

Style Analysis of Argumentative Texts by Mining Rhetorical Devices

Khalid Al-Khatib¹ Viorel Morari² Benno Stein¹

¹ Bauhaus-Universität Weimar, Weimar, Germany, <first>.<last>@uni-weimar.de

² Averbis, Freiburg, Germany, viorel.morari@averbis.com

Abstract

Using the appropriate style is key for writing a high-quality text. Reliable computational style analysis is hence essential for the automation of nearly all kinds of text synthesis tasks. Research on style analysis focuses on recognition problems such as authorship identification; the respective technology (e.g., n-gram distribution divergence quantification) showed to be effective for discrimination, but inappropriate for text synthesis since the “essence of a style” remains implicit. This paper contributes right here: it studies the automatic analysis of style at the knowledge-level based on *rhetorical devices*. To this end, we developed and evaluated a grammar-based approach for identifying 26 syntax-based devices. Then, we employed that approach to distinguish various patterns of style in selected sets of argumentative articles and presidential debates. The patterns reveal several insights into the style used there, while being adequate for integration in text synthesis systems.

1 Introduction

The decision for an adequate writing style plays a crucial role for an author who wants to achieve a particular goal, such as persuading the readers (Burton, 2007). “Style” is an elusive concept which covers a wide range of techniques an author can follow, including justifying a conclusion by anecdotal evidence, using regular repetition of the same phrase, or raising questions and then answering them. In the literature on the subject, these techniques are called *rhetorical devices* (Johnson, 2016).

The automatic analysis of style has been addressed mostly by developing a set of style features (aka style indicators) such as the percentage of function words (Ganjigunte Ashok et al., 2013; Bergsma et al., 2012). Those features have proven to be effective in various analysis tasks, such as genre classification and author recognition. However, they are not appropriate for typical text synthesis and writing assistance tasks, since they cannot reveal the “essence of a style” in an explicit and describable manner.

By contrast, analyzing the writing style based on rhetorical devices provides a mechanism to describe *where*, *what*, and *how* specific techniques are used. This kind of analysis is not only important for exploring content in social science (Niculae and Danescu-Niculescu-Mizil, 2014), but it can also serve text synthesis systems by improving the quality of automatically generated texts (Hu et al., 2017). Moreover, it can form the backbone of style suggestion tools. For example, when writing a text for which the desired specification (e.g., the genre) is given, adequate style techniques can be suggested to improve the text quality. In such a manner, new writers can learn to improve their texts and approach the quality of masterpieces written by top writers. Figure 1 illustrates the described connections.

Rhetoric has been the subject of investigation amongst scholars since the time of ancient Greece. Meanwhile, a considerable number of rhetorical devices were developed and discussed in the literature. The most well-known collected lists of devices contain more than 500 devices (Lawrence et al., 2017). Though various of them, such as irony and sarcasm, is hard to be computationally identified (Java, 2015), there is still a sufficiently large portion of popular and—for our purpose—highly useful devices whose identification can be tackled with the current state of the art. Basically, rhetorical devices can

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

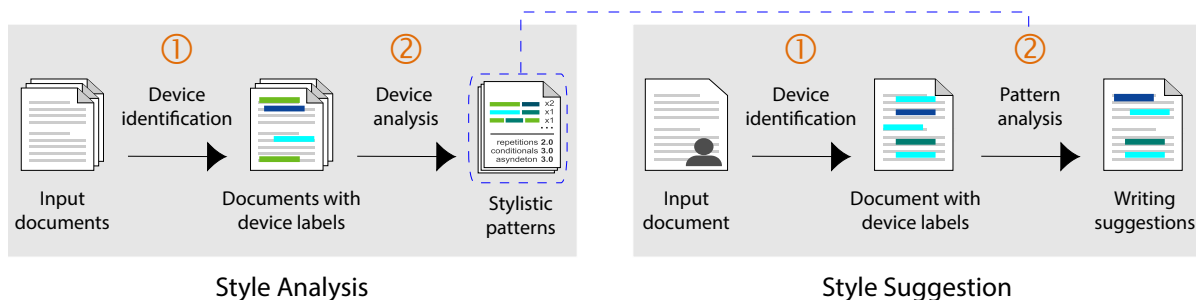


Figure 1: Envisioned tool for style checking and suggestion.

be categorized according to different principles, where an important one is a linguistic level (lexical, syntactic, semantic, and pragmatic). For the time being, we will deal with syntax-based devices.

Against the above background, this paper addresses three research questions. (1) How to identify syntax-based rhetorical devices in a text? (2) What are the most common patterns of using these devices? (3) To which degree differ these patterns across different monological and dialogical argumentative texts? Within and across the texts' genres, topics, and authors? And across different opponent debaters?

To answer these questions, we develop a grammar-based approach for the identification of 26 rhetorical devices. The grammars are built on top of the outputs of a probabilistic context-free grammar parser, PCFG. For evaluation purpose, we create a corpus of 1718 texts which are labelled for rhetorical devices. The evaluation results show that our approach is able to identify the devices with an average of 0.70 in terms of F_1 . Based on the developed approach, we quantify and discuss the usage of devices in monological texts within and across different genres, topics, and authors using a subset of the New York Times annotated corpus (Sandhaus, 2008). We also analyze the devices usage patterns in dialogical texts using a set of presidential debates from the American presidency project (Woolley and Gerhard, 2017).

We consider the gained qualitative and quantitative insights about the usage of rhetorical devices as step forward to a new generation of semi-automated argumentative text generation and writing tools. All developed resources in this paper are made publicly available at www.webis.de

2 Related Work

Recently, Investigating rhetorical devices for style analysis has been considered in computational linguistics. Various devices at the semantic and pragmatic levels have been addressed singly such as irony (e.g., (C. Wallace et al., 2014)), sarcasm (e.g., (Ghosh et al., 2015)), evidence (e.g., (Rinott et al., 2015)), and means of persuasion (e.g., (Duthie et al., 2016)). In a notable work, Strommer (2011) work on identifying 'epanaphora'. They try to distinguish between accidental and intentional use of this device.

Other studies target identifying a mix of syntax, and semantic devices. Gawryjolek et al. (2009) addressed four rhetorical devices: 'anaphora', 'isocolon', 'epizeuxis', and 'oxymorons'. These devices were utilized to recognize the author of a set of documents. Java (2015) identified the four devices mentioned above in addition to nine new devices belonging to parallelism, repetition, and trope. The primary purpose of that work is to use the presence of a rhetorical device as a feature in machine learning models for authorship attribution. Since the authors consider syntax-based devices, we already considered five of their devices in our study. Regarding argumentation, Lawrence et al. (2017) analyzed eight devices, six belong to the syntax and lexical levels, and two to the trope (i.e., semantic or pragmatic). Mainly, a pilot study was conducted to study the relation between argumentation structure and the identified devices.

Few resources for rhetorical devices are publicly free. Up to our knowledge, the code of the previous studies is not available anywhere on the web. Hence, researchers often have to write a new piece of code every time they need to analyze style based on rhetorical devices. This paper resolves this problem considerably by providing a tool for identifying 26 different rhetorical devices. Our developed resources, including the code, will be made freely available.

PCFG outputs have been employed for different tasks including response generation in dialogue (Yuan et al., 2015), multiword expression identification (Green et al., 2011), and the task at hand: identifying

rhetorical devices (Gawryjolek et al., 2009; Java, 2015). However, we develop a set of original heuristic rules that map the devices' definitions to PCFG grammars. As far as we know, many devices from the 26 we identified have not been considered in any other study.

Writing style analysis has been studied widely. The authorship recognition has been tackled in a large number of papers (e.g., (Sundararajan and Woodard, 2018)). Besides, quality assessment research has involved applying several style analysis features (e.g., (Ganjigunte Ashok et al., 2013)). In comparison to our analysis, we conducted a *controlled* analysis using 'matching' technique and we covered various aspects of monological and dialogical texts such as genre, topic, author, and debate opponent.

3 Identification of Rhetorical Devices

Rhetorical devices are the techniques of using the language to produce an effect on the target audience or readers (McKay and McKay, 2010). For example, repeating particular phrases can produce effects such as emphasizing a certain argument, or evoking a specific emotion (Corbett, 1990).

This paper targets syntax-based rhetorical devices. Particularly, we aim to identify 26 devices belonging to two main categories: (1) **figurative syntax**, which is referred as *schemes* in literature, and (2) **ordinary syntax**, which concerns the rules of well-formed structuring texts. The effect of the first attributed to using an artful deviation from the ordinary arrangement of words, while the effect of the second is coming from using a specific arrangement of words among other arrangements.

In the next subsections, we detail the figurative and ordinary syntax devices and describe our approach for identifying them.

3.1 Figurative Syntax Devices

Figurative devices center on arranging words artfully (Burton, 2007). They are divided into four types: *balance*, *inversion*, *omission* and *repetition*.

- The *balance* devices involve arranging the rhythm of thoughts. Hence, they can produce a sense of equivalence among the proposed ideas, or emphasize ideas' differences. For example, we can notice the contrast between ideas in the famous quote of Neil Armstrong: "That's one small step for man, one giant leap for mankind".
- The *inversion* devices concern changing the order of words, either to stress some ideas or to avoid the monotonous flow of a sentence. For example, "Everybody's got troubles" could be reordered to "Troubles, everybody's got."
- The *omission* devices deal with removing words that readers can reveal intuitively. They are often used to imply unfinished thoughts or to keep a fast rhythm, such as: "He came, he saw, he conquered."
- The *repetition* is the most frequent, and arguably, the most powerful. According to Aristotle, repetition is the key to a persuasive speech (Fahnestock, 2003). Typically, repetition devices aim at influencing the emotional state of the reader by emphasizing or implicating a specific idea (Burton, 2007; Corbett, 1990). An example which illustrates the emotional impact of repetitions is the famous line from King Lear written by Shakespeare: "Never, never, never, never, never." (Müller, 2006).

Table 1 shows an overview of the identified figurative devices in our work. The overview covers a definition, a formalization, and an example for each device belongs to balance, omission, or repetition¹. Our formalization is grounded on the devices' definitions which are taken from a set of reliable sources such as 'Silva Rhetoricae': a comprehensive source for rhetoric on the web (Burton, 2007).

The formalization elements are: 'Cl' for clause, 'Phr' for phrase, 'W' for word, 'N' for noun, 'Vb' for verb, 'CC' for conjunction, 'COMMA' for comma, ... for arbitrary intervening material, [...] for word boundaries, {...} for phrase or clause boundaries, $a = b$ for identity, and $a \neq b$ for nonidentity. The elements of formalization are adopted from Harris and DiMarco (2009).

Notice that we essentially concentrate on identifying the devices at the sentence-level, or across consecutive sentences. Besides, some rhetorical devices, according to their definitions, might overlap with

¹The inversion is left to future work due to its complexity.

Device definition Formalization Example	Device definition Formalization Example
<p>(B1) Enumeration: Lists a series of details, words or phrases. < ...W [CC COMMA] W ...> Diligence, talent and passion will drive anybody to success.</p>	<p>(C1) If-cond. Zero: Conditionals which express general truths. If [VB/VBP/VBZ], then [VB/VBP/VBZ] If you heat ice, it melts.</p>
<p>(B2) Isocolon: Similarly structured elements with the same length. < ...<Phr>_a <Phr>_a ...<Phr>_a ...> Fill the armies, rule the air, and pour out the munitions.</p>	<p>(C2) If-cond. One: Express situations which are very likely to happen in the future. If [VB/VBP/VBZ/VBG], then [MD+VB] If it rains, you will get wet.</p>
<p>(B3) Pysma: Asking multiple questions successively. < ...< Cl? > < Cl? > ...>? Ex: Who are you? Why are you doing here</p>	<p>(C3) If-cond. Two: Expresses consequences that will not likely happen in the future. If [VBD], then [MD+VB] If it rained, you would get wet.</p>
<p>(O1) Asyndeton: Omission of conjunctions between clauses. <Cl_a> COMMA <Cl_b> COMMA <Cl_c> ...) I came, I saw, I conquered.</p>	<p>(C4) If-cond. Three: Used to explain that present circumstances would be different if something different had happened in the past. If [VBD+VBN] , then [MD+VBN] If I had worked harder, I would have passed the exam.</p>
<p>(O2) Hypozeugma: Placing last, in a construction containing several elements of equal value, the word(s) on which all of them depend. < ... [W]_a , [W]_b , [W]_c ...Vb > Friends, Romans, countrymen, lend me your ears. . .</p>	<p>(C5) If-Counterf.: Statements that examine how a hypothetical change in a past experience could have affected the outcome of that experience. If [VBD+VBN], then [past modals] If I were you, I wouldn't come.</p>
<p>(O3) Epizeugma: Placing the verb that holds together the entire sentence either at the very beginning or the very ending of that sentence. < Vb ...> or, < ...Vb > Neither a borrower nor a lender be.</p>	<p>(C6) Unless-cond.: Restricted version of if-conditional (its intrinsic meaning is narrowed down to "Q in the case other than P"). <... unless ...> You can't go on vacation unless you save some money.</p>
<p>(R1) Epanalepsis: Repetition at the end of a line, the word(s) that occurred at the beginning of the same line. < [W]_a ... [W]_a > Believe not all you can hear, tell not all you believe.</p>	<p>(C7) Whether-cond.: Expresses alternative (disjunctive) conditions. <... whether ... or ...> Whether you are overweight or not, it is always better to watch your diet.</p>
<p>(R2) Mesarchia: Repetition of the same word(s) at the beginning and middle of successive sentences. < [W]_a ... [W]_b ...> < [W]_a ... [W]_b ...> I was looking for a paper. I was anxious for a paper.</p>	<p>(CS) Comp./Super. Adj. and Adv.: Used to compare differences between the two objects/states they modify. <... [JJR/JJS/RBR/RBS] ...> (comp. adjective): My house is larger.</p>
<p>(R3) Epiphzoa: Repetition of the same word(s) at the end of successive sentences. < ... [W]_a > < ... [W]_a > O apple! wretched apple! Miserable apple!</p>	<p>(PV) Passive Voice: Occurs when the object of an action is changed into the subject of a sentence. <... [to be] ... [VBN] ...> The problem is solved.</p>
<p>(R4) Mesodiplosis: Repetition of the same word(s) in the middle of successive sentences. < ... [W]_a ...> < ... [W]_a ...> There's no time like the future! There's no time like the past!</p>	
<p>(R5) Anadiplosis: Repetition of the last word(s) from the previous sentence at the beginning of the next. < ... [W]_a > < [W]_a ...> We ordered a pizza pie. A pizza pie that changed our lives.</p>	
<p>(R6) Diacope: Repetition of a word with one or more in between. < ... [W]_a ... [W]_a ...> The horror! Oh, the horror!</p>	
<p>(R7) Epizeuxis: Repetition of words with no others between. < [W]_a [W]_a > Awake, awake and stand up O Jerusalem.</p>	
<p>(R8) Polysyndeton: Several conjunctions in close succession (mainly between clauses). <Cl_a> CC <Cl_b> CC <Cl_c> ... He ran and jumped and laughed for joy.</p>	

Table 1: An overview of the (B) Balance, (O) Omission, and (R) Repetition figurative devices, and the (C) Conditionals, (CS) Comparatives and Superlatives, and (PV) Passive Voice ordinary devices.

other devices in some special cases. This overlap is rare and partial. Nevertheless, we consider minimizing the possible overlaps among devices as much as possible in our formalization.

3.2 Ordinary Syntax Devices

From the ordinary syntax devices, we select *conditionals*, *comparatives and superlatives*, and *passive voice*. This selection is based on the impact of these devices on the readers (Martinet, 1960).

- The *conditional* devices entail the causality aspect of the language, and causality, in turn, could imply explaining an event. But it can also be used to argue about positive/negative consequences of a specific action such as “If we were elected him, we would not have achievements”.
- The *comparatives and superlatives* devices might be used to emphasize the superiority of an entity or idea, e.g., “I will be the greatest jobs president that God ever created”.
- The *passive voice* might be used to hide the subject of a negative action, or to stress the importance of an event, e.g., “many mistakes were made, but the future will be great”.

Table 1 provides an overview of the conditionals, comparatives and superlatives, and passive voice devices. The overview is analog to the one of the figurative category. The formalization is based on definitions from the same set of resources used in the figurative category. The elements of formalization are taken from The Penn Treebank POS Tag Set (Marcus et al., 1993).

3.3 Experiments and Results

Device	Instances	Prec.	Recall	F1	Device	Instances	Prec.	Recall	F1
(B1) Enumeration	60	0.76	0.93	0.84	(C1) If-cond. Zero	60	0.71	0.76	0.73
(B2) Isocolon*	180	0.57	0.83	0.68	(C2) If-cond. One	60	0.78	0.78	0.78
(B3) Pysma	60	1	1	1.00	(C3) If-cond. Two	60	0.82	0.75	0.78
(O1) Asyndeton	60	0.25	0.93	0.39	(C4) If-cond. Three	60	0.86	0.65	0.74
(O2) Hypozeugma	60	0.61	0.8	0.69	(C5) If-Counterf.	60	0.84	0.87	0.85
(O3) Epizeugma	60	0.65	0.7	0.67	(C6) Unless-cond.	60	1	1	1.00
(R1) Epanalepsis	60	0.63	0.83	0.72	(C7) Whether-cond.	60	1	0.83	0.91
(R2) Mesarchia	20	0.45	0.85	0.59	(CS1) Comp. Adj.	68	0.51	0.61	0.56
(R3) Epiphiza	60	0.58	0.93	0.71	(CS2) Comp. Adv.	70	0.6	0.62	0.61
(R4) Mesodiplosis	40	0.27	0.68	0.39	(CS3) Super. Adj.	70	0.62	0.73	0.67
(R5) Anadiplosis	60	0.76	0.73	0.74	(CS4) Super. Adv.	70	0.63	0.5	0.56
(R6) Diacope	60	0.73	0.73	0.73	(PV) Passive Voice	60	0.78	0.98	0.87
(R7) Epizeuxis	60	0.79	0.77	0.78	Other †	60	0.23	0.23	0.23
(R8) Polysyndeton	60	0.77	0.7	0.73					

* including samples of bicolon (60), tricolon (60) and tetracolon (60).

† we applied the 26 classifiers on the ‘other’ instances. If any of them labels an instance with its device, we consider the instance as wrongly classified.

Table 2: The precision, recall, and F₁-score for identifying the 26 rhetorical devices.

Category	Instances	Prec.	Recall	F1
(B) Balance	300	0.67	0.88	0.76
(O) Omission	180	0.4	0.81	0.54
(R) Repetition	420	0.6	0.77	0.67
(C) Conditionals	420	0.85	0.8	0.82
(CS) Comp.&Super.	278	0.59	0.62	0.60
(PV) Passive voice	60	0.78	0.98	0.87

Table 3: The precision, recall, and F₁-score for identifying the rhetorical devices by category.

Here, we discuss the evaluation experiments of our approach for identifying the syntax-based rhetorical devices. First, we describe the newly created evaluation dataset. Then, we talk about the experimental settings and report on the obtained results. Finally, we address the limitations of our approach and perform an error analysis for its output.

Evaluation Dataset: Creating a dataset for rhetorical devices using manual annotation, even with crowd-sourcing, is extremely expensive and time consuming (Java, 2015); The reason behind this is the big number of devices, the potential overlaps between them, and the possibility for some devices to be spread across phrases, sentences, or even paragraphs. Thereby, we decided to follow a bunch of related research studies (e.g., (Java, 2015)) and build the evaluation dataset as follows: We first identify a set of trustworthy sources on the web, which address the rhetorical devices, and have credibility as being developed by experts in rhetoric. Most of the selected sources are either mentioned or already used in some research studies, which speaks for their trustworthiness. From those sources, we use meta-data information (e.g., “Example of Pysma:”) to collect a set of instances for our rhetorical devices. We found that targeting about 60 examples for each device is reasonable considering the size of content in the selected sources. We verified all the examples and ensured that there are no duplicated ones. Additionally, we accounted for the possible overlaps between the devices and minimized them adequately, i.e., all the examples for a device belong solely to this device. Unfortunately, two devices turned out to be considered only from few sources, and hence, we got less than 60 examples for them. We also collected 60 examples where none of the devices covered by our work is used. Overall, we collected 1718 examples: 1658 example distributed among the 26 devices and around 60 examples that belong to ‘other’. The distribution is shown in Table 2. This dataset, despite its relatively small size, is significantly larger than those that have been used for rhetorical devices in related work (Java, 2015).

Experimental Settings: The implementation of our approach was carried out using Apache RutaTM (Rule-based Text Annotation) (Kluegl et al., 2016). This tool provides a flexible language for identifying patterns in text spans intuitively. Thus, it facilitates identifying sophisticated patterns with a few lines of code. The implementation is performed on top of the outputs of Stanford Parser (Manning et al., 2014), the version of 3.8.0. We evaluated our approach using the one-vs.-rest classification. That means we performed one classification experiment for each device; The instances of this device in the evaluation dataset is considered as the *positive* class, and the instances of the remaining devices as well as the ‘other’ as the *negative* class. The classifiers’ effectiveness is reported in terms of precision, recall, and F₁-score.

Classification Results: Table 2 shows the results of our experiments. Overall, we manage to identify the 26 devices with an average of 0.70 F₁-score., which indicates a high effectiveness of our approach.

As for the “figurative” devices, the approach got high scores for the *balance* devices, including F₁-score of 1.00 for ‘pysma’. The ‘isocolon’ is the most challenging with F₁-score of 0.68. As for the *omission* devices, the F₁-scores range from 0.39 for ‘asyndeton’ and 0.69 for the ‘hypozeugma’. These results are a bit lower than the other types. Most of the *repetition* devices have F₁-score of about 0.73, except ‘mesarchia’ with 0.59, and ‘mesodiplosis’ with 0.39. Besides, “ordinary” devices got scores between 0.56 and 1.00. Interestingly, despite their simple syntax, *comparatives and superlatives* devices got the lowest scores.

Table 3 shows the results of our approach regarding the six rhetorical categories that group the 26 devices. The F₁ scores range from 0.54 to 0.87. The best result is obtained for passive voice (0.87) and conditionals (0.82). Omission and repetition are the hardest to identify with 0.54 and 0.67 F₁.

Figure 2 shows an excerpt from a news editorial along with several rhetorical devices that our approach manages to identify.

Error Analysis: Despite the high effectiveness of our approach, it is subject to fail in some cases.

Concerning the “figurative” category, identifying the *balance* devices seems to be precise except for ‘isocolon’. The identification of this device is based on the outputs of the syntax parser (i.e., POS tags) which are sometimes inaccurate, especially for long sentences. This has a negative impact on the precision score; for instance, “It looks like the Libertarian candidate is racking up the percentage points in recent polls. As far as I can see the Libertarian candidate has over . . .”. Here, the ‘Libertarian candidate’ makes the classifier of ‘isocolon’ treats it wrongly as a valid instance. For the *omission* devices, our approach manages to get 0.93 recall score for ‘asyndeton’ device, but only 0.25 for precision. We found that the abundance of commas, which we use as an indicator of the lack of conjunctions is insufficient to distinguish ‘asyndeton’ from other devices, especially ‘enumeration’. For example, “Old McDonald

Title: HEALTH; Salk's Injectable Polio Vaccine May Be Revived	Rhetorical Device
Health officials are considering a major change in the strategy of polio vaccination, using a new, more potent version of the injectable Salk vaccine that helped eradicate polio in the United States almost 30 years ago.	Comp. Adv.
The injectable "killed-virus" vaccine was largely replaced by an oral vaccine made from live viruses, which is still being given to millions of American children.	Passive Voice
The development of the new form of the Salk vaccine opens the way for it to be used in combination with other childhood vaccinations. Some health officials, noting that it has been used in Europe and tested in the developing world, believe that it can be an effective way to reduce immunizations and associated costs.	Passive Voice Passive Voice Asyndeton
However, other experts say the current reliance on the Sabin oral live-virus vaccine has worked so well that great care should be taken before changing policies.	Passive Voice
Study Under Way The Institute of Medicine, an adjunct of the National Academy of Sciences, is studying polio policy and is expected to submit recommendations to the Federal health authorities by April. At a recent public meeting in Washington, the committee heard suggestions for bringing back the inactivated-virus vaccine by combining it with the diphtheria, tetanus and pertussis shots.	Passive Voice Asyndeton Enumeration Hypozeugma Enumeration
Dr. Frederick C. Robbins of Case Western Reserve University, chairman of the panel, said early attempts at the combined vaccine were abandoned in this country because of potency problems. Later successes with this approach in Europe, using an enhanced polio vaccine, have rekindled the idea of a combination approach, including the possibility of using both types of polio vaccines to merge their benefits, he added.	Asyndeton Hypozeugma Passive Voice Asyndeton Hypozeugma

Figure 2: An excerpt from a NYT news editorial. The rhetorical devices in each sentence are identified using our approach.

had a pig, a dog, a cow and a horse.” is identified as ‘asyndeton’, while it is actually ‘enumeration’. As regards *repetition*, two devices there got low scores: the ‘mesarchia’ and ‘mesodiplosis’. These devices have the least number of instances in our evaluation dataset. We also observed that our heuristic rules for defining the beginning and middle of sentences are the reason for some errors.

For the “ordinary” category, the approach has promising results. However, the scores for the ‘comparatives and superlatives’ are moderate. Observing the errors there, we found that the main reason is again the inaccurate POS tags. For example, in the sentence ‘the airport is *further* than the train station.’, ‘further’ is tagged as comparative adverb instead of comparative adjective.

The ‘other’ class got a low F_1 - score. In addition to its restrictive way of evaluation that we followed, this score indicates that some devices’ classifiers tend to have a lot of false positives.

To have a better idea regarding the effectiveness of our approach, we performed a manual inspection of the classifiers’ outputs on a set of ten newspaper articles. We found that some devices such as ‘isocolon’ and ‘asyndeton’ indeed have many false positives. Besides, we found that the classifiers make more mistakes with very long sentences.

4 Analysis of Rhetorical Devices

We rely on our identification approach to analyze the usage patterns of rhetorical devices in argumentative newspaper articles and presidential debates. First, we describe the acquisition and sampling of the *analysis datasets*. Then, we discuss the distribution of rhetorical devices there along with different article and debate aspects. The computed distributions illustrate various patterns of rhetorical devices and lead to several interesting insights.

Analysis Datasets: To conduct insightful analysis, we constructed two datasets for newspaper articles and presidential debates.

(1) Newspaper dataset: to construct this dataset, we used the NYT annotated corpus (Sandhaus, 2008). The corpus comprises more than 1.8 million high-quality articles written by professional writers. It comes with many types of meta-data labeled by NYT staff, including the type of material (e.g., editorial), the author name, and the topic (e.g., sport). From this corpus, we sampled three subsets, each of which represents one of the three properties of genre, topic, and author. To conduct a *controlled analysis*, the sampling should account for the confounding variables. For example, studying the style in articles with

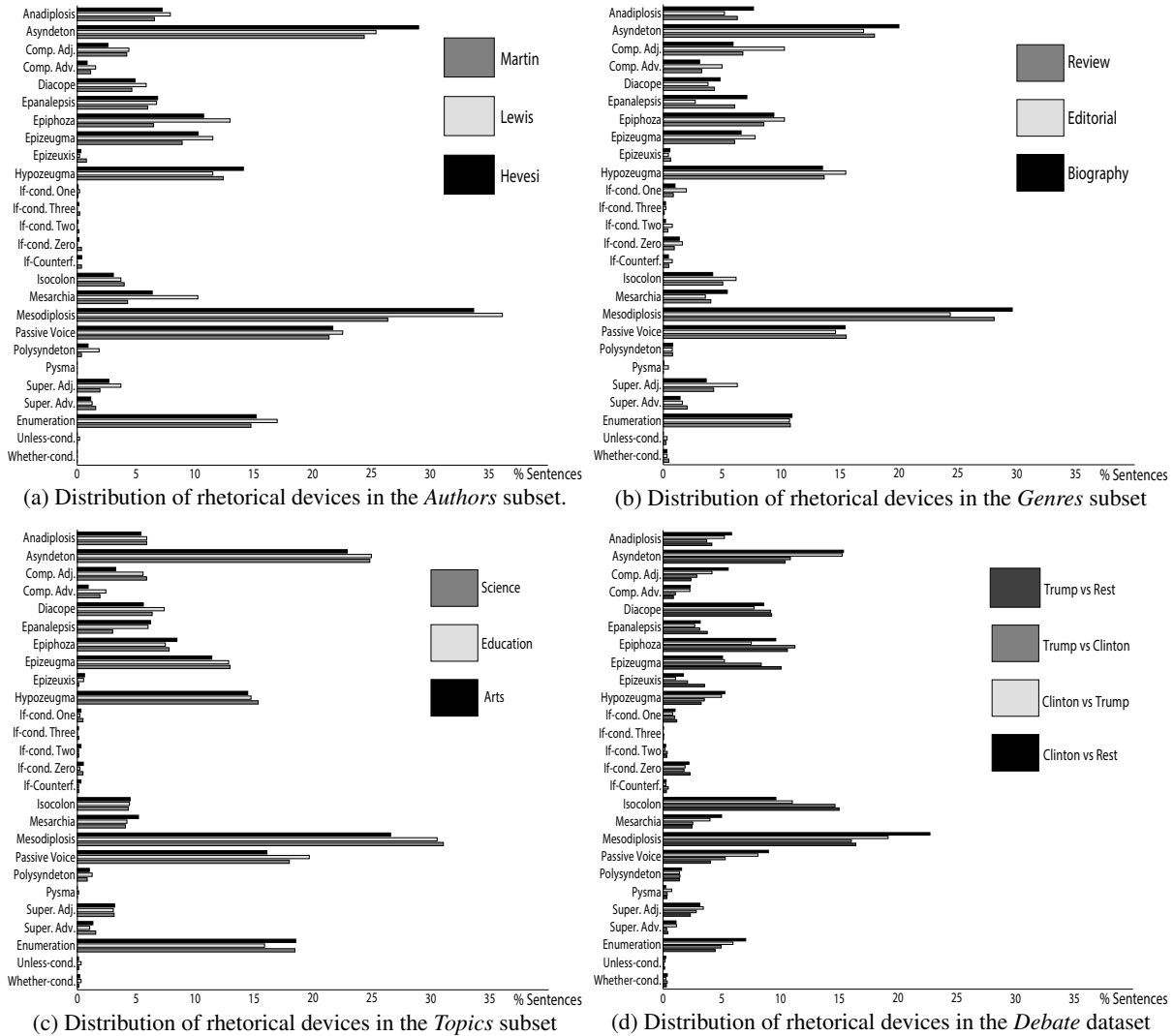


Figure 3: The distribution of rhetorical devices among the authors (a), the genres (b), and the topics (c) in the newspaper articles dataset, and the distribution of rhetorical devices in the debate dataset (d).

a specific ‘topic’ can be influenced by their genres and authors. Hence, we first tried to resolve this issue with the stratification method (Tripepi et al., 2010), which turned out to be not successful; despite the large size of the corpus, we found no information about the authors of about 40% of articles. Also, the distribution of articles in texts belong to the three properties are very skewed. The corpus includes much more reviews than editorials, for example. Many articles are written for ‘politics’ and few for ‘sport’, and some authors wrote tens of articles while others wrote only one. Therefore, we tried the matching technique (de Graaf et al., 2011), where we managed successfully to sample the three subsets. To preserve the balance between the subsets, we consider three *instances* for each propriety, i.e., the ‘topic’ subset includes 114 articles belong to science, education, and art. The ‘genre’ subset includes 89 articles belong to biography, editorial, and review. Finally, the ‘authors’ subset includes 159 articles written by Martin, Lewis, and Hevesi.

(2) Debate dataset: we acquired this dataset based on the presidential debates from the American presidency project (Woolley and Gerhard, 2017). In particular, we extracted the entire set of debates that involve Donald Trump or/and Hillary Clinton. We think that these two characters are different in many aspects such as ideology, background, experience, opinions on different topics, etc. This difference could be reflected in their styles leading to interesting patterns. We created three subsets of the dataset: ‘Trump vs. Clinton,’ ‘Trump vs. Not-Clinton’, and ‘Clinton vs. Not-Trump’. In this way, we can analyze the style of the two characters, and also address the question of whether they change their styles according to

the debate opponent. In total, Clinton has 226 turns in her debates with Trump, and 1216 in her debates with the other candidates. Trump, on the other side, has 342 turns in his debates with Clinton, and 778 in his debates with the rest of candidates.

Analysis Method: Basically, we applied our Identification approach (see Section 3) to the analysis datasets. In particular, a classifier for each device is applied to the articles or debates turns, resulting in the frequency of the device there. However, since our identification approach is not perfect, it is crucial to account for its errors. Hence, we followed the method used in (Al-Khatib et al., 2017): for the frequency n of a rhetorical device rd in an instance i in a dataset. We computed a confidence interval for n , where the $lowerbound = n * precision(rd)$, and the $upperbound = n/recall(rd)$. Ultimately, The mean of the upper and lower bounds is the new frequency, which is normalized by the number of sentences in the articles/debate-turns belong to i . Accordingly, we computed the distributions of rhetorical devices in the analysis datasets and their subsets. The chi-squared test with 0.01 significant level is used to check whether the difference in the usage of rhetorical devices in the datasets and across their instances is significant, and the Cramer’s V test is used to measure the effect-size of the distributions’ difference.

Analysis Results: Figure 3 shows the distribution of rhetorical devices among the three authors (a), the three genres (b), the three topics (c), and the debate subsets (d). As expected, the style in newspaper articles (monologue) are significantly different than in debates (dialogue). Some analysis results for each of the datasets are as follows.

(1) Newspaper dataset: In addition to the significant difference among the three properties under studied, the results show a significant difference among the three authors. For example, Lewis and Hevesi use more *repetition* than Martin. Also, Lewis barely considers conditionals, in contrast to the other two authors. The results also show a significant difference between ‘biography’ and ‘editorial’ as well as ‘editorial’ and ‘review’, but not between ‘review’ and ‘biography’. The reason might be that the articles in these two genres are written mainly to describe an entity. Interestingly, there is no significant difference between the three topics. Overall, our analysis suggests that the “style” identified by syntax-based rhetorical devices is primarily influenced by the ‘author’, and ‘genre’, while ‘topic’ has the least impact. (2) Debate dataset: Interestingly, the results show that Clinton is more fond of ‘comparatives’ and ‘passive voice’ than Trump, which actually contradicts a widespread assumption (Gingell, 2016; Raskin, 2016). However, our findings are mainly related to the debate genre. The style could be different in speeches, for example. We also found that Clinton uses ‘asyndeton’ more often than Trump. Since this device is very effective for making the turns easier to grasp, our finding this time is in line with (Raskin, 2016), where they find that Clinton’s language is 13% clearer and more direct than Trump’s. The results indicate a significant difference between Clinton and Trump styles. More interestingly, while Clinton’s style is significantly different when she debates with Trump than when she debates with the rest, Trump’s style has no significant difference between his debates with Clinton and his debates with the rest. Apparently, unlike Clinton, Trump does not change his style depending on the opponent.

5 Conclusion and Future Work

Writing style analysis has become a mature discipline, but it is mostly tackled from the recognition perspective. I.e., it can give strong classification results that, because of their intrinsic nature, cannot be transferred to constrained text generation or computational writing assistance. We address this shortcoming by proposing an approach for the *explicit* encoding and identification of rhetorical devices. In carefully designed experiments, we study the usage of these devices in different argumentative articles and presidential debates. The distributions show different patterns of style among three text’s properties and provide new insights regarding style usage within the studied topics. The achieved F_1 classification performance (0.70) can be considered as very good for concrete multi-class classification setting; it shows that the applied approach has the potential to find its way into real-world argumentative text synthesis tools. We plan in the future to improve our grammars to minimize mistakes and increase the number of devices considering the *inversion* type of figurative syntax devices.

References

- Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of Argumentation Strategies across Topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 17)*, pages 1362–1368. Association for Computational Linguistics.
- Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 12)*, pages 327–337. Association for Computational Linguistics.
- G. Burton. 2007. The forest of rhetoric (silva rhetoricae). Accessed on 16.08.2017.
- Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans Require Context to Infer Ironic Intent (so Computers Probably do, too). In *Proceedings of the 2014 Annual Meeting on Association for Computational Linguistics (ACL 14) - Volume 1*.
- E. P. J. Corbett. 1990. *Classical rhetoric for the modern student*. USA: Oxford University Press, 3 edition.
- M A de Graaf, K J Jager, C Zoccali, and F W Dekker. 2011. Matching, an Appealing Method to Avoid Confounding? *Nephron Clin Pract.*
- R. Declerck and S. Reed. 2001. *Conditionals: A Comprehensive Empirical Analysis*. Beitrage Zur Alexander-Von-Humboldt-Forschung. Mouton de Gruyter.
- Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining Ethos in Political Debate. In *6th International Conference on Computational Models of Argument (COMMA 16)*, pages 299–310.
- Jeanne Fahnestock. 2003. Verbal and Visual Parallelism. *Written Communication*, 20(2):123–152.
- Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with Style: Using Writing Style to Predict the Success of Novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1764. Association for Computational Linguistics.
- Jakub J. Gawryjolek, Randy A. Harris, and Chrysanne DiMarco. 2009. An annotation tool for automatically detecting rhetorical figures. In *Proceedings, CMNAIX (Computational Models of Natural Argument)*.
- Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 15)*.
- James Gingell. 2016. Why superlatives are the absolute worst (unless you’re Donald Trump). <https://www.theguardian.com/media/mind-your-language/2016/apr/15/>. visited on 24.10.17.
- Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. 2011. Multiword Expression Identification with Tree Substitution Grammars: A Parsing Tour De Force with French. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’11*, pages 725–735. Association for Computational Linguistics.
- R. Harris and C. DiMarco. 2009. Constructing a Rhetorical Figuration Ontology. In *In Symposium on Persuasive Technology and Digital Behaviour Intervention*.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward Controlled Generation of Text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.
- James Java. 2015. *Characterization of Prose by Rhetorical Structure for Machine Learning Classification*. Ph.D. thesis.
- R. Johnson. 2016. *The Alphabet of Rhetoric*. BiblioLife.
- Peter Kluegl, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe. 2016. Uima ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22:1–40.
- Johna Lawrence, Jackya Visser, and Chris Reed. 2017. Harnessing rhetorical figures for argument mining. *Argument & Computation*, 8:289–310.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

- Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.*, 19(2):313–330.
- André Martinet. 1960. *Elements of General Linguistics*. Faber and Faber Ltd., London.
- Brett McKay and Kate McKay. 2010. Classical Rhetoric 101. Accessed on 14.08.2017.
- Wolfgang G. Müller. 2006. Style. In Thomas O. Sloane, editor, *Encyclopedia of Rhetoric*. Oxford University Press, February.
- Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of EMNLP*, October.
- Robin Raskin. 2016. Hillary clinton’s acceptance speech as seen by the algorithms. the huffington post. <https://www.huffingtonpost.com/robin-raskin>. visited on 18.11.17.
- Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show Me Your Evidence - An Automatic Method for Context Dependent Evidence Detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 15)*, pages 440–450. Association for Computational Linguistics.
- Evan Sandhaus. 2008. The new york times annotated corpus ldc2008t19. dvd. *Philadelphia: Linguistic Data Consortium*.
- Claus W. Strommer. 2011. *Using rhetorical figures and shallow attributes as a metric of intent in text*. Ph.D. thesis, University of Waterloo, Waterloo, Ontario, Canada.
- Kalaivani Sundararajan and Damon L. Woodard. 2018. What represents "style" in authorship attribution? In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2814–2822.
- Giovanni Tripepi, Kitty J Jager, Friedo W. Dekker, and Carmine Zoccali. 2010. Stratification for confounding – part 1: The mantel-haenszel formula.
- John T. Woolley and Peters Gerhard. 2017. American Presidency Project. <http://www.presidency.ucsb.edu/>. visited on 18.11.17.
- Caixia Yuan, Xiaojie Wang, and Qianhui He. 2015. Response Generation in Dialogue Using a Tailored PCFG Parser. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 81–85. Association for Computational Linguistics.