either as is (✓) or needs to be reversed (✗). The correct matching is given in the GT column. The Shakespeare example was taken from Krishna et al. (2020). The rest were either inspired or taken from Xu et al. (2016), Rao and Tetreault (2018) and Baumgartner et al. (2020).



In this study, you will be expected to complete the following task:

Compare the linguistic style of text snippets.

Opposed to content, **style is not about "what" is said but about "how" it is said**.

The survey will take place in the form of ranking questions of the same setup. An example could be:

Question: Given the text snippets

1. It reminds me of an old song from the Beatles.
2. Reminds me of an old beatles song... cant remember which one tho.

rank the following text snippets to match the given order (1. then 2.) with respect to linguistic style.

Alternative A: KEVIN n nfnhfnigbubjbni.....I dunt really watch American Idol.....
Alternative B: Kevin, I am not exactly an 'American Idol' viewer.

Here, alternative B is more formal than alternative A (e.g., 'nfnhf' or 'dunt really' in A). We can also see that text snippet 1 is more formal than text snippet 2 (e.g., snippet 2 contains 'tho' and 'cant'). As a result, the ordering that is most consistent with the text snippets is alternative B then alternative A.

Another example could be:

Question: Given the text snippets

1. This stamp remained the standard letter stamp for the remainder of Victoria's reign, and vast quantities were printed.
2. This stamp stayed the standard letter stamp for the remainder of Victoria's reign, and a lot of them were printed.

rank the following text snippets to match the given order (1. then 2.) with respect to linguistic style.

Alternative A: Both names became defunct in 2007 when they were merged into The National Museum of Scotland.
Alternative B: Both names stopped being used in 2007 when they became a part of The National Museum of Scotland.

Here, alternative A is phrased in a more complex style than alternative B (e.g., 'became defunct' instead of 'stopped being used'). We can also see that text snippet 1 is more complex than 2 (e.g., 'vast quantities' instead of 'a lot of them'). As a result, the ordering that is most consistent with the text snippets is alternative A then alternative B.

The examples in the survey might be quite hard. In case you can not find a good reasoning for which ordering is more consistent in style, **try to compare and find the differences between alternative A and alternative B and match them to differences in 1. and 2.**

Figure B.3: Survey Task Description for the Quadruple Setup. This is a screenshot of what was shown to the crowd workers.

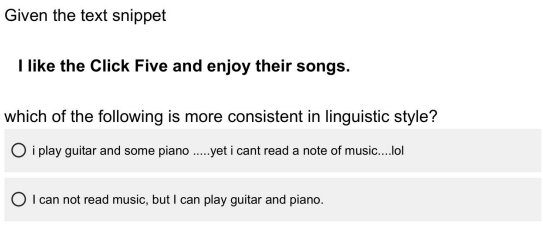


Figure B.4: Example Survey Question for the Triple Setup. This is an example of what was shown to the crowd workers.

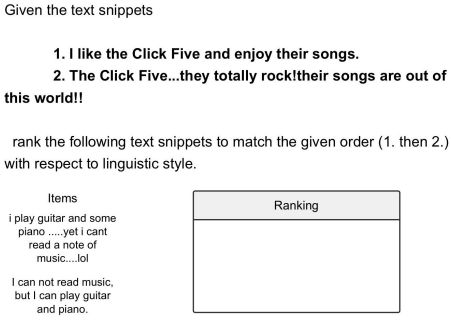


Figure B.5: Example Survey Question for the Quadruple Setup. This is an example of what was shown to the crowd workers.

B.3.2. Annotation Results

We provide additional results in Table B.2 for the annotations performed in Section 4.4.3 in Chapter 4. Specifically, we provide examples from the sample of 601 task instances that were annotated for both the quadruple and the triple setup.

For both style dimensions (301 instances respectively), the most common annotation combinations are ✓✓and ✗✓—i.e., annotated correctly for quadruple and triple setup and annotated correctly for quadruple, but incorrectly for the triple setup—totaling 68.1% and 88.7% for complex/simple and formal/informal respectively.

The Table shows ambiguous examples, where one could argue for both possible orders. After manual inspection, this seems to be more prevalent for the simple/complex dimension but it also happens for the formal/informal style dimension. E.g., for row (formal, ✗✗), anchor 1 could be understood as more formal (e.g., 'gentleman') or more informal (e.g., '!' and an unusual grammatical structure). Row (formal, ✗✓) is an example of the "Triple Problem" (cf. Section 4.4.2).

Component					Example			
d	**T**	**Q**	**%**	**GT**	**Anchor 1 (A1)**	**Anchor 2 (A2)**	**Sentence 1 (S1)**	**Sentence 2 (S2)**
f	✗	✓	59 ≈ 0.196	✓	List your best April Fools Pranks here	Please compile a list on here of your best April Fool pranks.	becuase in one of her songs she talks about saying no to sex pressure from her boyfriend	In one of her songs, she addresses the issue of not letting her boyfriend pressure her into having sexual intercourse.
c	✗	✓	94 ≈ 0.312	✓	The Book of Nehemiah is a book of the Hebrew Bible, historically seen as a follow-up to the Book of Ezra, and is sometimes called the second book of Ezra.	The Book of Nehemiah is a book of the Hebrew Bible, historically regarded as a continuation of the Book of Ezra, and is sometimes called the second book of Ezra.	All the bats look up to him, and he says he caught two tiger moths which everyone in the colony knows to be a difficult feat for such a young bat	All the bats admire him, and he claims to have caught two tiger moths which are known by all the others in the colony to be an extraordinary achievement by such a young bat.
f	✓	✗	14 ≈ 0.047	✗	pointsreaper is lame he cannot sue Yahoo for him cheating, what a cry baby	He is not smart. You can not sue a website because you cheated.	A woman did not perform the vocals.	A girl did not sing it.
c	✓	✗	42 ≈ 0.14	✓	Meanwhile the KLI has about 20 of those former Beginners' Grammarians.	Meanwhile, the KLI has about 20 of those past Beginner's Grammarians.	N-Dubz are a MOBO award winning hip hop group from London, based around Camden Town.	N-Dubz is a MOBO award winning hip hop group, based around Camden Town in London.
f	✗	✗	20 ≈ 0.066	✓	Gentleman, and I thank God everyday for the one that I have!	I thank God for each day that I have.	GIRLFRIENDS IS ONE OF MY FAVORITE SHOWS.	Girlfriends is one of my favorite shows on television.
c	✗	✗	54 ≈ 0.179	✓	Among the casualties were two fishers who were reported missing.	Two fisherman are missing among the people who may have been hurt or killed.	Baduhennna is solely attested by Tacitus' Annals where Tacitus records that a grove in Frisia was dedicated to her, and that near this grove 900 Roman prisoners were killed in 28 CE.	In Tacitus' Annals by Tacitus, it is recorded that a grove in Frisia was dedicated to Baduhennna, and near to this grove 900 Roman prisoners were killed in 28 CE.
f	✓	✓	208 ≈ 0.691	✓	im pretty sure that it was kiss	I am fairly certain it was a kiss.	Law and Order... it just has a clunk clunk	I like Law and Order, although it is a bit clunky lately.
c	✓	✓	111 ≈ 0.369	✓	Mifepristone is a synthetic steroid compound used as a pharmaceutical.	Mifepristone is a synthetic steroid compound which is used as a medicine.	The video was released on 7/14/06.	The video was premiered on MTV2 on July 14, 2006.

Table B.2: Annotation Analysis. In total 602 examples were annotated for both quadruple (Q) and triple (T) settings with 301 per style dimension, cf. Section 4.4.3. Annotations were compared to the automatically inferred ground truth (GT). For the simple/complex (c) and the formal/informal (f) dimensions, we give the number of occurrences of each combination of correct (✓) and wrong (✗) annotations for T and Q. For every combination an example is given.

	all		formal		complex		nb3r	c'tion
	filter	full	filter	full	filter	full		
BERT UNCASED	0	0	0	0	0	0	0	0
BERT CASED	0	0	0	0	0	0	0	0
ROBERTA	0	0	0	0	0	0	0	0
SBERT MPNET	0	0	0	0	0	0	0	0
SBERT PARA-MPNET	0	0	0	0	0	0	0	0
USE	0.00	0.00	0.00	0.00	0.00	0.01	0	0
BERT UNCASED NSP	0	0	0	0	0	0	0	0
BERT CASED NSP	0	0	0	0	0	0	0	0
char 3-gram	0.05	0.05	0.03	0.04	0.01	0.01	0.57	0.02
word length	0.08	0.08	0.04	0.05	0.04	0.04	0.91	0
punctuation	**0.38**	**0.39**	**0.31**	**0.31**	**0.42**	**0.42**	**0.97**	0.06
LIWC	0.09	0.09	0.01	0.01	0.12	0.13	0.53	0
LIWC (style)	**0.62**	**0.64**	**0.37**	**0.38**	**0.80**	**0.79**	**0.94**	**1.0**
LIWC (function)	0.28	0.28	0.14	0.14	0.38	0.38	0.81	0
deepstyle	0	0	0	0	0	0	0	0
POS Tag	0.20	0.20	0.02	0.02	0.24	0.24	0.64	**1.0**
share cased	0.08	0.08	0.02	0.02	0.05	0.05	0.91	0
edit dist	0.08	0.07	0.01	0.01	0.05	0.05	0.52	**0.33**
AVERAGE	0.11	0.11	0.05	0.05	0.13	0.13	0.38	0.15

Table B.3: Share of Random Decisions. The share of task instances for which a method can not decide between the two options is given per component. The performance on the set of task instances before (full) and after crowd-sourced filtering (filter) is displayed. The two highest shares of random decisions are **boldfaced**. The share of random decisions is highest for the nb3r and lowest for the formal dimension. LIWC (style) and punctuation similarity have the overall highest share of random decisions.

B.4. Additional STEL Results

In Table B.3, we display the share of task instances where models and methods could not decide between the two possible answers. This is adding more detail to the 'random' column of Table 5.3. The share of random decisions is lower for the more complex style dimensions (formal/informal: 0.05 and simple/complex: 0.13) and higher for the simpler style characteristics (nb3r substitution: 0.38 and contrac'tion usage: 0.15). This aligns with the intuition that the difference between the sentence pairs in the numb3r substitution and contrac'tion dimension is smaller. The neural methods have a lower share of random decisions overall.

B.5. Computing Infrastructure

The evaluation of the 18 (language) models and methods took 14 hours in total on a machine with 32 GB RAM and 8 intel i7 CPUs using Ubuntu 20.04 LTS. No GPU was used.

C

Additions to Chapter 5

		conversation		domain		no	
		CAV	AV	CAV	AV	CAV	AV
loss	**margin**	acc	auc	acc	auc	acc	auc
contrastive	0.4	0.63	0.63	**0.68**	**0.68**	**0.71**	**0.71**
	0.5	0.63	0.63	**0.68**	**0.68**	**0.71**	**0.71**
	0.6	0.62	0.63	**0.68**	**0.68**	**0.71**	**0.71**
triple	0.4	0.63	0.62	**0.68**	0.67	0.70	0.70
	0.5	**0.64**	**0.64**	**0.68**	**0.68**	0.70	0.70
	0.6	0.63	0.63	0.67	0.67	0.70	0.70
contrastive online	0.4	0.58	0.58	0.64	0.64	0.67	0.67
	0.5	0.58	0.58	0.64	0.64	0.67	0.67
	0.6	0.58	0.58	0.64	0.64	0.67	0.67

Table C.1: Hyperparameter-tuning Results on the Development AV and CAV Datasets with Varying Content Control. Results for BERT uncased trained on the contrastive authorship verification tasks (CAV). With different loss functions (contrastive, triple, contrastive online) and margin values (0.4, 0.5, 0.6). For each development set (conversation, domain and no content control), we display the accuracy of the models for the CAV task and the AUC for the authorship verification task (AV). For each development set and CAV/AV setup, the best performance is **boldfaced**. Contrastive and triple loss behave comparable. The margin value only has a small influence.

C.1. Results on the Development Set

C.1.1. Hyperparameter Tuning

Loss functions We evaluated contrastive (on the AV training setup), triple (on the CAV training setup) and online contrastive loss (on the AV training setup) using implementations from `sentence-transformers`. We experiment with the loss hyperparameter "margin" with values of 0.4, 0.5, 0.6 for the uncased BERT model (Devlin et al.,

		conv		sub		no	
		CAV acc	AV AUC	CAV acc	AV AUC	CAV acc	AV AUC
-	BERT	0.52	0.51	0.59	0.57	0.64	0.61
	BERT	0.53	0.52	0.59	0.57	0.63	0.60
	RoBERTa	0.53	0.53	0.58	0.57	0.63	0.61
v	BERT c 0.5	0.65	0.66	0.66	0.67	0.68	0.68
	BERT t 0.5	0.65	0.66	0.66	0.67	0.67	0.68
	BERT c 0.5	0.66	0.67	0.67	0.68	0.69	0.70
	BERT t 0.5	0.66	0.67	0.67	0.68	0.68	0.69
	RoBTa c 0.5	**0.69**	**0.70**	0.70	0.71	0.70	0.72
	RoBTa t 0.5	0.68	0.69	0.69	0.70	0.70	0.70
d	BERT c 0.5	0.63	0.63	0.68	0.68	0.71	0.71
	BERT t 0.5	0.64	0.64	0.68	0.68	0.70	0.70
	BERT t 0.5	0.65	0.65	0.68	0.68	0.71	0.71
	BERT c 0.5	0.64	0.65	0.69	0.69	0.71	0.72
	RoBTa c 0.5	0.67	0.68	**0.71**	**0.72**	0.73	0.74
	RoBTa t 0.5	0.68	0.68	0.70	0.70	0.72	0.73
n	BERT c-0.5	0.55	0.54	0.63	0.62	0.76	0.76
	BERT t-0.5	0.55	0.54	0.62	0.61	0.74	0.75
	BERT c 0.5	0.57	0.56	0.64	0.63	0.76	0.77
	BERT t 0.5	0.58	0.56	0.64	0.62	0.75	0.75
	RoBTa c 0.5	0.59	0.58	0.65	0.64	**0.77**	**0.78**
	RoBTa t 0.5	0.59	0.57	0.65	0.63	**0.77**	0.77

(a) CAV and AV Performance

		conv		sub		no	
		AV		AV		AV	
		thr	acc	thr	acc	thr	acc
-	BERT	0.82	0.51	0.70	0.55	0.69	0.58
	BERT	0.86	0.51	0.85	0.55	0.85	0.58
	RoBERTa	0.96	0.52	0.97	0.55	0.97	0.58
v	BERT c 0.5	0.72	0.61	0.73	0.62	0.73	0.63
	BERT t 0.5	0.27	0.61	0.27	0.62	0.29	0.63
	BERT c 0.5	0.24	0.62	0.28	0.63	0.26	0.64
	BERT t 0.5	0.72	0.62	0.73	0.63	0.73	0.64
	RoBTa c 0.5	0.72	**0.64**	0.72	0.64	0.73	0.65
	RoBTa t 0.5	0.30	0.63	0.31	0.64	0.32	0.64
d	BERT c 0.5	0.73	0.59	0.73	0.63	0.73	0.65
	BERT t 0.5	0.16	0.60	0.19	0.63	0.19	0.64
	BERT t 0.5	0.20	0.61	0.27	0.63	0.23	0.65
	BERT c 0.5	0.74	0.60	0.74	0.64	0.72	0.66
	RoBTa c 0.5	0.72	0.63	0.72	**0.65**	0.72	0.67
	RoBTa t 0.5	0.22	0.63	0.24	**0.65**	0.19	0.66
n	BERT c-0.5	0.76	0.53	0.77	0.58	0.74	0.69
	BERT t-0.5	0.14	0.53	0.37	0.57	0.24	0.68
	BERT c 0.5	0.40	0.54	0.35	0.59	0.23	0.69
	BERT t 0.5	0.74	0.54	0.76	0.59	0.74	0.69
	RoBTa c 0.5	0.80	0.56	0.77	0.60	0.74	**0.71**
	RoBTa t 0.5	0.38	0.55	0.34	0.59	0.19	0.66

(b) Details on the AV results

Table C.2: (Dev) Results. We display the accuracy of the models for the contrastive authorship verification (CAV) setup and the AUC for the authorship verification (AV) setup on each dev set (conversation, domain and no). We show results for the base models (-), and 18 fine-tuned models: BERT uncased (BERT), RoBERTa and BERT cased trained with the conversation (v), domain (d) and no content control (n). With different loss functions (contrastive - c, triple - t) and margin values (0.4, 0.5, 0.6). For the AV task, we also display the optimal threshold according to AUC (thr) and its matching accuracy. Generally, RoBERTa models perform the best with increasing performance from conversation to domain to random. Accuracies for CAV are higher than for AV. Models perform the best on the task they have been trained on. Contrastive and Triple loss seem to behave comparable. Best performance per dev set and CAV/AV task is boldfaced.

2019) on the domain training data. Results on the development sets are displayed in Figure C.1. Contrastive and triplet loss perform better than online contrastive loss. The margin value only has a small influence on the performance scores. Based on these results, we decided to run all further models with the contrastive and triplet loss functions and a margin value of 0.5.

C.1.2. Detailed Results on the Development Sets

We display the performance of the fine-tuned models on the development sets in Table C.2. RoBERTa (Liu et al., 2019) generally performs better than the uncased and cased BERT model (Devlin et al., 2019). Performance for the triplet and contrastive loss functions are comparable. We only use RoBERTa models in the main paper and both contrastive and triplet loss as a result.

train data	model	all		formal		complex		nb3r		c'tion	
		STEL	o-c	STEL	o-c	STEL	o-c	STEL	o-c	STEL	o-c
-	BERT	0.75	0.03	0.76	0.05	0.70	0.00	**0.93**	0.09	1.00	0.00
	BERT	**0.78**	0.05	0.80	0.10	**0.71**	0.00	0.92	**0.11**	1.00	0.00
conv.	BERT c 0.5	0.68	0.21	0.72	0.40	0.59	0.07	0.73	0.06	1.00	0.01
	BERT t 0.5	0.68	0.30	0.71	0.52	0.61	0.15	0.72	0.05	0.99	0.06
	BERT c 0.5	0.73	0.32	0.83	0.62	0.60	**0.19**	0.67	0.06	1.00	0.00
	BERT t 0.5	0.73	**0.37**	0.79	**0.66**	0.63	0.15	0.74	0.05	1.00	**0.15**
domain	BERT c 0.4	0.70	0.12	0.76	0.26	0.61	0.01	0.72	0.02	1.00	0.00
	BERT c 0.5	0.69	0.13	0.74	0.27	0.59	0.01	0.68	0.05	1.00	0.00
	BERT c 0.6	0.70	0.13	0.76	0.26	0.61	0.01	0.72	0.04	1.00	0.00
	BERT t 0.4	0.71	0.15	0.78	0.31	0.59	0.01	0.78	0.05	1.00	0.00
	BERT t 0.5	0.68	0.18	0.74	0.37	0.58	0.03	0.72	0.06	1.00	0.00
	BERT t 0.6	0.69	0.22	0.76	0.44	0.58	0.04	0.69	0.06	1.00	0.00
	BERT c-0.5	0.73	0.23	0.82	0.48	0.61	0.02	0.77	0.03	1.00	0.00
	BERT t-0.5	0.71	0.28	0.81	0.56	0.57	0.06	0.80	0.04	1.00	0.00
random	BERT c 0.5	0.69	0.09	0.77	0.20	0.58	0.01	0.68	0.02	0.98	0.00
	BERT t 0.5	0.70	0.13	0.75	0.26	0.61	0.03	0.79	0.06	1.00	0.00
	BERT c-0.5	0.72	0.21	**0.84**	0.44	0.55	0.02	0.75	0.07	1.00	0.01
	BERT t-0.5	0.73	0.23	**0.84**	0.48	0.59	0.03	0.68	0.05	1.00	0.00

Table C.3: Results on STEL and STEL-Or-Content. We display STEL accuracy for different language models and methods. BERT stands for uncased BERT base model and BERT stands for cased BERT base model. The performance on the set of STEL and STEL-Or-Content (o-c) task instances is displayed. The best performance is **boldfaced**. Performance for the trained models goes down for the original STEL framework in the complex/simple and nb3r substitution dimension. Performance generally increases for the STEL-Or-Content task.

C.2. Details on STEL Results

We display the STEL results on further trained models in Table C.3. Interestingly, cased BERT seems to be the better choice for the contraction STEL dimension.

	unlearned	**learned**
no ambiguity	$\frac{5}{55} \approx 9\%$	$\frac{12}{41} \approx 29\%$
typo simple	$\frac{21}{55} \approx 38\%$	$\frac{13}{41} \approx 32\%$
typo complex	$\frac{11}{55} \approx 20\%$	$\frac{6}{41} \approx 15\%$
error grammar simple	$\frac{15}{55} \approx 27\%$	$\frac{9}{41} \approx 22\%$
error grammar complex	$\frac{5}{55} \approx 9\%$	$\frac{3}{41} \approx 7\%$
changed content	$\frac{5}{55} \approx 9\%$	$\frac{3}{41} \approx 7\%$
word as/more complex	$\frac{16}{55} \approx 29\%$	$\frac{11}{41} \approx 27\%$
naturalness	$\frac{7}{55} \approx 13\%$	$\frac{3}{41} \approx 7\%$

Table C.4: Categories Error Analysis STEL Results. For the six fine-tuned ROBERTA models (AV/CAVxconversation/domain/no), we manually inspected at the 55 commonly learned as well as the 41 commonly unlearned simple/complex examples. We label the examples for the displayed ambiguity classes.

C.2.1. Error Analysis ROBERTA STEL results

We manually inspect the complex/simple STEL instances that were commonly learned and unlearned by the ROBERTA models and annotate if they contain ambiguities. In Table C.4, we display the results. Overall, the learned STEL instances contain fewer ambiguities. However, they still show considerable amounts of ambiguities.

n	avgerage silhouette score
2	0.23
3	0.21
4	0.23
5	**0.27**
6	**0.27**
7	0.26
8	0.23
9	0.19
10	0.20
11	0.19
12	0.18
13	0.19
14	0.17
15	0.16
16	0.16
17	0.16
18	0.17
19	0.17
20	0.17
21	0.16
22	0.16
23	0.15
24	0.15
25	0.15
26	0.15
30	0.15
40	0.15
50	0.15
100	0.13
150	0.13
200	0.12

Table C.5: Silhouette Values. We experiment with different numbers of clusters for one fine-tuned RoBERTa model (R CAV CONV 106). The highest silhouette score is reached for cluster sizes of 5–7.

C.3. Details on Cluster Parameters

We use agglomerative clustering for the RoBERTa model trained on the CAV setup with a margin of 0.5 and conversations as content control with seed 106 (R CAV CONV 106). We experiment with different numbers of clusters and display the results in Table C.5. The highest Silhouette scores are reached for cluster sizes of 5, 6, 7. We select a cluster size of 7 for evaluation.

C.4. Details on the Cluster Analysis

We give more examples of the seven clusters for our fine-tuned RoBERTa model in Table C.6 and for the base RoBERTa model in Table C.7. Refer to our Github repository for the complete clustering. We did not find obvious consistencies for clusters 1, 2 and 6. That does, however, not mean that more nuanced stylistic consistencies are not present. We recommend using a higher number of clusters, possibly different clustering algorithms and testing out statistics for known style features to pinpoint more consistencies.

Out of all utterance pairs that have the same author, 46.2% appear in the same cluster for the style embedding model. This is different from a random distribution among 7 clusters[1] which corresponds to 20.1% ± .00. As authors will have a certain variability to their style as well (e.g., Zhu and Jurgens, 2021), a perfect clustering according to writing style would not assign all same author pairs to the same cluster. For the RoBERTa base model the fraction of same author pairs in the same cluster is closer to the random distribution (75.4% vs. 76.1% for the random distribution[2]). The fraction of utterance pairs that appear in the same domain are close to the random distribution for both the style embedding model (23.6% vs. 20.1%) and the RoBERTa base model (77.6% vs. 76.0%). Results are similar for utterance pairs that appear in the same conversation.

[1] Calculated mean and standard deviation of 100 random assignments of utterances to the 7 clusters, with the same number of elements in each cluster.

[2] The share is high for RoBERTa base because the first cluster already contains 86.7% of all utterances. Random assignment of utterances across the 7 clusters, that keeps the clustering size would already lead to 76.1% same author pairs appearing in the same cluster (almost all of them in the first).

C	#	Consistency	Example 1	Example 2	Example 3
1	4065	citing previous comments, standard punctuation, URLs	Yes. Proportionally, this kid's feet are absolutely enormous.	> Please delete your account. Says the no life who always shits on anything Kanye or anti-Drake I can promise you that capitalism is very much alive in Norway.	[This should help.](YOUTUBE-LINK)
2	4016	short sentences?	Nice catch! Well done. cookies are in the back of this Grammar party. You can have two.	You can mute them we've been told!	Came here to post this only to find it's already the top voted comment. This is a good sub.
3	2165	no last punctuation mark	I am living in china, they are experiencing an enormous baby boom	Seems like sarcasm. But could also be Poe	[...] The earth probably has two or more degrees of symmetry, but less than infinite (like a sphere), but I'm honestly not too concerned about the minutiae of it
4	1794	punctuation / casing	huh thats odd i'm in the 97% percentile on iq tests, the sat, and the act	Its not a problem if you a got a full game. Whats the problem if a game didnt get expansions?	Fair point, I didnt know that. Just at glance I kind of went 'woah that doesnt seem right'
5	1555	' instead of ' apostrophe	I assume it's the blind lady?	Oh I wasn't really dismissing them. I'm saying Ford will try their own thing compared to Fiat	It's 4am in Brussels and I am still hyped
6	781	similar to 1?	Well, as your neighbors, I'd say Fuck you.. But we're not like that, see? We want to be part of the alliance, not part of the 'fuck you, we cant be competitive with jobs or innovate any more, so we're going to run massive tariffs against all our friendly nations	Hah, thus the one calf larger than the other issue. I have it too ;)	[So you are saying that current encryption falls apart as long as the quantum computer is large enough](URL). (for reference, the current highest qubit is 50)'
7	380	linebreaks	I admire what you're doing but [...] I know I'm in the minority. [...]	75% of the problems I run into are solved by [...] I work in live streaming.	All the suggestions others have given are excellent. RS7 makes the most sense to me. But [...] Meanwhile, [...]

Table C.6: Clusters for Fined-tuned RoBERTa Model. We display examples for each cluster of the 7 clusters that resulted from the agglomerative clustering of 14,756 randomly sampled texts from the conversation test set. We mention noticeable consistencies (Consistency) within the cluster and give three examples each. Consistencies that are not as clear are marked with a '?'.

VI

C	#	Consistency	Example 1	Example 2	Example 3
1	12798	wide variety	Just googled it, looks like a great device for the price! If I weren't so impatient I would have bought this online. Great battery life!	This is exactly why i believe iphone 5 body was perfect example of good balance with design(timeless) and utility	[...] The earth probably has two or more degrees of symmetry, but less than infinite (like a sphere), but I'm honestly not too concerned about the minutiae of it
2	1110	short utterances	here we go!!	And her good posture.	Not in California.
3	310	long utterances	I've never had the pleasure of seeing Neil live but I got on a big kick a few years ago after buying one of his live albums (can't remember which one) where I listened to all his live albums and then wanted to see as many of his live performance I could find on YouTube. [...]	> but the movie has the superior ending I think. [...] [...]	So .... heavily influenced by the social economics ... but still voluntary, got it. [...] Then how about this. [...] Everyone still keeps their child that way, you even promote child birth. No sterilization, no stigmatization of poor people, no poor people stuck with child with heavy needs requiring care that they can't pay for.
4	232	URLs	`https://youtu.be/GmULc5VANsw`	[This](`https://np.reddit.com/r/MakeupAddiction/comments/25hkqi/how_to_tell_if_your_foundationprimer_is_silicone/`) might help!	I thought there was 51 stars because of Puerto Rico `https://en.m.wikipedia.org/wiki/51st_state`

Table C.7: Clusters for RoBERTa Base. We display examples for 4 out of 7 clusters as a result of the agglomerative clustering of 14756 randomly sampled texts from the conversation test set. We mention noticeable consistencies (Consistency) within the cluster and give three examples each.

C.5. Computing Infrastructure

The training of 23 RoBERTa (Liu et al., 2019), 13 uncased BERT and 6 cased BERT models (Devlin et al., 2019) took about 846 GPU hours with one RTX6000 card with 24 GB RAM on a Linux computing cluster. Further analysis and clustering of two RoBERTa models took about 24 GPU hours. We used a machine with 32 GB RAM and 8 intel i7 CPUs using Ubuntu 20.04 LTS without GPU access to generate the training data.

We used `sentence-transformers` 2.1.0 (Reimers and Gurevych, 2019) and `numpy` 1.18.5 (Harris et al., 2020), `scipy` 1.5.2 (Virtanen et al., 2020) and `scikit-learn` 0.24.2 (Pedregosa et al., 2011).

We use previous work, including code and data, consistent with their specified or implied intended use (Reimers and Gurevych, 2019; Chang et al., 2020; Wegmann and Nguyen, 2021). The `ConvoKit` open-source Python framework invites NLP researchers and 'anyone with questions about conversations' to use it (Chang et al., 2020). The `sentence-transformers` Python framework can be used to compute sentence / text embeddings.[3] We comply with asking permission for part of the dataset for `STEL` and citing the specified works (Wegmann and Nguyen, 2021). Wegmann and Nguyen (2021) state the intended use of developing improved style(-sensitive) measures.

C.6. Intended Use

We hope our work will inform further research into style and its representations. We invite researchers to reuse any of our provided results, code and data for this purpose.

[3] `https://sbert.net/`

D

Additions to Chapter 6

D.1. Context-Dependent Paraphrases in Dialog

Should one include repetitions? Repetitions have been typically included in paraphrase taxonomies (Bhagat and Hovy, 2013; Zhou et al., 2025) even though, e.g., Kanerva et al. (2023) asked annotators to exclude such pairs as they considered them uninteresting paraphrases. However, distinguishing repetitions from paraphrases turns out to be especially hard in dialog: For example, speakers tend to leave words out when they repeat and adapt the pronouns to match their perspective (e.g., I -> you). We therefore include repetitions in our definition of context-dependent paraphrases. In fact, those mainly make up the "Clear Contextual Equivalence" Paraphrases (see Table 6.2).

D.2. Dataset

Utterance pair IDs We use unique IDs for utterance pairs. For example, for NPR-4-2, "NPR-4" is the ID used for interviews[1] as done in Zhu et al. (2021), "2" is the position of the start of the guest utterance in the utterance list as separated into turns by Zhu et al. (2021), in this case "Thank you.".

[1] In this case referring to `https://www.npr.org/templates/story/story.php?storyId=16778438`

D.2.1. Preprocessing

We give details on the three preprocessing steps (see Section 6.4.1).

1. Filtering for 2-person interviews We filter 49,420 NPR and 414,176 CNN interviews from Zhu et al. (2021) for 2-person interviews only. This can be challenging: In the speaker list, authors sometimes have non-unique identifiers (e.g., 'STEVE PROFFITT', 'PROFFITT' or 'S. PROFFITT' refer to the same speaker). If one author identifier string is contained in the other we assume them to be the same speaker.[2] We generally assume the first speaker to be the host. We remove 538 NPR and 1,917 CNN interviews because the identifier of the second speaker includes the keywords "host" or "anchor"—thus contradicting our assumption. This leaves 14,000 NPR and 50,301 CNN 2-person interviews.

2. Removing first and last turns of an interview The first turns in our 2-person interviews are usually (reactions to) welcoming addresses and acknowledgments by host and guest[3], while the last often contain goodbyes or acknowledgments[4]. We remove the first two and the last two (guest, host)-pairs. This step removes 2,409 NPR and 26,419 CNN interviews because they are fewer than 5-turns long. For the remaining interviews, this removes 34,773 NPR and 71,646 CNN (guest, host)-pairs.

3. Removing short and long utterances We further remove short guest utterances of 1–2 words as they leave not much to paraphrase.[5] 3,540 NPR and 12,675 CNN pairs are removed like this. We also remove pairs where the host utterance consists of only 1–2 words.[6]. 2,940 NPR and 11,389 CNN pairs are removed like this. We also remove pairs where guest or host utterance consist of more than 200 words.[7] Overall, this leaves 148,522 (guest, host)-pairs in 34,419 interviews for potential annotation, see Table 6.4.

[2] There might be other cases where different string identifiers in the dataset refer to the same speaker although they are not substrings of the other (e.g., 'S. PROFFITT' and 'STEVE PROFFITT'). For a randomly sampled selection of 44 interviews that were identified as more than 2 person interviews, 12 contained errors in the matching. 2/12 were the result of typos and 10/12 were the result of additions to the name like "(voice-over)" or "(on camera)".

[3] For example, "I'm Farai Chideya." "Welcome." "Thank you."

[4] For example the last 3 turns in the considered NPR-4 interview: "Well, Dr. Hader. Thanks for the information.", "Well, thank you for helping share that information [...]", "Well, thanks again. Dr. Shannon Hader [...]"

[5] We manually looked at a random sample of 0.3% ≈ 48 such pairs. The 1-2 token guest utterances are mostly (40/48) assertions of reception by the guest (e.g., "Yes.", "Exactly. Exactly.", "That's right"). Some are signals of protest (4/48) (e.g., "Hey, man.", "Yes, but...", "Hold on."). None of them were reproduced by the host in the next turn.

[6] We manually looked at a random sample of 0.3% ≈ 37 such pairs. The 1–2 tokens host utterances are mostly (28/37) assertions of reception by the host (e.g., "Yeah.", "Yes.", "Sure.", "Right.", "Right. Right.", "Ah, okay."). Some are requests for elaboration (5/37) (e.g., "How so?", "Like?", "Four?") or reactions (3/37) (e.g., "Wow!", "Oh, interesting."). Only one example "Four?" was reproducing content in the form of a repetition.

[7] 200 is the practical limit for the number of words for the chosen type of question (i.e., 'Highlight" Question) in the used survey hosting platform (i.e., `Qualtrics`). It also limits annotation time per question.

D.2.2. Lead Author Annotations

The share of paraphrases in randomly sampled (guest, host)-pairs was only at around 5-15% in initial pilots with lab members. As a result, and in line with previous work (Dolan and Brockett, 2005; Su and Yan, 2017; Dong et al., 2021; Kanerva et al., 2023), we opted to do a pre-selection of text pairs (so called *paraphrase candidates*) before proceeding with the resource-intensive paraphrase annotation (cf. Section 6.5.5). We use annotations by the lead author to select the paraphrase candidates that are then annotated by crowd workers (cf. Section 6.4.2). Here, we provide more details on why we used lead author annotations for selecting paraphrase candidates, how we selected paraphrase candidates using lead author annotations and how the lead author annotations overlap with crowd-majority annotations.

Deciding to do lead author annotations for paraphrase candidates instead of using crowdsourcing Commonly used automatic heuristics were not suitable for the highly contextual discourse setting as these are systematically biased towards selecting more obvious cases, e.g., text pairs that are lexically similar (cf. Section 6.4.2). We therefore experimented with discarding obvious "non-paraphrases" via crowd-sourced annotations and compared this to manual filtering by the lead author. Ultimately, we chose the latter. One key reason was that discarding obvious "non-paraphrases" was more resource intensive and difficult for crowd workers than expected, making the resources needed for discarding non-paraphrases too close to annotating paraphrases themselves—which defeats the purpose of doing a pre-selection in the first place.

Shifting from discarding obvious non-paraphrases to selecting paraphrases In an initial set of 750 randomly sampled (guest, host)-pairs, the lead author discarded obvious non-paraphrase pairs. The resulting set of paraphrase candidates was initially the set we wanted crowd workers to annotate. However, this approach resulted in a high proportion of uninteresting or improbable paraphrases among the paraphrase candidates. To address this, we shifted our strategy from filtering out non-paraphrases to explicitly classifying paraphrases vs. non-paraphrases. This approach includes a higher risk of discarding paraphrases, but brings the benefit of including more actual paraphrases in the set of paraphrase candidates. The lead author re-annotated the initial set of 750 paraphrase candidates and annotated 3700 additional (guest, host)-pairs for paraphrases. In the first batch, the lead author also labeled a variety of different paraphrase types/difficulties (e.g., high lexical similarity, missing context, unrelated), see also Table D.1a. In the second batch, annotations were restricted to repetition paraphrase, paraphrase and non-paraphrase.

Relation to the crowd-majority annotations We display the overlap between the lead author's paraphrase classifications and the released classifications of the crowd majority in Table D.2. On a random set (RANDOM), the overlap is 89%.

	#
Paraphrase	88
High Lexical Similarity	59
Repetition	45
Context-Dependent	
Perspective-Shift	10
Directional	17
Other Difficult Cases	16
Non-Paraphrase	519
Unrelated utterances	>103
More Difficult	
Topically Related	>83
High Lexical Similarity	>18
Partial	>24
Conclusion	46
Ambiguous	18
Missing Context	125

(a) Statistics Labels First Batch.

	#	acc.
Paraphrase	46	0.80
High Lexical Similarity	24	0.92
Repetition	16	0.88
Context-Dependent		
Perspective-Shift	10	0.90
Directional	12	0.67
Other Difficult Cases	12	0.58
Non-Paraphrase	54	0.81
Unrelated utterances	13	1.00
More Difficult	41	0.76
Topically related	24	0.67
High Lexical Similarity	11	0.64
Partial	10	0.80
Conclusion	11	0.55

(b) Statistics Labels BALANCED.

Table D.1: Overview of First Author Labels for First Batch and BALANCED. For (a), for the first batch of 750 manually reviewed pairs, the lead author also labeled several additional categories beyond the primary paraphrase/non-paraphrase distinction. We identified 88 paraphrases, 519 non-paraphrases, 18 ambiguous cases and 125 instances missing context prevented a clear decision. The lead author avoided labeling a pair as ambiguous if they leaned to one category over another. Additional labels include: "perspective-shift" (the perspective shifts between guest and host, e.g., "you" → "I"), "directional" (one utterance entails from or subsumed in the other), "partial" (a tiny subselection could be understood as a paraphrase, but every larger selection is clearly not a paraphrase, e.g., see Table D.3 "going to"/"to go" example), "related" (two utterances are closely related but not paraphrases), "conclusion" (host draws a conclusion or adds an interpretation that goes beyond the original meaning). Categories within Paraphrase and Non-Paraphrase can overlap. Some of these labels were only added for the last 200 annotations and therefore their counts are marked with ">" to indicate lower bounds. For (b), we display the label distribution for a BALANCED subset of 100 paraphrase candidates selected for detailed annotation. The sample was curated based on the assigned labels from the first batch (a). We display overlap with lead author annotations in "acc.".

Dataset	Overlap Lead Author and Crowd Majority
BALANCED	0.72
RANDOM	0.89
PARA	0.72

Table D.2: Lead vs. Crowd Classifications. We display the average overlap between the lead author's classifications and the majority vote of the crowd. The overlap is the highest on the RANDOM set. Probably because we keep all obvious non-paraphrases for classification and the annotators face less ambiguous (guest, host)-pairs to classify.

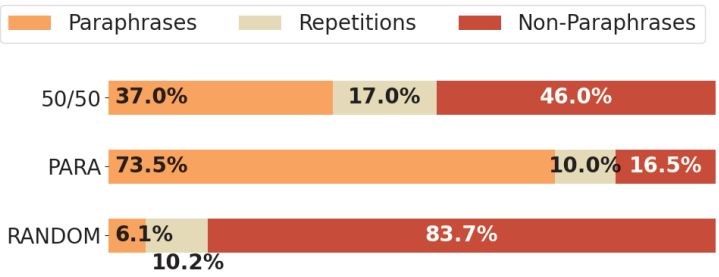


Figure D.1: Distribution of Labels by Lead Author. We display the estimated number of (non-)paraphrases from the lead author annotations for the random subsample (RANDOM), the BALANCED sample and the wider paraphrase variety sample (PARA). Note, RANDOM consists of 100 elements, however only 98 are included in this statistic here (leading to numbers like 6.1). 2 pairs were not classified by the lead author because they were too ambiguous or were missing context information to reach a decision. We exclude such pairs in all other samples.

VI

D.2.3. Paraphrase Candidate Selection

Based on the lead author classifications into paraphrases, non-paraphrases and repetitions, we build three datasets for annotation (main paper Section 6.4.2). We display the lead author classification distribution for the three datasets in Figure D.1.

BALANCED The BALANCED set is a sample of 100 (guest, host)-pairs that were randomly sampled based on the first batch of lead author annotations (Section D.2.2) to equally represent paraphrases and non-paraphrases. We had additional lead author labels available for this set, see Table D.1b for the distribution of these on the BALANCED set. Constraints were 50 paraphrases and 50 non-paraphrases. In order to include more complex cases, we sampled more difficult than unrelated non paraphrase pairs and we limited the number of repetition paraphrases (51% of paraphrases are repetitions in the full batch, but only 33% of paraphrases in BALANCED are repetitions). Due to a sampling error, we ended up with a 46/56 split. Later, we calculate the majority vote of the 20–21 annotations per (guest, host)-pair on this set, and then evaluate it by comparing it against the lead author classification, see "acc." column.

RANDOM The random set is a sample of 100 (guest, host)-pairs that was uniformly sampled from the second batch of lead author annotations (Section D.2.2).

PARA We sampled the PARA(PHRASE) set to reach a specified total 350 paraphrases and 150 non-paraphrases across both the RANDOM and PARA sets—thereby increasing the overall number of paraphrases in the final dataset. Specifically, the PARA set was selected to ensure that, when combined with RANDOM, there would be 300 non-repetition paraphrases and no more than 50 repetition paraphrases. Conversely, the number of non-paraphrases was adjusted to reach a total of 150. As a result, the PARA set included 334 paraphrases and 66 non-paraphrases as annotated by the lead author.

D.3. Annotations

We provide details on how the annotator training was developed (Section D.3.1) and how annotators were ultimately trained (Section D.3.2).

D.3.1. Development of Annotator Training.

The eventual study design used in this work (see Section 6.5) is the product of iterative improvement with lab members, other volunteers and `Prolific` annotators. The consecutive steps can roughly be separated into:

(1) The lead author repeatedly annotated the same set of (guest, host)-pairs with a time difference of one week. See an example of early self-disagreement in Table D.3.

(2) With insights from (1) and our definition of context-dependent paraphrases, we created annotator instructions. We iteratively improved instructions while testing them with volunteers, lab members and `Prolific` crowd workers. See examples of disagreements that led to changes in Table D.3.

(3) Based on insights from (2), we introduced an intermediate annotator training that explains paraphrase annotation in a "hands-on" way: Annotators have to correctly annotate a teaching example to proceed to the next page instead of just reading an instruction. As soon as the correct selection is made, an explanation is shown (e.g., Figures D.4 and D.8). After further testing rounds, we also require annotators to pass two attention checks (see Figure D.10) as well as two comprehension checks (see Figures D.3 and D.9).

(4) We test the developed training from (3) by training annotators with it and then asking them to annotate a selection of 20 (guest, host)-pairs. We selected 10 pairs that were classified by the lead author as clearly containing a paraphrase, and 10 as containing no paraphrase. Half of the examples were chosen to be more challenging to classify (e.g., paraphrase with a low lexical overlap, non-paraphrase with a high lexical overlap). Two lab members reached pairwise Cohen of 0.51 after receiving training. Two newly recruited `Prolific` annotators reached average pairwise Cohen of 0.42 after going through training. Due to the inherent difficulty of the task and the good annotation quality when manually inspecting the 20 examples for each annotator, we carry on with this training setup.

Who?	Example	see Instructions
Self-Disagreement	Guest: [..] So there was a consensus organization last year that people from genetics and ethics law got together and said, in theory, it should be acceptable to try this in human beings. The question will be, how much safety and evidence do we have to have from animal models before we say it's acceptable. Host: When it comes to this issue, let's face it, while there are the concerns here in the United States, it's happening in other countries.	**(C)** distinguish paraphrases from inferences, conclusions or "just" highly related utterances
Lab Members	Guest: Hey, it's going to be a long and a long week, and we're going to use every single minute of it to make sure that Americans know that Al Gore and Joe Lieberman are fighting for working families, right here in Los Angeles and across America. Host: And are you guys ready to go?	**(P)** short subselections of tokens might be "paraphrases" that do not adequately represent the content of the guest's utterance
	Guest: [...] There are militant groups out there firing against the military. And we just - we really don't know who is whom. Host: Why did the army decide today to move in and clear out the camp?	
	Guest: Police have indicated that they have been getting cooperation from the people involved, of course, they are looking at all of her personal relationships to see if there were any problems there. [...] Host: Well what have family members told you? I know you've talked to various members of her family. I understand she never missed her shifts at the restaurant where she worked. [...]	
	Guest: Yes, it is, all $640,000. Host: That's a lot of dough.	**(CD)** emphasize situational aspect to annotators, **(H)** ask for token-level accuracy of highlights
`Prolific` **Annotators**	Guest: [...] He was an employee that worked downtown Cleveland and saw it fall out of the armored car carrier, and pick it up, and took it, and placed it in his car. Host: And he's been holding it ever since?	similar to **(C)**
	Guest: [...] Would I ever thought that this would be happening, no, it is, it's crazy? Just enjoy the moment. Host: [...] , Magic Johnson was saying that when he first started taking meetings with investors or with business people, they didn't take him seriously, but he thought maybe they just wanted his autograph. [...]	**(AT)** use annotator screening to throw out annotators more likely to produce non-sensical pairs
	Guest: [...] they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is – Host: So you weren't even familiar?	**(AT)** throw out annotators that do not select obvious pairs

Table D.3: Examples of Disagreements in Paraphrase Annotation Pilots. All of the presented examples were highlighted by at least one annotator and selected as not showing any paraphrases at all by at least one other annotator. We show examples from three different conditions: Self-disagreement by the lead author, disagreements between volunteers/lab members and disagreements between `Prolific` annotators. These disagreements informed later training instructions: For (C)conclusion, see Figure D.4; for (P)artial, see Figure D.7; for (C)ontext(D)ependent, see Figure D.8; for (H)ighlighting, see Figure D.6; for (AT)tention, we chose a separate training setup with attention and comprehension checks, see Figures D.3, D.9 and D.10. Early on, we chose to include repetitions in our paraphrase definition since it turned out to be conceptually difficult to separate the two—especially in a context-dependent setting (e.g., is "You don't know." a repetition of "I do not know it." or not?), see Figure D.2.

D.3.2. Annotator Training.

We train participants to recognize paraphrases (see Figure D.2–D.10 for the exact instructions they received).[8] We presented (guest, host)-pairs with their respective interview summaries (extracted from Zhu et al. (2021)'s MediaSum corpus), the date of the interview and the interviewer names for context.[9] Participants were only admitted to the paraphrase annotation if they passed two attention checks (see Figure D.10) and two comprehension checks (see Figure D.3 and D.9).

Comprehension Checks Annotators are presented with a clear paraphrase pair (Appendix Figure D.3) and a less obvious context-dependent paraphrase pair (Appendix Figure D.9) that they have to classify as a paraphrase (similar to examples in Table 6.2). Additionally, they have to highlight the specific text spans that are a part of the paraphrase.

Training Statistics Of the initial 347 `Prolific` annotators who started the training, 95 aborted the study without giving a reason[10] and 126 were excluded from further studies because they failed at least one comprehension (29%) or attention check (24%) during training. Since annotators can perform annotations after training over a span of several days, we further exclude single annotation sessions, where the annotator fails any of two attention checks.

D.3.3. Annotation After Training.

Next, the trained annotators were asked to highlight paraphrases. See Figure D.11 for an example of the annotation interface. Annotators had access to a summary of their training at all times, see Figure D.12. We again included two attention checks. Answers failing either attention check are removed from the dataset.

D.3.4. Annotator Payment.

Via `Prolific`'s internal screening system, we recruited native speakers located in the US. Payment for a survey was only withheld if annotators failed two attention checks within the same survey or when a comprehension check at the very beginning of the study was failed[11] in line with `Prolific` guidelines.[12] Across all `Prolific` studies performed for this work (including pilots), we paid participants a median of $8.98£/h \approx 11.41\$/h$[13] which is above federal minimum wage in the US.[14]

[8] See it in action at `https://annawegmann.github.io/Paraphrases.html`

[9] The additional information of summary, date and speaker names increased reported understanding of context and eased difficulty of the task in pilot studies among lab members.

[10] Usually quickly, we assume that they did not want to take part in a multi-part study or did not like the task itself.

[11] Technically, in line with `Prolific` guidelines, we do not withhold payment but ask annotators to "return" their study in this case. Practically this is the same, as all annotators did return such a study when asked.

[12] `Prolific` Attention and Comprehension Check Policy

[13] on March 20th 2024

[14] Federal minimum wage in the US is $\$7.25/h \approx 5.71£/h$ according to `https://www.dol.gov/agencies/whd/minimum-wage` on March 20th 2024

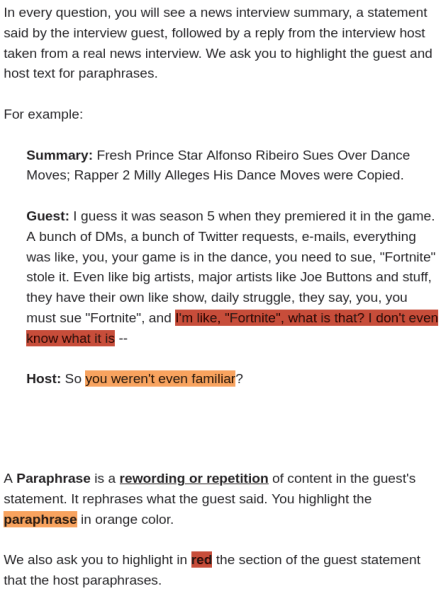


Figure D.2: Annotator Training (1). Definition Paraphrase

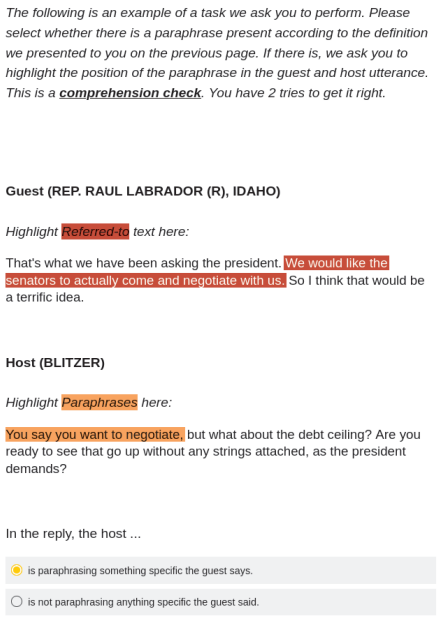


Figure D.3: Annotator Training (2). Comprehension Check Paraphrase. Variations of the the shown highlighting are accepted.

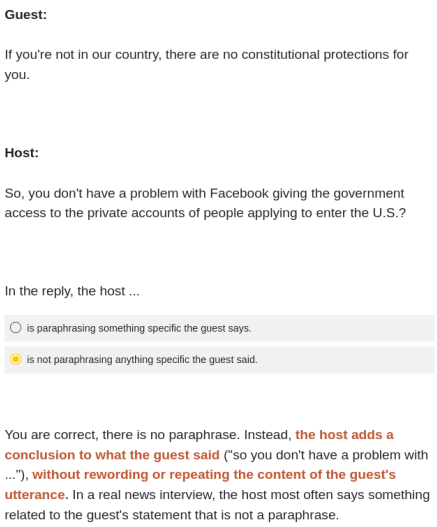


Figure D.4: Annotator Training (3). Related but not a Paraphrase

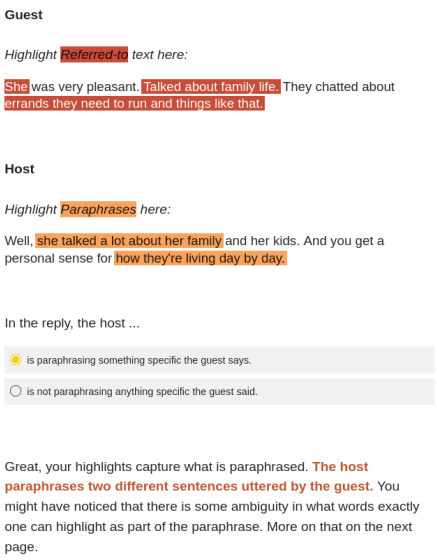


Figure D.5: Annotator Training (4). Multiple Sentences.

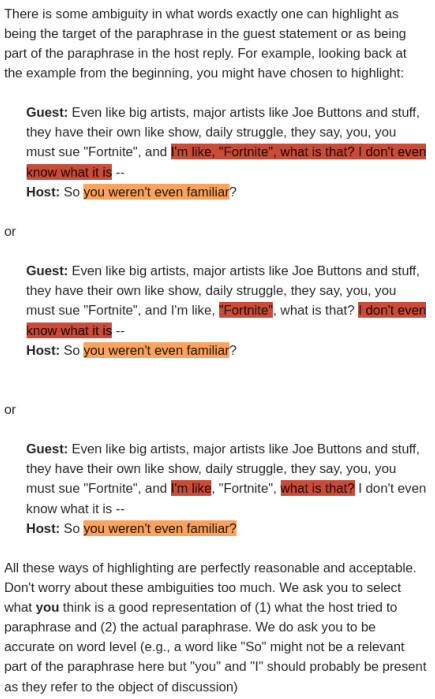


Figure D.6: Annotator Training (5). Highlighting

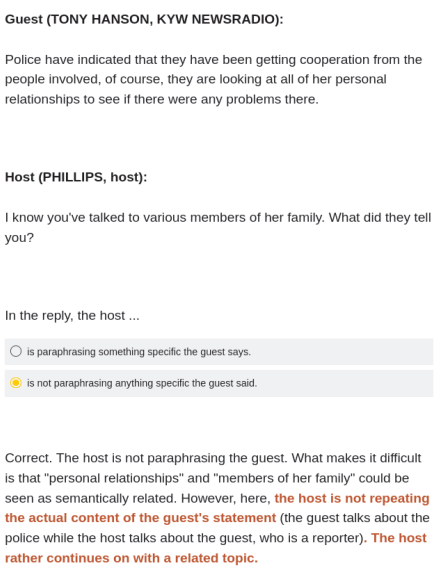


Figure D.7: Annotator Training (6). Partial vs actual paraphrase

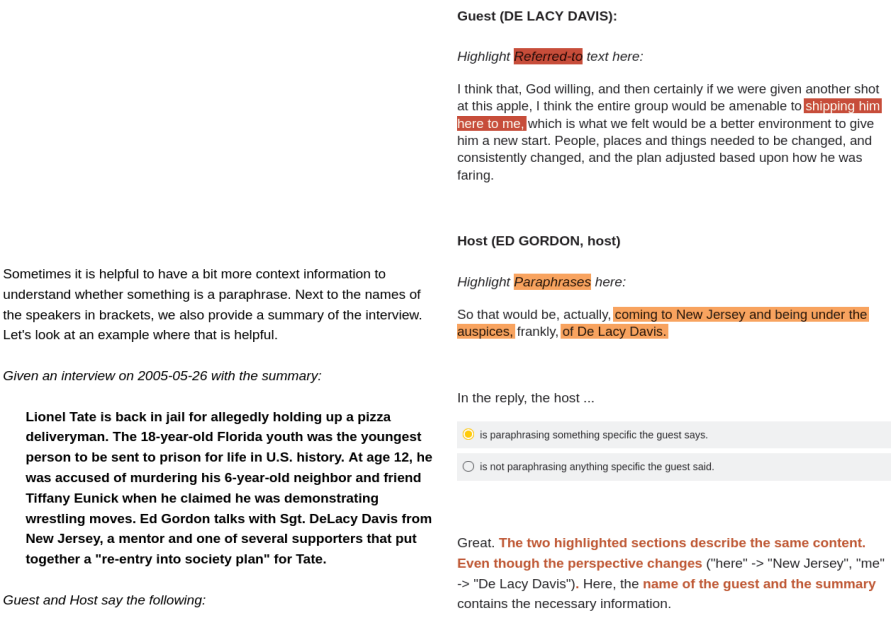


Figure D.8: Annotator Training (7). Using context information

Figure D.9: Annotator Training (8). Example of an accepted answer for the comprehension check at the end. Only annotators who highlighted similar spans are admitted to annotate unseen instances. Some of the admitted annotators additionally selected the pair "he's improved a lot" and "he's expected to make a full recovery".

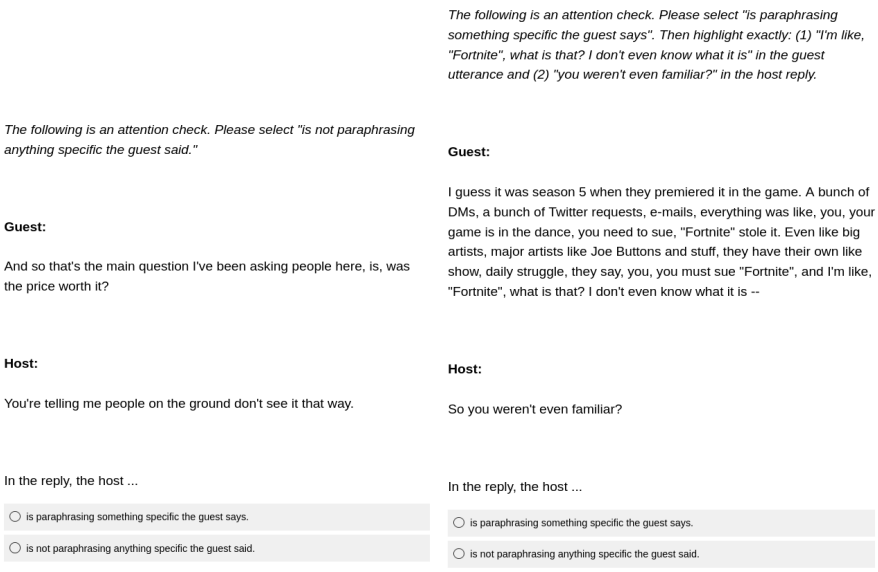


Figure D.10: Annotator Training (9). Two attention checks shown at different times during training.

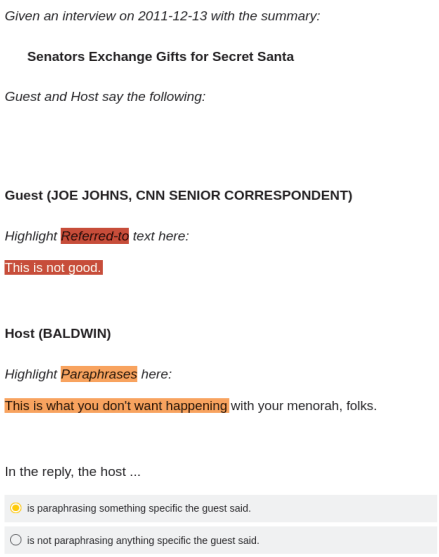
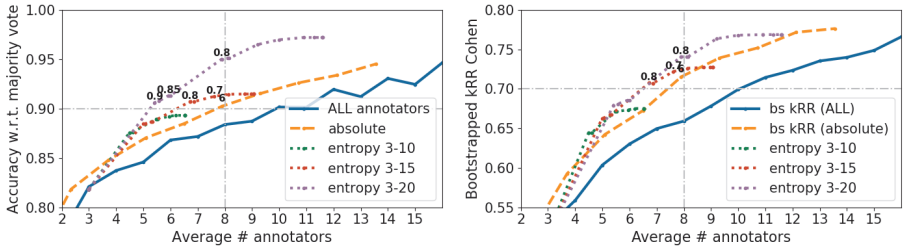


Figure D.11: Interface for Highlighting Categories. Annotators are asked to highlight the categories on word level.

	Short Description	Example
In the Guest Statement		
Highlight	Referred-To is the target of the paraphrase	see below
In the Host Reply		
Highlight	Paraphrases reword or repeat what the guest said Subtleties *- several ways of highlighting are acceptable*	**Guest:** they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is -- *or:* **Guest:** *they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is* -- **Host:** So you weren't even familiar?
		Guest: That's what we have been asking the senators to actually come and negotiate with us. **Host:** I know you want to negotiate ...
		Guest: She was very pleasant. Talked about family life. They chatted about errands they need to run and things like that. **Host:** Well, she talked a lot about her family and her kids. And you get a personal sense for how they're living day by day.
	- can include perspective shifts (here -> New Jersey, I -> you)	**Guest (De Lacy Davis):** I think the entire group would be amenable to shipping him here to us, which is what we felt would be a better environment to give him a new start. **Host:** So that would be, actually, coming to New Jersey and being under the auspices, frankly, of De Lacy Davis.
Do Not Highlight	additional content like consequences, inferences or facts	**Guest:** if you're not in our country, there are no constitutional protections for you **Host:** So, you don't have a problem with Facebook giving the government access to the private accounts of people applying to enter the U.S.? **Guest:** ... And so that's the main question I've been asking people here, is, was the price worth it? **Host:** You're telling me people on the ground don't see it that way.
	vague references or (short) terms that are only related in content but are not actually rephrasing the guest's point	**Guest:** ... they are looking at all of her personal relationships to see if there were any problems there. **Host:** Well what have family members told you? I know you've talked to various members of her family.

Figure D.12: Overview Table Shown to Annotators



(a) Accuracy w.r.t. 20 annotators

(b) kRR

Figure D.13: Annotator Recruitment Strategies. To decide the number of annotators for a specific item, we test three different strategies: (1) using a fixed number of annotators across all items (ALL), (2) increasing the number of annotators until at least n annotators agree for each item (absolute) and (3) increasing the number of annotators from 3 until the entropy is smaller than a given threshold (entropy) or a maximum of 10, 15 or 20 annotators is reached. We display the accuracy of the methods compared to using all 20 annotations in (D.13a) and the reliability measure kRR depending on the average number of annotators used (Wong and Paritosh, 2022) in (D.13b). We set a maximum average cost of 8 annotators per item and require a minimum accuracy of 90% as well as a minimum kRR of 0.70. When a strategy fulfills these requirements (i.e., falls in the upper left quadrants for (a) and (b)), we display the entropy thresholds for (3) and absolute number of annotators for (2).

VI

D.3.5. Annotator Allocation Strategy

To the best of our knowledge, what constitutes a "good" number of annotators per item has not been investigated for paraphrase classification.

Summary Using the 20–21 annotations per item collected for the BALANCED set, we simulate both fixed and dynamic strategies to recruit up to 20 annotators per item. We evaluate the different strategies based on their agreement with the majority vote derived from all 20–21 annotators. When considering resource cost and performance trade-offs, dynamic recruitment strategies performed better than allocating a fixed number of annotators for each item.

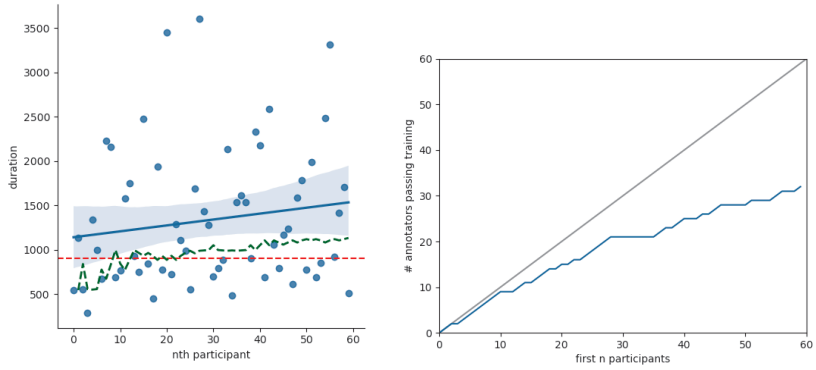
Details We consider three different strategies for allocating annotators to an item: (1) using a fixed number for all items, (2) for each item, dynamically allocate annotators until n of them agree and (3) similar to Engelson and Dagan (1996), for each item, dynamically allocate annotators until the entropy is below a given threshold t or a maximum number of annotators has been allocated. We simulate each of these strategies using the annotations on BALANCED. We evaluate the strategies on (a) cost, i.e., the average number of annotators per item and (b) performance via (i) the overlap between the full 20 annotator majority vote (i.e., we assume this is the best possible result) and the predicted majority vote for the considered strategy and (ii) k-rater-reliability (Wong and Paritosh, 2022)—a measure to compare the agreement between aggregated votes. Note, for the dynamic setup we need to change the original calculation of kRR (Wong

and Paritosh, 2022) by aggregating the votes of a varying instead of a fixed number of annotators.

Results See Figure D.13 for the results. We selected a practical resource limit of an average 8 annotators per items and the requirement of at least 90% overlap with the majority vote and 0.7 kRR (dotted lines). We decide on strategy (3) dynamically recruiting annotators (minimally 3, maximally 15) until entropy is below 0.8. Also with other min/max parameters this was a good trade-off between accuracy, kRR and average number of annotators. The average number of annotators needed per item is then about 6.8. In this way, most items receive annotations from 3 annotators, while difficult ones receive up to 15.

D.3.6. Anonymization

We replace all `Prolific` annotator IDs with non-identifiable IDs. We only make the non-identifiable IDs public.



(a) Duration **(b)** Quality Checks Passed

Figure D.14: On BALANCED, later training sessions take longer and pass fewer quality checks. In D.14a, we display the time (in seconds) the nth annotator needs to go through the training session. The annotators are ordered according to the dates they completed the training. Annotations were distributed across 6 different days in June 2023. The green line represents the median duration time of the first n participants. The red line displays the initially estimated completion time of 900 seconds according to pilot studies. The blue line is a linear regression estimate of the duration and it's 95% confidence interval. On average, participants participating on a later date need more time to finish. In D.14b, we display the summed number of the first n participants that passed the quality checks during training. The grey line represents the angle bisector, i.e., if every participant would pass all quality checks. Later participants are less likely to pass the quality checks.

VI

D.3.7. Varying Annotator Behavior over Time.

For the BALANCED set, we performed several rounds of training and annotation. Figure D.14 shows the completion times and share of passed quality checks among `Prolific` annotators during the training session. Notably, participants that were recruited later performed worse: They passed fewer quality checks and took more time. While this effect was clear, it is not quite clear to us why this happened. To mitigate potential issues with the age of the study, we recruited all participants at once in subsequent studies and not iteratively as for the BALANCED set. Overall, the time of recruitment should have minimal impact on the quality of the final annotations, as we always excluded any annotators who failed our quality checks.

D.3.8. Intra-Annotator Annotations Quality

We randomly sample ten annotators (with anonymized `Prolific` ids 60, 6, 86, 84, 47, 31, 68, 88, 41, 92) and manually analyze 42 of their annotations. Nine annotators consistently provide plausible annotations, while the other annotator chooses "not a paraphrase" a few times too often. We also noticed some other annotator-specific tendencies, for example, one annotator might tend to highlight fewer words, more words or prefer exact lexical matches.

Learning Rate	Epoch	F1 on development set (↑)
1e-5	8	0.61 ± 0.04
3e-5	8	0.64 ± 0.06
5e-5	8	0.52 ± 0.15
3e-5	4	0.65 ± 0.07
3e-5	12	**0.65** ± 0.00
3e-5	16	0.60 ± 0.10

Table D.4: Hyperparameter Tuning. We train token classifiers with varying learning rates and epochs. Results show mean and standard deviation over three seeds. The best learning rate and epoch are underlined, the best F1 score is **boldfaced**.

D.4. Modeling

VI

D.4.1. In-Context Learning

Hugging Face URLs VICUNA 7B: `https://huggingface.co/lmsys/vicuna-7b-v1.5`, MISTRAL 7B Instruct: `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`, OPENCHAT: `https://huggingface.co/openchat/openchat-3.5-0106`, GEMMA 7B: `https://huggingface.co/google/gemma-7b-it`, MIXTRAL 8X7B INSTRUCT: `https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1`, LLAMA 7B: `https://huggingface.co/meta-llama/Llama-2-7b-hf` and LLAMA 70B: `https://huggingface.co/meta-llama/Llama-2-70b-hf`.

Prompt We use a few-shot prompt closely aligned with the original annotator training and instructions, see Figure D.15. We use chain-of-thought, i.e., starting with "Let's think step by step." and ending with "Therefore, the answer is" (Kojima et al., 2022). The few-shot prompt includes all eight examples shown to human annotators (Figures D.2–D.10). For GPT-4, we use a temperature of 1, use self-consistency through prompting the model 3 times (Wang et al., 2023b), and use the default top_p nucleus sampling value of 1. The maximum number of new tokens is set to 512. For all Hugging Face models, we also use a temperature of 1, apply self-consistency through prompting the model 10 times (only 3 times for LLAMA 70B due to resource limits) and use top_k sampling of the top 10 tokens. The maximum number of new tokens is set to 400. Note, there are many more prompts and choices we could have tried that are out-of-the scope of this work. Further steps could have included separating the classification and highlighting task, experimenting with other phrasings and so on.

D.4.2. Token Classification

We use settings closely aligned with Wang et al. (2022a), and experimetn with different learning rates and number of epochs. For the results, see Figure D.4. We use a learning rate of 3e-5 and 12 epochs for further modeling. We use the model checkpoints that performed best on the development set.

```
A Paraphrase is a rewording or repetition of content in the guest's statement. It rephrases what the guest said.
Given an interview on – with the summary: Fresh Prince Star Alfonso Ribeiro Sues Over Dance Moves; Rapper 2 Milly Alleges His
        Dance Moves were Copied.
Guest and Host say the following:
Guest (TERRENCE FERGUSON, RAPPER): I guess it was season 5 when they premiered it in the game. A bunch of DMs, a bunch of
        Twitter requests, e-mails, everything was like, you, your game is in the dance, you need to sue, "Fortnite" stole it.
        Even like big artists, major artists like Joe Buttons and stuff, they have their own like show, daily struggle, they
        say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is --
Host (QUEST): So you weren't even familiar?
In the reply, does the host paraphrase something specific the guest says?

Explanation: Let's think step by step.
Terrence Ferguson says at the end of his turn that he didn't know Fortnite.
Quest, the host of the interview, repeats that the guest doesn't know Fortnite.
So they both say that the guest didn't know Fortnite. Therefore, the answer is yes, the host is paraphrasing the guest.
Verbatim Quote Guest: "I'm like, "Fortnite", what is that? I don't even know what it is"
Verbatim Quote Host: "you weren't even familiar?"
Classification: Yes.


Given an interview on 2013-10-1 with the summary: ...
Guest and Host say the following:
Guest (REP. RAUL LABRADOR (R), IDAHO): ...
Host (BLITZER): ...
In the reply, does the host paraphrase something specific the guest says?

Explanation: Let's think step by step. EXPLANATION Therefore, the answer is yes, host is paraphrasing the guest.
Verbatim Quote Guest: "We would like the senators to actually come and negotiate with us."
Verbatim Quote Host: "you want to negotiate"
Classification: Yes.


ITEM

Explanation: ...
Verbatim Quote Guest: None.
Verbatim Quote Host: None.
Classification: No.

ITEM

Explanation: ...
Verbatim Quote Guest: "She" "Talked about family life." "errands they need to run and things like that."
Verbatim Quote Host: "she talked" "about her family and her kids." "how they're living day by day."
Classification: Yes.

ITEM

Explanation: ...
Verbatim Quote Guest: None.
Verbatim Quote Host: None.
Classification: No.


ITEM

Explanation: ...
Verbatim Quote Guest: None.
Verbatim Quote Host: None.
Classification: No.


ITEM

Explanation: ...
Verbatim Quote Guest: "shipping him here to me"
Verbatim Quote Host: "coming to New Jersey and being under the auspices" "of De Lacy Davis."
Classification: Yes.


ITEM

Explanation: ...
Verbatim Quote Guest: "I'm to see him."
Verbatim Quote Host: "him" "have a visit from you"
Classification: Yes.


Given an interview on DATE with the summary: SUMMARY
Guest and Host say the following:
Guest (NAME): UTTERANCE
Host (NAME): UTTERANCE
Explanation: Let's think step by step.
```

Figure D.15: Prompt Template Close to Annotator Instructions. The used prompt template is based closely on our annotator training and instructions. Phrasings were adapted to match the prompt-setting but kept the same where possible. See the full prompt in our Github Repository.

D.4.3. Highlighting Analysis

We compare the highlights provided by DeBERTa AGGREGATED[15] and DeBERTa ALL[16] on 10 text pairs from the test set that were classified as paraphrases by both models. We provide examples in Table D.5. DeBERTa ALL highlights are shorter, often more on point and arguably more consistent than DeBERTa AGGREGATED highlights. We also manually analyzed 10 text pairs from the test set that GPT-4 classified as paraphrases. We provide examples of GPT-4 highlights in Table D.6. Generally, they seem of good quality, but have the tendency to span complete sub-sentences, even if not all is relevant.

Hallucinations One of the biggest problems for in-context learning are the extractions of the highlighting from the model responses which has errors in up to 71% of the cases in Table 6.9. Most of these errors can be split into two categories: (1) inconsistent highlighting, where the model classifies a paraphrase but does not highlight text spans in both the guest and host utterance and (2) hallucinations, where the model highlights spans that do not exist in the guest or host utterance. Hallucination is more prevalent than inconsistent highlighting for GPT-4, where in most cases it leaves out words (e.g., "coming back to a normal winter" vs. "coming back daryn to a normal winter"), in some other cases it adds or replaces words (e.g.,"he's a counterpuncher" vs. "he's counterpuncher"), uses morphological variation (e.g., "you've" vs. "you have") or quotes from the wrong source (e.g., from the host when considering the guest utterance). Most of these extraction errors appear to be resolvable by humans when looking at them manually, so it might be possible to address them in future work with a more advanced matching algorithm or by querying GPT-4 until one gets a parsable response. Notably, many GPT-4 classifications seem plausible, even when marked as incorrect by the crowd majority.

D.4.4. Computing Infrastructure

The fine-tuning of 18 DeBERTa token classifier, and running inference on 7 generative models took approximately 260 GPU hours on one single A100 card with 80GB RAM on a Linux computing cluster. We use `scikit-learn` 1.2.2 (Pedregosa et al., 2011), `statsmodels` 0.14.1 (Seabold and Perktold, 2010) and `krippendorff`[17] 0.6.1 for evaluation.

D.5. Use of AI Assistants

We used ChatGPT and GitHub Copilot for coding, specifically to look up commands and sporadically to generate functions. Generated functions are marked in our code.

[15] i.e., trained on the deduplicated set of annotations with seed 202 with F1 score of 0.76, precision of 0.72 and recall of 0.84, see `https://huggingface.co/AnnaWegmann/Highlight-Paraphrases-in-Dialog`

[16] i.e., trained on all annotations with seed 201 with F1 score of 0.72, precision of 0.84 and recall of 0.63, see `https://huggingface.co/AnnaWegmann/Highlight-Paraphrases-in-Dialog-ALL`

[17] `https://github.com/pln-fing-udelar/fast-krippendorff`

AGG	ALL	C	Shortened Examples
✓	✓	✗	**G:** There are people that are in that age range where we know they're high risk, why are they going to the supermarket to **buy** their **own groceries?** Get the community, the neighborhood to go and help them. **H:** if you're going to help somebody by **helping them maybe get their groceries**, how long does the coronavirus live on surfaces?
✓	✓	✓	**G:** And people always prefer, of course, to see the pope as the principal celebrant of the mass. So that's good. That'll be tonight. And it will be his 26th mass and it will be the 40th or, rather, the 30th time that this is offered in round the world transmission. And it will be **my 20th time in doing it** as a television commentator from Rome so. **H:** Yes, **you've been doing this for a while** now.
✓	✓	✓	**G:** Well, what happened was we finally waved down a Coast Guard helicopter. And what they were looking for were people with disabilities and medical conditions, which none of us really had. They didn't lift any of us into the helicopter or anything. What they told us was to basically **walk** out of our house, up **the street**, trying to **fight against the current** **that was** going the **opposite** way of where we needed to go. **H:** So **you walked through that current** **to get to the higher ground** or get to a drier spot?
✓	✗	✗	**G:** They've now spent $6 million on this Benghazi investigation. They keep coming up with more and more interviews. **H:** On Benghazi, Trey Gowdy now says your committee has interviewed 75 witnesses.

Table D.5: DeBERTa ALL vs. DeBERTa AGGREGATED Highlights. Paraphrase highlights predicted by the best DeBERTa ALL (**boldfaced**) and the best DeBERTa AGG model (underlined). Even though DeBERTa AGG gets better F1 scores on classification, the DeBERTa ALL highlights are arguably more on point. For comparison, we also display the human highlights if they exist. Note, highlights can exist even if the crowd-majority vote (C) did not predict a paraphrase.

Generated functions were tested w.r.t. expected behavior. We did not use AI assistants for writing.

VI

GPT-4	C	**Shortened Examples**
✓	✗	**G:** We also want to see what connections exist between pardons and potential gifts to the Clinton Library. **H:** Congressman, short of, though, having a thank-you note attached to a check that went to the Clinton Library, what is it exactly that is going to prove that there was a quid pro quo, that these pardons were actually bought?
✓	✗	**G:** They've now spent $6 million on this Benghazi investigation. They keep coming up with more and more interviews. **H:** On Benghazi, Trey Gowdy now says your committee has interviewed 75 witnesses.
✓	✓	**G:** [Trump] is appointing very young judges. **H:** [...] if you're 50-plus, you're probably too old for the Trump Administration to be seriously considered for a district court judgeship.

Table D.6: GPT-4 Highlights. Paraphrase highlights predicted by GPT-4. For comparison, we also display the human highlights if they exist. Note, highlights can exist even if the crowd-majority vote (C) did not predict a paraphrase.

Part VII

Backmatter

Contents

List of SIKS-dissertations

2016

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VU), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VU), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UVA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (TUE), Mining Social Structures from Genealogical Data
20 Daan Odijk (VU), Context & Semantics in News & Web Search
21 Alejandro Moreno Celleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Y. Kiseleva (TUE), Using Contextual Information to Understand Searching and Browsing Behavior

26 Dilhan J. Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale agent-based social simulation: A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak Reduction in Decentralised Electricity Systems - Markets and Prices for Flexible Planning
30 Ruud Mattheij (UVT), The Eyes Have IT
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UVT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UVA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UVA), Combinatorial and compositional aspects of bilingual aligned corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic assistants for database exploration
45 Bram van Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects: A Game-Theoretic Analysis
50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UVA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TUE), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UVA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VU), Logics for causal inference under uncertainty
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VU), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (UvT), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VU), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TUE), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TUE), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TUE), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech
30 Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VU), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming
36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills
37 Jian Fang (TUD), Database Acceleration on FPGAs
38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
04 Maarten van Gompel (RUN), Context as Linguistic Bridges
05 Yulong Pei (TUE), On local and global structure mining
06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components
08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10 Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12 Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15 Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20 Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21 Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25 Xin Du (TUE), The Uncertainty in Exceptional Model Mining
26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30 Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst
31 Gongjin Lan (VU), Learning better – From Baby to Better
32 Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising
33 Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation
34 Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
03 Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices
04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
07 Armel Lefebvre (UU), Research data management for open science
08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19 Roberto Verdecchia (VU), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (LIACS), Robust rules for prediction and description
24 Kaijie Zhu (TUE), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI & VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games

02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (LU), Optimal decision-making under constraints and uncertainty
26 Anne Dirkson (LU), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays

31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

2023

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TU Delft), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TU Delft), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (Leiden University), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity

23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TU Delft), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
29 Tim Draws (TU Delft), Understanding Viewpoint Biases in Web Search Results

2024

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
04 Mike Huisman (Leiden University), Understanding Deep Meta-Learning
05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
06 Azqa Nadeem (TU Delft), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
12 Peide Zhu (TU Delft), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TU Delft), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TU Delft), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA, Gran Sasso Science Institute), Ethics-aware Software Architecture Design
18 Laurens P. Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh S. Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (Leiden University), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution

24 Nirmal Roy (TU Delft), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TU Delft), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (Leiden University), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (Leiden University), Grip on Software: Understanding development progress of Scrum sprints and backlogs
29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TU Delft), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TU Delft), Designing for Appropriate Trust in Human-AI interaction
34 Robert Deckers (VUA), From Smallest Software Particle to System Specification MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a Fuzzy Set Approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TU Delft), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TU Delft), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TU Delft), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
48 Ioannis Petros Samiotis (TU Delft), Crowd-Assisted Annotation of Classical Music Compositions

2025

2025 01 Max van Haastrecht (Leiden University), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
03 Andra-Denis Ionescu (TU Delft), Feature Discovery for Data-Centric AI
04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
05 Nele Albers (TU Delft), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
06 Daniël Vos (TU Delft), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
08 Stefan Bloemheuvel (Tilburg University), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
10 Zhao Yang (Leiden University), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
11 Shahin Sharifi Noorian (TU Delft), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
12 Lijun Lyu (TU Delft), Interpretability in Neural Information Retrieval
13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
15 Michiel van der Meer (Leiden University), Opinion Diversity through Hybrid Intelligence
16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
18 Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
19 Fang Hou (UU), Trust in Software Ecosystems
20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
21 Mandani Ntekouli (Maastricht University), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making
23 Roderick van der Weerdt (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
24 Zhong Li (Leiden University), Trustworthy Anomaly Detection for Smart Manufacturing
25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions

26 Tom Pepels (Maastricht University), Monte-Carlo Tree Search is Work in Progress
27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback
28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning
29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline
32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms
33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction
34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering
35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment

English Summary

Consider two Dutch sentences “Ik ben een Utrechter” and “Ik ben een Utrechtenaar”. Even thought their surface level presentation is different, a translation tool like DeepL might translate both of these sentences to “I am an Utrecht resident”. This translation is perfectly reasonable as both “Utrechter” and “Utrechtenaar” refer to an inhabitant of the city of Utrecht. In this case, DeepL can be said to be robust to language variation: it treats both statements equally. However, there are also many cases in which NLP models benefit from being sensitive to language variation. The Utrecht example illustrates this: Historically, “Utrechtenaar” was the more common term. However, it has now been largely replaced by "Utrechter" in everyday language, as “Utrechtenaar” has been associated with gay men since the Utrecht sodomy trials (around 1730). Today, when someone uses “Utrechtenaar” over “Utrechter” to refer to themselves, we might know more about them—for example, that they are more likely part of the local queer community. Let’s imagine a newspaper article in which two people refer to themselves as “Utrechter” and “Utrechtenaar”: translating both terms as “resident of Utrecht” could obscure subtle differences in background and social identity—potentially leading to confusion or a loss of narrative nuance.

In this dissertation, I develop methods to make language models both more sensitive and more robust to language variation. In Chapter 3, I examine tokenizers—a fundamental building block of language models—with respect to their sensitivity and robustness to language variation. I show that it is important to take language variation into account at all stages of language model development. In Chapters 4 and 5, I develop vector representations that are sensitive to one particular aspect of language variation: the style of a text. In Chapter 4, I propose the STyle Evaluation Framework (STEL), the first systematic method for evaluating how sensitive NLP methods are to stylistic variation in text. In Chapter 5, I train neural text representations that—unlike previous approaches—capture linguistic style independently of content and achieve strong results on STEL. These resulting vector representations have already found a wide range of applications in the NLP community. In Chapter 6, I introduce a novel task: the detection of cross-speaker paraphrases in dialogue. For this, I train crowdworkers using my own iterative procedure for classifying paraphrases. My results show that both humans and NLP models face considerable challenges in robustly recognizing utterances that vary linguistically but have the same content.

I hope that this work will encourage the NLP community to take language variation into account more when developing NLP methods.

Nederlandse Samenvatting

In taal is variatie alomtegenwoordig. Mensen gebruiken veel verschillende uitdrukkingen om hetzelfde te zeggen. Echter, NLP-modellen hebben de neiging te kort te schieten in de omgang met taalvariatie. Dit kan van invloed zijn op taken waarbij het model robuust moet zijn voor taalvariatie (bij semantische taken, zoals zoeken naar relevante documenten voor een zoekopdracht, is het bijvoorbeeld niet van belang of een tekst Britse of Amerikaanse spelling gebruikt) en taken waarbij het model gevoelig moet zijn voor taalvariatie (op vorm gebaseerde taken, zoals verificatie van auteurschap, maken bijvoorbeeld onderscheid tussen Britse en Amerikaanse spelling). In dit proefschrift ontwikkel ik methoden die taalmodellen gevoeliger en robuuster maken voor taalvariaties.

In hoofdstuk 3 onderzoek ik tokenizers—een fundamenteel onderdeel van taalmodellen—op hun gevoeligheid en robuustheid voor taalvariaties. Daarbij laat ik zien dat het belangrijk is om in alle fasen van de ontwikkeling van taalmodellen rekening te houden met taalvariaties. In hoofdstuk 4 en 5 ontwikkel ik vectorrepresentaties die gevoelig zijn voor één aspect van taalvariatie: de taalstijl van een tekst. In hoofdstuk 4 identificeer ik een gebrek aan evaluatiebenaderingen van NLP-methoden in termen van hun gevoeligheid voor linguïstische stijl. Ik presenteer het STyle EvaLuation Framework (STEL), waarmee NLP-methoden voor het eerst systematisch kunnen worden beoordeeld op hoe gevoelig ze reageren op stilistische variaties in teksten. In hoofdstuk 5 train ik neurale tekstrepresentaties die—in tegenstelling tot eerdere benaderingen—de taalstijl onafhankelijk van de inhoud vastleggen en goede resultaten behalen op STEL. De resulterende vectorrepresentaties zijn al op verschillende manieren toegepast in de NLP-gemeenschap. In hoofdstuk 6 behandel ik de herkenning van parafrases tussen sprekers in dialogen. Hiervoor train ik crowdworkers in mijn eigen ontwikkelde iteratieve procedure om parafrases te classificeren. Mijn resultaten tonen aan dat zowel mensen als NLP-modellen aanzienlijke moeite hebben om taalkundig variërende, maar inhoudelijk identieke uitspraken op een robuuste manier te herkennen.

Ik hoop dat dit proefschrift de NLP-gemeenschap ook zal aanmoedigen om taalvariatie sterker te integreren in de ontwikkeling van NLP-methoden.

Deutsche Zusammenfassung

"Ik ben een Utrechter" und "Ik ben een Utrechtenaar"—obwohl sich diese beiden Äußerungen oberflächlich voneinander unterscheiden, übersetzt das Übersetzungstool DEEPL beide Sätze mit "Ich komme aus Utrecht". Diese Übersetzung ist völlig angemessen, da sowohl "Utrechter" als auch "Utrechtenaar" einen Einwohner der Stadt Utrecht bezeichnen. Man kann sagen, dass DEEPL in diesem Fall *robust* gegenüber Sprachvariationen ist: Es behandelt beide Äußerungen gleich. Es gibt aber auch viele Situationen, in denen es hilfreich sein kann, wenn NLP-Modelle *sensibel* gegenüber Sprachvariationen sind. Historisch war "Utrechtenaar" der üblichere Begriff. Er wurde jedoch heute im allgemeinen Sprachgebrauch weitgehend durch "Utrechter" ersetzt, da "Utrechtenaar" seit den Utrechter Sodomieprozessen (ca. 1730) mit schwulen Männern assoziiert wird. Wenn sich heute jemand als "Utrechtenaar" statt "Utrechter" bezeichnet, wissen wir möglicherweise mehr über diese Person—etwa, dass sie eher Teil der lokalen queeren Community ist. Stellen wir uns einen Zeitungsartikel vor, in dem sich zwei Personen als "Utrechter" bzw. "Utrechtenaar" bezeichnen: Die Übersetzung beider Begriffe mit "Einwohner von Utrecht" könnte subtile Unterschiede in Bezug auf Hintergrund und soziale Identität verschleiern—was möglicherweise zu Verwirrung oder zum Verlust narrativer Nuancen führt.

In dieser Dissertation entwickle ich Methoden, die Sprachmodelle sensibler und robuster gegenüber Sprachvariationen machen. In Kapitel 3 untersuche ich Tokenizer—ein Grundbaustein von Sprachmodellen—hinsichtlich ihrer *Sensibilität* und *Robustheit* gegenüber Sprachvariationen. Dabei zeige ich, dass es wichtig ist, Sprachvariationen in allen Phasen der Entwicklung von Sprachmodellen zu berücksichtigen. In den Kapiteln 4 und 5 entwickle ich Vektordarstellungen, die für einen Aspekt von Sprachvariation sensibel sind: den Sprachstil eines Textes. In Kapitel 4 schlage ich das STyle EvaLuation Framework (STEL) vor, mit dem sich NLP-Methoden zum ersten Mal systematisch daraufhin bewerten lassen, wie *sensibel* sie auf stilistische Variationen in Texten reagieren. In Kapitel 5 trainiere ich neuronale Textrepräsentationen, die—im Gegensatz zu vorherigen Ansätzen—den Sprachstil unabhängig vom Inhalt erfassen und auf STEL gute Ergebnisse erziehlen. Die daraus resultierenden Vektordarstellungen haben in der NLP-Community bereits vielfältige Anwendung gefunden. In Kapitel 6 befasse ich mich mit der Erkennung sprecherübergreifender Paraphrasen in Dialogen. Dazu trainiere ich Crowdworker in meinem eigens entwickelten iterativen Verfahren, Paraphrasen zu klassifizieren. Auf dieser Basis erstelle ich einen Datensatz, der von bis zu 21 Personen annotiert wurde. Meine Ergebnisse zeigen, dass sowohl Menschen als auch NLP-Modelle erhebliche Schwierigkeiten haben, sprachlich variierende, aber inhaltlich gleiche Äußerungen *robust* zu erkennen.

Ich hoffe, dass diese Arbeit die NLP-Community dazu anregt, Sprachvariationen stärker in die Entwicklung von NLP-Methoden einzubeziehen.

List of Publications

Research articles

6. **Anna Wegmann**, Dong Nguyen and David Jurgens (2025). Tokenization is Sensitive to Language Variation. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 10958—10983). https://doi.org/10.18653/v1/2025.findings-acl.572

5. Evgeny Vasilets, Tijs Broek, **Anna Wegmann**, David Abadi and Dong Nguyen (2024). Detecting Perspective-Getting in Wikipedia Discussions. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS 2024)* (pp. 1–15). https://doi.org/10.18653/v1/2024.nlpcss-1.1

4. **Anna Wegmann**, Tijs van den Broek and Dong Nguyen (2024). What's Mine becomes Yours: Defining, Annotating and Detecting Context-Dependent Paraphrases in News Interview Dialogs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 882–912). https://doi.org/10.18653/v1/2024.emnlp-main.52

3. **Anna Wegmann**, Marijn Schraagen and Dong Nguyen (2022). Same Author or Just Same Topic? Towards Content-Independent Style Representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP* (pp. 249–268). https://doi.org/10.18653/v1/2022.repl4nlp-1.26

2. **Anna Wegmann** and Dong Nguyen (2021). Does It Capture STEL? A Modular, Similarity-based Linguistic Style Evaluation Framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 7109–7130). https://doi.org/10.18653/v1/2021.emnlp-main.569

1. **Anna Wegmann**, Florian Lemmerich and Markus Strohmaier (2020). Detecting Different Forms of Semantic Shift in Word Embeddings via Paradigmatic and Syntagmatic Association Changes. In *Proceedings of the 19th International Semantic Web Conference* (pp. 619–635). https://doi.org/10.1007/978-3-030-62419-4_35

Curriculum Vitæ

2004 - 2013 Gauß Gymnasium Worms, Germany

Anna attended secondary school in her hometown. Outside of school, she joined the choir, played football, participated in several musicals, spent time with her three siblings and role-played as diplomats in Model United Nations debates. This led her to Leiden in the Netherlands twice. Anna finished high school with the best grades of her year.

2013 - 2017 RWTH Aachen University, Germany

Anna attended RWTH Aachen University to study mathematics for her bachelor's degree. Group work was important in the mathematics study program—and it led to the formation of Anna's tightly knit group of friends that enjoyed weekly episodes of Dungeons and Dragons together. Anna completed two funded computer science research internships during her bachelor's that led her to Sydney, Australia and Auckland, New Zealand. One of them included implementing a feed forward neural network from scratch in Python. Fun with Python and growing questions about the real-world applications of her study led her to switch the major for her master's degree from mathematics to computer science.

2017 - 2020 RWTH Aachen University, Germany

During her master's at RWTH Aachen, Anna studied computer science with a focus on machine learning and computational social sciences. For her master's thesis, she chose a topic in natural language processing with ties to sociolinguistics: measuring the temporal meaning shift of words in a word embedding space. This work became her first academic publication, which she presented at the International Semantic Web Conference. While writing her master's thesis, Anna practiced juggling in the hallway and got increasingly interested in doing a PhD in the field of natural language processing. With the encouragement of her thesis advisor, she attended the International Conference on Computational Social Science in Amsterdam, where she met Dong Nguyen. A few months later, when Anna reached out for possibilities for a PhD, Dong had just won a grant that would finance Anna's dissertation.

2020 - 2025 Utrecht University, the Netherlands

Anna began her PhD at Utrecht University in August 2020 with Dong Nguyen and Kees van Deemter. The first few months were tough—full on COVID times, a new country and remote work from a tiny Dutch room, with the desk next to her bed. Still, the topic was exciting, the rare in-person company was good and her advisor a great help. This led her to make progress in an important strand of her research career: measuring the linguistic style of texts. A year later, Anna had to undergo surgery for a football-related meniscus tear. Unfortunately, her recovery—including several months of not being able to walk without crutches—overlapped with the brief period when some COVID restrictions were lifted in the Netherlands. Later, with increasing popularity of and competition in the field of NLP, Anna found herself working overtime. To restore the balance between work and private life, she transitioned to a four-day workweek and found time for new hobbies: Lindy Hop dancing, improvisation theater, DJ-ing, singing, drumming, guiding city tours and fanfiction writing. Energized by a new sense of vitality, Anna handed in her thesis as planned on April 22nd, 2025. During her PhD, Anna attended several ACL and EMNLP conferences, enjoyed a research visit in Ann Arbor with David Jurgens, a research visit at the Vrije Universiteit Amsterdam with Tijs van den Broek, and a machine learning summer school in Lisbon.

2025 - present Utrecht University, the Netherlands

Anna joined Dong Nguyen's ERC project on data diversity as a postdoctoral researcher at Utrecht University with July 2025. Among others, Anna is working on measuring the diversity of styles present in textual datasets.

Acknowledgements

I would not have been able to complete this dissertation without the support of many people. This is one place where I want to thank them.

Colleagues

Dong, thank you for being a consistent supporter and believer in my abilities from the start. I appreciate that you always made time for me, provided detailed and invaluable feedback and held me to a high standard that helped me grow. Thank you for your occasional push to pull through when projects seemed to find no end and for opening up possibilities for me, among others, by including me in the DiLCo network. I learned a metric ton from and with you, including meticulous research design, how to ask even more critical questions, and how to develop research questions and methods as an independent researcher. Thank you for welcoming and encouraging me to take up an increasing amount of space in the NLP & Society Lab and for creating an environment where I feel valued when expressing my ideas and perspectives.

Tijs, thank you for being such a warm person that believes in the abilities of the people around him. In the short time I worked closely with you at the VU you managed to always do what I now think is key to being a successful manager: motivate and inspire to continue on and discover in a safe environment where trying and failing is part of the journey.

Kees, thank you for making time for short-term meetings when I needed them. You always provided an interesting and radically different perspective that I hadn't heard from anyone else to that point. You made my work better.

Assessment Committee Thank you for taking the time to read and evaluate my dissertation, Prof. dr. Antal van den Bosch, Dr. Stefania Degaetano-Ortlieb, Prof. dr. Jack Grieve, Prof. dr. Massimo Poesio and Prof. dr. Heike Zinsmeister. I look forward to the defense with you.

NLP & Society Lab It was great to see your development from careful speaker to someone that loudly expresses fascinating opinions and insights in our meetings, **Yupei**. I'm impressed by your dedication to your work and your understanding of many complex details. I learned much more about the current state of affairs across fields in NLP than I would have without you. Thank you for your support and kindness from start to finish of my PhD journey. **Qixiang**, you truly are the cheerleader of the lab, holding us all up with your encouraging and positive feedback and your belief in our contributions. I'm impressed by the seeming ease with which you produce fascinating work in so many different areas and push your career forward. **Sanne**, I'm so happy you joined the NLP & Society Lab, originally for a short visit, and now as our adopted longer

term co-worker. I rarely felt such a quick and effortless connection with someone as I felt with you. I'm impressed by your ability to meaningfully connect to others, your deep expertise in your research topic (expert goals!) and your skill to create space for well-deserved breaks. **Gianluca**, **Melody**, **Shane** & **Elize**, thank you for your valuable feedback in our lab sessions and on my papers. I truly appreciate the atmosphere in our lab and you are all a large part of that. **Tao** & **Menan**, we haven't been working together long, but I already know you will both do great things. **Tao**, you are a great researcher and a kind friend with the greatest compliments. **Menan**, you are an equally great and diligent researcher, and the most considerate person I have had the pleasure to meet. I look forward to getting to know you both better and hope to be able to support you in ways that might help you avoid some of the struggles I experienced in my PhD journey.

NLP Group, UU Thanks to the NLP Group for providing a space to discuss my work in progress and providing constructive feedback. Thanks **Edu**, for shining a light of positivity and ease on university life. I will forever be amazed by your ability to connect to people and see the best in them and their work. You truly are a gift for the group, the department and the university as a whole. **Pablo**, I always love talking to you and hearing about the detailed thoughts you have about the university's two factor authentication system, research practices or time management systems. **Marijn**, thank you for being a helpful and considerate office mate. I always appreciate your detailed and insightful feedback in group and project meetings. **Davide**, we only brushed paths briefly on my way out the door before handing in my dissertation, still you quickly made a lasting impression on my life and the life of those close to me. We hope to see you back in Utrecht soon. Special thanks also to **Albert**, **Guanyi**, **Yingjin**, **Ece**, **Hugh Mee** and **Daniil**.

ICS Department, UU Thank you to all the support staff for your continued support and for suffering my last minute requests, with special thanks to **Ella**, and **Christina**. Thank you also to the support from my other kind colleagues and friends in the broader department: **Thomas**, **Til**, **Jens**, **Evanthia**, **Heleen**, **Upal**, **Eduardo** and **Aditya**.

Blablablab, UMich Thank you, **David**, for inviting me to Ann Arbor and financing my stay. I enjoyed the insight into your research lab and discovering Ann Arbor. I loved the snacks in the lab meetings and the collaborative atmosphere you are fostering. I was impressed by how much time you make for weekly meetings with your students and postdocs and by the level of detail with which you are involved in all projects. Thank you also to **Dustin**, **Anders**, **Jiaxin**, **Kenan** and **Miriam** for enriching my stay in Ann Arbor.

VU, Amsterdam Thank you Tijs, for hosting me at the VU in Amsterdam. **Bianca**, during my stay, one of the things I appreciated the most was your positive and encouraging attitude. **Charlotte**, it's a bit sad we never got to work together as closely as originally intended. The timing didn't quite work out for us. Still, in the times we did exchange ideas I was impressed by your no nonsense attitude and your ability to rally people and resources. **Ana**, thank you for being so open and honest about your academic journey. I feel truly seen and understood by you. I learned from you how to lead a group of researchers with optimism, kindness and patience.

DiLCo Network Thank you for including me in the network and the regular financed meetings, among others, in Hamburg and Utrecht. **Janis**, I appreciate your kindness. Thank you **Christoph**, for last minute feedback on my background section on linguistic style. Errors remain mine of course. **Melissa**, **Jacques**, **Carla**, **Jenia**, **Cristos** & **Holly**, it was great to exchange bad and good experiences across countries and research labs as peers from the beginning of my PhD.

Friends

ABC Family: **Coco**, you are a great and supportive friend. I always appreciate your different perspectives which insightfully complement my own and more than once helped me navigate challenges at work and at home. Thank you for enduring all the loud hobbies I put you through in the past years. I love having you in my life. **Kiki**, you're a truly genuine and interested person. You always make me feel seen when we spend time together. **Mirte**, thank you for your expert help with my thesis cover design. I'm constantly amazed by your talents, creativity and versatile hobbies and deeply appreciate your kindness and calmness. You enrich my life with your presence. **Barbara**, I can't imagine what my PhD would have been like without you. You have been there every step of the way. You are the reason that even in the most difficult times, I kept thinking that there is at least one positive thing about doing my PhD: Becoming friends with you. I learned my most treasured adult life lessons from you. With you I feel that someone really and truly listens, deeply cares about understanding what I say and challenges me when I need it. Thank you for your support, among so many other things, cooking for me while I could not walk and making the distance from home to Ann Arbor easier with your regular letters.

Niek, you were one of the first people I met in Utrecht when you were willing to meet in spite of COVID restrictions for a Gluehwein in the cold. I treasure the time we spend together, be it at Lindy Hop dancing, tabletop RPG adventures or bouldering sessions. Thank you for being a voice of reason and calm and showing me that I can have a life that is fueled by being with friends, taking care of one's needs and being interested in but not dependent on success at work for my well-being. I'm impressed and inspired by your drive to stand for what's right and to care deeply about those around you.

Helen, danke für die regelmäßigen gegenseitigen Anrufe und Besuche. Zeit mit Menschen wie dir ist das, was das Leben für mich lebenswert macht. Ich bin jedes Mal fasziniert und inspiriert von deinen präzisen Gedanken über die Arbeit, Beziehungen, Psychologie und andere Dinge, die dich gerade bewegen.

Marlies, dank je wel voor je vriendschap. We hebben allebei soortgelijke struggles gehad in de afgelopen jaren en het helpt enorm dat ik er met jou over kan praten. Ik heb zelden iemand ontmoet die mensen zo totaal accepteert zoals ze zijn, zoals jij dat doet. Met jou voelt het altijd veilig.

Jeroen, thank you for all our walks and talks about things that move us. The discussions about dating, sense of life and work helped me put my struggles into perspective.

Jacco, je was een van de eerste mensen die ik in Utrecht heb ontmoet. Vanaf het begin voelde het heel gemakkelijk om met jou Nederlands te spreken en te leren. En het is altijd leuk om met je te wandelen, Doctor Who te kijken of gewoon te praten over wat ons bezighoudt.

Amelie, danke für die Wochenenden in Düsseldorf und Utrecht und unsere Spaziergänge zu Weihnachten. Mit dir zu sprechen, fühlt sich an als kennen wir uns schon ewig. ;)

Daniel, danke für deine undermüdliche Unterstüzung, deine Freundschaft, deine Bereitschaft für nächtliche verzweifelte Telefonate, und dass du dich regelmäßig bei mir meldest, wenn ich es versäume. Der Besuch bei dir und **Rian** in Wien war eines der Highlights während meiner Promotionszeit.

Danke auch an **Torben**, **Joana**, **Doro**, **Gereon**, **Alex** and **Esteban**.

Family

Papa & Mama, danke für eure unermüdliche und bedingungslose Unterstützung. Papa, danke, dass du alles stehen und liegen gelassen hast, um bei meiner Knie-OP in Utrecht zu sein. Danke für die Wochen der Pflege bei euch in Worms. Danke, dass ich euch all meine Vorträge zur Probe halten durfte und für das viele Feedback zu Lebensläufen, zur Einleitung meiner Dissertation und zu meinen Umfragen. Ich schätze unsere Telefonate und freue mich jedes Mal darüber, wenn ich dabei merke, wie sehr ihr an mich und meine Fähigkeiten glaubt. Ich habe keine andere Doktorandin getroffen, deren Eltern ein so großes Interesse an der Arbeit hatten und so viel dazu beigetragen haben wie ihr.

Johanna, Laura & Paul Liebe Geschwister, danke für eure Unterstützung in jedem meiner Lebensjahre, und auch insbesondere während meiner Promotion. Danke für unsere samstäglichen Videoanrufe, für die gegenseitigen Besuche, und gemeinsamen Urlaube. Ich bin ständig inspiriert und gefordert von euren Meinungen und Gedanken. **Paul**, du bist eine ständige Inspiration, Interessen zielstrebig zu verfolgen, aber auch zu verwerfen, wenn sie einem nichts mehr geben. **Laura**, deine emotionale Unterstützung ist unermüdlich und für mich unersetzlich. **Johanna**, danke für deine Hilfe bei der Einrichtung unserer Wohnung in Utrecht und für die vielen Second Hand Klamotten, die meine Garderobe so viel einfacher und einzigartiger machen. **Mehrdad**, **Alex** & **Janina**, danke, dass ich euch so oft in Karlsruhe und Zürich besuchen durfte.

Oma & Opa, danke für eure emotionale und finanzielle Unterstützung. Opa, danke für das Künstliche Intelligenz Buch von Manfred Spitzer, das mir eine interessante neue Perspektive auf KI gegeben hat. Oma, die Gespräche mit dir sind immer bereichernd. Danke, dass ich an deinen Gedanken teilhaben und von deinen Erfahrungen lernen darf.

Christof, Judith & Lilith, schön, dass ich euch bei meinen zwei Arbeitsaufenthalten in Hamburg besuchen durfte. Eure weihnachtliche Großzügigkeit hat mich außerdem ermutigt, eines der Hobbies zu beginnen, das mir viel Ausgleich gebracht hat: das Sch-

lagzeugen. **Barbara, Marie & Jule** Danke für eure konsistenten Weihnachtsgrüße und Aufmerksamkeiten in meiner PhD-Zeit.

Conrad, wir haben uns erst drei Wochen vor meiner Abgabe kennengelernt. Aber du warst schnell einer der Gründe, warum ich es nicht abwarten konnte, endlich die Arbeit abzugeben und einen neuen Lebensabschnitt zu beginnen. Du hast mir die Wochen nach meiner Abgabe versüßt und meine Perspektive auf die Zukunft stark ins Positive beeinflusst. Ich freue mich auf post-PhD Abenteuer mit dir.