Computational Analysis of Argumentation Strategies

by

Khalid Ibrahim Jamal Al Khatib

Dissertation to obtain the academic degree of
Dr. rer. nat.

Faculty of Media
Bauhaus-Universität Weimar, Germany

September 2019

Advisor: Prof. Dr. Benno Stein
Reviewer: Prof. Dr. Manfred Stede

Contents

Abstract

The computational analysis of argumentation strategies is substantial for many downstream applications. It is required for nearly all kinds of text synthesis, writing assistance, and dialogue-management tools. While various tasks have been tackled in the area of computational argumentation, such as argumentation mining and quality assessment, the task of the computational analysis of argumentation strategies in texts has so far been overlooked.

This thesis principally approaches the analysis of the strategies manifested in the persuasive argumentative discourses that aim for persuasion as well as in the deliberative argumentative discourses that aim for consensus. To this end, the thesis presents a novel view of argumentation strategies for the above two goals. Based on this view, new models for pragmatic and stylistic argument attributes are proposed, new methods for the identification of the modelled attributes have been developed, and a new set of strategy principles in texts according to the identified attributes is presented and explored.

Overall, the thesis contributes to the theory, data, method, and evaluation aspects of the analysis of argumentation strategies. The models, methods, and principles developed and explored in this thesis can be regarded as essential for promoting the applications mentioned above, among others.

Chapter 1

Introduction

What's the use of running if you are not on the right road.

People's lives are crowded with diverse situations in which they need to form an opinion, shape a belief, or make a decision on a certain topic. Typically, people satisfy such a need using one of the fundamental aspects of communication: *argumentation*. Argumentation is "a verbal activity that aims at increasing or decreasing the acceptability of a controversial standpoint" [van Eemeren *et al.*, 2014].

Argumentation is exposed within diverse forms (e.g., news, debate), genres (e.g., editorial, scientific publication), directionalities (dialogue and monologue), and modalities (spoken and written). Examples of spoken argumentation are presidential debates, sales pitches, and classroom discussions. Written argumentation arises in editorials, forum posts, and funding proposals, to name a few. Argumentation is used to achieve several goals including persuasion, consensus, and justification. Figure 1.1 illustrates some of the facets of argumentation.

Persuasion and consensus are two of the primary goals of argumentation [Mohammed, 2016; Walton, 2010]. Persuasion concerns changing other people's beliefs, attitudes, etc., while consensus entails agreeing on the best course of action among possible ones. Persuasion can be targeted in persuasive monologues or dialogues, while the consensus is mainly targeted in dialogues, and in particular, in deliberative discussions.

Ideally, the goal of persuasion or consensus can be achieved as long as the composition of arguments and textual information in the argumentation follows an effective *argumentation strategy*. An argumentation strategy, as this thesis argues, is a set of *principles* that guides the selection and arrangement of arguments (plus contextual information) in an argumentative discourse. In this regard, a principle is a rule that specifies a basis for the selection and the arrangement. This may be as simple as selecting all available arguments and

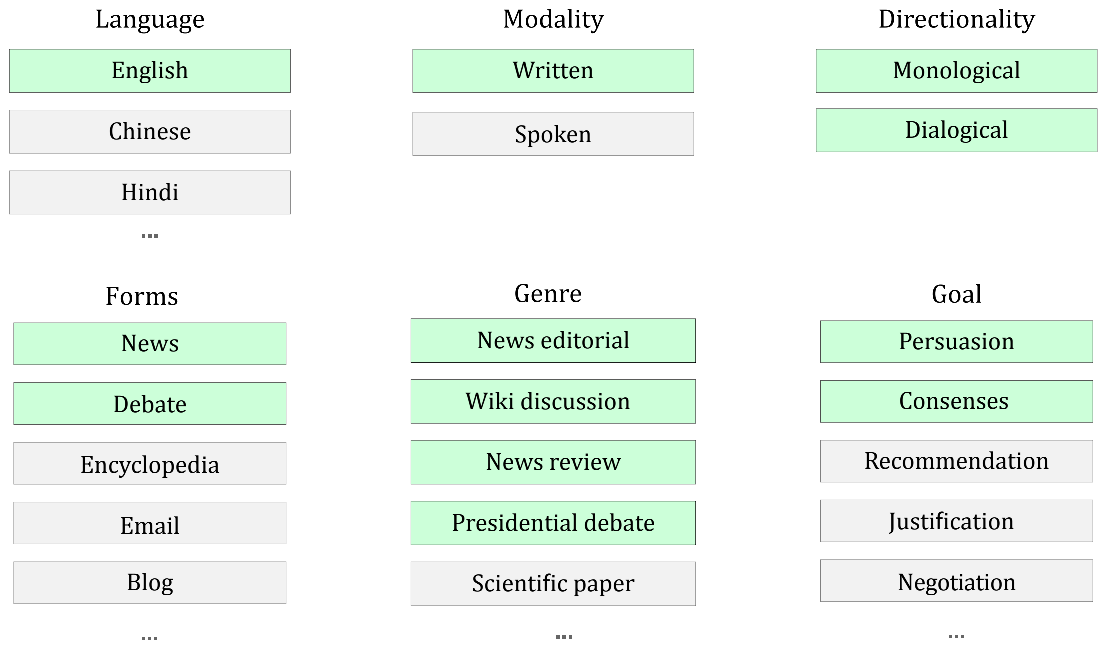


Figure 1.1: Facets of argumentation with examples. The examples which are highlighted are those which we consider in this thesis.

arranging them randomly. However, we assume that an effective strategy, typically, considers the attributes of arguments including dialectical (e.g., argument strength), pragmatic (e.g., argument evidence type), and stylistic (e.g, argument formality) ones.

Principles are usually defined in compliance with the goal of the argumentation and the target audience, considering the text's properties such as the domain and the topic. As a concrete example, the composition of an editorial that addresses abortion can be guided by the following principles:

- "Several arguments that are *acceptable* and *easy to understand* should be used in the *body* of the editorial". (Dialectical attribute)
- "An argument with *anecdotal evidence type* should be selected and used in the *first paragraph* of the editorial". (Pragmatic attribute)
- "A claim that is written as a *rhetorical question* should be selected and used in the *last paragraph* of the editorial". (Stylistic attribute)

Broadly speaking, principles are frequently responsible for producing certain effects on the audience. Persuasion may be achieved, for instance, if an argumentative discourse influences the audience with the demonstration of sound and logical reasons, the evocation of a certain emotion, or the establishment of authority and credibility, i.e., the modes of persuasion according to Aristotle [Aristotle and Roberts, 2004]. Consensus, for example, may be achieved if an argumentative discussion has the impact of setting a foundation for understanding the discussed topic among its participants.

Recently, the development of *computational models* for processing argumentation has attracted considerable attention, as such models yield ample benefits to human communities [Stede and Schneider, 2018]. Computational argumentation models are beneficial for diverse downstream applications such as writing assistants, summarisation systems, fact-checking tools, search engines, and decision-making machines. For instance, one could see the great benefit of using a tool that retrieves a large set of arguments regarding a particular topic, classifies these arguments into 'pro' and 'con' depending on how they relate to the topic, ranks these arguments according to their strength, and finally exposes them with an intuitive interface. Accomplishing these tasks is the ultimate goal of argument search engines such as `args.me` [Wachsmuth *et al.*, 2017b]. Moreover, it would actually be profitable to develop a writing assistant tool that not only reviews the grammar and style consistency of texts, but also delivers suggestions regarding the persuasiveness of texts, for instance. Such a tool could help to make a text more persuasive by recommending the best evidence type to be used at the start of an article, or the most powerful emotion type to be provoked at the end of an article, among others.

Driven by the current advances in artificial intelligence technologies, an emergent area in Natural Language Processing (NLP) that studies the automation of processing argumentation has evolved under the name of *computational argumentation* [Gurevych *et al.*, 2016]. Existing studies in computational argumentation focus on two core tasks: (1) Argumentation mining: distinguishing argumentative units [Ajjour *et al.*, 2017; Al-Khatib *et al.*, 2016a], determining the role of each unit such as premise and conclusion [Stab and Gurevych, 2014b], and finding the relations between their units (support vs. attack) [Peldszus and Stede, 2015]. (2) Argumentation assessment: quantifying the quality of a single argument [Wachsmuth *et al.*, 2017a], scoring argumentative discourses [Persing and Ng, 2015], and predicting the persuasiveness of arguments [Habernal and Gurevych, 2016]. Recent studies target the tasks of argument generation [Hua and Wang, 2018], argument search [Wachsmuth *et al.*, 2017b], and argument question answering [Panchenko *et al.*, 2019].

The thesis in hand introduces and tackles a new task in computational argumentation: *computational analysis of argumentation strategies* in texts. In the light of our view of argumentation strategies, as outlined earlier, this task comprises (1) the identification of arguments attributes in texts, (2) the exploration of the selection and arrangement principles regarding the identified attributes, and (3) the evaluation of the effectiveness of the explored principles at achieving persuasion, consensus, and other goals. More details are given in Chapter 2, Subsection 2.1.3.

The computational analysis of argumentation strategies is substantial for many down-stream applications. It is needed for nearly all kinds of text synthesis, writing assistance, and dialogue management tools. In text synthesis, generating a text following an adequate argumentation strategy may help to persuade readers towards a certain stance. Furthermore, the principles of argumentation strategies could form the essence of writing assistants' suggestions to authors, which would assist in the writing of top-quality texts. With regard to dialogue management, integrating argumentation strategies into deliberative discussions can play a decisive role in supporting discussion participants in reaching a consensus. Analysing strategies is key to building debating machines that are capable of arguing effectively. Such machines are envisioned by companies such as IBM within its 'debater project' [1] or by research organisations such as the German Research Foundation (DFG) within its priority programme on 'Robust Argumentation Machines' [2].

In relation to the computational analysis of strategies, different studies addressed the diverse attributes of an argument or an argumentative discourse that can be considered within strategy principles. The dialectical attributes studied include proposition verifiability [Park and Cardie, 2014], argumentation schemes [Feng and Hirst, 2011], and fallacies [Habernal *et al.*, 2018b]. The pragmatic attributes investigated cover the role of counter-arguments [Wachsmuth *et al.*, 2018b], argument evidence types [Rinott *et al.*, 2015], and many speech acts [Visser *et al.*, 2019], while the stylistic attributes covered include various rhetorical figures such as irony [C. Wallace *et al.*, 2014] and parallelism [Song *et al.*, 2016].

Furthermore, several studies investigated the question of whether and how an argumentative discourse achieves the goal of persuasion or consensus. The text's persuasiveness was examined by analysing Aristotle's modes of persuasion [Hidey *et al.*, 2017] as well as the linguistic and structural characteristics of texts [Tan *et al.*, 2016; Wei *et al.*, 2016]. Most notably in this context, the task of predicting the success of changing someone's view in the subreddit of 'Change My View' was approached in several papers [Duthie *et al.*, 2016; Hidey *et al.*, 2017]. As for standard debates, Cano-Basave and He [2016] explored how effective the semantic framing of arguments is for predicting a speaker's influence. In a similar vein, [Wang *et al.*, 2017] examined the appropriate shift of a debate's topic for predicting the winner of the debate. Recently,

[1] `https://www.research.ibm.com/artificial-intelligence/project-debater`

[2] `http://www.spp-ratio.de/home/`

some studies have considered the role of modelling the target audience in the persuasiveness of text [Durmus and Cardie, 2018; El-Baff *et al.*, 2018; Lukin *et al.*, 2017]. Compared to persuasion, fewer studies have addressed the consensus goal, especially in the NLP community. Most of the developed models, datasets, and methods for consensus in deliberative discussions have aimed to minimise the coordination effort amongst discussion participants. The studies in this direction focused mainly on Wikipedia discussions within talk pages [Ferschke *et al.*, 2012; Kittur *et al.*, 2007; Wang and Cardie, 2014].

Altogether, existing work in computational argumentation deals with argumentation strategies in either an *implicit*, *partial*, or a *superficial* manner. By the term 'implicit', we mean that the focus of the work is not on how to model a strategy, but on how to model the argumentation goal such as persuasion: i.e., the focus is on predicting how persuasive a given text is. By the term 'partial', we imply that the aim of the work is to determine only a specific attribute of an argument, which may contribute to the identification of argumentation strategies, whereas, by the term 'superficial', we indicate that the work may point to some principles of a strategy but that it lacks a detailed analysis of strategies and their connections to the different properties of text such as genre, topic, and author. More details can be found in Chapter 2, Section 2.2.

This thesis aims to overcome the outlined shortcomings in the current state of computational argumentation research. The thesis studies argumentation strategies with the following goals:

1. Modelling a set of dialectical, pragmatic, and stylistic argument attributes whose identification can be tackled within the current state of the art.

2. Operationalising the models through automatic identification of the modelled attributes. This includes building annotated corpora for the attributes and using these corpora to develop robust learning methods for attribute identification.

3. Mining the strategy principles in texts according to the identified attributes, while taking into consideration the different properties of texts such as genres and topics.

1.1 Scope of the Thesis

Drawing on the diverse tasks in computational argumentation, this thesis pays particular attention to the analysis of argumentation strategies in English texts. More specifically, it focuses on analysing strategies in persuasive argumentative discourses that aim for persuasion and in deliberative argumentative discourses that aim for consensus. The ultimate goal of the thesis is to employ such an analysis to supporting the development of writing assistant tools, among others.

Besides narration, description, and exposition, argumentation is one of the main modes of discourse [Braddock, 1963]. Argumentation thus exists in both monological and dialogical texts and it covers a broad range of text genres. In this thesis, we concentrate on a set of widespread and highly influential genres within monological and dialogical texts, namely, editorials, reviews, Wikipedia discussions, and presidential debates. In addition, this thesis addresses different argument attributes that belong mainly to the pragmatic (i.e., the function of a discourse act) and stylistic (i.e, phrasing) categories (see Chapter 2, Section 2.1.2).

Overall, we study argumentation strategies within three major classes:

Pragmatic Persuasive Strategies in Monological Texts Of the various monological genres whose primary goal is persuasion, we investigate news editorials since they allow for exploring diverse sets of strategies. Authors of editorials are known to follow different strategies to persuade their readers along with dynamic and rich argumentative discourses. Furthermore, editorials can generally influence the attitude of human communities by propagating specific ideologies [van Dijk, 1992]. Within our study of editorials, we concentrate on the argument attributes that belong to the pragmatic category. We mainly analyse the types of argumentative discourse units, including the evidence types of 'anecdote', 'statistics', and 'testimony', in addition to the types 'common ground' and 'assumption'.

Pragmatic Deliberative Strategies in Dialogical Texts Regarding dialogical deliberative texts that aim for consensus, we explore the argumentation strategies in Wikipedia discussions, the biggest source of deliberative discourses on the Web, with more than six million discussions. Each Wikipedia article has an associated page called a 'Talk' page. Each talk page comprises a number of discussions that discuss the development of the Wikipedia article. The discussions in the talk pages embody a dynamic environment with a broad spectrum of interactions among Wikipedia users. Such interactions reveal varied strategies that users follow to reach a consensus for a decision regarding the content of the article, such as the merging of two paragraphs. In Wikipedia discussions, we study various argument attributes from the pragmatic category. Two of these attributes concern the argument roles of 'supporting' and 'attacking' another argument. Seven other attributes including ‘recommending an act’, ‘asking a question’, and ‘providing evidence’, are related to dialogue acts, and four are related to the frames of a discussion, such as ‘writing quality’ and ‘neutral point of view’.

Stylistic Persuasive Strategies in Monological and Dialogical Texts As regards the persuasive monological and dialogical texts, we study the argumentation strategies in editorials, newspaper reviews, and presidential debates. The rationale for choosing presidential debates to represent dialogical texts is our assumption regarding their richness of persuasive strategies from, presumably, expert debaters. With regard to the three selected genres, unlike the previous two classes, we focus here on the stylistic category of argument attributes. In particular, we explore 26 syntax-based rhetorical figures such as ‘enumeration’, ‘asyndeton’, and ‘anadiplosis’.

1.2 Research Questions and Contributions

Together with the outlined classes in the previous section, we formulate research questions regarding the analysis of argumentation strategies and develop methods for answering these questions. The research questions and contributions are distributed among the following classes, where each chapter in this thesis targets one class in particular:

Pragmatic Persuasive Strategies in Monological Texts Chapter 3 studies the persuasive argumentation strategies in editorials. In particular, several types of argumentative discourse units are investigated. The investigation involves the question of how arguments with such types are selected and arranged in editorials that are derived from high-quality news portals. The investigation also considers how the selection and arrangement are varied across different topics such as economy and health.

Such an analysis of argumentation strategies in editorials is helpful in various scenarios. Consider for example an author who writes an editorial with a certain property (e.g., the topic is economy), an adequate argumentation strategy can be suggested in order to improve the persuasiveness of her text. In such a manner, new writers can learn how to improve their texts and approach the quality of masterpieces written by top writers.

Overall, within Chapter 3, the following research questions are tackled:

Research Question 1. (Modelling Argument Attributes) How can we select an appropriate set of pragmatic attributes for modelling the strategies in editorials?

Research Question 2. (Identifying Argument Attributes) (a) How can modelled attributes in editorials be effectively identified? (b) How can we build a reliable annotated dataset regarding the attributes in editorials? (c) Which method is effective for identifying the attributes, and how can the identification method be evaluated?

Research Question 3. (Exploring Strategy Principles) (a) How can we explore the selection and arrangement of arguments using the identified attributes? (b) To what degree do the selection and arrangement differ across editorials of different topics?

To begin answering these questions, we first introduce a new model of pragmatic attributes in editorials. This model covers several types of argumentative discourse units such as 'anecdote' and 'testimony'. We then build a new corpus of 300 editorials which are segmented into phrases and manually annotated according to the model. We provide a detailed review of the inter-annotator agreement and the reliability of the resulting annotations. Using the annotated editorials, we analyse the selection and arrangement of the types found there. The selection is determined using the distribution of the types and the arrangement using their sequential flows. The results of the analysis expose various principles for the selection and arrangement of the types across editorials from different news portals. [3]

In addition to the above, we develop a supervised classifier for automatically identifying the argument types in an editorial. The classifier is trained and evaluated using the annotated corpus, achieving a score of 0.77 in terms of the F_1-measure. We apply this classifier to a collection of about 29,000 editorials that we automatically classify by topic (e.g., economy, sport). Analysing the selection and the arrangement of the identified types in the 29,000 editorials on 12 different topics, we reveal substantial deviations in the distribution of types across topics. Furthermore, we differentiate various common structural flows of the types in editorials. This outcome affords valuable insights into what principles of argumentation strategies are present in editorials across different topics.

Pragmatic Deliberative Strategies in Dialogical Texts Chapter 4 addresses the question of how the argumentation strategies of participants in

[3] The corpus has been made freely available to encourage further research on computational argumentation.

deliberative discussions can be assisted computationally. Specifically, we take into account the following scenario: In an ongoing discussion, every participant should understand the current state of the discussion and come up with the move that *best* serves the discussion. This demands substantial effort and time from the newcomer, especially when the discussion expands with several disputes and back-and-forth arguments. For this context, we propose to support the development of a tool that is able to *recommend the best move* based on an effective argumentation strategy.

In Chapter 4, we approach two research questions:

Research Question 4. (Modelling Argument Attributes) How can deliberative discussions be modelled using an abstract and representative set of pragmatic attributes?

Research Question 5. (Identifying Argument Attributes) (a) Is it possible to identify the modelled attributes with reasonable effectiveness? (b) How can we build a large-scale annotated corpus for the attributes? (c) What methods are effective for identifying the attributes?

To answer these questions, we introduce a new model of argument attributes in deliberative discussions. While there have been previous models for such discussions, these models were derived manually by inspecting a small set of discussions. This, in turn, results in models with a level of abstraction that is not appropriate for the recommendation of best move. By contrast, our model is derived statistically from various types of metadata written by the participants in 6 million discussions in Wikipedia talk pages. The proposed model comprises the attributes of argument roles such as ‘support’ and ‘attack’, dialogue acts such as ‘asking a question’, and frames such as ‘writing quality’.

Based on the model, we automatically generate a large-scale corpus of about 200,000 discussions’ turns which are labelled according to the modelled attributes. The automatic generation is performed using a distant supervision method based on the metadata of the discussions. The resulted corpus is one of the largest for deliberative discussions. The model is operationalised by three supervised learning classifiers: one for identifying the argument roles, one for the dialogue acts, and one for the frames. The three classifiers are trained and evaluated using the generated corpus. As for the effectiveness of the classifiers, the results of the evaluation experiments reveal that the classifiers achieve scores between 0.71 and 0.13 in terms of the F_1-measure. Given such a high variance in the effectiveness of the classifiers, it appears that they need further improvement before they can be used for reliable analysis of the strategy principles in Wikipedia discussions.

Stylistic Persuasive Strategies in Monological and Dialogical Texts Chapter 5 discusses the analysis of stylistic strategies in persuasive editorials, newspaper reviews, and presidential debates. The analysis is conducted based on identifying a set of stylistic attributes and exploring their usage. In particular, we decide to deal with syntax-based rhetorical figures that can be effectively approached by the available technologies.

This analysis is intended to afford a mechanism to develop controlled text generation tools. Such tools are able to determine which, where, and how rhetorical figures should be manifested in generated texts in order to boost their persuasiveness.

In Chapter 5, three research questions are posed:

Research Question 6. (Modelling Argument Attributes) How can we model stylistic attributes in editorials, reviews, and presidential debates?

Research Question 7. (Identifying Argument Attributes) (a) How can we identify the modelled attributes in monological and dialogical texts? (b) How can we create a high quality annotated dataset of those attributes? (c) How can we approach the identification task with an appropriate methodology, and how can we evaluate such a methodology?

Research Question 8. (Exploring Strategy Principles) (a) What are the commonest principles of stylistic attributes in persuasive texts? To what degree do these patterns differ across monological and dialogical texts, within and across text genres, topics, and authors, and across different debaters?

We answer these questions by first defining a model of 26 syntax-based rhetorical figures grouped into four types: balance, inversion, omission, and repetition. Next, we develop a grammar-based method for classifying the 26 figures. The outputs of a probabilistic context-free grammar parser are employed to create rules for figure identification. The evaluation of the rules is conducted on a newly-built corpus. This corpus comprises a collection of 1718 examples of the 26 figures from a set of trustworthy sources on the Web, which has been developed by experts in rhetoric. The results of evaluating the identification method using the created corpus show an effectiveness of 0.70 in terms of the F_1-measure.

The proposed grammar-based method is used to analyse the distribution of the 26 figures within and across the genres of editorials and reviews; the topics of art, education, and science; and different authors. Furthermore, we explore the distribution of the figures in a set of presidential debates from the American presidency project, and especially the US election debates between Donald Trump and Hilary Clinton. The distributions reveal several insights into the strategy principles used in persuasive monological and dialogical texts, while being adequate for integration into text synthesis tools.

1.3 Publication Record

Most of the chapters in this thesis are based on one or more peer-reviewed publications from top conference venues, as outlined in Table 1.1. More specifically, Chapter 3 is written based on three publications at the conferences of COLING, EMNLP, and EACL respectively. Chapter 4 relies on one publication at the ACL conference, while Chapter 5 represents an in-progress work that is not yet published. Several other publications related to NLP (listed at the bottom of the table) can not be assigned to a specific chapter. However, valuable insights from these publications are employed implicitly in the content of the thesis chapters, especially Chapter 2. The reason for not using all of the publications is to maintain the focus on the promising topic of the analysis of argumentation strategies.

1.4 Thesis Structure

This thesis is organised as follows: Chapter 2 provides an overview of the background and related work on argumentation strategies. Chapter 3 proposes an analysis of the selection and arrangement of several types of argumentative discourse units in editorials. The analysis distinguishes principles of argumentation strategies across news-portals and topics. In Chapter 4, we model deliberative strategies in Wikipedia discussions and operationalise the model by identifying argument roles, dialogue acts, and frames. Chapter 5 addresses the identification of syntax-based rhetorical figures and the derivation of strategy principles in news editorials and presidential debates across different genres, topics, and authors. Chapter 6 summarises the thesis and proposes directions and research questions for future work.

Table 1.1: A selection of peer-reviewed publications by the author and their usage within this dissertation.

Used in	Venue	Type	Pages	Year	Publisher
Chap. 3	COLING	conference	9	2016	ACL
	***Khalid Al-Khatib**, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A News Editorial Corpus for Mining Argumentation Strategies.*				
Chap. 3	EMNLP	conference	6	2017	SIGDAT
	***Khalid Al-Khatib**, Henning Wachsmuth, Matthias Hagen, and Benno Stein. Patterns of Argumentation Strategies across Topics.*				
Chap. 3	EACL	demo	6	2017	ACL
	*Johannes Kiesel, Henning Wachsmuth, **Khatib Al-Khatib**, and Benno Stein. WAT-SL: A Customizable Web Annotation Tool for Segment Labeling.*				
Chap. 4	ACL	conference	10	2018	ACL
	***Khalid Al-Khatib**, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen and Benno Stein.Modeling Deliberative Argumentation Strategies on Wikipedia.*				
Chap. 5	Under submission	conference	10	2019	ACL
	***Khalid Al-Khatib**, Viorel Morary and Benno Stein. Style Analysis by means of Mining Rhetorical Devices.*				
–	INLG	conference	10	2018	ACL
	*Wei-Fan Chen, Henning Wachsmuth, **Khalid Al-Khatib**, and Benno Stein. . Argumentation Synthesis following Rhetorical Strategies.*				
–	CoNLL	conference	10	2018	ACL
	*Roxanne El Baff, Henning Wachsmuth, **Khalid Al-Khatib**, and Benno Stein. Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus.*				
–	ArgMining	workshop	10	2018	ACL
	*Henning Wachsmuth, Martin Potthast, **Khalid Al-Khatib**, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an Argument Search Engine for the Web.*				
–	COLING	conference	9	2017	ACL
	*Henning Wachsmuth, Manfred Stede, Roxanne El Baff, **Khalid Al-Khatib**, Maria Skeppstedt and Benno Stein. Argumentation Synthesis following Rhetorical Strategies.*				
–	NAACL	conference	10	2016	ACL
	***Khalid Al-Khatib**, Henning Wachsmuth, Matthias Hagen, Jonas Kohler, and Benno Stein. Cross-Domain Mining of Argumentative Text through Distant Supervision.*				
–	COLING	conference	10	2016	ACL
	*Henning Wachsmuth, **Khalid Al-Khatib**, and Benno Stein. Using Argument Mining to Assess the Argumentation Quality of Essays.*				
–	ArgMining	workshop	10	2015	ACL
	*Johannes Kiesel, **Khalid Al-Khatib**, Matthias Hagen, and Benno Stein. A Shared Task on Argumentation Mining in Newspaper Editorials.*				
–	COLING	conference	10	2012	ACL
	***Khalid Al-Khatib**, Hinrich Schutze, and Cathleen Kantner. Automatic Detection of Point of View Differences in Wikipedia.*				

Chapter 2

Background and Related Work

This chapter introduces the background and the related work on the analysis of argumentation strategies. Section 2.1 describes argumentative discourses along with their goals and directionalities. Furthermore, it proposes our view on argumentation strategies defining their elements as well as their formulation and evaluation processes. Section 2.2 concisely reviews the literature on computational argumentation strategies including the existing studies regarding argumentative genres, argument attributes, and argumentation goals.

2.1 Analysis of Argumentation Strategies

In this subsection, we briefly give an overview of argumentation and introduce the argumentative discourses including their main goals and directionalities. Later, we present our view on argumentation strategies, addressing their central elements and their formulation and evaluation. Before this, we explicate how our view of argumentation strategies is closely associated with key theories in argumentation: Aristotle's modes of persuasion and canons of rhetoric, the pragma-dialectical theory, and strategic manoeuvring.

2.1.1 On Argumentation

On a daily basis, people argue in order to derive reasonable conclusions. Such conclusions rule the set of people's beliefs, opinions, and decisions. It is for this reason that argumentation has been a subject of investigation for decades. The concept of argumentation probably emerged in Ancient Greek times, when philosophers and rhetoricians established the fundamental concepts related to logic and rhetoric. Since that time, argumentation has been studied by scholars from various disciplines. Within the field of philosophy and logic, reasoning, standards of proof, and fallacies have been explored intensely [Woods *et al.*, 2004]. Moreover, in the communication field, numerous theories have been developed to address the practical usage of argumentation with real examples from daily life [Maillat and Oswald, 2013]. Argumentation has also been studied in psychology, with investigations of questions such as how arguments are comprehended and generated in relation to age and educational level [F. Voss and Van Dyke, 2001]. The role of language in argumentation has also been explored in linguistics [Oswald *et al.*, 2018].

Over recent years, computer scientists have become actively engaged in studying argumentation. Advances in artificial intelligence technologies along with the contribution of interdisciplinary research are currently employed towards promoting robust argumentation systems. From the broad areas of artificial intelligence, the NLP community has paid remarkable attention to argumentation. The new area of computational argumentation in NLP investigates automatic methods for understanding how people argue in natural language, so addressing the linguistic side of argumentation.

As the thesis at hand belongs in the field of computational argumentation, we deal with argumentation from the communication perspective, where *language* is the primary tool of communication. We view argumentation as one of the four major modes of discourse [Stede and Schneider, 2018]. Basically, 'discourse' is a broad term with diverse definitions. Among these definitions, discourse can be viewed as a language-based communication tool. Put simply, discourse is a coherent language above the sentence level (more than one sentence), with a specific communication purpose [Bublitz *et al.*, 2012]. From the argumentation mode of discourse, we distinguish the term 'argumentative discourse'. Such a discourse goes beyond stating 'abstract' arguments by encoding how arguments, along with other textual units, are utilized for social interactions [Ellis, 2008].

Argumentative discourse, according to the pragma-dialectical theory, among others, is dialectical [Schwarz and Asterhan, 2010]. Implicitly or explicitly, two opposing stances for a certain issue contest each other. For example, a typical editorial for 'banning abortion' not only argues for its stance (e.g., pro), but also addresses the opposing stance using counter-arguments. Argumentative discourses can be observed from diverse sources, various genres and registers, and within monologues or dialogues. Moreover, they can be manifested in either spoken or written modalities, targeting diverse communication goals. In the following paragraphs, we elaborate further on two aspects of argumentative discourse that intensely influence our work: directionality and goal.

Argumentative discourses are typically presented using one of two directionalities: monologue and dialogue. In a monologue, the source (e.g., an author) uses the language (e.g., writes) to make a product (e.g., an editorial), which is perceived (e.g., read) by the target (e.g., a reader). Here, the argumentation is developed by the source and then delivered to the target, making the flow of information unidirectional. By comparison, in a dialogue, participants switch between being the source and the target during the discussion, and the flow of information in dialogues is thereby bidirectional.

An argumentative discourse, in light of its communication function, should target a particular goal. More specifically, the discourse might aim at persuading the audience, reaching a consensus for a collective decision, seeking information, or negotiating for an advantageous settlement, among others. Persuasion is one of the most studied goals, with various theoretical and computational models having been developed for the purpose. The rewards of influencing people's beliefs, stances, behaviours, or attitudes might be tremendous. Consensus, on the other hand, is of equal importance to persuasion. Usually, people need to reach a consensus regarding the best action to take when they tackle a problem or perform a task.

The directionality and goal of a discourse are closely connected. In the context of argumentation, persuasion can be approached in a monologue or a dialogue, while consensus is mainly approached in dialogues, and particularly, in deliberative discussions. Furthermore, a discourse that aims for consensus is different from one that aims for persuasion from various angles. In a deliberative discourse, the discussion participants share their knowledge, arguments, and preferences, while attempting to objectively address the pros and cons of each possible action, in order to reach a consensus. Contrarily, in a persuasive debate, debaters might subdue high-quality arguments in case these arguments counter their stances on the debated topic.

2.1.2 Argumentation Strategies

Delving into the rich legacy of argumentation theories, and in an attempt to underpin a view of argumentation strategies based on that respectable legacy, we visit the following influential theories: Aristotle's modes of persuasion and canons of rhetoric [Aristotle, 2007], the pragma-dialectical theory [Eemeren and Grootendorst, 1987; van Eemeren and Grootendorst, 2004], and strategic manouevring [van Eemeren and Houtlosser].

Aristotle discussed many years ago the question of how to persuade people, proposing four fundamental modes of persuasion: logos, ethos, pathos, and kairos. According to Aristotle, to persuade, logical arguments should be delivered (logos), the credibility of the speaker should be demonstrated (ethos), and specific emotions in the target audience should be evoked (pathos). In addition, the right time and place should be exploited (kairos). The kairos mode, though, attracts less attention compared to the other modes.

In addition to the four modes of persuasion, Aristotle proposed three genres of rhetoric (deliberative, forensic, and epideictic), some rhetorical topics, and the five canons of rhetoric. The latter play a key role in establishing a guide for producing an impressive and persuasive speech. The five canons are:

1. Inventio (invention): selecting arguments that suit the audience, acknowledge the powerful evidence, and follow the best available modes of persuasion. This canon is of great significance, since it acts as the basis for the subsequent canons.
2. Dispositio (arrangement): arranging arguments to raise the chance of persuasion. Basically, this canon deals with the organization of the arguments which are chosen in the former canon.
3. Elocutio (style): selecting the appropriate style for delivering the arguments. The chosen style should make the argumentative discourse clear and interesting for the audience.
4. Memoria (memory): memorising the speech and making it memorable.
5. Actio (delivery): using body language and adjusting the tone while giving a speech.

While the first three canons deal with both written and spoken discourses, the last two canons consider only the spoken type.

On the other hand, moving a large step forward through time, a relatively recent theory that deals with argumentation is the pragma-dialectical theory [Eemeren and Grootendorst, 1987]. Pragma-dialectical theory studies the ideal way to construct an argumentative discourse in order to resolve a conflict of opinions. In accordance with the theory, an optimal argumentative discourse should be approached from two angles: (1) a dialectal one, which concerns the reasonableness of a discourse, where the discourse should be guided by a set of rules, referred to as critical standards, and (2) a pragmatic one that concerns the functional moves in a discourse. Specifically, the theory covers the resolution process (of a conflict) and the different types of speech acts that are appropriate within certain stages of the discourse.

Gradually, van Eemeren and Grootendorst [2004]; van Eemeren and Houtlosser revisited their theory and clearly pointed to the importance of accounting for rhetorical effectiveness along with dialectal reasonableness. In particular, van Eemeren and Houtlosser proposed the concept of ‘strategic manoeuvring’ and defined it as making a choice from a space of potential topical options related to the issues discussed. The choice is made by selecting, what van Eemeren and Houtlosser call responsive adaptation to audience demand using appropriate presentational devices. The easiest to handle, the highly acceptable, and the carefully phrased arguments are concrete examples of such choices. Overall, the strategic manoeuvres that bring together the presentational devices and the audience demand in regard to 'topical potential' are the base of an effective 'rhetorical strategy', van Eemeren and Houtlosser stated.

Through a careful interpretation of the outlined theories, we observe that these theories imply different notions, which we rely on to further clarify argumentation strategies. The remaining texts in this subsection examine these notions in detail.

A strategy can be defined as a high-level plan to achieve a specific goal under the condition of uncertainty [Kvint, 2009]. Typically, a strategy is developed by formulating a set of *principles* (i.e., rules) that regulate how to use the possible *means* to reach the goal. The means could be the available resources, successful techniques, or any kind of actions, to name a few.

In the context of argumentation, and in close parallel with the above definition, we see argumentation strategy as a set of *principles* that govern the selection and arrangement of available arguments and contextual information. The principles regulate the selection and arrangement on the basis of arguments (plus contextual information) *attributes*. These attributes can be grouped into dialectical, pragmatic, and stylistic categories. In principle, the strategy is formulated to deliver an argumentative discourse that can achieve persuasion, consensus, or another goal within the target audience.

To clear up the essential elements of this definition, we elaborate on the argument attributes and strategy principles, after which, we discuss the formulation as well as the implementation and evaluation of a strategy.

Argument Attributes We consider argument attributes as the primary bases for strategy principles, and we distinguish three categories of them:

1. Dialectical attributes: Aristotle pointed to this category through the 'invention' canon of rhetoric. Likewise, the pragma-dialectical theory, notably, introduced the *dialectical* dimension of reasoning. This dimension demonstrates the presence of different opinions that have to be resolved by establishing a well-maintained reasoning process. One can consider various argument attributes that fall under the dialectical category. For example, easy to understand, sound, agreeable, and valid are attributes of arguments. If such attributes are employed properly in a discourse,

they may increase its approval and thereby lead to accomplishing its goal. Broadly speaking, this category deals with the logical attributes of argument(s) and of the reasoning process in argumentative discourses.

2. Pragmatic attributes: The pragma-dialectical theory described the *pragmatic* dimension of reasoning and modelled it based on the speech act theory [Searle, 1969]. According to this theory, a speech act is a technique of using utterance that serves a certain function in communication. Such a technique is supposed to express an intention, specify a function, or produce an effect on the target readers. Various speech acts can be seen as argument attributes. For example, writing a statement with the function of stating a fact, asking a question, or clarifying a misunderstanding can all be influential attributes in deliberative discussions. Moreover, writing statements with the function of providing expert evidence or countering an argument are powerful attributes that may develop the author's credibility in a persuasive discourse. Overall, this category covers the attributes related to the functions of stating an argument, which are usually represented as speech acts.

3. Stylistic attributes: Aristotle recognised 'style' as one of the rhetorical canons, with a clear emphasis on its role in persuasion. The strategic manoeuvrings theory also addressed the presentational devices in argumentative discourses, with an emphasis on the need for applying rhetorical techniques and using effective wording. An evident example here is repetition, which Aristotle described as key for persuasion. Repetition may evoke specific emotions in readers, leading to improvement in the chances of persuading them. In summary, this category considers the attributes related to the techniques of phrasing texts.

Besides the three categories discussed above, the discourse itself might have what we call logistical attributes. These attributes emerge within the kairos mode of persuasion following Aristotle's thoughts about clinching the right moment and place for delivering a discourse. Such attributes include where to publish a discourse; in the New York Times or on Fox News, for instance. Moreover, they involve the time of publishing. As an example, publishing an editorial about gun control after a school shooting incident is likely to make the target readers more emotional towards the topic, which probably makes the editorial more persuasive. Furthermore, a logistical attribute can be a visual aspect, such as including a graphic in a scientific publication.

Strategy Principles The strategy principles guild the method of composing arguments to maximise the probability of achieving the goal of the discourse.

Such principles can be drafted by applying two operators on the argument attributes: the *selection* and the *arrangement*. The operators are drawn from the canons of rhetoric, pragma-dialectical, and strategic manoeuvring theories. More particularly, the 'invention' canon of rhetoric talked explicitly about the 'selection' of arguments; and furthermore, the 'selection' operator is encoded in the definition of strategic manoeuvring, wherein van Eemeren and Houtlosser stressed that the choice from the topic potential is made by 'selecting' a responsive adaptation to audience demand. On the other hand, the 'arrangement' operator is discussed in the 'arrangement' canon of rhetoric. The pragma-dialectical theory also studied diverse types of speech acts and described the appropriate usage of these acts in connection with the 'stages' of a discourse.

The 'selection' operator considers which argument attributes should be considered in the discourse, and which should not. On many occasions, arguments with specific attributes should not be used in the same discourse. The 'arrangement' operator considers how to order the arguments in a discourse based on their attributes. This operator can be applied at different levels of a discourse's structure: sentence, paragraph, and lead-body-end.

Technically, principles can be delineated in several ways. An example of principles is that "an argument that provides statistical evidence should be used frequently in the experiment section of a scientific publication in biology". Another example is the principle of "using at least one argument with the function of counter-attack in the second paragraph of an argumentative essay". Undoubtedly, principles also concern the argument attributes that should not be used. For example, "a text with an ironic rhetorical figure should be avoided in deliberative discussions". Note that the principles can be more sophisticated if combinations of argument attributes and their arrangements are considered. For example, "arguments that provide anecdotal evidence should not be used in the same paragraph with other arguments that provide statistical evidence in editorials about religion".

The properties of a discourse play a considerable role in outlining principles, since they might restrict the selection and arrangement of argument attributes. For example, it is not adequate to use 'rhetorical questions' in scientific publications. Furthermore, persuasive essays have a standard structure that includes, for instance, stating the major claim in the essay's lead. Primarily, the properties of a discourse include the register, genre, and topic of the text, in addition to the directionality of the discourse.

In reality, it is necessary to understand that principles should, first and foremost, be devoted to the discourse's goal. Such devotion is usually reflected in specific effects that the principles bring forth on the target readers. This is certainly observed in Aristotle's modes (of persuasion), which are intended to strengthen the credibility of a writer in the eyes of their target readers, or to evoke the feelings of sadness, anger, or fear, among others, in the readers. On the whole, strategy principles are responsible for producing effects on the target readers in order to attain the goal of the discourse.

Principles in persuasive discourses are usually responsible for effects related to logos, ethos, and pathos. On the other hand, principles in deliberative discourses are typically responsible for effects related to enhancing the collaboration between participants in the discussion and to improving the mechanisms for sharing information among them. Various other effects are nevertheless highly effective in both persuasive and deliberative discourses, especially those related to logos.

Strategy Formulation In basic terms, the principles are the essence of a strategy. In practice, however, coming up with high-quality principles is a highly challenging task. This will be demonstrated in the following discussion about the outlining of principles.

First, a strategy is formulated for achieving a particular goal with regard to the 'target readers'. van Eemeren and Houtlosser bonded strategic manoeuvring to the 'audience demand', implying the need for modelling (aka understanding) the target of a discourse. In simple words, the way in which the target readers perceive the discourse impacts the discourse's chances of reaching its goal. However, the readers' perception is influenced by various aspects beyond the quality of the discourse, such as readers' prior beliefs, values, attitudes, and personality traits, while not overlooking the dilemma of human bias.

Secondly, strategies are either individual or group-based. In a monological text, an author formulates a strategy by encoding effects in their text aiming for persuasion. A reader, on the other hand, decodes the effects, aiming to form a stance or make a decision. This is considered as an individual-based strategy, as the author is the only one who is responsible for formulating and implementing the strategy. Typically, the interacting between the author and the readers is minimal in this type of strategy. Figure 2.1 illustrates a conceptual model of an individual-based strategy. By way of contrast, a participant in a deliberative discussion reads and comprehends the topic of the discussion as well as all previous turns, and strives to come up with the best moves that help to reach an agreement regarding the topic under discussion. In light of this scenario, the participants in the discussion formulate their strategy as a group. To state it differently, the strategy is formulated by the group while the group itself is the target reader. Accordingly, the group-based strategy usually pays close

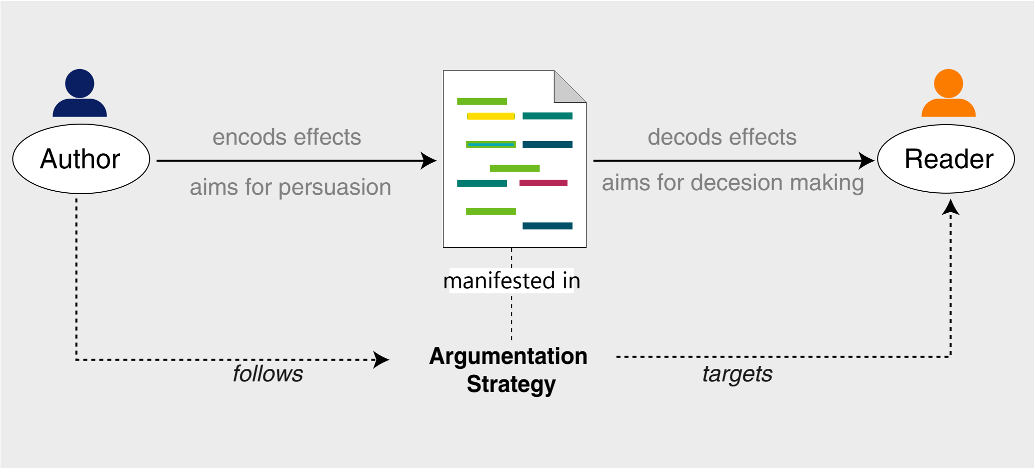


Figure 2.1: Conceptual model of an individual-based strategy. In a monological argumentative text, to persuade a target reader, an author intentionally follows an argumentation strategy, encoding the effects he or she thinks to be the best for influencing the target readers. A reader, on the other hand, decodes the effects, usually subconsciously, and tries to employ such effects to form an opinion or to make a decision.

attention to the coordination between the participants. Figure 2.2 portrays a conceptual model of a group-based strategy.

Thirdly, the strategy can be formulated based on top-down or bottom-up settings. In the top-down setting, the author of a monological text starts by determining the most powerful effects that the author aims to produce on the target readers. That is followed by outlining several principles that yield the chosen effects. For example, the author of an editorial about 'gun control' might decide to focus on provoking the emotions of fear and sadness in the target readers. The strategy's principles are then outlined according to the selected emotions. Alternatively, in the bottom-up setting, the author starts with observing the available content (arguments plus contextual information) regarding the topic of the discourse. The principles are then outlined in accordance with the content and considering the production of various effects. For instance, the author of an editorial about 'abortion' might start by collecting arguments regarding the topic, and then outline the principles based on the attributes of the collected argument, selecting those that are sound and avoid fallacies, for example.

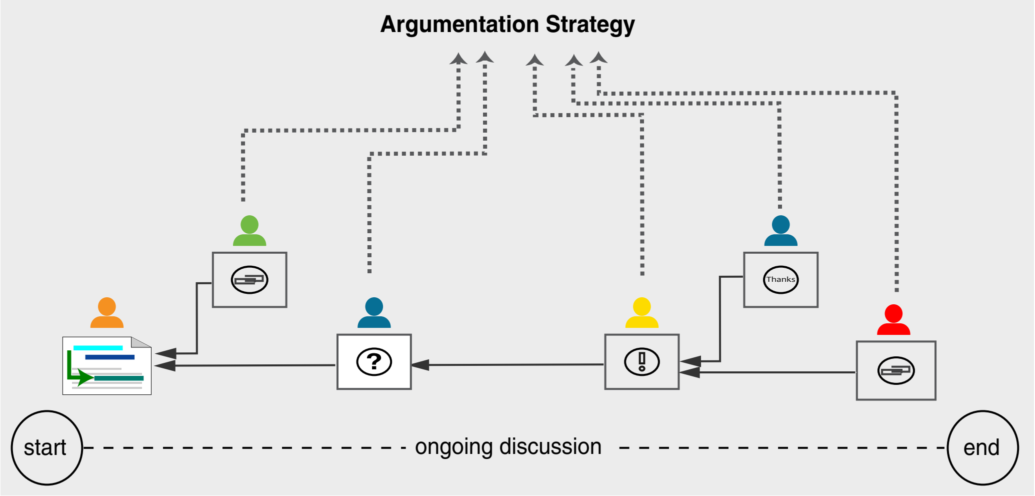


Figure 2.2: In a dialogical argumentative text, particularly in deliberative discussions, the discussion starts with a proposal regarding a particular action, decisions, or similar. Later, the discussion participants develop an argumentation strategy by coming up with the best deliberative moves that would lead to reaching a consensus on whether the proposal should be accepted or not.

Strategy Implementation and Evaluation The implementation of a strategy simply entails putting it into action. During the writing of an editorial, for example, an author follows a strategy that s/he formulates by applying the most important principles to the text. This means that the available arguments will be selected and arranged following the principles. After all, it is important to keep in mind that the effectiveness of a strategy is not known in advance, and there is no guarantee that applying the principles will lead to achieving the goal. This complies with the strategy definition, as strategies lie beneath the umbrella of uncertainty. For example, when an author writes an editorial, s/he is not 100% sure about the ideal selection and arrangement of arguments that will necessarily lead to persuading the target readers. Rather, an author utilises his or her knowledge and personal experience to come up with the principles that s/he thinks will be the best for achieving the goal.

The clear-cut method for evaluating a strategy is to observe whether and to what extent the discourse that the strategy was applied to achieved its goal. However, this means of evaluation is not straightforward as, usually, there is no easy way to get direct and comprehensive feedback from the readers.

Within the field of computational argumentation, quality assessment of argumentation is somehow related to the evaluation of argumentation strategies. Argumentation quality assessment is studied in a large collection of papers addressing various dimensions of argument and argumentation quality [Wachsmuth *et al.*, 2017a]. Wachsmuth *et al.* [2017a] proposed a new taxonomy that represents a unified view of studied quality dimensions. This taxonomy categorises many quality dimensions into three quality aspects: logical cogency, rhetorical effectiveness, and dialectical reasonableness. Regarding logical cogency, an argument is cogent if its premises are acceptable, relevant to the conclusion, and sufficient to support it. The cogency aspect is divided into local acceptability, local relevance, and local sufficiency. Rhetorical effectiveness mainly concerns whether an argumentative discourse is persuasive. Besides the credibility (ethos) and emotional appeal (pathos), these aspects cover the clarity, appropriateness, and the arrangement dimensions. As regards dialectical reasonableness, reasonable argumentation seeks an ultimate conclusion, provides acceptable arguments, and considers counter-arguments regarding the opposite view. This aspect is divided into global acceptability, global relevance, and global sufficiency.

Taking a closer look at this taxonomy, we observe that many of the dimensions described therein are actually effects produced by some strategy principles. For example, the local acceptability in the logical cogency aspect is the effect of using sound argument. With regard to rhetorical effectiveness, credibility can be seen as the effect of providing expert evidence. Lastly, for dialectical reasonableness, global sufficiency is the effect of providing counter-arguments.

2.1.3 Computational Argumentation Strategies

Based on our view of the argumentation strategies discussed above, the computational analysis of such strategies should consider three core steps:

1. (Identification of Argument Attributes) Identification of a selected set of argument attributes.
2. (Mining of Principles) Mining of the selection and arrangement principles for the identified attributes.
3. (Evaluation of Principles) Evaluation of the effectiveness of the mined principles.

Figure 2.3 illustrates the steps mentioned above for the analysis of argumentation strategies.

In the first step, conceptual models for the argument attributes should be designed. These models may build schemes of the attributes (e.g., taxonomy), depict the possible overlaps between the attributes, and define the granularity of the text in which the argument attributes are encoded (e.g., sentence-level).

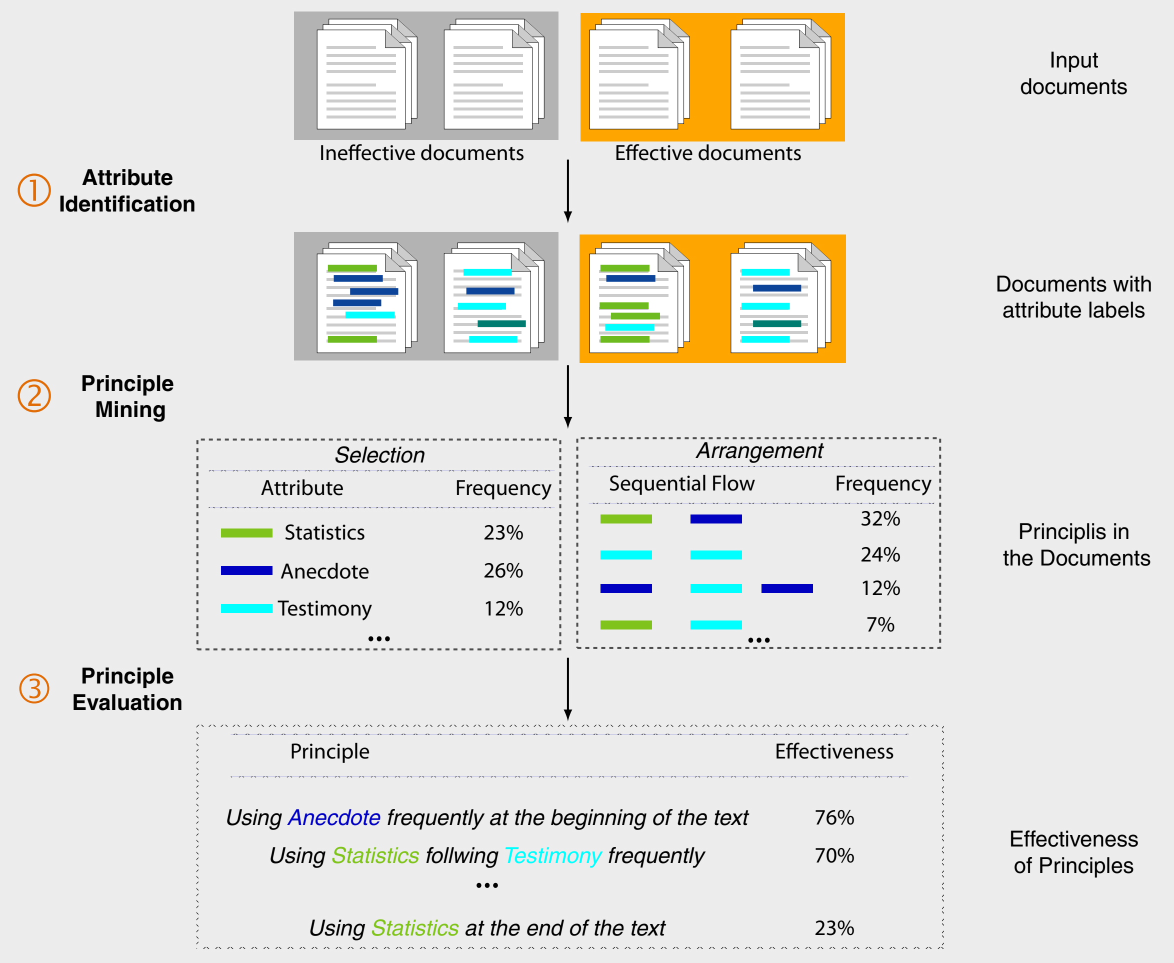

Ineffective documents
Effective documents
Input documents
① Attribute Identification
Documents with attribute labels
② Principle Mining
Selection
Attribute
Frequency
Statistics 23%
Anecdote 26%
Testimony 12%
•••
Arrangement
Sequential Flow
Frequency
32%
24%
12%
7%
•••
Principlis in the Documents
③ Principle Evaluation
Principle
Effectiveness
Using Anecdote frequently at the beginning of the text 76%
Using Statistics follwing Testimony frequently 70%
•••
Using Statistics at the end of the text 23%
Effectiveness of Principles

Figure 2.3: An overview of the main steps for the analysis of argumentation strategies.

The conceptual models are then employed for developing computational models for automatic identification of the attributes. Since supervised machine-learning models dominate the computational models in NLP, annotated datasets according to the conceptual models are constructed (or used, if already available). One part of the used dataset is utilised for training the supervised models, while the remaining part is used for evaluating them.

In the second step, the computational models for identifying the attributes (i.e., the output of the first step) can be applied to the collection of texts to be analysed. Various methods can be adopted for investigating the selection and arrangement of the identified attributes in the analysed texts. Following the frequent patterns theory proposed by Han *et al.* [2012], the selection can be modelled using the itemset pattern. This means that each identified attribute, along with its frequency in the texts, is detected. Arrangement, on the other hand, can be modelled using the sequential or structural patterns. Sequential patterns concern the order in which the identified attributes are used within the text discourse. The structural patterns can be more complicated, accounting for the relations between the attributes in the structure of trees, graphs, and the like.

In the third step, the patterns of the selection and the arrangement of the attributes (the output of the second step) are examined with respect to their effectiveness. Specifically, the patterns will be evaluated, examining their correlation to those in the discourses that achieved their goal compared to the discourses that did not. For example, a particular pattern might be observed frequently in effective discourses while being absent in ineffective ones. This step may involve ranking the patterns based on their effectiveness, inspecting conflicting patterns, and discovering connected patterns (those that occur together frequently).

Such an analysis should take into account the text's properties, the discourse's goals, and the target readers. This can be fulfilled by grouping the analysed texts based on a single property or goal. For example, the analysis can be carried out on a collection of persuasive editorials that belong to a single topic. Nevertheless, it is expected that general effective principles will be found (those that can be found in texts with diverse properties and goals). In view of this, it might be beneficial to also analyse heterogeneous texts.

The automatic strategy analysis is an essential step for automatic strategy formulation and implementation. The strategy formulation can be made based on the principles revealed in the analysis step. Put simply, effective principles should be respected, while the least effective ones should be avoided. To implement a strategy, attributes should be identified in the available arguments, and the principles regarding these attributes should be applied in the discourse.

How many principles can be mined from a collection of texts? How many principles should a strategy have? And how many effective strategies can be discovered? These are still open research questions. The intuitive way of answering these questions is to examine the data (i.e., using a data-driven approach). Apparently, the answers to these questions will be based on several determinants such as the texts' properties, the size of the analysed texts, and the number of recognised attributes, among others.

Manual strategy formulation is different from its automatic counterpart from different angles. Firstly, when strategy principles are drawn up manually, it is not expected that an author outlines a large number of principles, since this is very time- and effort-consuming. Automatic formulation, in contrast, can consider a large number of attributes and output many principles. Moreover, the manual strategy is usually formulated based on intuition and personal experience, which is often limited. Automatic formulation, on the other hand, can be done based on the analysis of big data, which allows for the adoption of a collective experience (the experience of all the people who produce the data). Furthermore, manual formulation of a strategy is generally restricted to the genre or the topic that suits the author's expertise. In contrast, automatic strategy formulation can be conducted for any genre or topic, as long as corresponding data is available.

Combining manual and automatic strategy formulation is still possible, and may even be desirable. Imagine if an author decides to write an editorial, and an intelligent writing assistant can help by suggesting various principles ranked by their effectiveness. The author can then use his/her expertise to select, adjust, or ignore some of the principles.

2.2 Related Work

Over the last few years, computational argumentation has gained considerable popularity in the NLP community. Computational argumentation contributes to the essential building blocks of various applications such as automated decision making [Bench-Capon *et al.*, 2009] and argument search engines [Cabrio and Villata, 2012]. At the time of writing this thesis, the sixth Annual Workshop of Argumentation Mining is taking place at the annual meeting of the Association for Computational Linguistics (ACL).

Computational argumentation encompasses a wide range of tasks such as argument mining and argument quality assessment. The task of argumentation strategy analysis, the topic of this thesis, was introduced in our work on building a corpus for argumentation strategies in editorials [Al-Khatib *et al.*, 2016b]. This work has been followed by several publications dealing with diverse aspects of argumentation strategies [Al-Khatib *et al.*, 2017a, 2018; Wachsmuth *et al.*, 2018a]. Besides our work, different computational argumentation approaches partly consider some elements of argumentation strategies, especially those related to the identification of argument attributes and the modelling of argumentation goals.

In this section, firstly, we report on some studies regarding the main argumentative genres. Next, we describe different approaches regarding argument attributes, and finally, we discuss some studies that target the goals of persuasion or consensus.

2.2.1 Genres of Argumentation

In this subsection, we briefly review some widespread genres of argumentation. Among the monological texts, we report on persuasive essays, editorials, legal texts, product reviews, scientific articles, and Wikipedia articles; while within dialogical texts, we discuss Wikipedia discussions and the content of debating portals.

Persuasive Essays Persuasive essays aim to persuade the reader with a certain stance towards a specific topic. They are used frequently in education to assist students in improving their writing skills. Essays are one of the well-studied genres in NLP [Dong and Zhang, 2016; Persing and Ng, 2016]. In particular, assessing the quality of essays is addressed in various studies. Such studies strive at scoring essays by examining their grammar, structure, and used vocabulary [Dikli, 2006].

Stab and Gurevych [2014b] were the first to identify the argumentation structure of persuasive essays. Their work involved distinguishing the roles of argumentative units (i.e., claims or premises) and classifying the relation between these units (i.e., support or attack). As the basis of the identification approach, Stab and Gurevych [2014a] built a new corpus of around 400 annotated essays.

With regard to argumentation, four dimensions of essay quality were investigated: organisation [Persing *et al.*, 2010], thesis clarity [Persing and Ng, 2013], prompt adherence [Persing and Ng, 2014], and argument strength [Persing and Ng, 2015]. Wachsmuth *et al.* [2016] employed argument mining to assess the quality of persuasive essays, reaching the state-of-the-art effectiveness in two dimensions: the organisation and the strength of the argument.

Editorials Compared to persuasive essays, news editorials are opinionated articles that are written to persuade their readers to take stances on controversial issues. The author of an editorial often states a thesis that declares a certain stance on a controversial topic and justifies this thesis through using specific arguments.

On a regular basis, editorials target timely controversial issues such as the relocation of the US embassy in Israel to Jerusalem. As such, editorials can be a powerful tool for propagating ideologies or recommending attitudes to different communities [van Dijk, 1992]. For instance, editorials can incite public opinion towards not recognizing the new US embassy in Jerusalem.

The open-ended expansion of online news portals highlights the need for an automatic analysis of editorials. Such an analysis would be profitable for persuasive writing assistance tools and qualitative media content research. Editorialism, however, is an understudied text genre in computational argumentation and in NLP in general. Beyond our work in editorials [Al-Khatib *et al.*, 2017a, 2018; El-Baff *et al.*, 2018], opinion mining and retrieval were applied to a set of editorials [Bal, 2009; Yu and Hatzivassiloglou, 2003], and argumentation analysis of editorials was discussed partly in [Bal and Dizier, 2010; Kiesel *et al.*, 2015].

Legal Texts Legal text is a primary source of argumentation. Arguments are observed in legislative texts, case law, and doctrinal texts, among others [Moens *et al.*, 2007]. Nevertheless, few studies have tackled the computational argumentation tasks by addressing legal texts. It is possible that the lack of publicly available datasets of annotated legal texts for argumentation has constricted the ability to approach the legal genre in depth. Notable early work on legal texts was conducted as reported in [Palau and Moens, 2009]. Within that work, the ECHR corpus, which comprises a set of documents derived from the European Court of Human Rights (ECHR), was exploited to develop an argumentation mining approach.

Product Reviews People read product reviews to make decisions as to whether they need to buy a product or not. By reading product reviews, people expect to not only explore the opinions of other people on a product, but also to learn arguments for and against the product. A few of the studies considered arguments in product reviews; in particular, Wachsmuth *et al.* [2014, 2015] identified sequential flows of sentiment and argumentative roles in product reviews exposing various patterns of the identified flows, while Rajendran *et al.* [2018] leveraged a large-scale weakly supervised dataset for the task of stance identification. Liu *et al.* [2017] developed a set of argument-based features to predict how helpful hotel reviews are, and Wyner *et al.* [2012] proposed a new scheme for argumentation structure in camera reviews.

Scientific Articles Authors of scientific articles aim at proposing original and beneficial studies. The scientific articles typically include a description of the study along with argumentative texts. These texts persuade the readers regarding the merit of the proposed study. Automatic identification of the argumentation structure of scientific articles is desirable for multiple purposes; for example, it facilitates obtaining knowledge about several aspects of scientific articles such as their objectives, research questions, methods, and results [Guo *et al.*, 2011].

A leading line of research in this regard is argumentative zoning [Teufel and Moens, 2002]. In argumentative zoning, each passage in a scientific article is labelled with a distinct role such as the 'aim', 'own', and 'background'. A collection of papers followed this line of research, suggesting various methods for classifying argumentative zones [Contractor *et al.*, 2012; Guo *et al.*, 2012]. Besides argumentative zoning, Lauscher *et al.* [2018] introduced an annotation study for identifying relations (supports, contradicts, etc.) in addition to argumentative components (e.g., 'own claim', 'background') in scientific articles; while Lauscher *et al.* [2018] identified argumentative components and rhetorical aspects of writing in scientific articles.

Wikipedia Articles Wikipedia is the most influential collaborative writing platform on the Web, with about six million articles on English Wikipedia. Many Wikipedia articles contain various arguments in connection with diverse topics. One of the first attempts to retrieve arguments with respect to particular claims was made by [Aharoni *et al.*, 2014]. Wikipedia was utilised as the source for collecting context-dependent claims [Levy *et al.*, 2014] and for identifying evidence regarding the collected claims [Rinott *et al.*, 2015]. Furthermore, for a set of predefined topics, Roitman *et al.* [2016] retrieved the Wikipedia articles that encode claims related to those topics.

Wikipedia Discussions In Wikipedia, as a policy for maintaining the quality of the generated content, each article is associated with a 'talk' page. In a talk page, Wikipedia users discuss the content of the associated Wikipedia article with the aim of enhancing its content. As the users try to find the best action regarding the article (deleting a statement, merging two paragraphs, etc.), most of the discussions in Wikipedia can be regarded as deliberative. Wikipedia discussions in talk pages are considered as the largest source of deliberative discussions on the Web, with around six millions discussions. As a result, various studies have addressed the argumentation aspect of Wikipedia discussions. For example, Ferschke *et al.* [2012] and Viegas *et al.* [2007] have proposed models of dialogue acts in Wikipedia discussions. The goal of these models is to reduce the coordination effort between discussion participants. In these models, several acts such as 'information seeking and providing' are closely related to the pragmatic category of argument attributes. In a similar vein, Biran and Rambow [2011] built a corpus of Wikipedia discussions with manual annotation of claims and premises therein.

Debating Portals Content Debate portals are highly exploited sources in argumentation research. Debate portals such as `Idebate.org` and `depatepedia.org` are frequently employed for identifying argumentative units (e.g., Al-Khatib *et al.* [2016a]), finding argumentative components (e.g., [Habernal and Gurevych, 2015]), and classifying the relations between arguments (e.g., Anand *et al.* [2011]; Boltužić and Šnajder [2014]).

2.2.2 Argument Attributes

Many computational argumentation studies deal with identifying diverse attributes of arguments and argumentation. In this context, we highlight a collection of such studies, organising them within the three categories of argument attributes we presented earlier: dialectical, pragmatic, and stylistic.

Dialectical Attributes This category covers the logical and reasoning attributes of arguments and argumentation. From the different attributes that fall under this category, we discuss here the proposition verifiability, semantic types of propositions, argumentation schemes, argument reasoning comprehension, and fallacies.

Proposition Verifiability: Park and Cardie [2014] proposed a schema that categorises propositions into 'unverifiable', 'verifiable', 'verifiable non-experimental', and 'verifiable experimental'. This schema aims to determine the appropriate type of support for a proposition (evidence, reason, etc.). More particularly, Park and Cardie [2014] first conducted an annotation study that resulted in around 10,000 propositions labelled according to the schema. Later, the authors employed a support vector machine with diverse linguistic features such as N-grams and sentiment clues for automatic classification of the propositions categorised.

Semantic Types of Propositions: Hidey *et al.* [2017] introduced a new annotation schema for propositions in online persuasive forums. The schema involves assigning one of the following semantic types to claims: 'interpretation', 'evaluation', 'agreement', or 'disagreement'. The evaluation type was split into 'evaluation-rational' and 'evaluation-emotional'. Around 2,600 propositions were annotated according to the schema. Based on the annotation output, an analysis study was performed to examine whether specific types of propositions are influential regarding the persuasiveness of online comments.

Argumentation Schemes: An argument scheme represents a pattern of the logical inference from an argument's premises to its claim. In a remarkable work, Walton *et al.* [2008] proposed 65 schemes including 'argument from example' and ‘argument from cause to effect’. Each scheme is attached with a set of critical questions aimed at evaluating the arguments that follow that scheme. Feng and Hirst [2011] proposed a supervised model with simple linguistic features to distinguish the five most frequent schemes. The training and evaluation of the supervised method were accomplished based on the Araucaria corpus [Reed and Rowe, 2004]. Araucaria comprises around 600 arguments gathered from different sources and annotated according to Walton schemes [Walton *et al.*, 2008]. Cabrio *et al.* [2013] analysed the relation between argumentation schemes and discourse relations in the Penn Discourse Treebank (PDTB) [Prasad *et al.*, 2008]. The purpose of that analysis was to promote the argumentation mining systems by utilising new rich data that can be used for training such systems. Musi *et al.* [2016] introduced new guidelines for annotating argument schemes. The guidelines were constructed based on the argumentum model of topics [Rigotti and Greco, 2010]. The authors applied the guidelines by annotating 40 essays in total. The annotation results showed that obtaining high inter-annotator agreement requires highly trained annotators and robust classification of argument types (claim vs premise). Beside Walton schemes, the ‘argumentative microtext’ corpus of Peldszus and Stede [2016] has been extended with scheme annotations according to Musi *et al.* [2018].

Argument Reasoning Comprehension: According to the Tolmin model for argumentation [Toulmin, 1958], a warrant explains the logical inference from the argument’s premise to the drawn conclusion. Habernal *et al.* [2018a] developed a new methodology for an automatic formation of warrants and applied it within a crowdsourcing study. The study produced warrants for 2,000 arguments. These warrants were then employed for establishing a new challenge task: Given the claim and premise of an argument, the correct implicit warrant between the given claim and premise should be identified out of two options. This task was widely approached from several researchers with various neural deep learning methods [Habernal *et al.*, 2018c].

Fallacies: A fallacy entails a plausible argument with an invalid inference. Identification of fallacies is one of the most challenging tasks. To our knowledge, ad hominem is the only studied fallacy in computational argumentation so far. Habernal *et al.* [2018b] conducted an empirical study regarding ad hominem, exploring their topology and potential causes. Wulczyn *et al.* [2017] also built a new corpus for ad hominem on Wikipedia discussions. The corpus comprises around 115,000 comments extracted from Wikipedia talk pages. Each comment is labelled as a personal ‘attack’ or ‘not-attack’. Several approaches used that corpus for identifying personal attacks on discussions including [Bodapati *et al.*, 2019] and [Pavlopoulos *et al.*, 2017].

Pragmatic Attributes Here, we review particular attributes that concern the function, purpose, or intention of stating an argumentative unit: i.e., the speech act of an argumentative text. Under this category, we point out the argument roles, argument evidence types, argument frames, and the general speech acts.

Argument Roles: Identifying the argumentative roles can be substantial for establishing a strategy. That is to say, vital strategy principles can be designed by arranging argumentative units along with their roles.

The identification of an argument role is a sub-task of argumentation mining. This sub-task was approached within numerous studies that concentrated mainly on the roles of claim and premise. Stab and Gurevych [2014b] developed a new method for identifying argumentative structures in essays. In particular, given segmented units of an essay, a support vector machine model with diverse linguistic features was employed to classify each segment as major claim, claim, premise, or non-argumentative. The identification method was implemented using the Essays corpus of Stab and Gurevych [2014a].

Peldszus and Stede [2015] identified the argumentative structures of argumentative texts through jointly modelling the different subtasks of argumentation mining (finding argumentative units, classifying their roles and their relations). The joint model was applied on a corpus of short texts written in German and translated professionally into English [Peldszus and Stede, 2016]. The corpus contains 112 short texts with 576 argumentative units. The texts were collected manually under controlled experiments. Selected conditions related to the length of the text and the consideration of roles were followed by the annotators.

Argumentation mining, including the identification of argumentative roles, was approached in an end-to-end manner. Eger *et al.* [2017] studied several neural models for that purpose. The neural models were applied to the Essays corpus [Stab and Gurevych, 2014a]. The results highlighted the importance of joint neural learning in a multi-task setting.

Determining the role of countering an argument (counter-argument) is an influential sub-task in argumentation mining. The impact of considering counter-arguments is manifest in various studies. For example, Habernal and Gurevych [2016] found that an argument can be perceived as more convincing than others if it counter-attacks the opposed position. Furthermore, Zhang *et al.* [2016] reported that counter-arguments are used frequently by the winning sides of debates. A notable work regarding retrieval of counter-arguments was performed by Wachsmuth *et al.* [2018b]. The work aimed at retrieving the best counter-argument when no prior knowledge about the topic of argumentation is available. Such a scenario is common in argument search engines. The proposed approach relied on the assumption that a counter-argument targets the same aspects that the argument targets but expresses the opposite stance. A new model was developed based on this assumption, and operationalisation steps were carried out relying on word embeddings and distinct similarity functions.

Argument Evidence Types: A significant line of research herein is the identification of the type of evidence in an argumentative unit: ‘statistics’, ‘anecdotal’, or ‘expert’. In this regard, Rinott *et al.* [2015] introduced a supervised learning model for identifying context-dependent evidence (evidence related to given claims) in Wikipedia articles. To this end, the authors proposed a pipeline of supervised learning modules, each of which targets a particular task, namely ‘coherence selection’, ‘evidence characteristics’, ‘context-dependent’, and ‘claim selection’.

Argument Frames: In the NLP community, Card *et al.* [2015] developed a new corpus of news articles regarding the topics of ‘same-sex marriage’, ‘immigration’, and ‘smoking’. The corpus covers 15 frames including ‘morality’, ‘economics’, and ‘legality’. Based on this corpus, Naderi and Hirst [2017] developed a neural-based method for identifying frames at sentence level. Closely linked to argumentation, Naderi and Hirst [2015] investigated the identification of frames in parliamentary speeches considering seven arguments related to ‘gay marriage’ as frames. Recently, Ajjour *et al.* [2019] introduced a new corpus of arguments annotated for their topics and frames. Evaluated by this corpus, a new unsupervised approach is proposed for identifying the frame of an argument. The approach is based on removing topical features from arguments before clustering them into frames.

General Speech Acts: Speech act theory concerns the utterances that serve a function in communication [Searle, 1969]. The theory is one of the broadly accepted theories in pragmatics. Argumentation is well connected to speech acts (including discourse and dialogue acts) since an argument can be viewed as a complex speech act [Stede and Schneider, 2018]. Relatively few studies have explicitly explored speech acts theory in argumentative texts. Niven and Kao [2019] conducted a preliminary study investigating how argumentative discourse acts are associated with linguistic alignment.

Visser *et al.* [2019] built a new annotated corpus of dialogical argumentation considering argumentative relations and dialogue acts. The corpus covers the 2016 presidential election debates in the United States and reactions to these debates in social media (particularly in Reddit). The annotated set of dialogue acts comprises ‘arguing’, ‘agreeing’, ‘questioning’, and others.

Zhang *et al.* [2017a] introduced a new method for classifying the discourse acts of comments in online discussions. The classification was performed based on an annotated corpus. The corpus comprises annotations for nine discourse acts including ‘elaboration’, ‘appreciation’, ‘question’, and ‘answer’. The experiments on the discourse act classification illustrated that structured prediction models perform well in such tasks.

Stylistic Attributes: This category concerns how to phrase the arguments and other texts in argumentative discourse. A remarkable line of research here is the identification of rhetorical figures.

Rhetorical Figures: Rhetorical figures have been studied thoroughly in terms of humanity and communication [Craig, 2006]. In the NLP community, Gawryjołek *et al.* [2009] identified the rhetorical figures of ‘anaphora’, ‘isocolon’, ‘epizeuxis’, and ‘oxymorons’ for tackling the task of authorship attribution. Strommer [2011] focused on the ‘epanaphora’ figure, distinguishing between the accidental and intentional usage of this figure. Java [2015] proposed a framework for identifying 12 rhetorical figures including ‘parallelism’, ‘repetition’, and ‘trope’. As for studying rhetorical figures in argumentation, Lawrence *et al.* [2017] studied the connection between rhetorical figures and argumentation structure for supporting argument mining systems. In the MM2012c corpus [1], which comprises annotations of argumentative transcripts from BBC Radio 4’s Moral Maze discussion programme, eight rhetorical figures, including ‘anadiplosis’, ‘epanaphora’, and ‘epistrophe’, were identified and analysed regarding their correlation to the argument structures of the transcripts, considering different relations such as ‘support’, ‘incoming’, ‘conflict’, and others. The study stressed the presence of the investigated connection.

2.2.3 Argumentation Goals

Persuasion and consensus are highly respected goals of argumentation. In this subsection, we report on existing work regarding the two goals.

Persuasion Various studies approached the persuasion following Aristotle’s modes of persuasion. Duthie *et al.* [2016] developed a corpus and applied a new approach for identifying linguistic expressions that encode ethos from political debates. In the same vein, Hidey *et al.* [2017] built a corpus of annotated argument premises with the labels of logos, ethos, and pathos. Moreover, Habernal and Gurevych [2015] developed a corpus of annotated user comments

[1] corpora.aifdb.org/mm2012c

and forum posts. The annotation covers pathos and logos labels at document level. Carlile *et al.* [2018] built a new corpus for essays, annotating their persuasiveness along with pathos, logos, and ethos labels. Wang *et al.* [2019] aimed at developing persuasive conversational agents. To this end, the authors constructed a new corpus of dialogues with annotations regarding several persuasive strategies including pathos, logos, and ethos. A baseline classifier was built to identify the modelled strategies.

In addition to the above, the persuasiveness of texts was explored within the task of predicting the success of changing someone's view in the subreddit of 'Change My View'. This task was approached in [Tan *et al.*, 2016] and [Wei *et al.*, 2016]. The former demonstrated the high impact of the number of interactions between debaters on persuasiveness. In a like manner, the latter pointed out the importance of social interactions and argumentation-based features on persuasiveness. In political debates, Cano-Basave and He [2016] investigated the effectiveness of semantic framing of arguments for predicting the influence of a speaker. Moreover, Wang *et al.* [2017] showed that the winners of political debates use strong arguments and properly shift the topic of the debate. Yang *et al.* [2019] proposed neural networks for identifying persuasive strategies and their success in advocacy requests.

How modelling the target audience or readers of argumentation theory influences its persuasiveness was explored in several studies. Among these studies, how the background and beliefs of a discussion's participants influence their ability to be persuaded was studied in [Durmus and Cardie, 2018]. Similarly, Lukin *et al.* [2017] disclosed how the persuasiveness of logos-oriented and pathos-oriented arguments generally depends upon the personality of the target readers. Durmus and Cardie [2019] created a new corpus of debates that includes comprehensive profiles of the debates' participants. The dataset was used to examine the role of participants' traits for predicting the winner of a debate. Beyond that, how to challenge or reinforce the stance of different target readers of editorials was addressed in [El-Baff *et al.*, 2018]. Longpre *et al.* [2019] investigated the impact of modelling the audience in online debates regarding their prior stances on the debate's topic: namely, the decided vs undecided stances.

Consensus Particularly in the NLP community, relatively fewer studies have been conducted in connection with the consensus goal compared to the persuasive one. Most of the developed models, datasets, and methods for deliberative discussions aim to minimise the coordination effort among discussion participants. Several studies in this direction focused on Wikipedia discussions in talk pages, such as [Ferschke *et al.*, 2012] and [Kittur *et al.*, 2007]. In addition, Wang and Cardie [2014] attempted to differentiate between disputed vs undisputed Wikipedia discussions. Im *et al.* [2018] performed a qualitative and quantitative analysis for resolving disputes in Wikipedia discussions focusing on the Requests for Comments (RFCs) template. In addition to the analysis, the authors developed a new model for predicting the closed RFCs from those that went stale. In addition, Walker *et al.* [2012] built a dataset derived from debate platforms to understand how people argue in deliberative discussions.

2.3 Summary

This chapter has briefly presented the background of argumentation strategies, in addition to their related work in NLP. It has established the ground for understanding the subsequent chapters in this thesis.

In the first part of this chapter, we focused on the notion of argumentative discourse, elaborating on the discourse properties of goal and directionality. Later, we explained a novel view of argumentation strategies, demonstrating strategy elements as well as strategy formulation and implementation process. In the second part of this chapter, we listed and reviewed the major related work concerning computational argumentation strategies, including the current work on the identification of strategy attributes as well as the investigation of the argumentation goals of persuasion and consensus.

Chapter 3

Pragmatic Persuasive Strategies

A wise man proportions his belief to the evidence. (David Hume)

This chapter describes our analysis of argumentation strategies in persuasive editorials with respect to the first three research questions of this thesis (see Chapter 1, Section 1.2). The analysis is based on the exploration of the selection and arrangement of argumentative discourse units. Such an exploration is performed on the basis of the pragmatic attribute of the *types of argumentative discourse units* according to their roles in the discourse. The analysis results reveals several strategy principles in editorials from different portals as well as across different topics.

News editorials define a genre of written argumentative discourse whose main goal is persuasiveness. In a news editorial, an author states and defends a thesis that conveys his or her stance on a controversial topic usually related to the public interest. Editorials do not only persuade readers of some opinion, but they often also propagate particular ideologies or recommend certain attitudes to the community, e.g., a specific action towards an upcoming event [van Dijk, 1992].

To achieve persuasion, a news editorial follows a particular *argumentation strategy* that the author expects to be most suitable for the target audience, i.e., the author outlining principles regarding the composition of a series of claims, assumptions, and different types, while using argumentative language and structure [van Dijk, 1995]. This does not only cover the resort to quantitative features of text (e.g., related to lexical style, cohesion, or rhetorical structure), but it also refers to the pragmatic attributes such as the types of argumentative discourse units.

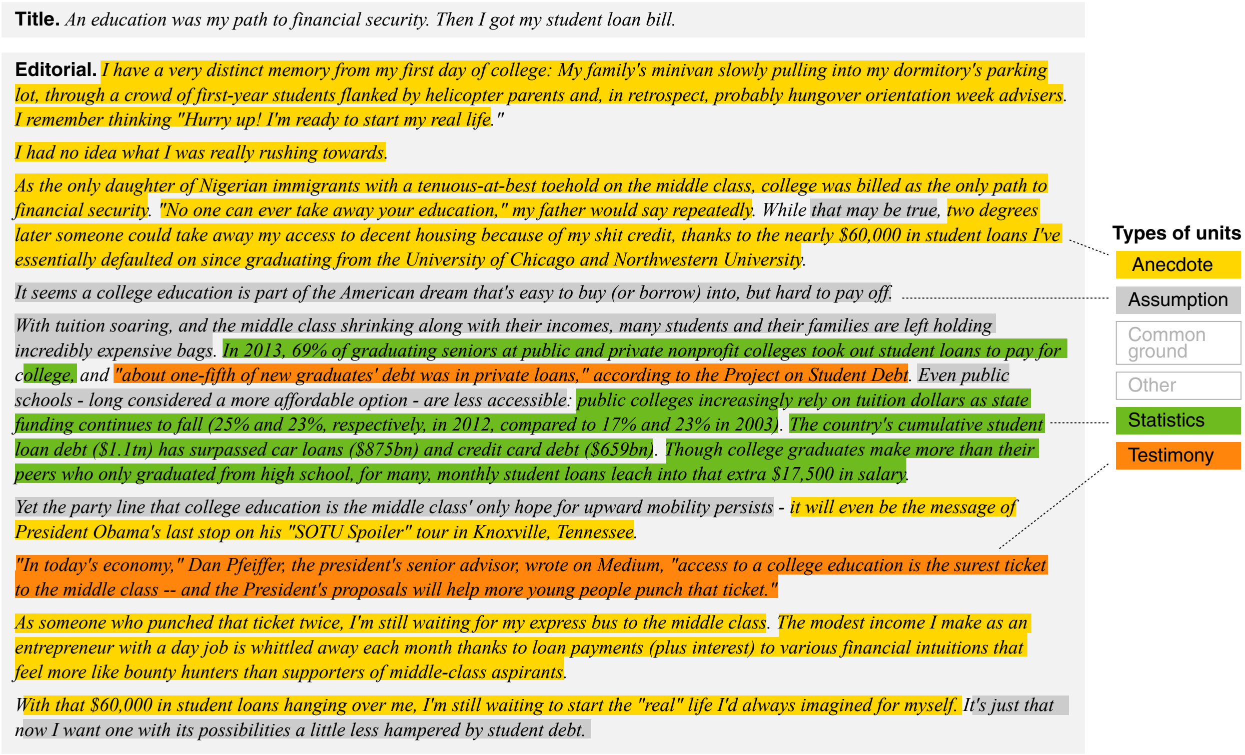
The rapid expansion of online news portals increases the need for algorithms that can analyse an editorial's discourse *automatically.* The needed analyses include argumentation mining and evidence detection, both of which are studied in computational argumentation. However, computational approaches that study how to deliver the arguments *persuasively* are still scarce — despite the importance of such studies for envisaged applications that deal with the synthesis of effective argumentation, such as debating systems.

In this chapter, we first introduce a new model regarding the pragmatic attributes of argumentative discourse units in editorials. This model covers the evidence types of ‘testimony’, ‘statistics’, and ‘anecdote’, as well as the types of ‘assumption’ and ‘common ground’. According to this model, we build a novel corpus with 300 news editorials evenly selected from three diverse online news portals: *Al Jazeera*, *Fox News*, and *The Guardian*. Basically, each editorial contains manual type annotations of all units that capture the role (aka the function) that a unit plays in the argumentative discourse. The corpus consists of 14,313 units of six different types (the five types mentioned above plus ‘other’), each annotated by three professional annotators from the crowdsourcing platform `upwork.com`. Based on the results of the annotation process, we analyse the agreement between annotators in order to scrutinize the major cases of disagreement as well as to designate the complex issues that humans face in classifying types of argumentative discourse units in editorials. Considering the number and complexity of the types, the obtained inter-annotator agreement of 0.56 in terms of Fleiss' κ can be seen as high. Then, in a first brief statistical analysis of the corpus, we investigate differences in the type distribution between the three portals, which indicate the presence of divergent argumentation strategies there.

Next, the built corpus is employed for the development of a new supervised learning method for the identification of the considered types. To explore the strategies across topics, experiments are conducted on around 29,000 editorials extracted from the *New York Times (NYT) Annotated Corpus* [Sandhaus, 2008]. We automatically categorize these editorials into 12 coarse-grained topics such as ‘economics’, ‘arts’, and ‘health’. Then, we apply the identification method to the 29,000 editorials. The results of the experiments expose significant differences in the distribution of the evidence types across the 12 topics. Furthermore, the results discriminate a number of sequential patterns of the evidence types which are common in editorials. Both results provide insights into what principles of argumentation strategies exist in editorials across different topics.

To foster future research on pragmatic persuasive argumentation strategies, the developed corpus, the developed classifier, and the topic categorization of all editorials are publicly available at `http://www.webis.de`.

The remainder of this chapter is structured as follows: Section 3.1 describes a new model for a pragmatic attribute of argumentative units in editorials. Section 4.2 discusses the construction of a new corpus regarding the proposed model. Section 3.3 reports on the discovered strategy principles across newspaper



Title. *An education was my path to financial security. Then I got my student loan bill.*

Editorial. *I have a very distinct memory from my first day of college: My family's minivan slowly pulling into my dormitory's parking lot, through a crowd of first-year students flanked by helicopter parents and, in retrospect, probably hungover orientation week advisers. I remember thinking "Hurry up! I'm ready to start my real life."*

I had no idea what I was really rushing towards.

As the only daughter of Nigerian immigrants with a tenuous-at-best toehold on the middle class, college was billed as the only path to financial security. "No one can ever take away your education," my father would say repeatedly. While that may be true, two degrees later someone could take away my access to decent housing because of my shit credit, thanks to the nearly $60,000 in student loans I've essentially defaulted on since graduating from the University of Chicago and Northwestern University.

It seems a college education is part of the American dream that's easy to buy (or borrow) into, but hard to pay off.

With tuition soaring, and the middle class shrinking along with their incomes, many students and their families are left holding incredibly expensive bags. In 2013, 69% of graduating seniors at public and private nonprofit colleges took out student loans to pay for college, and "about one-fifth of new graduates' debt was in private loans," according to the Project on Student Debt. Even public schools - long considered a more affordable option - are less accessible: public colleges increasingly rely on tuition dollars as state funding continues to fall (25% and 23%, respectively, in 2012, compared to 17% and 23% in 2003). The country's cumulative student loan debt ($1.1tn) has surpassed car loans ($875bn) and credit card debt ($659bn). Though college graduates make more than their peers who only graduated from high school, for many, monthly student loans leach into that extra $17,500 in salary.

Yet the party line that college education is the middle class' only hope for upward mobility persists - it will even be the message of President Obama's last stop on his "SOTU Spoiler" tour in Knoxville, Tennessee.

"In today's economy," Dan Pfeiffer, the president's senior advisor, wrote on Medium, "access to a college education is the surest ticket to the middle class -- and the President's proposals will help more young people punch that ticket."

As someone who punched that ticket twice, I'm still waiting for my express bus to the middle class. The modest income I make as an entrepreneur with a day job is whittled away each month thanks to loan payments (plus interest) to various financial intuitions that feel more like bounty hunters than supporters of middle-class aspirants.

With that $60,000 in student loans hanging over me, I'm still waiting to start the "real" life I'd always imagined for myself. It's just that now I want one with its possibilities a little less hampered by student debt.

Figure 3.1: Example of a news editorial from The Guardian. Each argumentative discourse unit of the editorial has been assigned one of six types, four of which are shown here.

portals. Section 3.4 proposes a method for the identification of the modelled pragmatic attribute. Section 3.5 illustrate the strategies discovered across topics. Section 3.6 reviews the related work, and finally, Section 4.6 briefly summarizes the chapter.

3.1 Model

In this section, we propose a new model, in terms of annotation scheme, for analysing the argumentation strategy of a news editorial. Primarily, we separate an editorial into argumentative discourse units of six different types where each type represents a particular role in the discourse. While our model is in line with related work on evidence types [Rinott *et al.*, 2015], we assign a type to each unit in order to capture an editorial's overall argumentation strategy.

In particular, we see argumentative discourse units as the smallest elements of the argumentative discourse of an editorial. They represent the propositions stated by the editorial's author to discuss, directly or indirectly, his or her thesis. In general, propositions affirm or deny that certain entities have certain attributes. An entity may be an object (e.g., *milk*), a being (e.g., *Obama* or *we*), or an abstract concept (e.g., *learning to cooperate*). Technically, we define a unit based on the notion of propositions as follows:

> **Argumentative Discourse Unit:** An argumentative discourse unit is the minimum text span that completely covers one or more propositions. It always includes a subject (or a placeholder, such as "which") and a verb, and it needs to include an object if grammatically required. It spans at most one sentence.

The following list shows typical examples of units to illustrate the given definition. Units are denoted with ***[bold font in square brackets]***.

1. ***[We should tear the building down,]***$^{(1)}$ ***[it is full of asbestos.]***$^{(2)}$
   Subsequent propositions within a sentence become separate units; they may be connected explicitly (e.g., with connectives like *and* or *but*) or implicitly (as shown).

2. ***[That guy confesses his mistakes,]***$^{(1)}$ ***[which makes me believe in him.]***$^{(2)}$
   In the second unit, the subject is replaced by a placeholder word.

3. ***[The virus was not created to make money or to play jokes.]***
   If two or more propositions overlap (here: *The virus was not created to make money* and *The virus was not created to play jokes*), they considered to belong to the same unit.

4. *In both cases, we see that* ***[learning to cooperate helps.]***$^{(1)}$ ***[She knew this]***$^{(2)}$*, too.*
   Leading or trailing transition words and phrases do not belong to units. Note that a unit is not necessarily understandable on its own, such as Unit 2 here.

5. ***[Many people are – you sure know some of them – nicest in the morning.]***
   A unit—although split sometimes—always spans a proposition entirely.

6. ***[Prof. Miller said in his talk that drinking milk strengthens the bones.]***
   While the span after *that* is a proposition alone, the unit covers the whole sentence, because the construction *said ... that* requires an object to be grammatically correct.

7. *"Trust me!* ***[You should drink as much milk as possible!", he concluded.]***
   A quotation that begins or ends within another sentence is assumed to belong to that sentence. *Trust me* is not a unit, because it misses a subject.

8. *What others exist? And* ***[who would be a better candidate than Obama?]*** Many questions contain no proposition, but rhetorical questions do.

The units of a news editorial play different roles in the editorial's argumentative discourse. For example, some represent knowledge or beliefs of the author or other people, and some serve as evidence in favor or against the truth of other units. We assume each unit to refer to exactly one of six types:

1. **Common Ground:** The unit states common knowledge, a self-evident fact, an accepted truth, or similar. It refers to general issues, not to specific events. Even if not known in advance, it will be accepted without proof or further support by all or nearly all possible readers.

   Example: *"History warns us what happens when empires refuse to teach known values that strengthen societies and help protect them from enemies intent on their destruction."*

2. **Assumption:** The unit states an assumption, conclusion, judgment, or opinion of the author, a general observation, possibly false fact, or similar. To make readers accept it, it is, or it would need to be supported by other units.

   Example: *"For too long young people have relied on adults who have done too little to stop the violation of the rights of the children for whom they were responsible."*

3. **Testimony:** The unit gives evidence by stating or quoting that a proposition was made by some expert, authority, witness, group, organization, or similar.

   Example: *"According to The Yazidi Fraternal Organization (YFO), thousands of young Yazidi women and children are being used by ISIL as sex slaves."*

4. **Statistics:** The unit gives evidence by stating or quoting the results or conclusions of quantitative research, studies, empirical data analyses, or similar. A reference may but needs not always be given.

   Example: *"Of the total of 779 men and boys that have been detained at Guantanamo Bay since 2002, only nine have been convicted of any crime."*

5. **Anecdote:** The unit gives evidence by stating personal experience of the author, an anecdote, a concrete example, an instance, a specific event, or similar.

   Example: *"In 1973, it deployed 18,000 troops with 300 tanks to save Damascus during the 'October War'."*

6. **Other:** The unit does not or hardly adds to the argumentative discourse or it does not match any of the above classes.

   Example: *"Happy New Year!"*

Our hypothesis is that these six types suffice to capture an important pragmatic attribute of argumentation strategies in news editorials. At the same time, they define an annotation scheme for a fine-grained mining of argumentative discourse units.

Figure 3.1 shows the type annotations of an editorial. The editorial has been taken from The Guardian.

3.2 Corpus Construction

While several text corpora for the evidence type analyses have recently been published for different domains and genres, a respective resource with news editorials is missing to this day. Moreover, existing corpora stick to coarse-grained and/or incomplete annotations of the units of an argumentative discourse (see Section 3.6 for details), which renders the mining of an author's argumentation strategy impossible.

This section describes the construction of a news editorial corpus based on the proposed model in the previous section. The purpose of the corpus is to study different argumentation strategies in news editorials in terms of the types of argumentative discourse units used there.

Data Acquisition and Preparation

Before the annotation, the editorials are selected from three diverse news portals and decomposed into clause-like segments in order to ease the annotation process to achieve scale.

Selection of Argumentative News Editorials The corpus consists of editorials from *aljazeera.com*, *foxnews.com*, and *theguardian.com*. This selection of news portals cover diverse cultures and styles. These portals are internationally well-known and have separate editorial sections. We randomly selected 100 editorials from each portal that (1) are published within the same short time interval (December 2014 and January 2015) to facilitate a topical overlap, (2) sparked at least a small discussion (had at least 5 comments), and (3) contain at least 250 words (to filter out texts that just pose a question instead of arguing).

Pre-Segmentation of Argumentative Discourse Units To allow for an annotation at larger scale, we automatically segmented the editorials before the annotation but then allowed annotators to merge adjacent segments to discard incorrect unit boundaries. In this setup, the annotators do not have to choose the exact unit boundaries, which simplifies the annotation process while making the evaluation of the annotator-agreement more intuitive. A similar manual approach was used by Park and Cardie [2014].

In detail, the applied segmentation algorithm, which we make publicly available, starts a new segment at the beginning or end of every clause not preceded by a relative pronoun. Clauses were identified using a state-of-the-art dependency parser [Manning *et al.*, 2014] and the clause tags from the Penn Treebank Guidelines [Bies *et al.*, 1995]. The heuristic behind the segmentation was chosen based on a careful analysis of news editorials as well as of the persuasive essays corpus from Stab and Gurevych [2014a], since essays resemble editorials in the way they compose argumentative discourse units. An evaluation of the segmentation algorithm on that corpus yielded very satisfying results: The algorithm segmented the 90 essays into 5132 segments. Only nine of these segments should have been split further, as they overlapped with several ground-truth units from the essay corpus. On the other hand, the segmentation was somewhat too fine-grained, namely, the 1552 ground-truth units were split into 3637 segments. In our setup, however, the annotators then perform the necessary segment merges. Table 3.1 shows statistics about the size of the corpus and its three sub-corpora after segmentation.

Annotation Process

Given the 300 selected news editorials, an annotation process was performed in order to identify all argumentative discourse units in each newspaper editorial, including an assignment of one of the six types from Section 3.1 to each unit.

The main steps of this process are summarized in the following.

Task Definition First, each editorial had to be read as whole in order to understand the main topic and to follow the stance of the editorial's author.

As the annotation task, one out of eight classes had to be chosen for each segment of each editorial (see pre-segmentation above): (1–6) Any of the six types of argumentative discourse units of our model from Section 3.1, (7) *no unit*, when the segment does not belong to a unit, and (8) *continued*, when the segment needs to be merged with subsequent segments in order to obtain a unit. In case (8), the class assigned to the last segment determines the class of the merged unit.

Type	**Editorials**	**Total**	**Mean**	**Std. dev.**	**Median**	**Min**	**Max**
Tokens	**All editorials**	**287364**	**957.88**	**257.28**	**932**	**298**	**1894**
	Al Jazeera	106430	1064.30	236.05	1033	440	1671
	Fox News	86415	864.15	226.36	855	298	1613
	Guardian	94519	945.19	267.13	906	481	1894
Sentences	**All editorials**	**11754**	**39.18**	**13.00**	**37**	**12**	**114**
	Al Jazeera	3962	39.62	10.55	38	16	75
	Fox News	3912	39.12	13.45	39	12	104
	Guardian	3880	38.80	14.65	36	18	114
Paragraphs	**All editorials**	**4664**	**15.55**	**6.48**	**15**	**2**	**45**
	Al Jazeera	1896	18.96	5.15	19	7	33
	Fox News	1689	16.89	6.71	16	2	45
	Guardian	1079	10.79	4.29	10	5	31
Segments	**All editorials**	**35665**	**118.88**	**38.21**	**116**	**28**	**309**
	Al Jazeera	11521	115.21	31.68	113	32	218
	Fox News	11315	113.15	35.4	112	28	231
	Guardian	12829	128.29	44.58	122	59	309

Table 3.1: Distribution of tokens, sentences, paragraphs, and segments in the corpus before annotation.

The annotation guidelines given to the annotators contained the type definitions from Section 3.1 and a few clear and controversial examples for each type. In addition, we pointed out that the correct classification of a segment may require looking at surrounding segments. Also, no distinction should be made between true and false propositions. For example, a wrong testimony should still be classified as testimony.

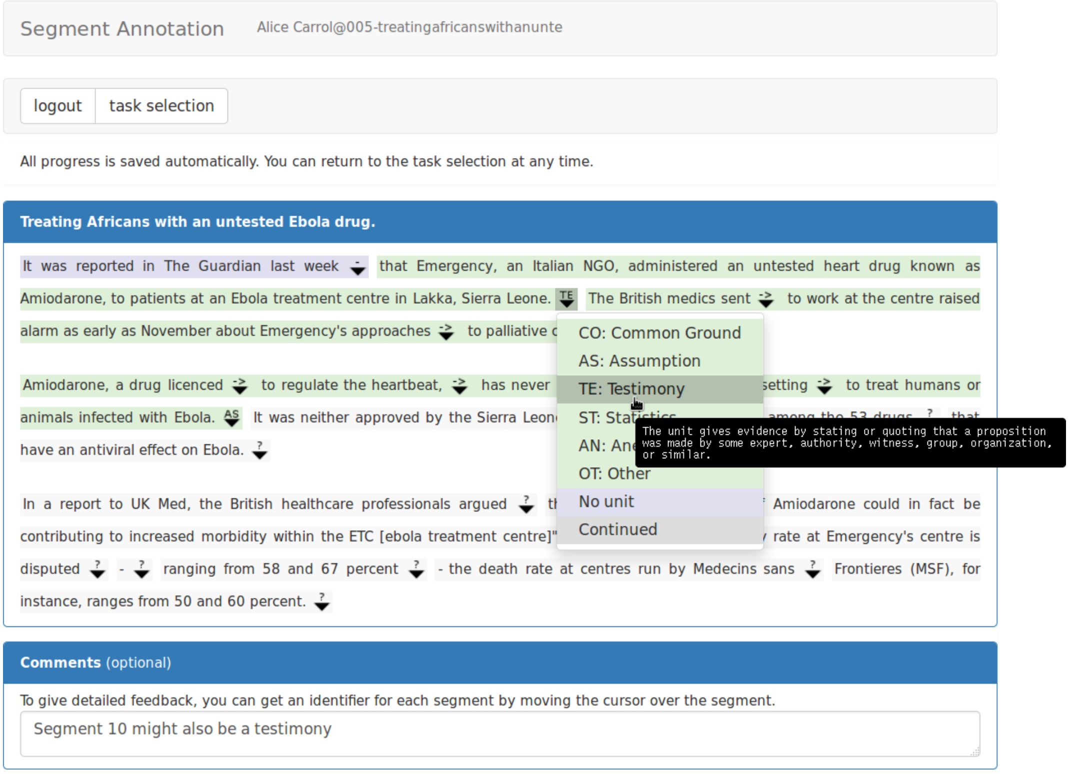
Annotation Tool To conduct our annotation study, we used *WAT-SL* (Web Annotation Tool for Segment Labeling) [Kiesel *et al.*, 2017]. *WAT-SL* is an open-source web-based annotation tool dedicated to segment labeling. WAT-SL provides all functionalities to efficiently run and manage segment labeling projects. Its self-descriptive annotation interface requires only a web browser, making it particularly convenient for remote annotation processes. The interface can be easily tailored to the requirements of the project using standard web technologies in order to focus on the specific segment labels at hand and to match the layout expectations of the annotators. At the same time, it ensures that the texts to be labeled remain readable during the whole annotation process. This process is server-based and preemptable at any point. The annotator's progress can be constantly monitored, as all relevant interactions of the annotators are logged in a simple key-value based plain text format.

Figure 3.2 depicts a screenshot of *WAT-SL* annotation tool.

Pilot Annotation Study A pilot study was conducted on nine editorials to evaluate the guidelines and to select the annotators. For this purpose, three editorials were chosen from each portal. These nine editorials are not part of the corpus, but were acquired and segmented in the same fashion. In total, the nine editorials comprised 1079 segments, with 119.9 segments on average in each editorial.

We decided to conduct the annotation process via the professional crowdsourcing platform `upwork.com` in order both to increase scalability and to obtain independent annotators. The list of candidate annotators comprised ten freelancers (six males and four females). All of them were native English speakers, had at least a bachelor's degree in one of the subjects of philosophy, (applied) linguistics, psychology, politics, and economics, and had already knowledge about argumentation theories from their education.

While most of the annotators were not experienced in the analysis of argumentative text, all had some or much previous knowledge about argumentation theories from their education. Similarly, all had much experience in writing, but not specifically in the writing of journalistic texts.



All progress is saved automatically. You can return to the task selection at any time.

Treating Africans with an untested Ebola drug.

It was reported in The Guardian last week that Emergency, an Italian NGO, administered an untested heart drug known as Amiodarone, to patients at an Ebola treatment centre in Lakka, Sierra Leone. The British medics sent to work at the centre raised alarm as early as November about Emergency's approaches to palliative c

Amiodarone, a drug licenced to regulate the heartbeat, has never setting to treat humans or animals infected with Ebola. It was neither approved by the Sierra Leon among the 53 drugs that have an antiviral effect on Ebola.

In a report to UK Med, the British healthcare professionals argued th Amiodarone could in fact be contributing to increased morbidity within the ETC [ebola treatment centre]" rate at Emergency's centre is disputed - ranging from 58 and 67 percent - the death rate at centres run by Medecins sans Frontieres (MSF), for instance, ranges from 50 and 60 percent.

The unit gives evidence by stating or quoting that a proposition was made by some expert, authority, witness, group, organization, or similar.

Comments (optional)

To give detailed feedback, you can get an identifier for each segment by moving the cursor over the segment.

Segment 10 might also be a testimony

Figure 3.2: Screenshot of the annotation interface of WAT-SL. In particular, the screenshot illustrates how the annotator selects a label (*Testimony*) for one segment of the news editorial.

The annotation process was controlled carefully. We directly contacted the annotators if possible problems were observed to resolve them. Also, the annotators were advised to contact us once they have any comments. The annotation tool included a comments area where editorial-specific comments could directly be made during the annotation. In total, we paid US-$ 5 per editorial for each annotator.

Seven annotators completed the nine editorials, taking around 30 minutes per editorial on average. The Fleiss' κ agreement score for all seven annotators was a moderate 0.433 [J. Richard Landis, 1977]. As we observed remarkable drops in the agreement caused by either of three specific annotators, we decided to exclude those annotators and keep the remaining four for the main annotation study.

An error analysis of the annotation of the four annotators revealed insightful regarding hard cases. For instance, the annotators had difficulties to distinguish

	Common ground	**Assumption**	**Anecdote**	**Testimony**
Fleiss' κ	0.114	0.613	0.399	0.591

	Statistics	**Other**	**No unit**	**Continued**
Fleiss' κ	0.582	0.152	0.365	0.684

Table 3.2: Inter-annotator agreement in the main annotation study, quantified in terms of Fleiss' κ.

between 'common ground' and 'anecdote' for units discussing a specific event that is well-known universally. Also, there was notable disagreement between 'common ground' and 'assumption'. This was expected, though, since the distinction of these two types appears more subjective than for other type combinations. Nevertheless, the agreement between the four annotators for all types was substantial with $\kappa = 0.606$. Therefore, we decided not to modify our scheme, but only to clarify the type definitions and to add some additional examples that clarify these hard cases.

Main Annotation Study The 300 selected editorials were evenly distributed among the four annotators. Each annotator got 225 editorials to annotate, 75 from each news portal. Accordingly, each editorial was annotated by three annotators. Analogous to the pilot study, the annotators received US-$ 5 per editorial. In total, the annotation process took about two months with a total cost of US-$ 4815.

3.2.1 Annotation Results

We analysed the results of the main annotation study in order to examine (1) the reliability of the corpus and (2) the major disagreements in units and types annotations between the annotators. Our main findings are as follows:

Inter-Annotator Agreement In terms of Fleiss' κ, the overall agreement is 0.56. As broken down in Table 3.2, however, the types 'common ground' and 'other' have only a slight agreement, while the annotators achieved fair agreement for 'no unit' and 'anecdote' as well as moderate or substantial agreement for the remaining four types. The class 'continued' obtained the highest agreement, which indicates that the annotators were able to identify the boundaries of argumentative discourse units successfully. Moreover, for 94.4% of all segments at least two of three annotators agree on one type, suggesting that a resort to majority agreement is very adequate. Considering that the annotators had to decide among eight different classes for every segment, such agreement seems high in overall terms. Therefore, we conclude that the annotations of the corpus can be seen as reliable.

Disagreement Analysis To analyse the disagreement between the annotators, we created the confusion probability matrix (CPM, [Cinková *et al.*, 2012]) for all classes shown in Table 3.3. Each matrix cell shows the probability of choosing the column's class, given that another annotator chose the row's class. Table 3.3 reveals the five class-pairs where annotators are most confused between:

1. Disagreement between 'other' and 'assumption' (0.324). An explanation may be that the annotators interpreted the intention of the author of a respective editorial differently in some segments.

   An example unit that led to confusion is *"I just don't get it"* after another unit *"the rave reviews for the first episode make me feel like a teetotaller at a lock-in"*. The first unit could be interpreted as an implicit assumption about the reviews in the second unit, say, that the review is corrupt or hard to understand. However, it could also simply be seen as an interjection not belonging to any argument.

2. Disagreement between 'common ground' and 'assumption' (0.562). Although we revised our guidelines to resolve the ambiguity of these types, their distinction still seems to be hard in practice.

   For example, the unit *"To see a movie legally you must leave your house, queue up, ask someone for a ticket and then sit down in the company of others"* can be viewed as 'common ground' if the annotator believes that most people agree with this statement, meaning there is no need for justification. In contrast, it is viewed as an 'assumption' if people are assumed to disagree to some extent, because a DVD can be bought and watched legally at home, for example.

3. Disagreement between 'common ground' and 'anecdote' (0.163). Confusion between these types occurred in cases where there was a distinct fact that the editorial's author uses to support his stance.

   For example, *"Iraq's Sunnis were the leading force within the Iraqi army since its foundation on January 6, 1921"*. This declaration was used to support the author's claim that the Sunnis respect their army and see it as a national institution of unrivalled prestige.

	Common ground	**Assumption**	**Testimony**	**Statistics**	**Anecdote**	**Other**	**No unit**	**Continued**
Common ground	0.129	**0.562**	0.012	0.005	**0.163**	0.012	0.075	0.042
Assumption	0.035	0.701	0.017	0.010	0.075	0.014	0.066	0.083
Testimony	0.006	0.134	0.618	0.016	0.087	0.002	0.034	0.104
Statistics	0.006	0.195	0.042	0.603	0.074	0.002	0.037	0.040
Anecdote	0.037	0.277	0.041	0.013	0.451	0.016	0.059	0.104
Other	0.018	**0.324**	0.006	0.003	0.101	0.166	**0.310**	0.073
No unit	0.008	0.114	0.008	0.003	0.027	0.023	0.440	**0.377**
Continued	0.001	0.036	0.006	0.001	0.012	0.001	0.094	0.849

Table 3.3: Probability confusion matrix for all pairs of annotated types of argumentative discourse units.

4. Disagreement between ‘other’ and ‘no unit’ (0.310). Without clear reason, these classes seem to have been used interchangeably sometimes.

5. Disagreement between ‘ no unit’ and ‘continued’ (0.377). The main reason for such disagreement was that the annotators dealt with discourse markers and connectives inconsistently.

   For example, in case of the subsequent segments (1) *“According to the administration”* and (2) *“the film by Nakoula Basseley Nakoulahad sparked spontaneous riots to defend Muhammad’s honor”*, the first was partly seen as ‘no unit’, although our guideline specified to consider such segments as one unit.

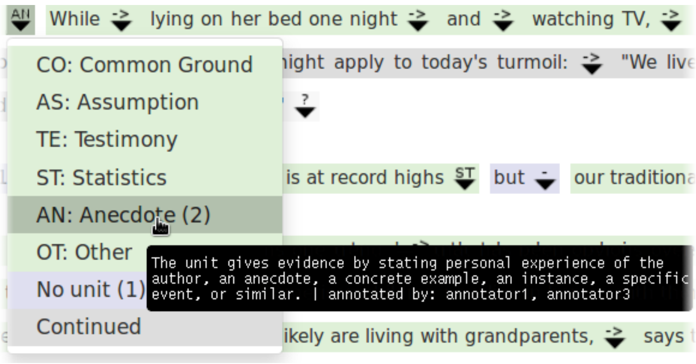
Post-Processing of the Annotations For the final version of the corpus, the corpus segments were consolidated using the *majority vote* for each segment. That means if at least two workers agreed on the class of a segment, the segment was classified accordingly. Else, an external expert selected one of the three suggested classes.

Based on the disagreement analysis and a manual inspection of the annotations, we found a few general misclassifications that could be fixed semi-automatically. While overruling some decisions of the annotators, we thereby achieve a more consistent annotation, which is crucial for learning based on the corpus. In particular, we conducted the following post-processing steps:

- A considerable number of segments was annotated as ‘no unit’, although it should have been merged with the next segment. We reviewed several instances of this problem, such as conditional statements (e.g., of the form “if A then B”) or relations that are not argumentative but temporal or spatial (e.g., of the form “when A then B”). Where necessary, we then merged the respective segments.

- According to our definitions, only non-rhetorical questions should be labelled as ‘no unit’. However, many rhetorical questions were also classified as ‘no unit’, even though they had, in our view, a clear argumentative function: most times implicitly conveying claims, recommendations, or similar. Following our definitions, we reclassified them as ‘assumption’.

- Second person voice segments were often classified as ‘no unit’, possibly due to the unintended interpretation that a unit requires an *explicit* subject. Nearly all of them are appeals, recommendations, or similar. As above, we thus reclassified them as ‘assumption’.



The unit gives evidence by stating personal experience of the author, an anecdote, a concrete example, an instance, a specific event, or similar. | annotated by: annotator1, annotator3

Figure 3.3: Screenshot of the post-processing step, illustrating how the label (*Anecdote*) is selected based on counts of all labels the annotators selected for the respective segment (*Anecdote (2)*, *No unit (1)*).

Type	Total	Mean	Std. dev.	Median	Min	Max	Percent
Common ground	241	0.80	1.53	0	0	13	1.7%
Assumption	9792	32.64	12.42	32	3	86	68.4%
Anecdote	2603	8.68	9.12	7	0	77	18.2%
Testimony	1089	3.63	5.42	2	0	44	7.6%
Statistics	421	1.40	2.76	0	0	19	2.9%
Other	167	0.56	1.64	0	0	24	1.2%
All units	14313	47.71	14.28	46	14	132	100%

Table 3.4: The distribution of types of argumentative discourse units in the created corpus.

In addition to the corrections above, we excluded periods, commas, or similar punctuation at the end of segments and put them in separate 'no unit' segments. This is important to prevent unit type classifiers from misleadingly learning to identify particular types based on these characters.

We used *WAT-SL*, again, for performing the postprocessing step. Figure 3.3 depicts a screenshot for using *WAT-SL* to perform the post-processing step.

Webis-Editorials-16 Corpus

Table 3.4 presents some statistics of the final corpus, which we call *Webis-Editorials-16*, obtained after post-processing. We observe that the most frequent type of argumentative discourse unit is 'assumption' covering almost 68.4% of all units. The 'anecdote' type represents about 18.2%, surpassing the 'testimony' (7.6%), 'statistics' (2.9%), and 'common ground' (1.7%). 'Other', finally, only refers to a very low percentage of units (1.2%). On one hand, this supports the hypothesis that editorials are a rich source for argumentation. On the other hand, it serves as strong evidence that the six proposed types of units cover most units found in editorials.

3.3 Argumentation Strategies across Portals

The *Webis-Editorials-16* corpus serves the investigation of how authors argue in news editorials in order to persuade the readers. In this section, we present some basic findings regarding the *selection* of the types of argumentative discourse units across the three news portals of *Al Jazeera*, *Fox News*, and *The Guardian*.

In particular, Table 3.5 shows detailed statistics about the types of argumentative discourse units in the corpus. Overall, we see that the length of news editorials is quite stable across the three news portals, with a mean between 48.76 (*The Guardian*) and 52.34 units (*Fox News*). Some very short (minimum 14 units) and very long editorials (maximum 132 units) exist, though.

Regarding the distribution of the types, some general tendencies as well as some insightful differences can be observed. Generally, more than two third of an editorial usually comprises assumptions. This is not surprising, as the type 'assumption' covers both claims and any other propositions that may require justification. While *The Guardian* has the highest proportion of assumptions (71.7%), it represents the median for most other types. *Fox News* more strongly relies on 'common ground', with more than one unit of that type on average. Even more clearly, 8.7% of all units in *Fox News* editorials is 'testimony' evidence, about twice as many on average as in *The Guardian* (4.55 vs 2.53). In contrast, *Al Jazeera* seems to put more emphasis on 'anecdote'. At least, it spreads anecdotes across more units (21.0% of all). Interestingly, all three portals behave very similar in their resort to 'statistics' at the same time.

3.4 Identification Method

This section describes our method for the automatic identification of the evidence type of argumentative discourse units in an editorial.

The identification method was performed based on our built corpus *Webis-Editorials-16* (see Section 3.2). To recall, the corpus contains 300 editorials. Each of these editorials is separated into argumentative segments, and every segment is labelled with one of six types. Three types refer to evidence: (1) 'statistics', where the segment states or quotes the results or conclusions of quantitative research, studies, empirical data analyses, or similar, (2) 'testimony', where the

Type	Editorials	Total	Mean	Std. dev.	Med.	Min	Max	Percent
Common ground	**All editorials**	**241**	**0.80**	**1.53**	**0**	**0**	**13**	**1.7%**
	Al Jazeera	59	0.59	0.97	0	0	5	1.2%
	Fox News	104	1.04	2.04	0	0	13	2.0%
	Guardian	78	0.78	1.36	0	0	10	1.6%
Assumption	**All editorials**	**9792**	**32.64**	**12.42**	**32**	**3**	**86**	**68.4%**
	Al Jazeera	3294	32.94	10.79	33	3	65	66.9%
	Fox News	3002	30.02	13.16	30	5	86	57.4%
	Guardian	3496	34.96	12.68	32	6	73	71.7%
Anecdote	**All editorials**	**2603**	**8.68**	**9.12**	**7**	**0**	**77**	**18.2%**
	Al Jazeera	1036	10.36	10.19	8	0	71	21.0%
	Fox News	727	7.27	6.67	6	0	37	13.9%
	Guardian	840	8.40	9.82	6	0	77	17.2%
Testimony	**All editorials**	**1089**	**3.63**	**5.42**	**2**	**0**	**44**	**7.6%**
	Al Jazeera	381	3.81	4.61	3	0	22	7.7%
	Fox News	455	4.55	7.42	2	0	44	8.7%
	Guardian	253	2.53	3.09	2	0	16	5.2%
Statistics	**All editorials**	**421**	**1.40**	**2.76**	**0**	**0**	**19**	**2.9%**
	Al Jazeera	141	1.41	2.60	0	0	17	2.9%
	Fox News	143	1.43	3.23	0	0	19	2.7%
	Guardian	137	1.37	2.37	0	0	12	2.8%
Other	**All editorials**	**167**	**0.56**	**1.64**	**0**	**0**	**24**	**1.2%**
	Al Jazeera	12	0.12	0.41	0	0	2	0.2%
	Fox News	83	0.83	1.20	0	0	5	1.6%
	Guardian	72	0.72	2.49	0	0	24	1.5%
All units	**All editorials**	**14313**	**47.71**	**14.28**	**46**	**14**	**132**	**100.0%**
	Al Jazeera	4923	49.23	12.23	48	21	81	100.0%
	Fox News	5234	52.34	15.64	50	17	123	100.0%
	Guardian	4876	48.76	16.55	46	22	132	100.0%

Table 3.5: Distribution of types of argumentative discourse units in the complete corpus and in the subcorpus of each news portal. Percentages refer to the proportions of units in the respective (sub-) corpus.

segment states or quotes that a proposition was made by some expert, authority, witness, group, organization, or similar, and (3) ‘anecdote’, where the segment states personal experience of the author, a concrete example, an instance, a specific event, or similar. In our identification step, we use the labels of all three evidence types, whereas we consider all remaining types in the corpus (i.e., ‘assumption’, ‘common ground’, and ‘other’) as belonging to the type ‘other’. Moreover, we split the editorials in the corpus into training (60%), validation (20%), and test sets (20%).

Each segment in the corpus spans one sentence or less. Accordingly, it is possible that a sentence includes multiple types (e.g., ‘testimony’ and ‘statistics’), although the proportion of such sentences is very low (less than 5%). We hence decided to simplify the task by identifying only one type for each sentence. In case a sentence has more than one type, we favor evidence types over ‘other’, and less frequent evidence types over more frequent ones. Thereby, we avoid dealing with argumentative text segmentation and multi-type classification.

For identifying evidence types, we relied on supervised learning. The task is relevant to the tasks which are concerned with the pragmatic level of text, such as language function analysis [Wachsmuth and Bujna, 2011] or speech act classification [Ferschke *et al.*, 2012]. We employed several features that capture the content, syntax, style, and semantics of a sentence. Some of these features have been used for the mentioned tasks, others are tailored to our task—based on our inspection of the training set of the corpus.

Lexical Features Previous work on speech acts classification showed a strong positive impact of lexical features [Jeong *et al.*, 200]. In case of evidence types, words such as “study” and “find” are indicators for ‘statistics’, “according” and “states” for ‘testimony’, and “example” and “year” for ‘anecdote’, for instance. We represent this feature type as the frequency of word unigrams, bigrams, and trigrams. We also consider punctuation and digits in our features; quotes play an important role for ‘testimony’, and numbers for ‘statistics’.

Style Features We hypothesize that texts with different evidence types show specific style characteristics. To test this, we use character 1–3-grams, chunk 1–3-grams, function word 1–3-grams, and the first 1–3 tokens in a sentence. Similarly, we expect ‘anecdote’ and ‘testimony’ sentences to be longer than ‘statistics’, which we capture by the number of characters, syllables, tokens, and phrases in a sentence. Moreover, we assess whether a sentence is the first, second, or last within a paragraph.

Syntactic Features Syntax plays a role in different linguistic tasks. For evidence type identification, narrative tenses may be indicators of anecdotes,

#	Feature Type	Accuracy	F_1-Score
1	Lexical features	0.76	0.73
2	Style features	0.74	0.70
3	Syntactic features	0.74	0.71
4	Semantics features	0.71	0.67
1 – 4	**Complete feature set**	**0.78**	**0.77**
	Majority baseline	0.69	0.56

Table 3.6: Effectiveness of each feature type and the complete feature set in identifying evidence types.

for instance. We model syntax simply via the frequencies of part of speech tag 1–3-grams.

Semantic Features We use the frequency of person, location, organization, and misc entities, as well as the proportion of each of these entity types. In many cases, a sentence with evidence refers to specific entities (e.g., a scientific lab in 'statistics'). Also, we use the mean SentiWordNet score of the words in a sentence, once for the word's first sense and once for its average sense [Baccianella *et al.*, 2010]. Moreover, we compute the frequency of each word class of the *General Inquirer* (`http://www.wjh.harvard.edu/~inquirer`).

In our experiments, the sequential minimal optimization (SMO) implementation of support vector machines from Weka performed best among several models on the validation set of the given corpus. There, SMO achieved the highest results for a cost hyperparameter value of 5, which we then used to evaluate SMO on the test set.

Results Table 3.6 shows the effectiveness of our method in terms of accuracy and weighted average F_1-score for each single feature type as well as for the complete feature set. In general, lexical features are the most discriminative, closely followed by the syntax features. All feature types contribute to the effectiveness of the complete feature set. Table 3.7 shows the precision, recall, and F_1-score values for classifying each of the three evidence types as well as the class 'other'. The classifier achieved the highest F_1-score for 'other', followed by 'testimony', ' anecdote', and 'statistics' respectively.

Error Analysis The identification method has a small tendency towards labelling sentences with the majority class 'other'. However, sampling the training set yielded worse results for all classes. Overall, the task is challenging,

Type	**Precision**	**Recall**	**F_1-Score**
Statistics	0.69	0.40	0.50
Testimony	0.63	0.55	0.59
Anecdote	0.55	0.47	0.51
Other	0.84	0.90	0.87

Table 3.7: Precision, recall, and F_1-Score for all four classes in the identification of evidence types.

Evidence Type		**All**	**Arts**	**Econ.**	**Edu.**	**Envir.**	**Health**	**Law**
AN	Anecdote	24.9	**31.6**	22.1	24.1	25.7	21.9	27.5
TE	Testimony	7.7	**11.3**	6.2	9.6	5.1	5.7	7.4
ST	Statistics	3.0	1.5	**5.0**	4.4	3.4	4.9	2.7
OT	Other	64.4	55.6	66.7	62.0	65.8	67.5	62.4
Editorials		28986	1274	3158	1977	1687	2524	2327

Evidence Type		**All**	**Polit.**	**Relig.**	**Science**	**Sports**	**Style**	**Tech.**
AN	Anecdote	24.9	24.4	31.1	24.9	31.1	29.7	23.7
TE	Testimony	7.7	8.4	10.8	6.3	6.5	7.1	6.3
ST	Statistics	3.0	2.1	1.8	3.0	2.8	2.3	2.3
OT	Other	64.4	65.1	56.3	65.8	59.6	60.9	**67.7**
Editorials		28986	12912	243	455	953	960	516

Table 3.8: Distribution of the four evidence types in all editorials and in those of each topic, given in percent. The bottom line shows the number of editorials of each topic.Values discussed in Section **??** are in bold.

and the results we obtained are in line with those that have been reported in speech act classification. Also, the decision to classify each sentence with one of the evidence classes (to avoid segmentation) may render the type identification itself harder. For example, some features such as quotation marks can be helpful to identify 'testimony'. However, if some testimony evidence covers several sentences, the ones which are between the first and the last sentences might be difficult to be identified as part of the testimony.

3.5 Argumentation Strategy across Topics

This section presents an analysis of the argumentation strategies in news editorials within and across topics. Given nearly 29,000 argumentative editorials from the *New York Times*, we developed a new machine learning method for determining an editorial's topic. Then, we applied the developed evidence type identification method (discussed in the previous section) on the editorials. Based on the distribution and the sequential flows of the identified types, we analysed the usage patterns of argumentation strategies within and across 12 different topics. We detected several common patterns that provide insights into the manifestation of strategy principles in editorials. Also, our experiments revealed clear correlations between the topics and the detected patterns.

3.5.1 Analysis Approach

The analysis of argumentation strategies across topics is rooted in our hypotheses that: (1) effective strategies for synthesizing an argumentative text can be derived from the analysis of existing strategies that humans use in high-quality texts, and (2) the decision for preferring one strategy over another is affected by several text characteristics such as genre, provenance, and *topic*.

We approach our study within three steps. Starting from a collection of argumentative news editorials, we (1) categorize the editorials into n topics, (2) identify the evidence types (*statistics, testimony, anecdote*) in each editorial, and (3) analyse the selection and arrangement of evidence types within editorials across topics.

The output of these steps are beneficial for synthesizing an effective argumentative text for a given topic (see Figure 3.4). The third step quantifies the distribution of evidence types and their *flows* [Wachsmuth *et al.*, 2015].

Since we already discussed our method for identifying evidence types in Section 3.4, we only discuss our method for topic categorization in the following.

Topic Categorization

The NYT Annotated Corpus comprises about 1.8 million articles published by the *New York Times* between 1987 and 2007. The corpus covers several types of articles that mainly categorized into 12 topics (the topics are given in Table 3.8) according to which section or sub-section the article is placed into in the news portal's hierarchy. Each article comes with 48 metadata tags that were assigned manually or semi-automatically by employees of the NYT. The tags cover several types of information such as *types of material* (e.g., review,

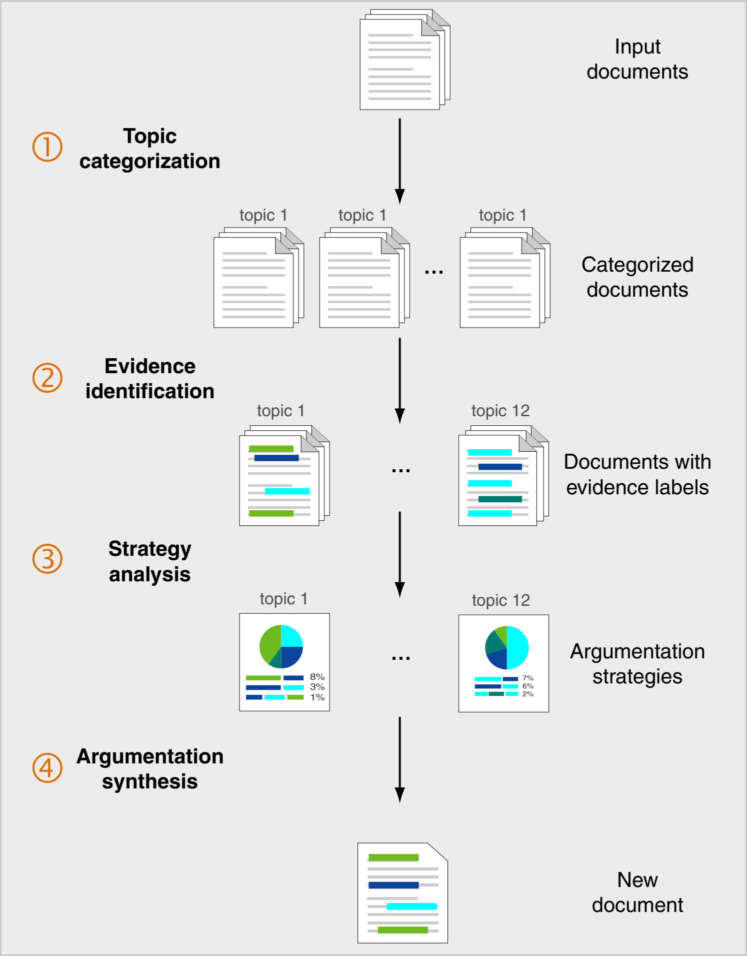


Figure 3.4: Four major steps of an envisioned system for synthesizing argumentative text with a particular strategy. This section presents approaches to the first three steps.

editorial, etc.) and *taxonomic classifiers* (the hierarchy of articles section), among others.

All 28,986 articles tagged as "editorial" are used in our analysis. However, identifying an editorial's topic is not straightforward. While the NYT classifies the topic of most non-editorial articles, only 6% of all editorials are provided with topic information. The remaining 94% are labelled as "opinion". Analysing the corpus, we observed that several tags include terms that describe the content of an article, such as "global warming". Some terms even include the topic itself, such as "Politics and Government". Thus, we exploited these tags to develop a standard supervised classifier for the topic categorization of editorials. In particular, we trained the classifier on all 1.29 million non-editorial articles that are assigned a topic, and then used it to classify editorials with unknown topic.

We used the default configuration of the Weka Naïve Bayes multinomial model with unigram features [Hall *et al.*, 2009], as related studies suggest that this classifier performs particularly well in topic categorization [Husby and Barbosa, 2012]. Since articles may have more than one topic, we label each article with all topics given a probability of at least 0.3 by the classifier. This threshold has been selected based on the training data.

The 6% of editorials, which are provided with "topic" labels in the corpus, were used for testing the effectiveness of our topic classifier. The classifier obtained an accuracy of 0.82 on these articles.

3.5.2 Argumentation Strategy Principles

In this subsection, we analyze strategy patterns across editorials of 12 topics, exploring the selection and the arrangement based on the distribution and the sequential flows of evidence types respectively.

To this end, we applied our topic and evidence type identification methods to all given 28,986 NYT editorials. As the analysis of argumentation strategies depends strongly on the effectiveness of evidence type identification, we considered the impact of the classification errors in the analysis results as follows: For each evidence type t in dataset d, we compute a confidence interval [$\mathit{lower\,bound}, \mathit{upper\,bound}$] for the n sentences that the method labels with t. The interval is derived from the precision and recall of our method for type t (determined on the ground truth): We compute the lower bound as $n \cdot \mathit{precision}(t)$ and the upper bound as $n/\mathit{recall}(t)$.

Based on the mean of the *lower bound* and the *upper bound*, we performed a significance test among the evidence type distribution across topics. In particular, we used the chi-square statistical method with a significance level of 0.001. For the sequential flows, however, a consideration of the impact of misclassified sentences seems unreliable. As each editorial is represented by only one flow, the 60 editorials in the test set of our editorial corpus (see Section 3.2) are not enough for computing precision and recall. In contrast, we again used the chi-square with a significance level of 0.001 for specifying significant differences among the flows.

Distribution of Evidence Types Altogether, the given 28,986 editorials contain 669,092 sentences whose type was classified. As Table 3.8 shows, the most frequent type is ‘other’ (64.4%) according to our method, followed by ‘ anecdote’ (24.9%), ‘testimony’ (7.7%), and ‘statistics’ (3.0%).

In terms of the performed chi-squared tests, all pairs of topic-specific type distributions in Table 3.8 are significantly different from each other with only one exception: ‘arts’ and ‘religion’. This results strongly support the hypothesis that topic influences the usage of evidence types. For anecdotes, the values of both ‘science’ and ‘technology’ differ not significantly from ‘all’. For testimony, ‘law’ does not differ significantly from ‘ all’, and for statistics, the analogue holds for ‘science’ and ‘sports’.

The highest relative frequency of anecdotes is observed for ‘arts’ (31.6%) and ‘religion’ (31.1%), followed by ‘sports’ (31.1%). Matching intuition, authors of arts and religion editorials add much testimony evidence (11.3% and 10.8% respectively). In contrast, anecdotes and testimony are clearly below the average for ‘health’, while statistics play a more important role there with 4.9%, the second highest percentage after ‘economy’ (5.0%).

Sequential Flows of Evidence Types Following related research [Wachsmuth *et al.*, 2015], we designated the *flow* here as a sequential representation of all evidence types in an editorial. Following one of the flow generalizations proposed by Wachsmuth *et al.* [2015], we abstracted flows considering only the changes of evidence types. For example, the flow (AN, AN, TE) for an editorial will be abstracted into (AN, TE). Such an abstraction produces more frequent and thus reliable patterns. Table 3.9 lists the resulting *evidence change flows* that are most common among all editorials.

The most frequent flow is (AN), representing 16.6% of all editorials across topics. This means that about one sixth of all editorials contain only this evidence type. The frequency of (AN) ranges from 9.3% (‘education’) to 26.7% (‘style’), revealing the varying importance of anecdotes in editorials of different topics. The frequency of (AN, TE, AN) is more stable across topics; only ‘health’ and ‘technology’ show notably lower values there (8.8% and 9.5% respectively). For ‘technology’, the percentage is much above the average for some other flows based on AN and TE, such as (AN, TE) (10.7% vs. 6.9%) and (TE, AN) (4.3% vs. 2.6%). Hence, the ordering of evidence seems to make a difference.

In accordance with literature on argumentation in editorials [van Dijk, 1995], many common flows start with an anecdote and end with one. While testimony occurs most often between the anecdotes, the fourth most frequent flow is (AN, ST, AN) (5.3%). This flow occurs particularly often in editorials about environment (8.6%), even though statistics are not that frequent in these editorials (see Table 3.9) — and similar holds for (AN, ST). Such observations

# Evidence Change Flow	All	Arts	Econ.	Edu.	Envir.	Health	Law	Polit.	Relig.	Science	Sports	Style	Tech.
1 (AN)	**16.6**	16.0	13.5	**9.3**	21.3	17.4	17.4	16.2	11.9	20.4	21.8	**26.7**	20.9
2 (AN, TE, AN)	13.2	13.5	10.3	10.2	11.6	**8.8**	14.7	15.1	14.0	13.2	14.1	15.5	**9.5**
3 (AN, TE)	**6.9**	7.9	4.6	7.5	5.9	6.7	8.1	7.0	7.8	7.7	7.2	7.0	**10.7**
4 (AN, ST, AN)	**5.3**	3.6	6.7	4.1	**8.6**	7.2	6.2	4.2	4.9	6.8	7.3	5.8	4.7
5 (AN, TE, AN, TE, AN)	5.3	8.4	3.4	4.3	3.9	2.4	6.4	6.3	7.0	3.1	4.6	3.5	6.6
6 (AN, TE, AN, TE)	4.9	6.2	3.3	4.9	3.5	3.2	5.3	5.7	8.2	4.0	4.8	4.0	4.3
7 (TE, AN)	**2.6**	2.4	2.2	2.3	1.7	2.5	1.8	3.0	<0.5	2.2	2.4	2.3	**4.3**
8 (AN, ST)	2.2	0.7	3.8	1.9	3.1	4.3	2.4	1.5	1.2	3.1	1.9	2.0	1.6
9 (AN, TE, AN, TE, AN, TE)	2.2	2.9	1.3	1.8	1.2	1.1	2.8	2.8	2.9	0.7	1.3	1.7	1.4
10 (AN, TE, AN, TE, AN, TE, AN)	2.0	4.3	1.5	1.8	0.8	1.0	1.9	2.3	5.8	1.3	1.6	1.7	1.4
11 (TE, AN, TE, AN)	1.8	2.2	0.9	2.5	0.7	1.0	1.5	2.3	2.1	0.7	1.8	1.4	1.0
12 (AN, ST, AN, TE, AN)	1.4	0.9	1.8	1.6	2.4	1.2	0.9	1.3	0.8	2.2	1.8	1.3	0.6
13 (ST, AN)	1.3	<0.5	2.2	1.0	1.3	2.8	1.2	0.9	<0.5	2.0	0.7	1.4	2.3
14 (TE, AN, TE)	1.3	1.4	0.8	1.4	0.7	0.9	1.1	1.6	2.1	1.5	<0.5	0.6	1.9
15 (AN, ST, AN, TE)	1.2	0.7	1.6	1.5	2.0	1.5	1.4	1.0	1.2	0.9	1.3	0.9	1.4

Table 3.9: Relative frequency of the top 15 evidence change flows in all editorials and in those of each topic, given in percent. In the flows, the type *Other* is ignored. Values discussed in Section **??** are in bold.

emphasize the role of topic on the arrangement decisions in argumentation strategies.

3.6 Related Work

Many recent publications in the area of computational argumentation are concerned with the construction of annotated resources, which is a fundamental step towards building automatic systems that analyze the argumentative structure of texts. Unlike most previous corpora, the annotations of our *Webis-Editorials-16* corpus provide a classification of argumentative units based on their content. In the most simple case, other works distinguish only argumentative discourse units from other text, as in [Al-Khatib *et al.*, 2016a]. Some corpora contain unit annotations based on the role units take in arguments such as premise or conclusion [Habernal and Gurevych, 2015; Stab and Gurevych, 2014a]. Such annotations represent the general argumentative structure, but they do not encode the means an author uses to persuade the readers. Previous corpora that contain content-based unit annotations are not suitable for analysing argumentation strategies due to their annotation method or to the genre of the contained texts. Similar to our corpus, the CE corpus of Wikipedia articles [Aharoni *et al.*, 2014; Rinott *et al.*, 2015] also considers types of evidence (anecdotes, statistics, and testimony). However, evidence is annotated only where it relates to a set of given topics, so no complete annotation of the article's discourse is provided.

Most existing work on news editorials in computational linguistics studies sentiment and opinions [Bal, 2009; Wilson and Wiebe, 2003; Yu and Hatzivassiloglou, 2003]. A first conceptual study of the relation between the opinions in a news editorial and its argumentative structure is described in Bal and Saint-Dizier [2009]. Besides, up to our knowledge, the only work in this regard is the work of Chow [2016] on Chinese editorials. Unfortunately, the annotation of their corpus is restricted to the argumentativeness of paragraphs as a whole, which makes the corpus unsuitable for analysing argumentation strategies.

Regarding evidence type classification, Rinott *et al.* [2015] have proposed a supervised learning model for identifying context-dependent evidence in Wikipedia articles. While the authors target the same evidence types that we consider in our work, they approach a different task. In particular they classify only evidence that is *related to given claims*. Hence, a comparison of their effectiveness results with ours would be meaningless. Moreover, some of their features rely on resources that are not publicly available (e.g., lexicons), which is why could not resort to their approach or compare it to ours.

The NYT Annotated Corpus has been analysed in several papers. Among others, Li *et al.* [2016] and Hong and Nenkova [2014] have used the metadata tag "abstract", which contains a manually created article summary. Other tags, such as those for people, locations, and organizations mentioned in an article, have been used by Dunietz and Gillick [2014].

3.7 Summary

Although news editorials are considered as one of the purest forms of argumentative text, still, few works exist in computational linguistics that study them. In this chapter, we presented a new model for argumentation strategies in editorials based on the pragmatic attribute of the type of argumentative units. Then, we described the development of an annotated corpus for the mining of an editorial's argumentative discourse and the analysis of its argumentation strategy. We expect that such an analysis will contribute to the computational text generation systems and the writing assistance tools through improving the persuasiveness of the generated monological argumentation.

We proposed the first model for exploring argumentation strategies that captures the type of each unit of an argumentative text. Our proposed model, unlike previous ones, allows for a complete annotation of the text on a fine-grained level. According to this model, we built a new corpus with 300 news editorials which are manually annotated with six types of argumentative units. Despite the resort to editorials, the so far under-resourced text genre, the corpus contains a considerably larger number of unit annotations than comparable existing corpora. The corpus was developed carefully in which the inter-annotator agreement and the reliability of the resulting annotations were closely examined. We used the new corpus to conduct an analysis study of argumentation strategies in editorials and to present insightful findings that indicate a collection of strategy principles in news editorials within and across different news portals.

Moreover, we presented an analysis of argumentation strategies in news editorials within and across topics. Given nearly 29,000 argumentative editorials from the *New York Times*, we developed two machine learning methods, one for determining an editorial's topic, and one for identifying evidence types in the editorial. Based on the distribution and the sequential flows of the identified types, we analysed the usage patterns of argumentation strategies among 12 different topics. We detected several common patterns that provide insights into the manifested principles of argumentation strategies. Also, our experiments revealed clear correlations between the topics and the detected patterns.

Overall, in this chapter, we acknowledged the first, second, and third research questions of this thesis. Though we were generally successful, there were some limitations in our study. While not being considered in our proposed model, we are aware that an argumentative unit sometimes may have more than one type. For example, a unit may represent both testimonial and statistical evidence. To create a clear classification setting, we decided to assign exactly one type to each unit, though, and to give the annotators clear instructions about what type to prefer in what context. Aside from that, we observed rather low agreement for the infrequent type 'common ground'. Still, the resulting annotations may be valuable for research questions related to argumentation quality or persuasiveness. For instance, some authors use specific terms such as "in fact" or "for sure" before assumptions to let them appear as common ground.

Regarding the argumentation strategies in editorials, and in particular, in relation to the pragmatic attributes of argumentative units, our work forms a constructive step towards an automatic formulation of high-quality principles, yet, there are rooms for various research directions here, such as accounting for more pragmatic attributes including the roles of claims and conclusion, formulating principles accounting for the interactions between the argument types (e.g., if two types should not appear together in the same paragraph), and ranking principles on the basis of their quality.

Chapter 4

Pragmatic Deliberative Strategies

> *Deliberation and debate is the way you stir the soul of our democracy.*
> *(Jesse Jackson)*

Deliberation is the type of discussions in which the aim is to reach a consensus on the best choice from a set of possible actions [Walton, 2010]. Deliberation is influential for making decisions in different processes including *collaborative writing*. Studies have shown the positive impact of deliberation on the quality of several document types, such as scientific papers, research proposals, political reports, and Wikipedia articles, among others [Kraut *et al.*, 2012].

However, deliberative discussions may fail, either by agreeing on the wrong action, or by failing to reach a consensus. While the former is hard to measure, the latter is, for example, clearly reflected in the number of disputed discussions on Wikipedia [Wang and Cardie, 2014].

Although a consensus can never be guaranteed, a deliberative argumentation strategy of a discussion's participants makes it more likely [Kittur *et al.*, 2007]. With *strategy*, we here mean how the moves that participants can take during the discussion is selected and arranged. By a move, we mean a discussion's turn with a particular attribute. Overall, such a selection and arrangement is effective if it leads to a consensus among participants.

To reach consensus, every participant has to understand the current state of a discussion and to come up with a next deliberative move that *best* serves the discussion. For newcomers, this requires substantial effort and time, especially when a discussion grows due to conflicts and back-and-forth arguments. Here, automated tools can help by annotating ongoing discussions with a label for each move or by providing a textual summary of past moves [Zhang *et al.*, 2017a,b]. A way to go beyond that is to let the tool *recommend how the best possible moves* should look like according to an effective strategy. The recommendation may consider various pragmatic attributes of the moves.

As the base for developing such recommendation tools, two fundamental steps are addressed in this chapter: (1) modelling deliberative discussions in light of the aim of consensus, and (2) operationalising the model in order to identify different argumentation strategies and to learn about their effectiveness. These steps are in line with the fourth and fifth research questions of this thesis (see Chapter 1, Section 1.2).

Different models of deliberative discussions have been proposed in previous studies. These models were developed based on expert analyses of a *small* set of sampled discussions (see Section 4.5). However, the small size, in fact, confines the ability to develop a *representative* model, which should ideally cover a wide range of moves while being abstract to fit the majority of discussions.

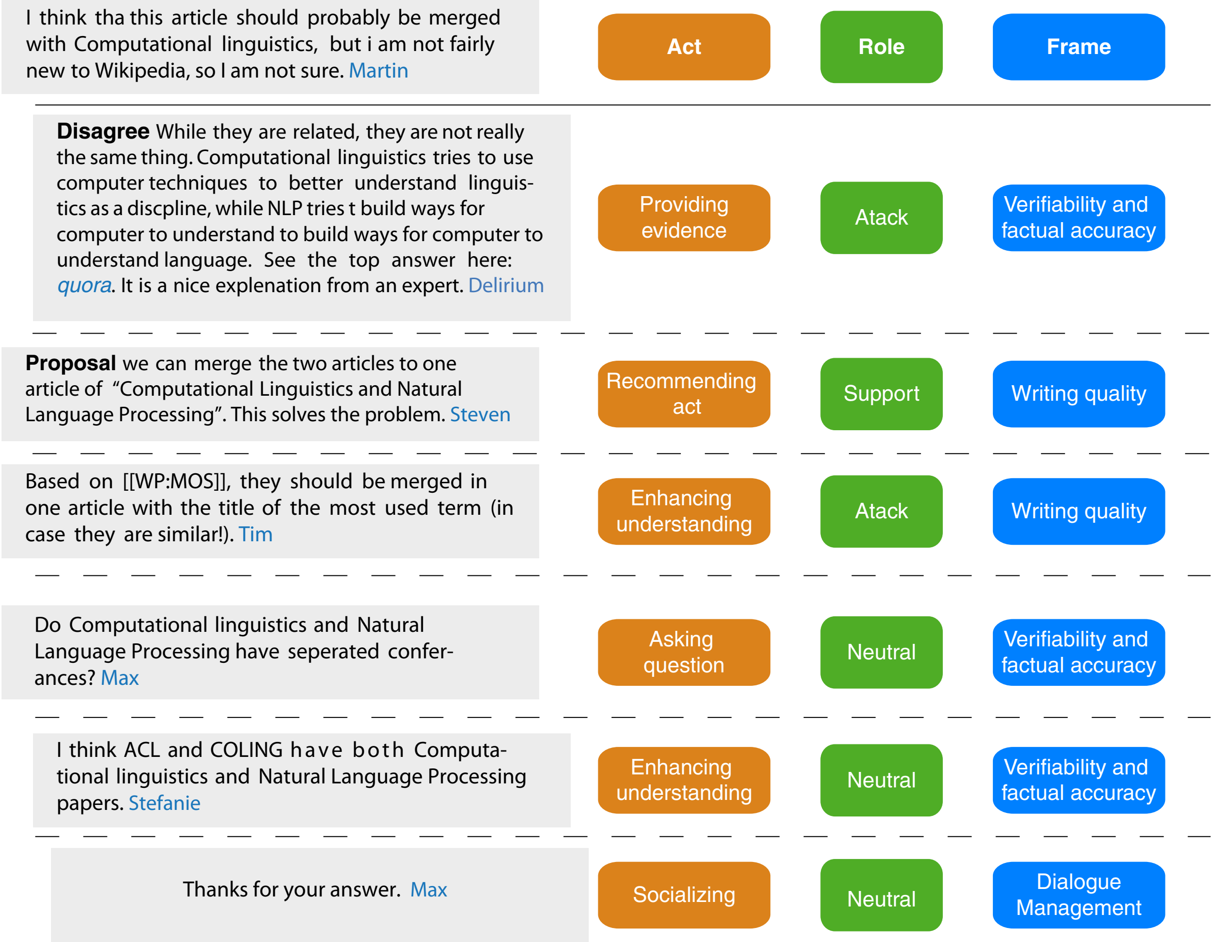
To overcome this limitation, we propose to derive a model statistically from a large set of discussions. We approach this based on different types of metadata that people use to describe their moves on Wikipedia talk pages, the richest source of deliberative discussions on the web.

Particularly, we extract the entire set of about six million discussions from all English Wikipedia talk pages. We parse each discussion to identify its structural components such as turns, users, and time stamps. Also, we store four types of *metadata* from the turns: the user tag, a shortcut, an in-line template, and links. To learn from the metadata, we cluster the types' instances based on their semantic similarity. Then, we map each cluster to a specific concept (e.g., 'providing a source'), and the related concepts into a set of categories (e.g., 'providing evidence'). Table 4.2 shows the categories of our model.

Analysing the distribution of these categories, we find that each turn ideally have (1) one of six categories that we call *dialog acts*, (2) one of three categories that we call *argumentative roles*, and (3) one of four categories that we call *frames*. As such, our model is in line with three well-established theories in pragmatic: *speech act theory* [Searle, 1969], *argumentation theory* [Peldszus and Stede, 2013], and *framing theory* [P. Levin *et al.*, 1998]. A model instance is sketched in Figure 4.1.

Based on the model, we generate a new large-scale corpus using the metadata automatically: *Webis-WikiDebate-18* corpus. Basically, if a turn in a discussion has metadata that belongs to a specific category according to the above-mentioned analysis, it is labelled with that category. The corpus includes 2400 turns labelled with a dialogue act, 7437 turns labelled with a role, and 182,321 turns labelled with a frame.

To operationalize our model, we train three supervised classifiers for acts, roles, and frames on the built corpus. The classifiers employ a rich set of linguistic features that has been shown to be effective in similar tasks [Ferschke *et al.*,



I think tha this article should probably be merged with Computational linguistics, but i am not fairly new to Wikipedia, so I am not sure. Martin	Act	Role	Frame
Disagree While they are related, they are not really the same thing. Computational linguistics tries to use computer techniques to better understand linguistics as a discpline, while NLP tries t build ways for computer to understand to build ways for computer to understand language. See the top answer here: *quora*. It is a nice explenation from an expert. Delirium	Providing evidence	Atack	Verifiability and factual accuracy
Proposal we can merge the two articles to one article of "Computational Linguistics and Natural Language Processing". This solves the problem. Steven	Recommending act	Support	Writing quality
Based on [[WP:MOS]], they should be merged in one article with the title of the most used term (in case they are similar!). Tim	Enhancing understanding	Atack	Writing quality
Do Computational linguistics and Natural Language Processing have seperated conferances? Max	Asking question	Neutral	Verifiability and factual accuracy
I think ACL and COLING have both Computational linguistics and Natural Language Processing papers. Stefanie	Enhancing understanding	Neutral	Verifiability and factual accuracy
Thanks for your answer. Max	Socializing	Neutral	Dialogue Management

Figure 4.1: Left: An excerpt of a discussion in a Wikipedia talk page. Right: The labels of each turn in the discussion according to our proposed model.

2012]. The results of our experiments suggest that we are able to predict the labels with a comparable performance to the one achieved in similar tasks.

The remainder of this chapter is structured as follows: Section 4.1 proposes a new model for deliberative moves in Wikipedia discussions. Section 4.2 details the construction of a large-scale corpus for the proposed model. Section 4.3 explains the methodology for identifying the attributes of turns within Wikipedia discussions. Section 4.4 discusses the possible analysis of strategies in Wikipedia discussion. Section 4.5 reviews the related work and Section 4.6 concisely summarizes the chapter.

4.1 Model

The web is full of platforms where users can share and discuss opinions, beliefs, and ideas. In case of deliberative discussions, in particular, participants try to agree on the best action from several choices. Apparently, the participants there follow a strategy to achieve an effective discussion, i.e., each participant tries to come with the best deliberative move that leads to achieve consensus.

The numerous deliberative discussions on these platforms do not only include user-written text, but also different types of metadata that users add to benefit the coordination between them. For example, users vote for specific posts, summarize texts, include references to the sources they use, refer to the discussion policies of a platform, or report bad behaviour of others. Overall, the available metadata represents a valuable resource that provides insights into three main aspects of a discussion: The functions of users' moves, the users' roles, and the discussion topics along with their flows. We propose to exploit the metadata for modelling argumentation strategies in deliberative discussions.

To this end, we proceed in four general steps: (1) *metadata inspection*, which includes investigating the used metadata and its functions, (2) *concept origination*, where clusters of similar metadata are created and mapped to corresponding concepts, (3) *concept categorization*, where similar concepts are abstracted into a defined set of categories, and (4) *category composition*, where possible overlaps between categories should be identified.

The idea of this approach is not only to model deliberative discussion, but also to allow for an operationalisation of the resulting model by providing a dataset for training classifiers. In particular, the metadata can also be used to label discussions based on distant supervision [Mintz *et al.*, 2009]. In the following, we describe how we implemented our approach to derive a new model of Wikipedia discussions, using the metadata provided by the participants.

4.1.1 Discussion Parsing

As part of the management policies of Wikipedia, each article has an associated page called 'Talk'. The main purpose of the talk page is to allow users to discuss how to improve the article through specific actions that they agree on. Most of these discussions can be seen as deliberative, since all participants share the same goal: the consensus on the best action to improve the article.

When a user has a proposal on how to improve an article, she can open a discussion on the article's talk page, specifying a title and the main topic of discussion. Usually, the topic denotes a suggestion to perform a specific action, such as adding, merging, or deleting certain content of the article, among others. Ideally, multiple users then participate in the discussion about whether the action would improve the article or not.

Each single comment written by a user at a specific time is called a ‘turn’. A turn may reply directly to the main topic of the discussion or to any other turn. Overall, a discussion consists of the title, the main topic, and a number of turns written by users with attached time stamps (see Figure 4.2). Based on a manual inspection of the turns’ texts of 50 discussions, we found four general types of metadata used by the participants: *user tags*, *shortcuts*, *inline-templates*, and *external links*.

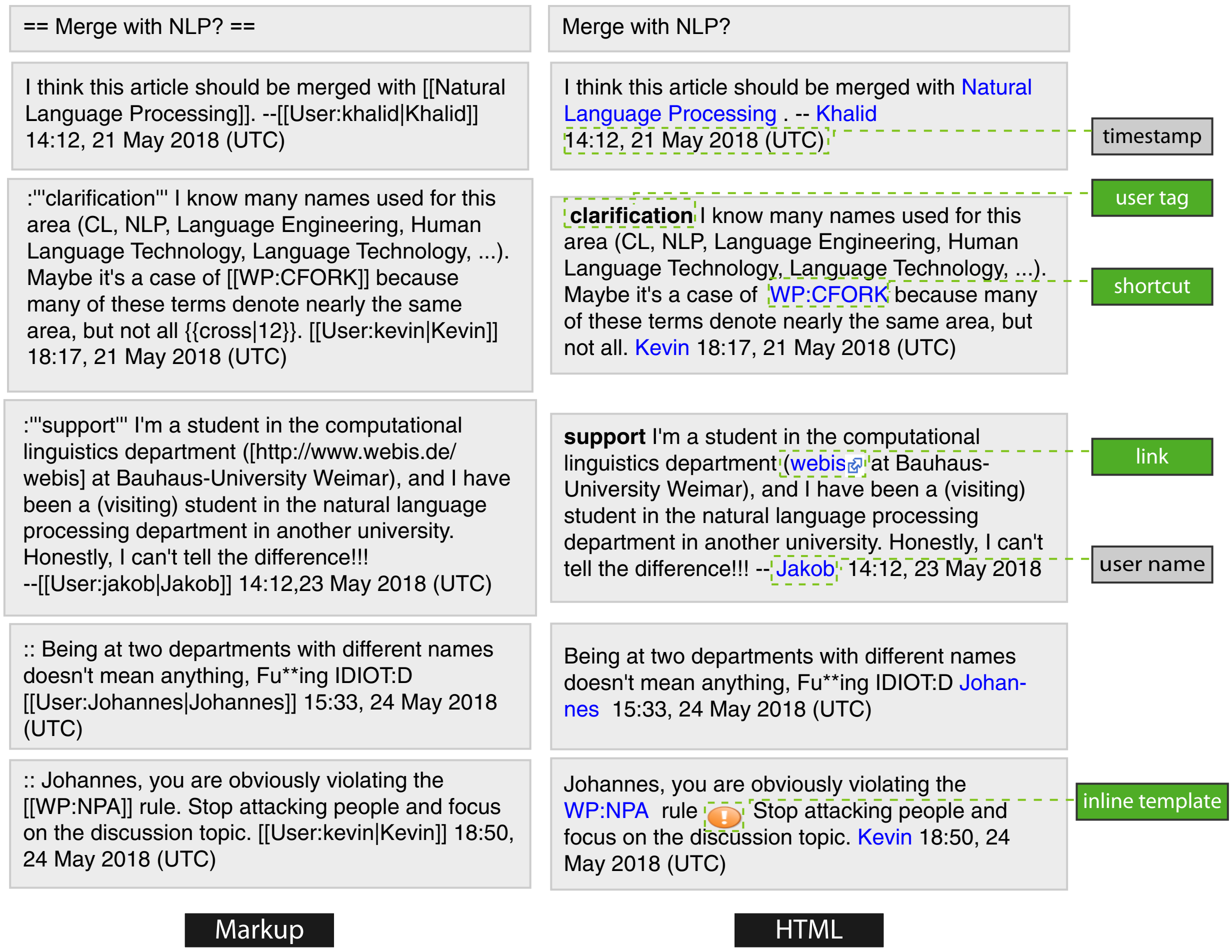
To derive a model from Wikipedia, we need to extract and parse the whole set of discussions on all talk pages, including both ongoing and closed ones. Particularly, the creation of a discussion is solely done by the users following the Wiki markups, which is a particular syntax for formatting Wikipedia articles and their discussions in the talk pages. This syntax is compiled by Wikipedia and converted to HTML. Figure 4.2 shows a discussion in the markup format and the resulted HTML highlighting several instances of the metadata there. However, parsing the discussion is all but trivial. While Wikipedia describes the required format of the different parts of a discussion in detail, not all users follow this format, often forgetting required symbols or mistakenly confusing a symbol with another one.

In the implementation of our approach, we built upon the English Wikipedia dump created on March 1st, 2017. Given a Wikipedia dump, we parsed it in the following steps:

Extraction of Talk Pages First, we obtained the talk pages. We used the Java Wikipedia Library (JWPL) from Zesch *et al.* [2008], which converts a Wikipedia dump into a database that provides an easy-to-use access to the dump components.

Extraction of Discussions Next, we extracted the discussions from the talk pages. To this end, we developed several regular expressions that capture the format for starting and ending a discussion.

Identification of Structure Given the discussion, we identify their structure. We created a specific template to mine the title. The topic of the discussion is simply given by the first turn. To identify and correctly segment all users’ turns, we use several indicators, for instance, indentations.



== Merge with NLP? ==

I think this article should be merged with [[Natural Language Processing]]. --[[User:khalid|Khalid]] 14:12, 21 May 2018 (UTC)

:'''clarification''' I know many names used for this area (CL, NLP, Language Engineering, Human Language Technology, Language Technology, ...). Maybe it's a case of [[WP:CFORK]] because many of these terms denote nearly the same area, but not all {{cross|12}}. [[User:kevin|Kevin]] 18:17, 21 May 2018 (UTC)

:'''support''' I'm a student in the computational linguistics department ([http://www.webis.de/ webis] at Bauhaus-University Weimar), and I have been a (visiting) student in the natural language processing department in another university. Honestly, I can't tell the difference!!! --[[User:jakob|Jakob]] 14:12,23 May 2018 (UTC)

:: Being at two departments with different names doesn't mean anything, Fu**ing IDIOT:D [[User:Johannes|Johannes]] 15:33, 24 May 2018 (UTC)

:: Johannes, you are obviously violating the [[WP:NPA]] rule. Stop attacking people and focus on the discussion topic. [[User:kevin|Kevin]] 18:50, 24 May 2018 (UTC)

Merge with NLP?

I think this article should be merged with Natural Language Processing . -- Khalid 14:12, 21 May 2018 (UTC)

clarification I know many names used for this area (CL, NLP, Language Engineering, Human Language Technology, Language Technology, ...). Maybe it's a case of WP:CFORK because many of these terms denote nearly the same area, but not all. Kevin 18:17, 21 May 2018 (UTC)

support I'm a student in the computational linguistics department (webis at Bauhaus-University Weimar), and I have been a (visiting) student in the natural language processing department in another university. Honestly, I can't tell the difference!!! -- Jakob 14:12, 23 May 2018

Being at two departments with different names doesn't mean anything, Fu**ing IDIOT:D Johannes 15:33, 24 May 2018 (UTC)

Johannes, you are obviously violating the WP:NPA rule ⚠ Stop attacking people and focus on the discussion topic. Kevin 18:50, 24 May 2018 (UTC)

Figure 4.2: Left: An excerpt of a discussion within a Wikipedia talk page in markup format. Right: the discussion in HTML format, with different types of metadata highlighted.

Identification of Turn Metadata Finally, we identified the metadata of each turn. We analysed how users include the tags in their turns, finding that they usually start a turn with a user tag in triple quotation marks. A shortcut starts with 'WP:', followed by a name for the shortcut, together encapsulated by brackets. Also templates are placed between double parentheses, but they do not start with 'WP:'. Links are simply identified by either of the affixes 'www.' and 'http:'.

Corpus Component	**Instances**
Page	5 807 046
Discussion	5 941 534
Discussion template	144 824
Turn	20 816 860
Registered users	739 244
Turns by registered users	10 926 670
Turns by anonymous user	9 890 190
Tag	99 889
Shortcut	425 583
Inline template	3 382 443
Links	4 824 085
Turns with tag and shortcut	2 347
Turns with tag and inline template	61 521
Turns with shortcut and inline template	170 065

Table 4.1: Instance counts of the different components of the Webis-WikiDiscussions-18 corpus.

4.1.2 The Webis-WikiDiscussions-18 Corpus

The result of the parsing process is a large-scale corpus of Wikipedia discussions. In particular, the *Webis-WikiDiscussions-18* corpus we created contains about six million discussions, consisting of about 20 million turns. The turns comprise around 74,000 different tags with a total of about 100,000 instances, around 7000 different shortcuts with about 400,000 instances, and around 51,000 different inline templates with about 3.3 million instances. Half of the turns are written by registered users. Table 4.1 lists the exact instance counts.

4.1.3 Model Derivation

We now explain how we derived a model of deliberative discussions from the metadata obtained in the previous subsection. The derivation process includes the four steps outlined in the beginning of this section.

Metadata Inspection As mentioned before, a turn on Wikipedia includes up to four types of metadata: user tag, shortcut, inline template, and external link. Each type has a specific definition, a suggested usage, and properties that we discuss in the following paragraphs.

A *user tag* is a short text that a discussion participant uses to describe or summarize her contribution. Most tags indicate the main function of the contribution, such as 'proposal' and 'question'. Users can define any free-text tag they want using a noun, verb, etc. Analysing the tags in the crawled discussions, we found the most frequent tags to be rather general and meaningful, whereas less frequent tags often capture aspects of the topic of discussion, such as 'Israel-Venezuela relations' in the discussion about 'Foreign relations of Israel'. Sometimes, tags are used to get the attention of specific users, such as 'For who reverted my change'. Unfortunately, many users also misuse tags, for example, by including the whole turn's text there or by encoding meaningless information.

A *shortcut* is an abbreviation text link that redirects the user to some page on Wikipedia. Although shortcuts may link to any Wikipedia page, they are often used to link to rules or policies. The respective pages belong to one of five categories:

(1) Behavioural guidelines: Pages that describe how users should interact with each other (e.g., during a discussion). This includes that users should be "good-faith" (WP:AGF), among others.

(2) Content guidelines: Pages that describe how to identify and include information in the articles, such as those about how an article should have reliable and accepted sources (WP:RELIABLE).

(3) Style guidelines: Pages that contain advice on writing style, formatting, grammar, and similar. This includes how to write the introduction (WP:LEAD) and headings (WP:HEADINGS), and what style to use for the content (WP:MOS).

(4) Notability guidelines: Pages that illustrate the conditions of testing whether a given topic warrants its own article. The most common shortcut in this category is (WP:N).

(5) Editing guidelines: Pages that provide information on the metadata of articles, such as the articles' categories (WP:CAT).

Overall, we found that shortcuts are used particularly frequently for style, content, and behavioural guidelines in Wikipedia discussions. The participants mainly use them to discuss the impact of applying an action that has been proposed to be performed on a Wikipedia article. For example, adding a lot of content to the introduction of an article may violate the style guidelines. A user can indicate this by referring to the style rules using the shortcut (WP:LEAD).

An *inline template* is a Wikipedia page that has been created to be included in other pages. Inline templates usually comprise specific patterns that are used in many articles, such as standard warnings or boilerplate messages. For example, there are templates for including a quotation, citation, or code, among others. Templates are used frequently in Wikipedia discussions, with the objective of writing readable and well structured turns.

An *external link*, finally, points to a web page outside Wikipedia. External links occur both in Wikipedia articles and in Wikipedia discussions. While there are some restrictions for using them in articles, they can be used without restriction in discussions. We found that these links are used in Wikipedia discussions to point to evidence on the linked web pages. In particular, they often link to research, news, search engines, educational institutions, and blogs.

Concept Origination We analysed the usage of the four types of metadata in Wikipedia discussions and identified a set of concepts. Each concept primarily describes the turn that a participant writes:

User tags: We explored all 376 tags that occurred at least 35 times. As discussed before, the tags could be seen as a keywords that describe the turns. Often, different tags refer to the same concept, for example, ‘conclusion’, ‘summary’, and ‘overall’ all capture the concept of ‘summarisation’, i.e., the main function of the respective turns is to summarize the discussion. As a result, we identified 32 clusters. We examined some turns belonging to each cluster, and mapped each cluster to a specific concept that describes it.

Shortcuts: Analogously, we explored all 99 shortcuts that occurred at least 900 times. Since the shortcuts themselves do not describe the turn, but rather the policy pages they refer to, we analysed these pages by reading their first paragraphs and by checking their relation to the pages of the five shortcut categories we discussed before (e.g., ‘behavioural’). This resulted in the identification of 12 concepts. We found that each shortcut concept describes the main quality aspect that a turn addresses. For example, ‘writing content’ specifies how a proposed action influences the quality of the writing of the associated article.

Inline-templates: Our investigation of this type led only to concepts that we already found before for the tags and shortcuts, such as ‘stating a fact’.

External links: Similar to the templates, we identified concepts in the links that we also observed in the tags, such as ‘providing source’.

Concept Categorization The concepts that we identified in the user tags can be grouped into six categories that we see as 'dialogue acts':

1. *Socializing:* All concepts related to social interaction, such as thanking, apologizing, or welcoming other users.

2. *Providing evidence:* All concepts concerning the provision of evidence. Evidence may be given in form of a quote, an example, a fact, references, a source, and similar.

3. *Enhancing the understanding:* All concepts related to helping users understand the topic of discussion or a discussion itself. This can be done by giving background information, by clarifying misunderstandings, or by summarizing the discussion, among others.

4. *Recommending an act:* All concepts proposing to add a new aspect to the discussion, to ask more users to participate in the discussion, or to come up with an alternative to the proposed action.

5. *Asking a question:* All concepts related to questions serving different purposes, such as obtaining information on the topic of discussion, requesting reasons of specific decisions, and similar.

6. *Finalizing the discussion:* All concepts related to the decision of a discussion, including reporting the decision, committing it, or closing the discussion to move it to the archive.

In addition, we identified three further categories based on the user tags, which we see as relevant to 'argumentation theory'. Each represents a role that specifies a relation between the turn and the topic of discussion or between the turn and another turn:

1. *Support:* The turn agrees with or supports another turn or the topic of discussion, for instance, by providing an argument in favor of the one in the 'supported' turn.

2. *Attack:* The opposite of the 'support relation', i.e., the turn disagrees or attacks another turn or the topic of discussion.

3. *Neutral:* The turn has a neutral relation to another turn or the topic of discussion when it neither support nor attack it.

Finally, we identified four categories based on the shortcuts that we see as relevant to 'framing theory'. They target a quality dimension of the article or of the discussion itself:

1. *Writing quality:* Turns that mainly address issues related to the quality of writing of an article, such as whether adding new content complies with the style guidelines for lead sections, the layout, or similar.

2. *Verifiability and factual accuracy:* Turns that address issues related to the quality of references, the reliability of sources, copyright violations, plagiarism, and similar.

3. *Neutral point of view:* Turns that focus on a fair representation of viewpoints and on how to avoid bias.

4. *Dialog management:* Turns that concentrate on issues related to managing the discussion, such as reporting abusive language, preserving respect between users, encouraging newcomer participants, and similar.

Category Composition Given these categories, we investigated the interaction between them in 20 discussions, for instance, to see whether the categories are orthogonal. We found that each turn may have one dialogue act, one role, and one frame at the same time. For example, a turn may support another turn by providing evidence (say, of the type 'source'), while focusing on the writing quality frame. Table 4.2 shows the categories of our model and their concepts.

Unlike previous approaches to the modelling of discussions on Wikipedia, our model decouples the three principle dimensions of discussions: *dialogue acts*, *argumentative roles*, and *frames*. We argue that the distinction of these dimensions is key to develop an tool for supporting discussion participants.

4.2 Corpus Construction

To create a corpus for our model, we decided to rely again on the metadata. In particular, for each category in our model, we retrieved the metadata instances that had been used to derive the category, and then labeled any turn that included any metadata with this category. For example, the user tag 'overall' was used to originate the concept 'summarisation', which was abstracted into the category 'enhancing the understanding'. Accordingly, all the turns that included this tag were labelled with the category 'enhancing the understanding'. This process is in line with the distant supervision paradigm. In case a turn contained metadata belonging to two categories, we excluded it from the corpus. This happened with some shortcuts in particular. Basically, such cases indicate that some turns address more than one frame.

Dimension	Category	Concepts
Dialogue act	Socializing	(1) Thank a user, (2) Apologize from a user, (3) Welcome a user, (4) Express anger
	Providing evidence	(1) Provide a quote, (2) Reference, (3) Source, (4) Give an example, (5) State a fact, (6) Explain a rational
	Enhancing the understanding	(1) Provide background info, (2) Info on the history of similar discussions, (3) Introduce the topic of discussion, (4) Clarify a misunderstanding, (5) Correct previous own or other's turn, (6) Write a discussion summary, (7) Conduct a survey on participants, (8) Request info
	Recommending an act	(1) Propose alternative action on the article, (2) Suggest a new process of discussion, (3) Propose asking a third party
	Asking a question	(1) Ask a general question about the topic, (2) Question a proposal or arguments in a turn
	Finalizing the discussion	(1) Report the decision, (2) Commit the decision, (3) Close the discussion
Argumentative role	Support	(1) Agree, (2) Support
	Neutral	(1) Be neutral.
	Attack	(1) Disagree, (2) Attack, (3) Counter-attack
Frame	Writing quality	(1) Naming articles, (2) Writing content, (3) Formatting, (4) images, (5) Layout and list
	Verifiability and factual accuracy	(1) Reliable sources, (2) Proper citation (3) Good argument
	Neutral point of view	(1) Neutral point of view
	Dialogue management	(1) Be bold. (2) Be civil, (3) Don't game the system

Table 4.2: The concepts covered by each category of each of the three principle dimensions of our model.

Dimension	**Category**	**Turns**	**Prec.**
Dialogue act	Socializing	83	0.71
	Providing evidence	781	0.49
	Enhancing the understanding	671	0.56
	Recommending an act	137	0.82
	Asking a question	106	0.71
	Finalizing the discussion	622	0.71
Argumentative role	Support	2895	1.00
	Neutral	1937	0.63
	Attack	2605	1.00
Frame	Writing quality	19893	0.51
	Verifiability and factual ac.	72049	0.89
	Neutral point of view	60007	0.89
	Dialogue management	30372	0.74

Table 4.3: Number of turns in each category of Webis-WikiDebate-18 corpus and the precision of sampled turns for each category according to an expert.

Webis-WikiDebate-18 Corpus Overall, the corpus comprises 2400 turns labelled with one of the six dialogue act categories, 7437 turns with one of the role categories, and 182,321 turns with one of the frame categories. In order to verify the reliability of the corpus, we randomly sampled about 100 turns from each category, ensuring that all the category's concepts are taken into consideration. The turns in the samples were verified regarding whether they belong to the assigned category by a worker hired from the freelancing platform `upwork.com`. The worker was a native speaker of English with deep expertise in writing. Table 4.3 shows statistics of the corpus, including the percentage of turns in each sample that belong to the assigned category according to the expert. In general, this verification result is comparable to the inter-annotator agreement achieved in some related studies [Ferschke *et al.*, 2012].

4.3 Identification Method

Based on the Webis-WikiDebate-18 corpus, we developed three supervised methods: one for identifying the dialogue acts, one for the roles, and one for the frames. Since we do not aim at proposing a novel approach for the identification tasks, but rather at showing the ability to operationalise the proposed model, we follow existing work that has proposed methods for the tasks at hand. Particularly, we implement a rich set of features that have been used by others before. These features capture lexical, semantic, style, and pragmatic properties of turns.

In short, we used the following features: The frequency of word 1–3-grams, character 1–3-grams, chunk 1–3-grams, function word 1–3-grams, and of the first 1–3 tokens in a turn. The number of characters, syllables, tokens, phrases, and sentences in a turn. the frequencies of part-of-speech tag 1–3-grams. The mean SentiWordNet score of the words in a turn (`http://sentiwordnet.isti.cnr.it`). The frequency of each word class of the General Inquirer (`http://www.wjh.harvard.edu/~inquirer`). The depth level of turns in the discussion. For the role identification, we had additional features that consider the target of the support or attack role (the parent turn), namely, the cosine, euclidean, manhattan, and jaccard similarity between turn and parent turn.

Experiments As a preprocessing step, we cleaned the turns in the *Webis-WikiDebate-18* Corpus by removing all the metadata: user tags, shortcuts, user and time stamps, etc. Then, we grouped the turns that belong to the dialogue act categories in a single dataset (say, the 'dialogue act dataset'). The same was performed for the turns belonging to roles and frames. We then split each of the three datasets randomly into training (60%), development (20%), and test (20%) sets. We ensured that turns from the same discussion should appear only in either of the split sets, in order to avoid biasing the classifiers by topical information.

We trained different machine learning models on the training sets and evaluated them on the development sets. The models included those which had been used before in similar tasks, such as naive bayes, logistic regression, support vector machine, and random forest. We tried both under and over-sampling on the training sets. The best results in the three tasks were achieved by using support vector machine without sampling the training sets. We used the support vector machine implementation from the LibLinear library [Fan *et al.*, 2008] on the test sets and report the results in Table 4.4.

Results Overall, the three identification methods achieved results that are comparable to the results of previous methods on the corresponding tasks [Ferschke *et al.*, 2012; Zhang *et al.*, 2017a]. We obtained the best results in the frame task, followed by roles and then dialogue acts. Apparently, the results

Dimension	Category	Prec.	Rec.	F_1
Dialogue act	Socializing	0.14	0.11	0.13
	Providing evidence	0.63	0.77	0.69
	Enhancing the understand.	0.62	0.55	0.58
	Recommending an act	0.13	0.09	0.10
	Asking a question	0.80	0.19	0.31
	Finalizing the discussion	0.67	0.74	0.71
Argumentative role	Support	0.53	0.59	0.56
	Neutral	0.55	0.50	0.52
	Attack	0.50	0.49	0.50
Frame	Writing quality	0.74	0.47	0.57
	Verifiability and factual ac.	0.62	0.74	0.67
	Neutral point of view	0.59	0.56	0.58
	Dialogue management	0.64	0.56	0.60

Table 4.4: The precision, recall, and weighted average F_1-score of our identification methods for all categories of the dimensions of dialog act, argumentative role, and frame.

correlate with the size of the datasets. In case of dialog acts, the method achieves low F_1-scores for ‘socializing’, ‘recommending an act’, and ‘asking a question’. These categories have a significantly smaller number of turns compared to other categories, which makes identifying them harder. The effectiveness of identifying the role and frame categories, on the other hand, appears promising given the difficulty of these tasks.

We point that we considered mainly the turns’ texts in our experiments. In principle, this helps to get an idea about the effectiveness of our methods in Wikipedia as well as other registers for discussions. Nevertheless, including the metadata and structural information of the analysed discussions is definitely worthwhile in general, and will naturally tend to lead to notably higher effectiveness.

4.4 Discussion of Strategy Analysis

While our approach to modeling argumentation strategies in deliberative discussions may seem Wikipedia-specific, the derivation of concepts and categories from metadata can be transferred to other online discussion platforms. We expect the general derivation steps to be the same, whereas the techniques applied within each step may differ depending on the types, frequency, and quality of metadata. For example, the consistent usage of the most common user tags in Wikipedia discussions helps originating concepts manually. In contrast, other metadata might require the use of computational methods, such as clustering, keyphrase extraction, and textual entailment.

Also, our model helps analysing the influence of user interaction and behavior on the effectiveness of discussion decisions. For example, some Wikipedia users focus on the frame 'well written' while ignoring others, which may negatively affect the accuracy of an article's content. Also, users often attack other turns, instead of considering neutral acts such as clarifications of misunderstandings.

Many categories in our model apply to deliberative discussions in general, particularly the dialogue acts and argumentative roles. While the found frames are more Wikipedia-specific, they still can play a role on collaborative writing platforms. For example, when writing a scientific paper, possible frames are the 'writing quality' or the 'verifiability of content and citations'.

Regarding the analysis of argumentation strategies in deliberative discussions, we found that the effectiveness of the identification method regarding the proposed models demonstrate a high variance. The obtained F_1-scores range between 0.13 and 0.71. Such a high variance makes using the method to analyse strategies questionable. Put simply, the discovered principles using the proposed identification method would have been taken with a grain of salt. However, Webis-WikiDebate-18 corpus is a promising base for developing robust identification methods in the future.

Following the development of a robust method that can distinguish the acts, roles, and frames reliably, the argumentation strategies in Wikipedia discussions can be analysed in various ways. The distribution of the acts, roles, or frames in the discussions may indicate diverse principles that the participants usually use there. This can be exhibited, for example, if most of the participants concentrate on the 'writing quality' frame when discussing merging two paragraphs in a Wikipedia article. The sequential flows of the acts, roles, or frames are also powerful for deriving principles. For instance, participants may tend to focus on the act of 'providing evidence' with the role of 'support' at the beginning of a discussion.

To explore the effectiveness of principles, discussions can be classified into successful and unsuccessful (i.e., disputed). Then, the correlation between the principles and both of the two classes should be examined. The principles which are highly correlated mainly with successful discussions should be the grounds for recommending the best moves to discussion participants.

4.5 Related Work

Modelling deliberative discussions in Wikipedia has been already addressed in different studies. The central goal of these studies is to minimize the coordination effort among discussion participants. In particular, Ferschke *et al.* [2012] have proposed a model of 17 dialogue acts, each belonging to one of four categories: article criticism, explicit performative, information content, and interpersonal. The model was derived by performing a manual analysis of 30 talk pages in the Simple English Wikipedia. Based on the model, a new corpus of 1367 turns has been created and used to train and evaluate a multi-label classifier for predicting the model's acts. Another model is the one proposed by Viegas *et al.* [2007]. The model consists of 11 different dialogue acts. These acts have been used to manually label 25 talk pages from the English Wikipedia. Furthermore, Bender *et al.* [2011] have developed a model for authority claims and alignment moves in Wikipedia discussions. The model then has been used to label 47 talk pages.

Rooted in the limitation of being derived from a small sample, these models obtain low coverage and/or are over-abstracted. This is indicated by labels such as 'other' [Viegas *et al.*, 2007] or by a very abstract 'information providing' act [Ferschke *et al.*, 2012], which covers 78% of the turns. We argue that recommending moves for new participants based on such labels will not be useful. On the other hand, the model of Ferschke *et al.* [2012] does not include anything similar to 'propose alternative action', for example, although such a concept was shown to be important in deliberative discussions [Walton, 2010].

Moreover, no existing model distinguishes the three dimensions of turns: act, role, and frame. They either consider only one dimension or mix an act with a role, such as in the label: 'criticizing unsuitable or unnecessary content' [Ferschke *et al.*, 2012]. This is a problem for predicting the next best deliberative move. For example, consider a discussion about adding new content to an article, where the participants support the action with different acts (e.g., 'providing evidence'), but all of them consider the 'writing quality' frame. A new turn attacks the action by providing evidence that the action would violate the 'neutral point of view'. The best next move should actually consider this frame, since no content that violates 'neutral point of view' policy should be added, regardless of its adherence to the 'writing quality'.

In contrast, our approach of deriving the model using thousands of different 'descriptions' of moves written by the numerous Wikipedia users is, in our view, more likely to give a representative picture of how people argue in deliberative discussions. This, in turn, leads not only to high coverage, but also to better abstraction. Our model is in line with three well-known theories, which we summarize in the next paragraph.

Speech act is a widely accepted theory in pragmatics [Searle, 1969]. Based on this theory, many research papers have been proposed for modelling different domains, such as one-on-one live chat [Kim *et al.*, 2010], persuasiveness in blogs [Anand *et al.*, 2011], twitter conversations [Zarisheva and Scheffler, 2015], and online dialogs [Khanpour *et al.*, 2016]. In the context of *argumentation theory* [Peldszus and Stede, 2013], agreement detection is a related direction of work which has been studied in discussions [Rosenthal and McKeown, 2015]. Notably, Andreas *et al.* [2012] annotated 822 turns from 50 talk pages with three labels: 'agreement', 'disagreement', and 'non'. Anyhow, over the last few years, argumentation mining became a hot topic in our community, where several studies have went beyond the agreement detection to investigate the identification of the 'support' and 'attack' roles in argumentative discourses [Peldszus and Stede, 2013]. Finally, *framing* is one of the important theories in discourse analysis [Entman, 1993]. This theory has been studied widely in different domains, such as news article [Naderi and Hirst, 2017] and political debates [Tsur *et al.*, 2015]. These three theories back up the essence of our proposed model. We found that a participant in a discussion writes her text considering a specific act, an argumentative role, and a frame.

The metadata in Wikipedia have been used for different tasks. The 'infobox' has been exploited in the tasks of question answering [Morales *et al.*, 2016] and summarisation [Ye *et al.*, 2009], among others. Moreover, Wang and Cardie [2014] have used specific discussion templates to identify discussions that are disputed. Besides Wikipedia, metadata such as 'point for', 'point against', and 'introduction' have been used successfully for modeling argumentativeness in debate platforms [Al-Khatib *et al.*, 2016a]. Also, The metadata for user interactions, such as the 'delta indicator' and users votes in the ChangeMyView subReddit discussions have been used to model the persuasiveness of a text [Tan *et al.*, 2016].

4.6 Summary

This chapter studies how the argumentation strategies of participants in deliberative discussions can be supported computationally. Our ultimate goal is to predict how the best next deliberative move of each participant should be according to an effective strategy. In this chapter, we present a new model for deliberative discussions and we illustrate its operationalization. Previous models have been built manually based on a small set of discussions, resulting in a level of abstraction that is not suitable for move recommendation. In contrast, we derive our model statistically from several types of metadata that can be used for move description. Applied to six million discussions from Wikipedia talk pages, our approach results in a model with 13 categories along three pragmatic attributes of discussion's turns: dialogue act, argumentative role, and frame. On the basis of the model, we automatically generate a corpus with about 200,000 turns labelled for the 13 categories. Next, we operationalise the model with three supervised learning methods.

The operationalisation of our model demonstrate high variance in terms of the effectiveness, which we thought that it would restrain the ability to explore the argumentation strategies in deliberative discussions with a high degree of reliability. Nevertheless, the model and its operationalisation method are anticipated to be an influential step towards highly successful strategy analysis. In the future, employing the modern developed state-of-the-art methods which rely on the deep learning techniques for text classification and, in particular, speech act classification, can be crucial for considerable advance in the identification of the elements of our model.

Chapter 5

Stylistic Persuasive Strategies

> *A good style must have an air of novelty, at the same time concealing its art. (Aristotle)*

The decision for (and the adherence to) 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 using irony, repeating the same phrase, or rhetorically answering a proposition. In the literature on the subject, these techniques are called *rhetorical figures* [Johnson, 2016].

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

The analysis of writing style based on rhetorical figures, however, provides a mechanism to describe *which*, *where*, and *how* specific techniques are used. This kind of analysis is important for deriving the principles in strategies, and hence, it can serve the text synthesis and writing assistants tools. Concretely, the analysis of strategies based on rhetorical figures can benefit text synthesis systems by improving the quality of the automatically generated texts [Hu *et al.*, 2017]. Furthermore, it can form the backbone of style suggestion tools. For example, in case of writing a text for which the desired property (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 5.1 illustrates the described connections.

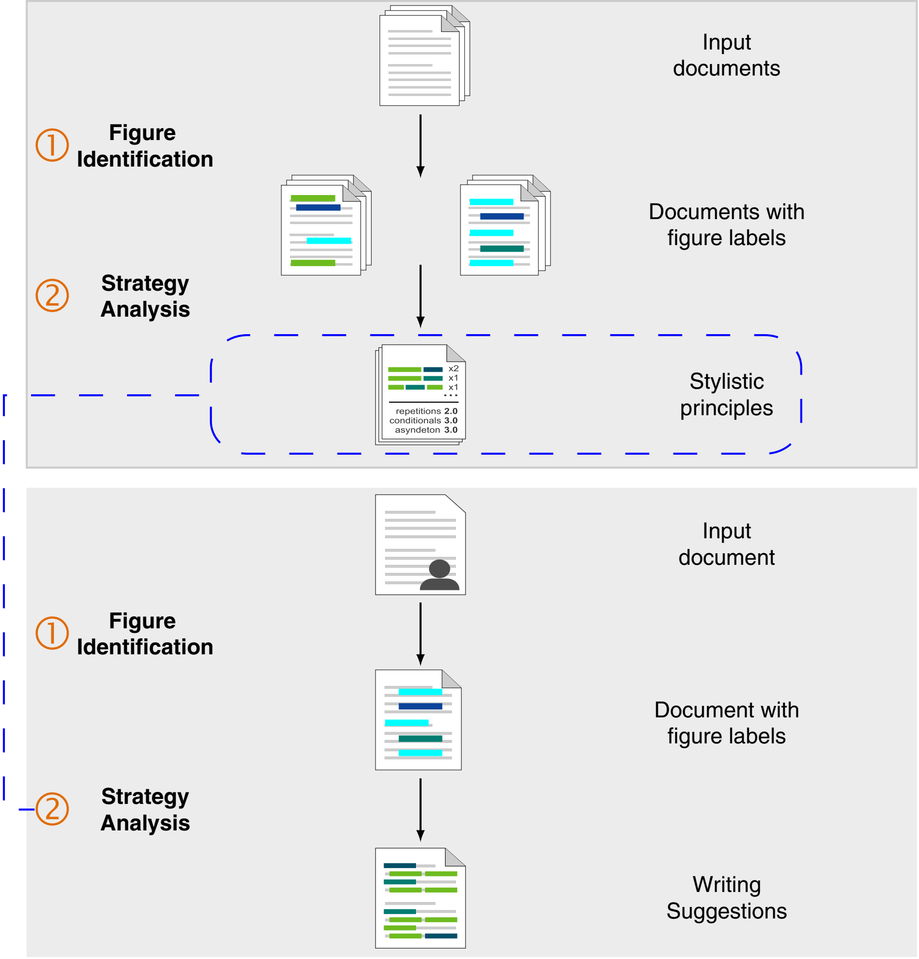


Figure 5.1: Envisioned tool for style checking and suggestion. The stylistic strategies obtained by exploring a collection of texts are used to recommend stylistic suggestions to a writer regarding his or her input document.

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

Against the above background, and in relation to the sixth, seventh, and eighth research questions of this thesis (see Chapter 1, Section 1.2), this chapter addresses the following core steps: (1) Modelling a set of syntax-based rhetorical figures. (2) Identification of the modelled figures in a text. (3) Exploration of the most common patterns regarding the usage of the identified figures. The exploration includes exposing how the patterns vary across monological and dialogical texts, within and across the texts' genres, topics, and authors, and across different opponent debaters.

To carry out these steps, we introduce a new model that comprises 26 rhetorical figures. Then, we develop a rule-based method for the identification of the 26 figures. The rules are built on top of the outputs of a probabilistic context-free grammar parser, PCFG. To evaluate the method, we create a corpus of 1718 texts which are labelled regarding the figures. The results of the evaluation experiments demonstrate that our method is able to identify the figures with an average of 0.70 in terms of F_1 measure. Based on the developed method, we explore and expose the usage of the figures in monological texts within and across different genres, topics, and authors, using a subset of the New York Times annotated corpus [Sandhaus, 2008]. In addition, we discover the usage patterns of figures in dialogical texts using a set of the presidential debates from the American presidency project [Woolley and Gerhard, 2017].

We consider the gained qualitative and quantitative insights about the usage of rhetorical figures as valuable stylistic principles for argumentation strategies. Such principles can help to devise a new generation of semi-automated text generation and writing tools. [1]

5.1 Model

A rhetorical figure is a 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].

In this section, we propose a model regarding the syntax-based rhetorical figures. In particular, we select 26 figures that belong to two 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 is usually attributed to using an artful deviation from the ordinary arrangement

[1] All the developed resources in this chapter are publicly available.

of words, while the effect of the second is produced by selecting a specific arrangement of words among other possible arrangements.

In the following, we describe the figurative and ordinary syntax figures.

5.1.1 Figurative Syntax

The figures in this category mainly focus on arranging words artfully [Burton, 2007]. They are divided into four types: balance, inversion, omission and repetition.

- The *balance* figures concern 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* figures 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* figures 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* figures are one of the most frequent, and probably, the most powerful. According to Aristotle, repetition is a key to a persuasive speech [Fahnestock, 2003]. Typically, repetition figures aim at influencing the emotional state of the readers by emphasizing or implicating a specific idea [Burton, 2007; Corbett, 1990]. An example which illustrates the emotional effect of repetitions is the famous line from King Lear written by Shakespeare: "Never, never, never, never, never." [Müller, 2006].

In the following, we present an overview of the considered figures in this category. The overview covers a definition, a formalization, and an example for each figure belongs to balance, omission, or repetition [2] type. Our formalization is grounded on the figures' definitions which are taken from a set of well-known and reliable sources: the 'Silva Rhetoricae', which is one of the most comprehensive sources for rhetoric on the web [Burton, 2007], and the reputable sources of `Literarydevices.net`, `Grammarly.com`, and the guide by Declerck and Reed [2001].

[2] The inversion is left to future work due to its complexity.

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, $_{\mathrm{a}\,=\,\mathrm{b}}$ for identity , and $_{\mathrm{a}\,\neq\,\mathrm{b}}$ for nonidentity.

Notice that we essentially concentrate on defining the figures on the sentence-level, or across consecutive sentences. Besides, some rhetorical figures, according to their definitions, may overlap with other figures in some special cases. Though such an overlap is rare and partial, we consider minimizing the possible overlaps among figures as much as possible in our formalization.

(1) (B) Balance Figures

Three rhetorical figures are introduced here: enumeration, isocolon, and pysma.

(B1) Enumeration: is used mainly to list a series of details, words, or phrases.

Example: *Diligence*, *talent* and *passion* will drive anybody to success.

Formally: < ... W [CC | COMMA] W ... >

Details: the window size between the first comma and the last conjunction ranges between 2 and 5.

(B2) Isocolon: a series of similarly structured elements with the same length (kind of parallelism).

Example: *Fill the armies*, *rule the air*, and *pour out the munitions.*

Formally: < ... <Phr>$_a$ <Phr>$_a$ ... <Phr>$_a$ ... >

Details: the considered sub-types here are ‘bicolon’ (two grammatically equal structures), ‘tricolon’ (three grammatically equal structures), and ‘tetracolon’ (four grammatically equal structures).

(B3) Pysma: is asking multiple questions successively (which would together require a complex reply).

Example: *In what place did he speak with them? with whom did he speak? did he hire them? whom did he hire, and by whom? To what end, or how much did he give them?*

Formally: < ... < Cl? > < Cl? > ... >

Details: the number of consecutive questions is between 2 and 10.

(2) (O) Omission Figures

Three rhetorical figures are introduced here: asyndeton, hypozeugma, and epizeugma.

(O1) Asyndeton: denotes the omission of conjunctions between clauses.

Example: *I came, I saw, I conquered.*

Formally: { <Cl$_a$> COMMA <Cl$_b$> COMMA <Cl$_c$> ... }

Details: a valid instance of this figure is sequence of more than two commas separated by clauses, words, or phrases.

(O2) Hypozeugma: is placing last, in a construction that contains several words or phrases of equal value, the word or words on which all of them depend.

Example: *Friends, Romans, countrymen, lend* me your ears ...

Formally: < ... [W]$_a$, [W]$_b$, [W]$_c$ ... Vb > or ,
< ... <Phr$_a$> , <Phr$_b$> , <Phr$_c$> ... Vb >

Details: based on the governor and its dependents, out of all the detected relations, only the nominal and clausal subject relations are considered here.

(O3) Epizeugma: is placing the verb that holds together the entire sentence (made up of multiple parts that depend upon that verb) either at the very beginning or the very ending of that sentence.

Example: Neither a borrower nor a lender *be.*

Formally: < Vb ... > or , < ... Vb >

Details: for a sentence with a single governor, if the governor is placed in the first or the last one-fifth of the sentence, it is considered as a valid instance of this figure.

(3) (R) Repetition Figures

Eight rhetorical figures are introduced here: epanalepsis, mesarchia, epiphoza, mesodiplosis, anadiplosis, diacope, epizeuxis, and polysyndeton.

(R1) Epanalepsis: the repetition of a word or a sequence of words at the beginning and at the end of a sentence, phrase, or clause.

Example: *Believe* not all you can hear, tell not all you *believe.*

Formally: < [W]$_a$ ... [W]$_a$ >

Details: the beginning of a sentence is denoted to be the first one-fifth and the ending to be the last one-fifth of that sentence. The sentence length is based on the number of its words (excluding the stop words).

(R2) Mesarchia: the repetition of the same word(s) at the beginning and middle of successive sentences.

Example: *I was* looking for *a piece* of paper. *I was* anxious for *a piece* to write on. *I was* in need of *a piece* to start my butterfly census project.

Formally: $< [W]_a \dots [W]_b \dots > < [W]_a \dots [W]_b \dots >$

Details: this figure is defined in a similar fashion to the previous one, except that consecutive sentences are considered here. Also, a sentence are divided by a factor of four for determining its beginning and middle. The single instances of stop words repeated in the considered spans are filtered out.

(R3) Epiphoza: denotes the repetition of the same word or words at the end of successive sentences.

Example: O *apple*! wretched *apple*! Miserable *apple*!

Formally: $< \dots [W]_a > < \dots [W]_a >$

Details: the consecutive sentences which have the same word or words within their last quarter are considered here.

(R4) Mesodiplosis: the repetition of the same word or words in the middle of successive sentences.

Example: There's *no time* like the future! There's *no time* like the past!

Formally: $< \dots [W]_a \dots > < \dots [W]_a \dots >$

Details: similar to the previous two figures, but considering the middle of sentences.

(R5) Anadiplosis: the repetition of the last word (or phrase) of a clause or a sentence at the beginning of the next clause or sentence.

Example: I will life my eyes unto the hills, from whence cometh *my help. My help* cometh from the . . . .

Formally: $< \dots [W]_a > < [W]_a \dots >$

Details: similar to the previous two figures, but considering the end of a sentence, and the beginning of its adjacent sentence.

(R6) Diacope: denotes the repetition of a word or phrase with one or more other words or phrases in between.

Example: *The horror*! Oh, *the horror*!

Formally: $< \dots [\mathrm{W}]_a \dots [\mathrm{W}]_a \dots >$

Details: punctuation marks, non-alphanumeric characters, numbers, and stop words are filtered out. For each word, the repetition in the neighbor words in a window that ranges between 1 and 5 should be scanned.

(R7) Epizeuxis: the repetition of words while no other words in between.

Example: *Awake, awake* and stand up O Jerusalem.

Formally: $< [\mathrm{W}]_a \ [\mathrm{W}]_a >$

Details: similar to the previous figure, except that the repetition is in the immediate succession neighbor.

(R8) Polysyndeton: is a rhetorical term which employs many conjunctions between clauses.

Example: He pursues his way, and swims, *or* sinks, *or* wades, *or* creeps, *or* flies.

Formally: $\{ <\mathrm{Cl}_a> \mathrm{CC} <\mathrm{Cl}_b> \mathrm{CC} <\mathrm{Cl}_c> \dots \}$

Details: similar to asyndeton, sequences of clauses, phrases, and words split by conjunctions should be considered. If at least two conjunctions follow in immediate succession, the instance is considered as valid.

5.1.2 Ordinary Syntax

In regard to the ordinary syntax figures, we decide to deal with conditionals, comparatives and superlatives, and passive voice. The selection of these figures is inspired by their potential impact on the readers [Martinet, 1960].

- The *conditional* figures entail the causality aspect of the language, and causality, in turn, could imply the explanation of an event. However, conditional figures can also be used to argue about positive or negative consequences of a specific action such as “If we had not joined the EU, we would be better off now.”.

- The *comparatives and superlatives* figures might be used to emphasize the superiority of an entity or idea. For example, "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. For example, "many mistakes were made, but the future will be great".

In the following, we overview the considered figures in this category. The overview includes a definition, a formalization, and an example for each figure belongs to *conditionals*, *comparatives and superlatives*, and *passive voice.* The formalization is based on definitions in 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].

(1) (C) Conditionals

We consider the four types of conditionals: zero, one, two, and three, in addition to counterfactual. The formalization of conditionals relies on the sentence Part of Speech (POS) tags according to the Penn Treebank tag set. Since if-conditional sentences comprise the P-clause and the Q-clause, recognizing those clauses is essential to determine the type of the conditional. To this end, we consider the text span between 'if' and the next governor as the P-clause, then, we find the next closest governor and identify the text span between it and the token placed at most 4 tokens backwards, and consider it as the Q-clause.

(C1) Zero conditionals: express general truths, events in which the premise always causes the conclusion to happen.

Example: *If* you don't *brush* your teeth, you *get* cavities.

```
Formally: If [VB / VBP / VBZ], then [VB / VBP / VBZ]
```

Details: a sentence belongs to this figure if its P and Q clauses include present tense verbs and if the Q-clause does not contain modal verbs as this conflicts with other conditionals.

(C2) First conditionals: is used to refer to situations which are very likely (yet not guaranteed) to happen in the future.

Example: *If* it *rains*, you *will get* wet.

```
Formally: If [VB / VBP / VBZ / VBG], then [MD + VB]
```

Details: similar to the previous figure, but the Q-clause here includes simple future tense combined with a modal verb.

(C3) Second conditionals: express consequences that are totally unrealistic or will not likely to happen in the future.

Example: *If* it *rained*, you *would get* wet.

```
Formally:          If [VBD], then [MD + VB]
```

Details: the P-clause here includes a verb in the past tense.

(C4) Third conditionals: are used to explain that present circumstances would be different if something different had happened in the past.

Example: *If* I *had worked* harder, I *would have passed* the exam.

```
Formally:          If [VBD + VBN], then [MD + VBN]
```

Details: unlike second conditionals, the P-clause includes past tense and past participle.

(C5) Counterfactuals: examine how a hypothetical change in a past experience could have affected the outcome of that experience.

Example: *If* I *were* you, I *wouldn't* come.

```
Formally:          If [VBD + VBN], then [past modals]
```

Details: this type is not always grammatically separable from other conditionals. However, as a heuristic rules: the governor verb of the P-clause is often in the past tense, and the Q-clause includes a past tense modal verb such as 'should have', 'would have', and 'could have'.

(C6) Unless conditional: is a restricted version of if-conditional, in a sense that its intrinsic meaning is narrowed down to "Q in the case other than P".

Example: You can't go on vacation *unless* you save some money.

```
Formally:          < ... unless ... >
```

Details: as far as we know, 'unless' is always used in the context of conditionals. Therefore, any sentence which includes this word belongs to this type.

(C7) Whether...or conditional: used to express alternative (disjunctive) conditions.

Example: *Whether* you are overweight *or* not, it is always better to watch your diet.

```
Formally:          < ... whether ... or ... >
```

Details: a sentence includes the word ‘whether’ followed by an ‘or’ that occurs in the same sentence.

(2) (CS) Comparative/Superlative Adjectives and Adverbs : comparatives are used to compare differences between two objects or states, while superlatives are used to describe an object which is at the upper or lower degree of a quality.

Example (comp. adjective): My house is *larger* than yours.

Example (super. adverb): Mrs. Smith talks *most quietly.*

```
Formally:          < ... [JJR / JJS / RBR / RBS] ... >
```

Details: an instance is considered to belong to this figure if it follows the corresponding POS tags.

(3) (PV) Passive voice: a type of a clause or sentence in which the focus is put on the main action or object of the said sentence rather than in its subject.

Example: The problem *is solved.*

```
Formally:          < ... [to be] ... [VBN] ... >
```

Details: any sentence that includes a verb ‘to be’ and a past participle at most four words to the right is considered a passive voice. The window size of four is specified considering that punctuation marks and stop words might occur in between the two constituents of a passive voice instance.

5.2 Corpus Construction

The manual construction of an annotated corpus for rhetorical figures, even using the crowdsourcing setting, is extremely expensive and time consuming [Java, 2015]. This is due to the big number of figures, the potential overlaps between them, and the possibility for some figures to be spread across phrases, sentences, or even paragraphs. Thus, we decided to follow a number of related studies (e.g., [Java, 2015]) and build a new corpus for the modelled rhetorical figures as follows: First, we listed a set of trustworthy sources on the Web. The listed sources address the rhetorical figures while demonstrating a high credibility as being developed by experts in rhetoric. Next, we employed the meta-data information (e.g., “Example of”) in the listed sources to collect a set of instances regarding

Figure	#Instances	Precision	Recall	F1
(C1) If-cond. Zero	60	0.71	0.76	0.73
(C2) If-cond. One	60	0.78	0.78	0.78
(C3) If-cond. Two	60	0.82	0.75	0.78
(C4) If-cond. Three	60	0.86	0.65	0.74
(C5) If-Counterf.	60	0.84	0.87	0.85
(C6) Unless-cond.	60	1	1	1.00
(C7) Whether-cond.	60	1	0.83	0.91
(CS1) Comp. Adj.	68	0.51	0.61	0.56
(CS2) Comp. Adv.	70	0.6	0.62	0.61
(CS3) Super. Adj.	70	0.62	0.73	0.67
(CS4) Super. Adv.	70	0.63	0.5	0.56
(PV) Passive Voice	60	0.78	0.98	0.87
Other	60	0.23	0.23	0.23

Table 5.1: The precision, recall, and F_1-score for identifying each of the ordinary rhetorical figures as well as the ‘other’ class.

our studied rhetorical figures. We found that targeting around 60 examples for each figure is reasonable considering the selected sources. We verified all the collected examples and eliminated any duplication. Additionally, we examined the possible overlaps between the figures and minimized them adequately, i.e., all the examples for a figure belong solely to this figure. Unfortunately, two figures turned out to be considered only in few sources, and hence, we obtained less than 60 examples for them. Furthermore, we collected 60 examples in which none of the studied figures is used.

Webis-RhetoricalFigures-19 Corpus Overall, we collected 1718 examples: 1658 example for the 26 figures and 60 examples without any figure (i.e., the ‘other’ class). The distribution is shown in Table 5.1 and Table 5.2. This corpus, despite its relatively small size, is larger than those which have been built for rhetorical figures [Java, 2015].

5.3 Identification Method

In this section, we discuss the identification of the 26 syntax-based rhetorical figures. In particular, we developed grammar-based rules in accordance with

Device	#Instances	Precision	Recall	F1
(B1) Enumeration	60	0.76	0.93	0.84
(B2) Isocolon*	180	0.57	0.83	0.68
(B3) Pysma	60	1	1	1.00
(O1) Asyndeton	60	0.25	0.93	0.39
(O2) Hypozeugma	60	0.61	0.8	0.69
(O3) Epizeugma	60	0.65	0.7	0.67
(R1) Epanalepsis	60	0.63	0.83	0.72
(R2) Mesarchia	20	0.45	0.85	0.59
(R3) Epiphoza	60	0.58	0.93	0.71
(R4) Mesodiplosis	40	0.27	0.68	0.39
(R5) Anadiplosis	60	0.76	0.73	0.74
(R6) Diacope	60	0.73	0.73	0.73
(R7) Epizeuxis	60	0.79	0.77	0.78
(R8) Polysyndeton	60	0.77	0.7	0.73

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

Table 5.2: The precision, recall, and F$_1$-score for identifying each of the figurative rhetorical figures.

Category	#Instances	Precision	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 5.3: The precision, recall, and F$_1$-score for identifying each of the figurative and ordinary categories.

the formulations we discussed in Section 5.1. The rules are constructed based on the output of a PCFG parser (Stanford Parser [Manning *et al.*, 2014]). Each rule is tailored for one figure: The input is a sentence, and the output is weather the figure is used in this sentence or not.

Technically, the implementation of the rules was carried out using Apache Ruta^{TM} (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 evaluation of the rules was performed using the one-vs.-rest classification setting, employing the *Webis-RhetoricalFigures-19* corpus discussed in Section 4.2. In this setting, we performed one classification experiment for each figure. In an evaluation experiment for a figure, the instances of this figure (in the corpus) is considered as the positive class, and the instances of the remaining 25 figures as well as the instances in the 'other' as the negative class. The rule of identifying the figure is applied to all the instances in the positive and negative classes. The rule output is correct if it considers an instance in the positive class as valid, or considers an instance in the negative class as invalid. Otherwise, the output is incorrect.

The effectiveness of identifying each figure is measured and reported using the precision, recall, and F_1 measures.

Classification Results: Overall, our rules were able to identify the 26 figures with an average of 0.70 F_1-score. Such a score indicates the high effectiveness of the rules.

Table 5.1 shows the results of our experiments regarding the 'figurative' category. The rules achieved high scores for the *balance* figures, including F_1-score of 1.00 for 'pysma'. The 'isocolon' is the most challenging figure with F_1-score of 0.68. As for the *omission* figures, the F_1-scores ranged between 0.39 for 'asyndeton' and 0.69 for the 'hypozeugma'. These results were a bit lower than the other types. Most of the *repetition* figures had F_1-score of about 0.73, except 'mesarchia' with 0.59, and 'mesodiplosis' with 0.39.

Table 5.2 shows the results of our experiments regarding the 'ordinary' category. In general, the rules achieved scores between 0.56 and 1.00. Strangely enough, despite their simple syntax, *comparatives and superlatives* figures were difficult to be identified.

Table 5.3 shows the results for identifying the figures organized in six groups (three for the figurative category and three for the ordinary category). There, the F_1 scores ranged between 0.54 and 0.87. The best result was obtained for *passive voice* (0.87), following by *conditionals* (0.82). *Omission* and *repetition* were the hardest to identify with 0.54 and 0.67 F_1 respectively.

Error Analysis: Despite the high effectiveness of our rules, they were subject to fail in some cases.

Concerning the ‘figurative’ category, identifying the *balance* figures seemed to be precise except for ‘isocolon’. The identification of this figure was based on the outputs of the syntax parser (i.e., POS tags) which were sometimes inaccurate, especially for long sentences. This had a negative impact on the precision score. In the example of: “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 . . . .”, the “Libertarian candidate” made the rule of ‘isocolon’ wrongly considered this example as a valid instance. Regarding the *omission* figures, our rules managed to achieve a recall of 0.93 for ‘asyndeton’ figure, but only 0.25 precision. We found that the abundance of commas, which we used as an indicator of the lack of conjunctions was insufficient to completely distinguish ‘asyndeton’ from other figures, especially ‘enumeration’. For example, “Old McDonald had a pig, a dog, a cow and a horse.” was identified as ‘asyndeton’, while it is actually an ‘enumeration’. Concerning *repetition*, the rules achieved low scores in identifying the ‘mesarchia’ and ‘mesodiplosis’. These two figures have the least number of instances in our corpus. We also observed that our heuristic rules for defining the beginning and middle of sentences were the reason for some errors.

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

To have a better idea regarding the effectiveness of our rules, we performed a manual inspection of the rules’ outputs on a set of ten newspaper articles. We found that the rules of some figures such as ‘isocolon’ and ‘asyndeton’ output many false positives. Moreover, we found that the rules made mistakes in case the input sentence is very long.

5.4 Stylistic Argumentation Strategies

Relying on our figure identification rules, we conducted an in-depth analysis study for the usage patterns of rhetorical figures in newspaper articles and presidential debates. Such patterns are regarded as stylistic principles of argumentation strategies. In this section, we first describe the acquisition and sampling of the *analysis datasets*. Then, we discuss the distribution of rhetorical figures there along with different text and debate properties. The discovered distributions illustrated various patterns of the usage of rhetorical figures and led to several interesting insights.

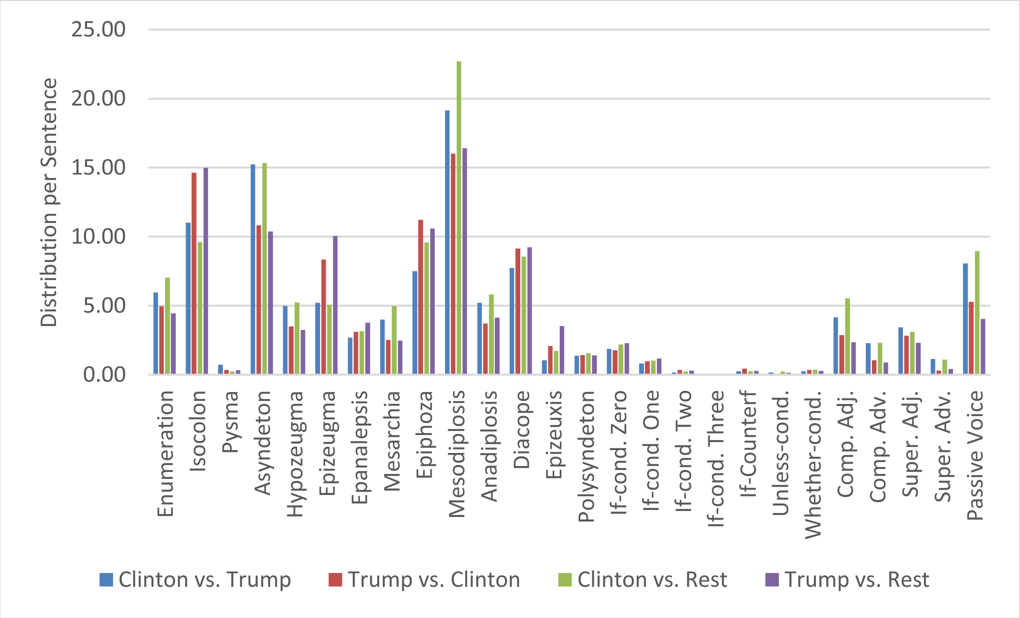


Figure 5.2: The distribution of the rhetorical figures in Clinton's turns when she debates with Trump, Trump's turns when he debates with Clinton, Clinton's turns when she debates with a candidate other than Trump, and Trump's turns when he debates with a candidate other than Clinton.

Analysis Datasets

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

(1) Newspaper dataset: to construct this dataset, we used the NYT annotated corpus [Sandhaus, 2008]. This corpus comprises more than 1.8 million high-quality articles written by professional writers. It comes with many types of meta-data labelled 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 has been accounted for the confounding variables. For example, studying the style in articles belong to a specific topic may be influenced by their genres and authors. Hence, we first tried to resolve this issue with the stratification method [Tripepi *et al.*, 2010], but this try was 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

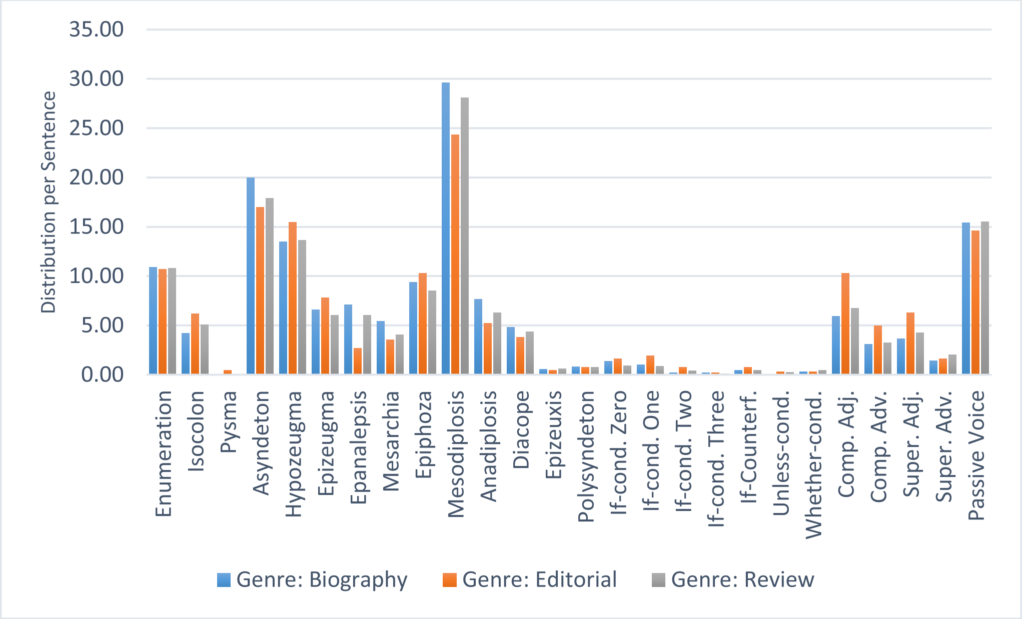


Figure 5.3: The distribution of the rhetorical figures among the genres of biography, editorials, and reviews.

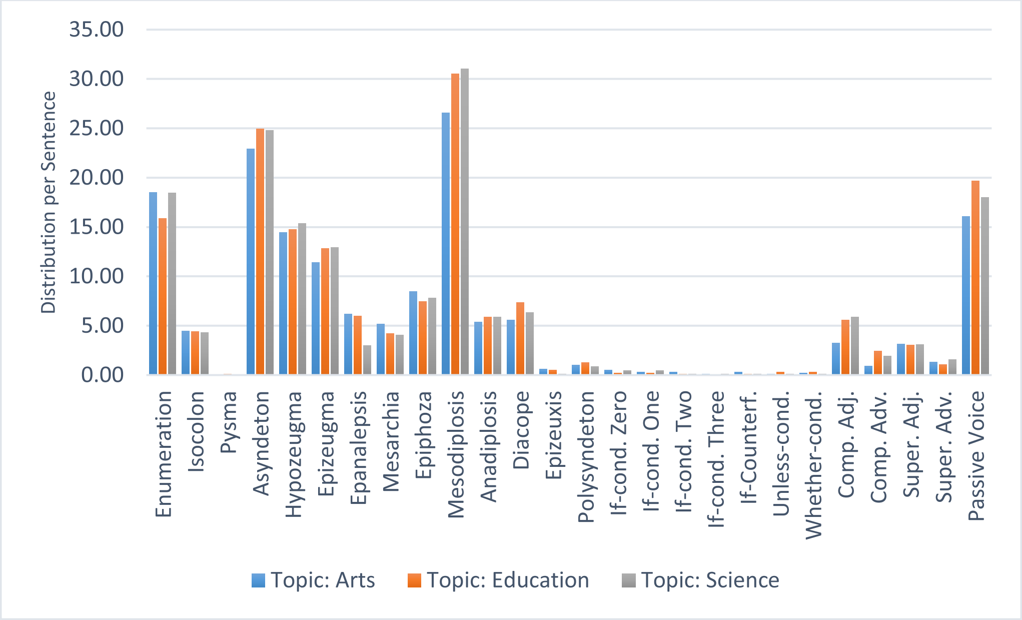


Figure 5.4: The distribution of the rhetorical figures among the topics of art, education, and science.

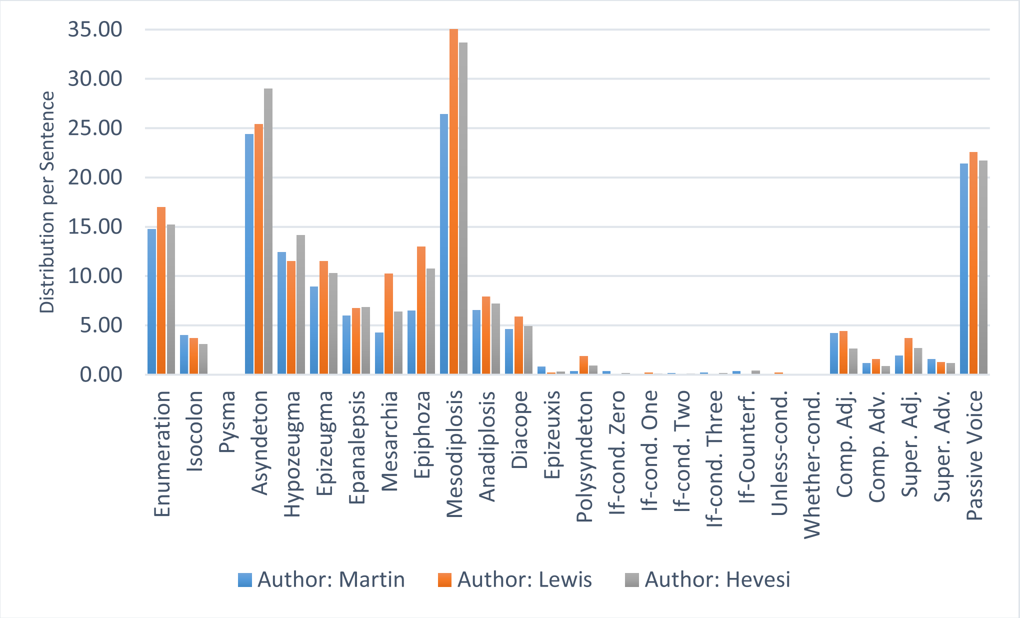


Figure 5.5: The distribution of the rhetorical figures among the authors of Martin, Lewis, and Hevesi.

tens of articles while others wrote only one. Therefore, we tried the matching technique [de Graaf *et al.*, 2011], and this time, we managed to successfully sample the three subsets. To preserve the balance between the subsets, we considered 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 using 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, and opinions on different topics. This differences 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 the candidates.

Dimension	Datasets	P-value	Independence
Authors	Hevesi vs. Lewis	0.015	TRUE
	Lewis vs. Martin	≈0	TRUE
	Martin vs. Hevesi	0.017	TRUE
Genres	Biography vs. Editorial	≈0	TRUE
	Editorial vs. Review	≈0	TRUE
	Review vs. Biography	0.68	FALSE
Toopics	Science vs. Education	0.70	FALSE
	Education vs. Arts	0.26	FALSE
	Arts vs. Science	0.19	FALSE

Table 5.4: The results of the significance test regarding the authors, genres, and topics subsets.

Analysis Method: Basically, we applied our rules (see Section 5.3) to the texts in the analysis datasets. In particular, the rule for each figure is applied to the articles and debates, resulting in the frequency of that figure there. However, since our identification rules are not perfect, it is crucial here to account for their errors. Hence, we followed the method used in [Al-Khatib *et al.*, 2017b]: for the frequency n of a rhetorical figure rf in an instance i in a dataset. We computed a confidence interval for n, where the $lowerbound = n * precision(rf)$, and the $upperbound = n/recall(rf)$. Ultimately, The mean of the upper and lower bounds is the new frequency of the figure, which is normalized by the number of sentences (in the articles) or the number of turns (in the debates) belong to i. Accordingly, we computed the distributions of rhetorical figures in the analysis datasets and their subsets. The chi-squared test with 0.01 significant level was used to check whether the difference in the usage of the rhetorical figures in the datasets and across their instances is significant. Moreover, the effect-size of the difference in the distributions is measured using Cramer's V test.

Analysis Results: Figures 5.3, 5.4, and 5.5 show the distribution of the rhetorical figures among the three genres, topics, and authors respectively. In addition, Figure 5.2 shows the distribution of the rhetorical figures regarding Trump and/or Clinton debates. Table 5.4 shows the results of the significance test regarding the three genres, topics, and authors.

In the following, we report some results for the analysis study in the newspaper and debate datasets.

(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 used more *repetition* than Martin. Also, Lewis barely considered 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 were written mainly to describe an entity. Interestingly, there was no significant difference between the three topics. Overall, our analysis suggested that the "style" identified by syntax-based rhetorical figures is primarily influenced by the 'author', and 'genre', while 'topic' has the least impact.

(2) Debate dataset: Interestingly, the results show that Clinton was more fond of 'comparatives' and 'passive voice' than Trump, which actually contradicts a widespread assumption [Gingell, 2016; Raskin, 2016]. However, our findings are limited to the debate genre, and they might be different in speech, for example. We also found that Clinton used 'asyndeton' more often than Trump. Since this figure is very effective for making the turns easier to grasp, our finding this time is in line with [Raskin, 2016], where they found that Clinton's language is 13% clearer and more direct than Trump's. The results indicated a significant difference between Clinton and Trump styles. More interestingly, while Clinton's style was significantly different when she debated with Trump than when she debated with the rest of debaters, Trump's style had 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.

5.5 Related Work

Recently, The investigation of rhetorical figures for style analysis has been considered in the computational linguistics community. Many figures in the semantic and pragmatic levels have been addressed singly such as irony [C. Wallace *et al.*, 2014] and sarcasm [Ghosh *et al.*, 2015].

Other studies have targeted identifying a mix of syntax, and semantic figures. Gawryjołek *et al.* [2009] addressed four rhetorical figures: 'anaphora', 'isocolon', 'epizeuxis', and 'oxymorons'. These figures were utilized to recognize the author of a set of documents. Java [2015] identified the four figures mentioned above in addition to nine new figures that belong to parallelism, repetition, and trope. The primary goal of their work is to use the presence of a rhetorical figure as a feature in machine learning models for authorship attribution. More recently, Lawrence *et al.* [2017] analysed eight figures, six belong to the syntax and lexical levels, and two to the trope (i.e., semantic or pragmatic). Mainly, they conducted a pilot study to investigate the relationship between argumentation structure and the identified figures.

Sadly, few resources for rhetorical figures 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 analyse style based on rhetorical figures. We resolve this problem considerably by providing rules for identifying 26 different rhetorical figures. Our developed resources, including the code, will be made freely available.

PCFG outputs have been employed in different tasks including response generation in dialogue [Yuan *et al.*, 2015], multi-word expression identification [Green *et al.*, 2011], and identifying rhetorical figures [Gawryjołek *et al.*, 2009; Java, 2015]. However, we developed a set of original heuristic rules that map the figures' definitions to PCFG grammars. As far as we know, many figures 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]). Also, in quality assessment research, several style analysis features were applied successfully (e.g., [Ashok *et al.*, 2013]). In comparison to our analysis, on the one hand, we conducted a *controlled* analysis using 'matching' technique. On the other hand, we covered strategy principles in various properties of monological and dialogical texts such as genre, topic, author, and debate opponent.

5.6 Summary

The analysis of stylistic attributes become a mature discipline, but it is mostly tackled from the recognition perspective. This means that such an analysis can give strong classification results that, because of their intrinsic nature, cannot be transferred to describable form which we need to explore argumentation strategies and expos stylistic principles. We addressed this shortcoming by proposing a new model that allows for an *explicit* encoding of style principles based on rhetorical figures. We developed a rule-based method for identifying 26 different syntax-based rhetorical figures that belong to two categories. Later, in carefully designed experiments, we study the usage of these figures in a set of newspaper articles and presidential debates. The distributions of the figures show different patterns of style among three text's properties and provide new insights regarding style usage.

The 0.70 F_1score achieved in the identification of the figures can be considered as very good for concrete multi-class classification setting. It shows that the applied method has the potential to find its way into real-world text synthesis tools. Moreover, the patterns we identified using this method can form the ground for formulating high-quality principles that account for different properties of texts.

We plan in the future to improve our grammars to minimize mistakes and increase the number of figures considering the *inversion* type of figurative syntax. We also plan to integrate the found style patterns in constrained text generation and computational writing assistance.

Chapter 6

Conclusions

This thesis strives to advance the state of computational argumentation by performing a thorough analysis of the *strategies* that are manifested in argumentative discourses. The strategies, in this context, are high-level plans that aim at achieving the goal of persuasion or consensus.

To this end, we propose a new view of argumentation strategies, defining their elements as well as their formulation and evaluation processes. Based on this view, we model a set of pragmatic and stylistic argument attributes, develop methods for automatic identification of the modelled attributes, and explore strategy principles in texts according to the identified attributes.

The models, methods, and principles this thesis has developed and explored are considered to be key for improving many downstream applications such as text synthesis, writing assistance, and dialogue-management tools.

6.1 Contributions

Overall, the contribution of this thesis to its academic field can be grouped according to the following four high-level dimensions:

Theory: The thesis in hand proposes three new models for studying argumentation strategies in many monological and dialogical texts. The models consider the various pragmatic and stylistic attributes of an argument and the contextual information in an argumentative discourse.

In the following, we present a summary of the proposed models:

1. *A model for the pragmatic attributes of argumentative units in editorials*: in accordance with our assumption that each argumentative unit in an argumentation discourse represents a particular role, including asserting an assumption or presenting evidence, we suggest that each argumentative

unit belongs to one of six types: common ground, assumption, testimony, statistics, anecdote, or other. By implementing this model on 300 editorials, we confirm the feasibility of adopting the model with sufficient agreement among human annotators. Furthermore, the results reveal that barely around 1% of the units belong to the type of 'other', which denotes a high level of coverage by the model.

2. *A model for the pragmatic attributes of turns in Wikipedia discussions*: In a novel setting, we derive a new model of deliberative discussions statistically by the use of several types of metadata in Wikipedia discussions. We inspected the metadata, considering those that describe the primary function of a turn. The model's derivation was performed by examining around six million discussions from Wikipedia talk pages. This is in contrast to previous models, which were derived based on a manual inspection of a small set of discussions. Our model comprises 13 categories in accordance with three types of pragmatic attributes of discussions' turns: dialogue acts, argumentative roles, and frames.

3. *A model for the stylistic attributes of texts in different monological and dialogical genres*: On the basis of the rich literature of rhetorical figures, we focused on the syntax-based rhetorical figures, defining 26 figures that belong to the categories of figurative and ordinary syntax. The selection of the figures was based on their potential effects in argumentative discourses.

This thesis has introduced the notion of argumentation strategies and proposed models for exploring the principles of strategies. Bearing in mind that argumentation strategies have not been studied before in the NLP community, this contribution is seen as substantial, and we expect that the models will be widely employed in many beneficial applications.

Data: In this thesis, we have developed four corpora, all of which contribute to the analysis of argumentation strategies. Three of the corpora were built in accordance with the three models described above, while the other corpus supplemented our studies regarding the derivation of the model for deliberative discussions.

The four created corpora are listed and briefly described in the following:

1. *Webis-Editorials-16* corpus: While several argument mining corpora have been published, they do not allow the study of argumentation strategies due to incomplete or coarse-grained unit annotations. In response to this, we built a new corpus with 300 editorials from three diverse news portals, which provides the basis for the mining of argumentation strategies. Each

unit in each of the editorials was assigned to one of six types by three annotators with a high Fleiss' κ agreement of 0.56. We investigate various challenges of the annotation process and we conduct a first corpus analysis.

2. *Webis-WikiDiscussions-18* corpus: This corpus includes the entire set of Wikipedia discussions (six million discussions with about 20 million turns, at the time of parsing) with an annotated discussion structure and different types of metadata such as user tags, shortcuts, and inline templates.

3. *Webis-WikiDebate-18* corpus: This corpus comprises around 200,000 labelled turns distributed as follows: 2,400 turns labelled for their dialogue acts, 7,437 turns labelled for their argumentative roles, and 182,321 turns labelled for their frames. The labels were generated automatically based on metadata on the turns. However, an expert also verified a sample of the corpus to confirm the high quality of the labels.

4. *Webis-RhetoricalFigures-19* corpus: This benchmark corpus comprises 1,718 instances labelled pertaining to 26 rhetorical figures. The corpus was constructed manually by collecting example sentences for the figures written by experts in rhetoric. This corpus, despite its relatively small size, is larger than those that have previously been built for rhetorical figures [Java, 2015].

Throughout the writing of this thesis, the computational argumentation community has suffered from a lack of annotated corpora, and evidently still does so to date. In view of this, we are sure that the carefully developed and verified corpora we have developed in this thesis are of great significance for promoting research in computational argumentation. Besides their importance for analysing argumentation strategies, the developed corpora can be employed for many tasks in computational argumentation, such as argumentation mining and quality assessment.

Method: This thesis studies argumentation strategies following the successive steps of introducing a model, building a corpus for that model, and operationalising it. We do not aim to develop novel approaches for the models' operationalisation; rather, our goal is to adopt effective methods that could successfully identify the models' elements, and thus help in the exploration of strategy principles.

The methods we have developed in this thesis can be categorised as follows:

1. *Supervised Learning*: Within this thesis, we developed two supervised machine-learning methods: one for identification of evidence type and one for topic categorisation. We employed a wide variety of linguistic features for training these methods, achieving a high level of effectiveness.

2. *Distant Supervision Learning*: We developed three distantly supervised methods for the identification of dialogue acts, argumentative roles, and frames. The training process of these methods was conducted based on a weakly labelled corpus, in which the labels were derived automatically based on particular metadata in Wikipedia discussions.

3. *Rule-Based Learning*: Rule-based methods tend to be used only rarely as machine learning methods have shown impressive results in a wide range of tasks. Nevertheless, rule-based methods are necessary if there are no available annotated corpora for the studied task, or if creating such corpora is extremely difficult, or even impossible, which was actually the case for the identification of the rhetorical figures. In view of this, we developed a rule-based approach that identified 26 rhetorical figures with high effectiveness.

The methods developed can play a major role in the analysis of argumentation strategies, as their output has been utilised for the exploration of strategy principles. All of the methods developed in this thesis have been made publicly available to other researchers.

Evaluation: The ultimate goal of this thesis is to employ the output of the computational analysis of argumentation strategies in the development of writing-assistant tools, text-generation systems, and similar applications. What we have managed to accomplish in this regard, primarily, are sets of various patterns of strategies among diverse monological and dialogical texts with different properties. We consider these patterns as a basic representation of principles.

To explore the strategy principles, we rely on two methods for pattern analysis:

1. *Item Distribution*: We used the distribution of evidence types and rhetorical figures to explore the selection patterns of the pragmatic and stylistic attributes in different monological and dialogical texts.

2. *Sequential Flows*: We used the sequential flows to explore the arrangement pattern of the pragmatic attributes of evidence types in editorials.

The principles discovered with regard to attribute distribution and flows are simple, yet they could be integrated successfully into downstream applications. While we have not reached this step within this thesis, we have succeeded in building the foundation for such integration.

6.2 Future Research Directions

This thesis is, to our knowledge, the first to study the computational analysis of argumentation strategies. While it represents a major step forward in this direction, there are still various relevant research questions that are definitely worth following up with thorough investigation.

Here, we discuss some future research directions along with the four high-level dimensions we mentioned in the previous section:

Theory: Within this thesis, we have proposed several models for the pragmatic and stylistic attributes of an argument. However, an essential category has still not been addressed adequately, namely, the dialectical one. It is quite clear that the identification of dialectical attributes is extremely challenging. This is due to many factors such as the need to approach the logical side of argumentation (e.g., for recognising sound arguments) as well as the potential of dealing with a high degree of subjectivity (e.g., for assessing an argument's strength). However, recent advances in artificial intelligence could open the door for approaching many dialectical attributes. In particular, the identification of argumentation schemes may present a deeply estimable ground for the analysis of strategies.

Moreover, various pragmatic attributes such as the argument roles of claim and conclusion, and those of support and attack, are deserving of in-depth investigation. As regards stylistic attributes, diverse rhetorical figures can be identified and utilised for discovering new strategy principals.

Data: Webis-WikiDebate-18 corpus was built using a distant supervision method. The method employed a set of metadata information in Wikipedia discussions to derive the labels for turns. There are still many types and instances of metadata that could help to expand the corpus. Webis-RhetoricalFigures-19 is relatively a small corpus. One way to expand it would be by applying the automatic identification rules to a large set of unlabelled sentences and then verifying the output of the rules for each sentence.

Method: There is no doubt that deep learning techniques have shown outstanding effectiveness in various tasks in NLP. We did not use such techniques in this thesis, but we assume that promising results will be accomplished by using deep learning in the identification of the pragmatic, stylistic, and dialectical attributes of arguments. Furthermore, as we expect that some attributes, especially in the stylistic category, are quite complicated to annotate, 'few-shot' learning methods could be investigated in this context, since such methods are able to learn based on a limited number of annotated instances.

Evaluation: In this thesis, the analysis of the selection and arrangement principles of strategies is restricted to exploring the distribution and sequential flows of the studied attributes. However, more complicated structures such as trees and graphs can be used for deriving principles regarding many attributes such as the argumentative roles of claim and conclusions. Moreover, an important step in the analysis of argumentation strategies is determining the effectiveness of principles. This step can be accomplished by examining discourses that achieved their goals and those that did not. Moreover, as we stated earlier, the patterns we explored in this thesis form a basic representation of principles. More complicated representation could be investigated, and such representations may take into account the relation between the principles, including the possible overlaps and conflicts.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68. Association for Computational Linguistics, 2014. 29, 62

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128. Association for Computational Linguistics, 2017. 3

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. Modeling Frames in Argumentation. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th Internationl Joint Conference on Natural Language Processing (EMNLP 2019)*. ACL, 2019. 33

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Kohler, and Benno Stein. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*, pages 1395–1404. Association for Computational Linguistics, 2016. 3, 30, 62, 82

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443. The COLING 2016 Organizing Committee, 2016. 26

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357. Association for Computational Linguistics, 2017. 26, 28

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 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, 2017. 103

Khalid Al-Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. Modeling Deliberative Argumentation Strategies on Wikipedia. In *56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 2545–2555. Association for Computational Linguistics, 2018. 26, 28

P. Anand, J. King, Jordan Boyd-Graber, E. Wagner, C. Martell, Douglas Oard, and Philip Resnik. Believe Me—We Can Do This! Annotating Persuasive Acts in Blog Text. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011. 30, 82

Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. Annotating Agreement and Disagreement in Threaded Discussion. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC*, pages 818–822, 2012. 82

Aristotle and W.R. Roberts. *Rhetoric*. Dover thrift editions. Dover Publications, 2004. 3

Aristotle. *On Rhetoric: A Theory of Civic Discourse (George A. Kennedy, Translator)*. Clarendon Aristotle series. Oxford University Press, 2007. 15

V.G. Ashok, S Feng, and Y Choi. 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 (EMNLP 13)*, pages 1753–1764. Association for Computational Linguistics, 2013. 85, 105

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA), 2010. 55

Bal Krishna Bal and Patrick Saint Dizier. Towards building annotated resources for analyzing opinions and argumentation in news editorials. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA), 2010. 28

Bal Krishna Bal and Patrick Saint-Dizier. Towards and Analysis of Argumentation Structure and the Strength of Arguments in News Editorials. In *AISB Symposium on Persuasive Technologies*, pages 55–63, 2009. 62

Bal Krishna Bal. Towards an Analysis of Opinions in News Editorials: How Positive Was the Year? In *Proceedings of the Eighth International Conference on Computational Semantics (IWCS 2009)*, IWCS-8 '09, pages 260–263. Association for Computational Linguistics, 2009. 28, 62

Trevor Bench-Capon, Katie Atkinson, and Peter McBurney. Altruism and Agents: An Argumentation Based Approach to Designing Agent Decision Mechanisms. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2, (AAMAS 2009)*, pages 1073–1080, 2009. 26

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. Annotating Social Acts: Authority Claims and Alignment Moves in Wikipedia Talk Pages. In *Proceedings of the Workshop on Languages in Social Media*, pages 48–57. Association for Computational Linguistics, 2011. 81

Shane Bergsma, Matt Post, and David Yarowsky. 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, 2012. 85

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. Bracketing Guidelines for Treebank II Style Penn Treebank Project. Technical report, University of Pennsylvania, 1995. 43

Or Biran and Owen Rambow. Identifying Justifications in Written Dialogs by Classifiying Text as Argumentative. *International Journal of Semantic Computing*, 5(4):363–381, 2011. 29

Sravan Bodapati, Spandana Gella, Kasturi Bhattacharjee, and Yaser Al-Onaizan. Neural word decomposition models for abusive language detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 135–145. Association for Computational Linguistics, 2019. 31

Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58. Association for Computational Linguistics, 2014. 30

R.R. Braddock. *Research in Written Composition.* National Council of Teachers of English, 1963. 6

Wolfram Bublitz, Timo Müller, Christina Wald, and Hubert Zapf. *Introducing Linguistics*, pages 367–369. J.B. Metzler, 2012. 14

G. Burton. The forest of rhetoric (silva rhetoricae), 2007. Accessed on 16.08.2017. 85, 88

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 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*, 2014. 4, 104

Elena Cabrio and Serena Villata. Natural Language Arguments: A Combined Approach. In *20th European Conference on Artificial Intelligence (ECAI 2012)*, 2012. 26

Elena Cabrio, Sara Tonelli, and Serena Villata. A natural language account for argumentation schemes. In *Proceeding of the XIIIth International Conference on AI*IA 2013: Advances in Artificial Intelligence - Volume 8249*, pages 181–192. Springer-Verlag New York, Inc., 2013. 31

Amparo Elizabeth Cano-Basave and Yulan He. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413. Association for Computational Linguistics, 2016. 4, 35

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 438–444, 2015. 33

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631. Association for Computational Linguistics, 2018. 35

Marisa Chow. Argument Identification in Chinese Editorials. In *NAACL Student Research Workshop 2016*. Association for Computational Linguistics, 2016. 62

Silvie Cinková, Martin Holub, and Vincent Kríž. Managing Uncertainty in Semantic Tagging. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, EACL '12, pages 840–850. Association for Computational Linguistics, 2012. 48

Danish Contractor, Yufan Guo, and Anna Korhonen. Using argumentative zones for extractive summarization of scientific articles. In *Proceedings of COLING 2012*, pages 663–678. The COLING 2012 Organizing Committee, 2012. 29

E. P. J. Corbett. *Classical rhetoric for the modern student.* USA: Oxford University Press, 3 edition, 1990. 87, 88

Robert T. Craig. Communication. In *Encyclopedia of Rhetoric.* Oxford University Press, 2006. 34

M A de Graaf, K J Jager, C Zoccali, and F W Dekker. Matching, an Appealing Method to Avoid Confounding? *Nephron Clin Pract*, 2011. 102

R. Declerck and S. Reed. *Conditionals: A Comprehensive Empirical Analysis.* Beitrage Zur Alexander-Von-Humboldt-Forschung. Mouton de Gruyter, 2001. 88

Semire Dikli. An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment*, 5(1), 2006. 27

Fei Dong and Yue Zhang. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077. Association for Computational Linguistics, 2016. 27

Jesse Dunietz and Daniel Gillick. A New Entity Salience Task with Millions of Training Examples. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (ACL 2014) volume 2: Short Papers*, pages 205–209. Association for Computational Linguistics, 2014. 63

Esin Durmus and Claire Cardie. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT) 2018, Volume 1 (Long Papers)*, pages 1035–1045, 2018. 5, 35

Esin Durmus and Claire Cardie. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607. Association for Computational Linguistics, 2019. 35

Rory Duthie, Katarzyna Budzynska, and Chris Reed. Mining ethos in political debate. In *6th International Conference on Computational Models of Argument (COMMA 16)*, pages 299–310, 2016. 4, 34

Frans H. Van Eemeren and Rob Grootendorst. Fallacies in pragma-dialectical perspective. *Argumentation*, 1(3):283–301, 1987. 15, 16

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22. Association for Computational Linguistics, 2017. 32

Roxanne El-Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus. In *22nd Conference on Computational Natural Language Learning (CoNLL 2018)*, pages 454–464. Association for Computational Linguistics, 2018. 5, 28, 35

Donald G. Ellis. *Argumentative Discourse.* American Cancer Society, 2008. 14

Robert M. Entman. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, 43(4):51–58, 1993. 82

James F. Voss and Julie Van Dyke. Argumentation in psychology: Background comments. *Discourse Processes*, 32:89–111, 2001. 14

Jeanne Fahnestock. Verbal and Visual Parallelism. *Written Communication*, 20(2):123–152, 2003. 88

Rong En Fan, Kai-Wei Chang, Cho-Jui Hsieh, X.-R. Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *JMLR*, 2008. 78

Vanessa Wei Feng and Graeme Hirst. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996. Association for Computational Linguistics, 2011. 4, 31

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 777–786. Association for Computational Linguistics, 2012. 5, 29, 35, 54, 66, 77, 78, 81

Jakub J. Gawryjołek, Randy A. Harris, and Chrysanne DiMarco. An annotation tool for automatically detecting rhetorical figures. In *Proceedings, CMNAIX (Computational Models of Natural Argument)*, 2009. 34, 104, 105

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 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)*, 2015. 104

James Gingell. Why superlatives are the absolute worst (unless you're Donald Trump). `https://www.theguardian.com/media/mind-your-language/2016/apr/15/`, 2016. visited on 24.10.17. 104

Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. 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, 2011. 105

Yufan Guo, Anna Korhonen, and Thierry Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283. Association for Computational Linguistics, 2011. 29

Yufan Guo, Ilona Silins, Roi Reichart, and Anna Korhonen. CRAB reader: A tool for analysis and visualization of argumentative zones in scientific literature. In *Proceedings of COLING 2012: Demonstration Papers*, pages 183–190. The COLING 2012 Organizing Committee, 2012. 29

Iryna Gurevych, Eduard H. Hovy, Noam Slonim, and Benno Stein. Debating Technologies (Dagstuhl Seminar 15512). *Dagstuhl Reports*, 5(12):18–46, 2016. 3

Ivan Habernal and Iryna Gurevych. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137. Association for Computational Linguistics, 2015. 30, 34, 62

Ivan Habernal and Iryna Gurevych. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223. Association for Computational Linguistics, 2016. 3, 32

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. 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 1930–1940. Association for Computational Linguistics, 2018. 31

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. 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 386–396. Association for Computational Linguistics, 2018. 4, 31

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. SemEval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 763–772. Association for Computational Linguistics, 2018. 31

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18, 2009. 59

Jiawei Han, Micheline Kamber, and Jian Pei. 6 - mining frequent patterns, associations, and correlations: Basic concepts and methods. In *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 243 – 278. Morgan Kaufmann, third edition edition, 2012. 24

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, 2017. 4, 30, 34

Kai Hong and Ani Nenkova. Improving the Estimation of Word Importance for News Multi-Document Summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (ECAL 2014)*, pages 712–721. Association for Computational Linguistics, 2014. 63

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596, 2017. 85

Xinyu Hua and Lu Wang. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230. Association for Computational Linguistics, 2018. 3

Stephanie Husby and Denilson Barbosa. Topic Classification of Blog Posts Using Distant Supervision. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 28–36. Association for Computational Linguistics, 2012. 59

Jane Im, Amy X. Zhang, Christopher J. Schilling, and David Karger. Deliberation and resolution on wikipedia: A case study of requests for comments. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):74:1–74:24, 2018. 36

Gary G. Koch J. Richard Landis. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. 46

James Java. *Characterization of Prose by Rhetorical Structure for Machine Learning Classification*. PhD thesis, Nova Southeastern University, 2015. 34, 86, 95, 96, 104, 105, 109

Minwoo Jeong, Chin-Yew Lin, and Geunbae Gary Lee. Semi-supervised Speech Act Recognition in Emails and Forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 1250–1259. Association for Computational Linguistics, 200. 54

R. Johnson. *The Alphabet of Rhetoric*. BiblioLife, 2016. 85

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. In *Proceedings of 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, COLING*, pages 2012–2021, 2016. 82

Johannes Kiesel, Khalid Al-Khatib, Matthias Hagen, and Benno Stein. A Shared Task on Argumentation Mining in Newspaper Editorials. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 35–38, 2015. 28

Johannes Kiesel, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. WAT-SL: A Customizable Web Annotation Tool for Segment Labeling. In *Software Demonstrations at the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 13–16, 2017. 45

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. Classifying Dialogue Acts in One-on-one Live Chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 862–871. Association for Computational Linguistics, 2010. 82

Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI*, pages 453–462. ACM, 2007. 5, 35, 65

Peter Kluegl, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe. Uima ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22:1–40, 2016. 97

Robert E. Kraut, Paul Resnick, Sara Kiesler, Yuqing Ren, Yan Chen, Moira Burke, Niki Kittur, John Riedl, and Joseph Konstan. *Building Successful Online Communities: Evidence-Based Social Design.* The MIT Press, 2012. 65

V. Kvint. *The global emerging market: strategic management and economics.* Routledge, London, 2009. 17

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46. Association for Computational Linguistics, 2018. 29

John Lawrence, Jacky Visser, and Chris Reed. Harnessing rhetorical figures for argument mining. *Argument & Computation*, 8(3):289–310, 2017. 34, 86, 104

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500. Dublin City University and Association for Computational Linguistics, 2014. 29

Jessy Junyi Li, Kapil Thadani, and Amanda Stent. The Role of Discourse Units in Near-Extractive Summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147. Association for Computational Linguistics, 2016. 63

Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. Using argument-based features to predict and analyse review helpfulness. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1363. Association for Computational Linguistics, 2017. 28

Liane Longpre, Esin Durmus, and Claire Cardie. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*, pages 167–176. Association for Computational Linguistics, 2019. 35

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753. Association for Computational Linguistics, 2017. 5, 35

Didier Maillat and Steve Oswald. Biases and constraints in communication: Argumentation, persuasion and manipulation. *Journal of Pragmatics*, 59:137 – 140, 2013. Biases and constraints in communication: Argumentation, persuasion and manipulation. 14

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. 43, 97

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. 93

André Martinet. *Elements of General Linguistics*. Faber and Faber Ltd., 1960. 92

Brett McKay and Kate McKay. Classical Rhetoric 101, 2010. Accessed on 14.08.2017. 87

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing, ACL*, pages 1003–1011. Association for Computational Linguistics, 2009. 68

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230. ACM, 2007. 28

Dima Mohammed. Goals in argumentation: A proposal for the analysis and evaluation of public political arguments. *Argumentation*, 30(3):221–245, 2016. 1

Alvaro Morales, Varot Premtoon, Cordelia Avery, Sue Felshin, and Boris Katz. Learning to Answer Questions from Wikipedia Infoboxes. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1930–1935. Association for Computational Linguistics, 2016. 82

Wolfgang G. Müller. Style. In *Encyclopedia of Rhetoric*. Oxford University Press, 2006. 88

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 82–93. Association for Computational Linguistics, 2016. 31

Elena Musi, Manfred Stede, Leonard Kriese, Smaranda Muresan, and Andrea Rocci. A multi-layer annotated corpus of argumentative text: From argument schemes to discourse relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Languages Resources Association (ELRA), 2018. 31

Nona Naderi and Graeme Hirst. Argumentation mining in parliamentary discourse. In *Principles and Practice of Multi-Agent Systems - International Workshops: IWEC 2014, Gold Coast, QLD, Australia, December 1-5, 2014, and CMNA XV and IWEC 2015, Bertinoro, Italy, October 26, 2015, Revised Selected Papers*, pages 16–25, 2015. 33

Nona Naderi and Graeme Hirst. Classifying Frames at the Sentence Level in News Articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*, pages 536–542, 2017. 33, 82

Timothy Niven and Hung-Yu Kao. Detecting argumentative discourse acts with linguistic alignment. In *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*, pages 104–112, 2019. 33

Steve Oswald, Thierry Herman, and JÃ©rÃ´me Jacquin. *Argumentation and Language – Linguistic, Cognitive and Discursive Explorations*. 2018. 14

Irwin P. Levin, Sandra L. Schneider, and Gary J. Gaeth. All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects. *Organizational Behavior and Human Decision Processes*, 76(2):149 – 188, 1998. 66

Raquel Mochales Palau and Marie-Francine Moens. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL 2009*, pages 98–107, 2009. 28

Alexander Panchenko, Alexander Bondarenko, Mirco Franzek, Matthias Hagen, and Chris Biemann. Categorizing Comparative Sentences. In *Proceedings* *of the 6th Workshop on Argument Mining*. Association for Computational Linguistics, 2019. 3

Joonsuk Park and Claire Cardie. Identifying Appropriate Support for Propositions in Online User Comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38. Association for Computational Linguistics, 2014. 4, 30, 43

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135. Association for Computational Linguistics, 2017. 31

Andreas Peldszus and Manfred Stede. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31, 2013. 66, 82

Andreas Peldszus and Manfred Stede. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948. Association for Computational Linguistics, 2015. 3, 32

Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Vol. 2*, pages 801–815. College Publications, 2016. 31, 32

Isaac Persing and Vincent Ng. Modeling Thesis Clarity in Student Essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics - Volume 1: Long Papers*, pages 260–269, 2013. 27

Isaac Persing and Vincent Ng. Modeling Prompt Adherence in Student Essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics - Volume 1: Long Papers*, pages 1534–1543, 2014. 27

Isaac Persing and Vincent Ng. Modeling argument strength in student essays. 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 543–552. Association for Computational Linguistics, 2015. 3, 27

Isaac Persing and Vincent Ng. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394. Association for Computational Linguistics, 2016. 27

Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239. Association for Computational Linguistics, 2010. 27

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn discourse TreeBank 2.0. In *LREC 2008*, 2008. 31

Pavithra Rajendran, Danushka Bollegala, and Simon Parsons. Is something better than nothing? automatically predicting stance-based arguments using deep learning and small labelled dataset. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 28–34. Association for Computational Linguistics, 2018. 28

Robin Raskin. Hillary clinton's acceptance speech as seen by the algorithms. the huffington post. `https://www.huffingtonpost.com/robin-raskin`, 2016. visited on 18.11.17. 104

Chris Reed and Glenn Rowe. Araucaria: Software for Argument Analysis, Diagramming and Representation. *International Journal on Artificial Intelligence Tools*, 13, 2004. 31

Eddo Rigotti and Sara Greco. Comparing the argumentum model of topics to other contemporary approaches to argument schemes: The procedural and material components. *Argumentation*, 24:489–512, 2010. 31

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 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, 2015. 4, 29, 33, 39, 62

Haggai Roitman, Shay Hummel, Ella Rabinovich, Benjamin Sznajder, Noam Slonim, and Ehud Aharoni. On the retrieval of wikipedia articles containing claims on controversial topics. In *Proceedings of the 25th International Conference on World Wide Web, Companion Volume*, pages 991–996, 2016. 29

Sara Rosenthal and Kathy McKeown. I Couldn't Agree More: The Role of Conversational Structure in Agreement and Disagreement Detection in Online Discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL*, pages 168–177, 2015. 82

Evan Sandhaus. The new york times annotated corpus ldc2008t19. dvd. *Philadelphia: Linguistic Data Consortium*, 2008. 38, 87, 100

Baruch Schwarz and Christa Asterhan. *Argumentation and reasoning*, pages 137–176. 2010. 14

J.R. Searle. *Speech Acts: An Essay in the Philosophy of Language.* Cam: Verschiedene Aufl. Cambridge University Press, 1969. 18, 33, 66, 82

Wei Song, Tong Liu, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. Learning to identify sentence parallelism in student essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 794–803. The COLING 2016 Organizing Committee, 2016. 4

Christian Stab and Iryna Gurevych. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014*, pages 1501–1510, 2014. 27, 32, 43, 62

Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56. Association for Computational Linguistics, 2014. 3, 27, 32

Manfred Stede and Jodi Schneider. *Argumentation Mining*, volume 40 of *Synthesis Lectures in Human Language Technology.* Morgan & Claypool, 2018. 3, 14, 33

Claus W. Strommer. *Using rhetorical figures and shallow attributes as a metric of intent in text.* PhD thesis, University of Waterloo, 2011. 34

Kalaivani Sundararajan and Damon L. Woodard. 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, 2018. 105

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International World Wide Web Conference*, pages 613–624, 2016. 4, 35, 82

Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Comput. Linguist.*, 28(4):409–445, 2002. 29

Stephen E. Toulmin. *The Uses of Argument.* Cambridge University Press, 1958. 31

Giovanni Tripepi, Kitty J Jager, Friedo W. Dekker, and Carmine Zoccali. Stratification for confounding – part 1: The mantel-haenszel formula. 2010. 100

Oren Tsur, Dan Calacci, and David Lazer. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL-IJCNLP*, pages 1629–1638. Association for Computational Linguistics, 2015. 82

Teun A. van Dijk. Racism and argumentation: Race Rio Rhetoric in Tabloib Editorials . In *In van Emeren et al. (Eds.), Argumentation illuminated*, 1992. 6, 27, 37

Teun A. van Dijk. Opinions and Ideologies in Editorials. In *Proceedings of the 4th International Symposium of Critical Discourse Analysis, Language, Social Life and Critical Thought*, 1995. 37, 60

Frans H. van Eemeren and Rob Grootendorst. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach.* Cambridge University Press, 2004. 15, 16

Frans H. van Eemeren and Peter Houtlosser. 15, 16, 17, 19, 20

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe nad A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. *Handbook of Argumentation Theory.* Springer Netherlands, 2014. 1

Fernanda B. Viegas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. Talk Before You Type: Coordination in Wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, HICSS '07, pages 78–. IEEE Computer Society, 2007. 29, 81

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 2019. 4, 33

Henning Wachsmuth and Kathrin Bujna. Back to the Roots of Genres: Text Classification by Language Function. In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 632–640. Asian Federation of Natural Language Processing, 2011. 54

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. Modeling Review Argumentation for Robust Sentiment Analysis. In *Proceedings of the 25th International Conference on Computational Linguistics (COLOING 2014)*, pages 553–564, 2014. 28

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. Sentiment Flow — A General Model of Web Review Argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 601–611. Association for Computational Linguistics, 2015. 28, 57, 60

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Using Argument Mining to Assess the Argumentation Quality of Essays. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, 2016. 27

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics, 2017. 3, 23

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59. Association for Computational Linguistics, 2017. 3

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. Argumentation Synthesis following Rhetorical Strategies. In *The 27th International Conference on Computational Linguistics (COLING 2018)*. Association for Computational Linguistics, 2018. 26

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251. Association for Computational Linguistics, 2018. 4, 32

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *Proceedings of the Eighth*

International Conference on Language Resources and Evaluation (LREC-2012), pages 812–817. European Languages Resources Association (ELRA), 2012. 36

Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008. 30, 31

Douglas Walton. Types of Dialogue and Burdens of Proof. In *Frontiers in Artificial Intelligence and Applications*, volume 216, pages 13–24, 2010. 1, 65, 81

Lu Wang and Claire Cardie. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. In *Proceedings of the52nd Annual Meeting of the Association for Computational Linguistics, ACL*, volume 2, pages 693–699. Association for Computational Linguistics, 2014. 5, 35, 65, 82

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association of Computational Linguistics*, 5:219–232, 2017. 4, 35

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649. Association for Computational Linguistics, 2019. 35

Zhongyu Wei, Yang Liu, and Yi Li. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200. Association for Computational Linguistics, 2016. 4, 35

Theresa Wilson and Janyce Wiebe. Annotating opinions in the world press. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, 2003. 62

John Woods, Andrew Irvine, and Douglas Walton. *Argument - critical thinking, logic and the fallacies (2. ed.)*. 2004. 14

John T. Woolley and Peters Gerhard. American Presidency Project. `http://www.presidency.ucsb.edu/`, 2017. visited on 18.11.17. 87, 102

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1391–1399. International World Wide Web Conferences Steering Committee, 2017. 31

Adam Wyner, Jodi A Schneider, Katie Atkinson, and Trevor Bench-Capon. Semi-automated argumentative analysis of online product reviews. In *Computational Models of Argument - Proceedings of COMMA 2012*, number 1 in Frontiers in Artificial Intelligence and Applications, pages 43–50. IOS Press, 2012. 28

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. 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 3620–3630. Association for Computational Linguistics, 2019. 35

Shiren Ye, Tat-Seng Chua, and Jie Lu. Summarizing Definition from Wikipedia. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing, ACL*, pages 199–207. Association for Computational Linguistics, 2009. 82

Hong Yu and Vasileios Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, EMNLP '03, pages 129–136. Association for Computational Linguistics, 2003. 28, 62

Caixia Yuan, Xiaojie Wang, and Qianhui He. 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, 2015. 105

Elina Zarisheva and Tatjana Scheffler. Dialog Act Annotation for Twitter Conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL*, pages 114–123, 2015. 82

Torsten Zesch, Christof Muller, and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC*. European Language Resources Association (ELRA), 2008. 69

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational flow in oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141. Association for Computational Linguistics, 2016. 32

Amy X. Zhang, Bryan Culbertson, and Praveen Paritosh. Characterizing Online Discussion Using Coarse Discourse Sequences. In *Proceedings of the 11th International AAAI Conference on Weblogs and Social Media, ICWSM*, pages 357–366, 2017. 33, 65, 78

Amy X. Zhang, Lea Verou, and David Karger. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW*, pages 2082–2096. ACM, 2017. 65

About the Author

Khalid Al-Khatib is a Ph.D. student at the Bauhaus-Universitat Weimar in Germany. His central research interests include computational argumentation, detection of bias and offensive language, and knowledge extraction from the web. During his study, he investigated various research questions related to our society such as: How do people argue on the web? What are the strategies that people use to achieve the goal of persuasion? and which strategies are the most successful in some particular contexts?

Khalid Al-Khatib received one of the prestigious IBM PhD Fellowship Awards.